Local Projections, Autocorrelation, and Efficiency

Lusompa, Amaze

University of California, Irvine

14 November 2019

Local Projections, Autocorrelation, and Efficiency

Amaze Lusompa*
University of California, Irvine
alussomp@uci.edu

Click here for the most recent draft.

April 11, 2020

Abstract

It is well known that Local Projections (LP) residuals are autocorrelated. Conventional wisdom says that LP have to be estimated by OLS with Newey and West (1987) (or some type of Heteroskedastic and Autocorrelation Consistent (HAC)) standard errors and that GLS is not possible because the autocorrelation process is unknown. I show that the autocorrelation process of LP is known and that autocorrelation can be corrected for using GLS. Estimating LP with GLS has three major implications: 1) LP GLS can be substantially more efficient and less biased than estimation by OLS with Newey-West standard errors. 2) Since the autocorrelation process can be modeled explicitly, it is possible to give a fully Bayesian treatment of LP. That is, LP can be estimated using frequentist/classical or fully Bayesian methods. 3) Since the autocorrelation process can be modeled explicitly, it is now possible to estimate time-varying parameter LP.

*I thank Regis Barnichon, Rhys Bidder, Bill Branch, Ivan Jeliazkov, Òscar Jordà, Fabio Milani, Eric Swanson, Mike West, Jonathan Wright, and seminar participants at several venues for helpful comments, discussions, and/or suggestions. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1106401, the Federal Reserve Bank of San Francisco's Thomas J. Sargent Dissertation Fellowship, and the Federal Reserve Bank of Boston under the American Economic Association Summer Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation, the Federal Reserve Bank of San Francisco, or the Federal Reserve Bank of Boston.

1 Introduction

Vector Autoregressions (VARs) were proposed in Sims (1980) as an alternative to the large scale simultaneous equation models of the time. Since then, VARs have been a major tool used in empirical macroeconomic analysis, primarily being used for causal analysis and forecasting through the estimation of impulse response functions. In a seminal paper, Jordà (2005) argued that impulse response functions could be estimated directly using linear regressions called Local Projections (LP) and that LP are more robust to model misspecification than VARs.[1,2] LP have been growing in popularity ever since, and the two methods often give different results when applied to the same problem (Ramey, 2016, Nakamura and Steinsson, 2018). If the true model is a VAR, then a correctly specified VAR is more efficient than LP because VARs impose more structure than LP (Ramey, 2016).[3] If the true model is not a VAR or if the lag length of the VAR is not sufficiently long, then LP can outperform VARs (Plagborg-Møller and Wolf, 2019). Being that LP impulse responses nest VAR impulse responses, the choice of whether to use impulse responses from LP or VARs can be thought of as the bias-variance tradeoff problem with VARs and LP lying on a spectrum of small sample bias variance choices.

It is well known that LP residuals are autocorrelated. Practitioners exclusively estimate LP via OLS with Newey-West standard errors (or some type of Heteroskedastic and Autocorrelation Consistent (HAC) standard errors) (Ramey, 2016). Jordà (2005) argues that since the true data generating process is unknown, Generalized Least Squares (GLS) is not possible and HAC standard errors must be used. Lazarus et al. (2018) claim that LP have to be estimated with HAC standard errors because GLS estimates would be inconsistent.[4] I show that under standard time series assumptions, the autocorrelation process is known and autocorrelation can be corrected for using GLS. Moreover, I show the consistency and asymptotically normality of the LP GLS estimator, as well as the asymptotic efficiency of LP GLS relative to LP OLS.

Being able to specify the autocorrelation process for LP has 3 major implications. First, LP GLS can be substantially more efficient than LP estimated via OLS with Newey-West standard errors. Moreover, once autocorrelation is corrected for, it can be shown that if the data is persistent and the true model is a VAR, LP GLS impulse responses can be approximately as efficient as VAR impulse responses. Whether or not LP GLS impulse responses are approximately as efficient depends on the persistence of the system, the horizon, and the dependence structure of the system. All else equal, the more persistent the system, the more likely

[1] As noted in Stock and Watson (2018), LP are direct multistep forecasts. However, the goal of direct multistep forecast is an optimal multistep ahead forecast, whereas the goal of LP is a consistent estimate of the corresponding impulse responses.

[2] In the case of stationary time series, Plagborg-Møller and Wolf (2019) show if the sample size is infinite, linear time-invariant VAR(∞) and LP(∞) estimate the same impulse responses. This equivalence does not hold if the models are augmented with non-linear terms.

[3] If one is willing to assume a likelihood function for the model, this is just the Cramer Rao Lower Bound argument.

[4] Lazarus et al. (2018) assume strict exogeneity (which neither LP or VARs satisfy) is necessary for GLS. Even though strict exogeneity is often assumed for GLS, it is not a necessary condition for GLS (see Hamilton (1994), Stock and Watson (2007) for discussions on the strict exogeneity assumption for GLS).

LP GLS impulse responses will be approximately as efficient for horizons typically relevant in practice. It follows that the efficiency of the VAR relative to the LP has been overstated in the literature.

Second, since the autocorrelation process is known, LP GLS can be estimated using fully Bayesian methods.[5] Bayesian LP have many advantages such as allowing the researcher to incorporate prior information for impulse responses at each horizon. Prior information can be used to shrink impulse responses at any horizon to prevent overfitting. Economic theory can be incorporated into the prior to inform the shape of the impulse responses (e.g. the impulse response is monotonic or hump shaped) and to discipline the long-run behavior. Priors can be used to shrink parameter estimates when the number of parameters is large relative to the number of observations making it possible to use LP to estimate systems with big data or panel data with large cross sections over relatively short time frames (e.g. the Eurozone). Moreover, methodologies used for Bayesian VARs (i.e. big data, sparsity, and variable selection methods) can now be carried over to LP. Lastly, Bayesian methods do not need to do anything special to take into account unit roots.

Third, since autocorrelation is explicitly modeled, it is now possible to estimate time-varying parameter LP. Time-varying parameter models are useful for several reasons. Researchers are often interested in whether there is parameter instability in regression models. As noted in Granger and Newbold (1977), macro data encountered in practice are unlikely to be stationary. Stock and Watson (1996) and Ang and Bekaert (2002) show many macroeconomic and financial time series exhibit parameter instability. It is also commonplace for regressions with macroeconomic time series to display heteroskedasticity of unknown form (Stock and Watson, 2007), and in order to do valid inference, the heteroskedasticity must be taken into account. Parameter instability can occur for many reasons such as policy changes, technological evolution, changing economic conditions, etc. If parameter instability is not appropriately taken into account, it can lead to invalid inference, poor out of sample forecasting, and incorrect policy evaluation. Moreover, as shown in Granger (2008), time-varying parameter models can approximate any non-linear model (non-linear in the variables and/or the parameters), which makes them more robust to model misspecification. Bayesian methods are the primary methods used to estimate time-varying parameter models, and since autocorrelation is explicitly corrected for in Bayesian LP, it is straightforward to apply time-varying parameters to LP.[6]

In this paper, I make several contributions. I show that the autocorrelation process of LP is known and that autocorrelation can be corrected for using GLS. Estimating LP with GLS has three major implications: First, LP GLS can be substantially more efficient and less biased than estimation by OLS with Newey-West standard errors. Second, LP GLS can be estimated using fully Bayesian or frequentist methods. Third, it is now possible to estimate time-varying parameter LP. The paper is outlined as follows: section 2 contains

[5] Miranda-Agrippino and Ricco (2018) introduce a method called Bayesian LP, but the method is not fully Bayesian, because they replace the estimated scale matrix in the inverse-Wishart posterior with the Newey-West variance covariance matrix. Using plug-in estimates for hyper–parameters is well known to cause probability intervals to underrepresent uncertainty (Koop and Korobilis, 2009, Hoff and Wakefield, 2013). Furthermore, I will show why autocorrelation should be explicitly corrected for in LP.

[6] Time-varying parameter LP do not have to implemented using Bayesian methods.

the core result showing that the autocorrelation process of LP is known and illustrates why GLS is possible. Section 3 explains how to estimate LP GLS using both frequentist and Bayesian methods. Section 4 discusses the relative efficiency of LP estimated by OLS with Newey-West standard errors vs LP GLS. Section 5 contains Monte Carlo evidence of the small sample properties of LP GLS. Section 6 discusses issues in regards to non-stationarity. Section 7 explains how time-varying parameter LP can be estimated and illustrates a Bayesian procedure to do so. Section 8 concludes.

Some notation: $N(\cdot,\cdot)$, $IW(\cdot,\cdot)$, are the normal, and inverse-Wishart distributions, respectively. $T_n(\cdot,\cdot)$ is the T-distribution with n degrees of freedom. $y_{1:T} = \{y_1, ..., y_T\}$. $\xrightarrow{p}$ is converges in probability, and $\xrightarrow{d}$ is converges in distribution.

2 The Autocorrelation Process, OLS, and GLS

2.1 LP and Newey-West Standard Errors

To illustrate how LP work, take the simple VAR(1) model

$$y_t = A_1 y_{t-1} + \varepsilon_t, \tag{1}$$

where y_t is a demeaned $r \times 1$ vector of endogenous variables and ε_t is an $r \times 1$ vector white noise process and $var(\varepsilon_t) = \Sigma_\varepsilon$.[7] Assume that the eigenvalues of A_1 have moduli less than unity and $A_1 \neq 0$. Iterating forward leads to

$$y_{t+h} = A_1^{h+1} y_{t-1} + A_1^h \varepsilon_t + \ldots + A_1 \varepsilon_{t+h-1} + \varepsilon_{t+h}.$$

To estimate the impulse responses of a VAR, one would estimate A_1 from equation (1) and then use the non-linear delta method, bootstrapping, or Monte Carlo integration to perform inference on the impulse responses: $\{A_1, A_1^2, \ldots, A_1^{h+1}\}$. To estimate impulse responses using LP, one would estimate the impulse responses directly at each horizon with separate regressions

$$y_t = B_1^{(1)} y_{t-1} + e_t^{(0)},$$

$$y_{t+1} = B_1^{(2)} y_{t-1} + e_{t+1}^{(1)},$$

$$\vdots$$

$$y_{t+h} = B_1^{(h+1)} y_{t-1} + e_{t+h}^{(h)},$$

[7] Without loss of generality, y_t is demeaned in order to remove the constant and simplify notation.

where h is the horizon, and when the true data generating process is a VAR(1), $\{B_1^{(1)}, B_1^{(2)}, \ldots, B_1^{(h+1)}\}$ and $\{A_1, A_1^2, \ldots, A_1^{h+1}\}$ are equivalent. Even if the true data generating process is not a VAR(1), $B_1^{(1)} = A_1$ because the horizon 0 LP is a VAR. In practice, it is common for more than one lag to be used. A VAR(k) and the horizon h LP(k) can be expressed as

$$y_t = A_1 y_{t-1} + \ldots + A_k y_{t-k} + \varepsilon_t,$$

and

$$y_{t+h} = B_1^{(h+1)} y_{t-1} + \ldots + B_k^{(h+1)} y_{t-k} + e_{t+h}^{(h)},$$

respectively. Bear in mind that any VAR(k) can be written as a VAR(1) (companion form), so results and examples involving the VAR(1) the can be generalized to higher order VARs.

LP have been advocated by Jordà (2005) as an alternative to VARs. There are several advantages of using LP as opposed to VARs. First, LP do not constrain the shape of the impulse response function like VARs, so it can be less sensitive to model misspecification (i.e. such as insufficient lag length) because misspecifications are not compounded in the impulse responses when iterating forward.[8] Second, LP can be estimated using simple linear regressions. Third, joint or point-wise analytic inference is simple. Fourth, LP can easily be adapted to handle non-linearities (in the variables or parameters).

LP do have a couple of drawbacks. First, because the dependent variable is a lead, a total of h observations are lost from the original sample when estimating projections for horizon h. Second, the error terms in LP for horizons greater than 0 are inherently autocorrelated. Assuming the true model is a VAR(1), it is obvious that autocorrelation occurs because the LP residuals follow an VMA(h) process of the residuals in equation (1). That is

$$e_{t+h}^{(h)} = A_1^h \varepsilon_t + \ldots + A_1 \varepsilon_{t+h-1} + \varepsilon_{t+h},$$

or written in terms of LP

$$e_{t+h}^{(h)} = B_1^{(h)} \varepsilon_t + \ldots + B_1^{(1)} \varepsilon_{t+h-1} + \varepsilon_{t+h}.$$

Frequentists account for the inherent autocorrelation using Newey-West standard errors, which will yield asymptotically correct standard errors in the presence of autocorrelation and heteroskedasticity of unknown forms.[9] Autocorrelation can be corrected for explicitly by including $\{\varepsilon_t, \varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$ in the conditioning set of the horizon h LP. Obviously $\{\varepsilon_t, \varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$ are unobserved and would have to be estimated, but this issue can be ignored for now and is addressed later.

There are two major advantages of correcting for autocorrelation explicitly. The first is that it fixes what

[8] In the case of the linear time-invariant estimators, VAR(∞) and LP(∞) estimate the same impulse responses asymptotically (Plagborg-Møller and Wolf, 2019). This result does not hold if the models are augmented with nonlinear terms.

[9] This is assuming that a large enough lag truncation parameter for the autocorrelation is chosen. There is a major line of research indicating that Newey-West standard errors perform poorly in small samples with persistent data (Müller, 2014).

I dub the "increasing variance problem". To my knowledge, the increasing variance problem has not been noticed in the literature. If the true model is a VAR(1), then $var(e_{t+h}^{(h)}) = \sum_{i=0}^{h} A_1^{i'} \Sigma_\varepsilon A_1^i$, which is increasing in h.[10] Newey-West standard errors are valid in the presence of autocorrelation because they take into account autocorrelation is present when estimating the covariance matrix; they do not, however, eliminate autocorrelation.[11,12] To illustrate, let the true model be an AR(1) with

$$y_t = .99 y_{t-1} + \varepsilon_t,$$

where $var(\varepsilon_t) = 1$. The $var(e_{t+h}^{(h)}) = \sum_{i=0}^{h} A_1^{i'} \Sigma_\varepsilon A_1^i = \sum_{i=0}^{h} .99^{2i}$. The table below presents the asymptotic variance of the residuals for different horizons when estimated by OLS with Newey-West standard errors vs LP estimated with GLS.

Table 1: Asymptotic Variance of Residuals for LP Horizons

Horizons	5	10	20	40
LP NW	5.7093	9.9683	17.3036	28.2102
LP GLS	1	1	1	1

Even if Newey-West standard errors are used, the increasing variance problem persists. In terms of the MLE and OLS, correcting for autocorrelation explicitly is asymptotically more efficient because $var(\varepsilon_t) \leq var(e_t^{(h)})$, where the equality only binds when $A_1 = 0$.

The second major advantage of correcting for autocorrelation explicitly is that it helps remedy what I dub the "increased small sample bias problem". When LP are estimated with OLS and Newey-West standard errors, the small sample bias from estimating dynamic models increases relative to the model with no autocorrelation. To see why, let us first review the finite sample bias problem with VARs (see (Pope, 1990) for detailed derivations). Assume the true model is a VAR(1). The OLS estimate for the VAR is

$$\hat{A}_1 = A_1 + \sum_{t=2}^{T} \varepsilon_t y_{t-1}' (\sum_{t=2}^{T} y_{t-1} y_{t-1}')^{-1}.$$

This estimate is biased in finite samples because $E(\sum_{t=2}^{T} \varepsilon_t y_{t-1}' (\sum_{t=2}^{T} y_{t-1} y_{t-1}')^{-1}) \neq 0$ because ε_t and $(\sum_{t=2}^{T} y_{t-1} y_{t-1}')^{-1}$ are not independent. The stronger the correlation between ε_t and $(\sum_{t=2}^{T} y_{t-1} y_{t-1}')^{-1}$, the larger the bias. In macroeconomic applications, the bias is typically downward. The bias disappears asymptotically since ε_t would be correlated with an increasingly smaller share of $(\sum_{t=2}^{T} y_{t-1} y_{t-1}')^{-1}$.

If one were to estimate a LP via OLS with Newey-West standard errors at horizon h, the OLS estimate

[10] Since A_1 has moduli less than unity, geometric progression can be used to show that the sum is bounded asymptotically.

[11] This is a major reason why Kilian and Kim (2011) found that LP had excessive average length relative to the bias-adjusted bootstrap VAR interval in their Monte Carlo simulations. I provide Monte Carlo evidence of this in section 5.

[12] Macro variables tend to be persistent, so A_1^i will more likely decay slowly leading to the increase in the variance to be pretty persistent as h increases.

would be

$$\hat{B}_1^{(h+1)} = B_1^{(h+1)} + \sum_{t=2}^{T-h} e_{t+h}^{(h)} y_{t-1}' (\sum_{t=2}^{T-h} y_{t-1} y_{t-1}')^{-1}.$$

If one were to correct for autocorrelation by including $\{\varepsilon_t, \varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$, the estimate would be

$$\hat{B}_1^{(h+1)} = B_1^{(h+1)} + \sum_{t=2}^{T-h} \varepsilon_{t+h} y_{t-1}' (\sum_{t=2}^{T-h} y_{t-1} y_{t-1}')^{-1}.$$

The absolute value of the correlation between $e_{t+h}^{(h)}$ and $(\sum_{t=2}^{T-h} y_{t-1} y_{t-1}')^{-1}$ is larger than the absolute value of the correlation between ε_{t+h} and $(\sum_{t=2}^{T-h} y_{t-1} y_{t-1}')^{-1}$ because $e_{t+h}^{(h)} = A_1^h \varepsilon_t + \ldots + A_1 \varepsilon_{t+h-1} + \varepsilon_{t+h}$ is correlated with a larger share of $(\sum_{t=2}^{T-h} y_{t-1} y_{t-1}')^{-1}$.[13] To illustrate, I conduct a simple Monte Carlo simulation where I generate 1,000 samples of length 200 for the following AR(1)

$$y_t = .99 y_{t-1} + \varepsilon_t,$$

where $var(\varepsilon_t) = 1$. I then estimate the impulse responses using a VAR, LP estimated with OLS, and LP estimated with GLS. To correct for autocorrelation using GLS, I include the estimated residuals. Below is the table of the mean impulse responses at different horizons for the different methods.

Table 2: Mean Impulse Response Estimates for T=200

Horizons	5	10	20	40
True	.951	.9044	.8179	.6690
VAR	.8355	.7072	.5231	.3148
LP NW	.8259	.6713	.4223	.0787
LP GLS	.8347	.7045	.5160	.2965

All of the estimated can be substantially biased, but not correcting for autocorrelation can make the bias substantially worse. Even if autocorrelation is corrected for in LP, there can still be a small sample bias due to the correlation between ε_{t+h} and $(\sum_{t=2}^{T-h} y_{t-1} y_{t-1}')^{-1}$ not being 0 in finite samples, but additional bias due to not explicitly correcting for autocorrelation would be eliminated.[14]

2.2 The Autocorrelation Process of LP

This subsection presents the core result: the autocorrelation process of LP is known under standard time series assumptions and can be corrected for via GLS. First, I will show that even when the true data

[13] This is probably a major reason why Kilian and Kim (2011) found that LP impulse responses were more biased than the VAR impulse responses in their Monte Carlo simulations.

[14] LP GLS tends to be a little more biased than the VAR because LP estimated at horizon h loses h observations at the end of the sample.

generating process is not a VAR, including the horizon 0 LP residuals (or equivalently, VAR residuals), $\{\varepsilon_t, \varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$, in the horizon h conditioning set will eliminate autocorrelation as long as the data are stationary and the horizon 0 LP residuals are uncorrelated. Second, I will show that the autocorrelation process of $e_{t+h}^{(h)}$ is known.

Assumption 1. *The data $\{y_t\}$ are stationary and purely non-deterministic so there exists a Wold representation*

$$y_t = \varepsilon_t + \sum_{i=1}^{\infty} \Theta_i \varepsilon_{t-i}.$$

Assumption 1 implies that by the Wold representation theorem, there exists a linear and time-invariant Vector Moving Average (VMA) representation of the uncorrelated one-step ahead forecast errors $\{\varepsilon_t\}$. It follows from the Wold representation theorem that $\varepsilon_t = y_t - Proj(y_t|y_{t-1}, y_{t-2}, \ldots)$ where $Proj(y_t|y_{t-1}, y_{t-2}, \ldots)$ is the (population) orthogonal projection of y_t onto $\{y_{t-1}, y_{t-2}, \ldots\}$.

Consider for each horizon $h = 0, 1, 2, \ldots$ the infinite lag LP

$$y_{t+h} = B_1^{(h+1)} y_{t-1} + B_2^{(h+1)} y_{t-2} + \ldots + e_{t+h}^{(h)}.$$

Proposition 1. *Under Assumption 1, including $\{\varepsilon_t, \varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$ in the conditioning set of the horizon h LP will eliminate autocorrelation in the horizon h LP residuals.*

Proof. I first show that

$$Proj(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_t, y_{t-1}, y_{t-2}, \ldots) = Proj(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_t, y_{t+h-1}, y_{t+h-2}, \ldots).$$

From the Wold representation we know that $\varepsilon_{t+h-1} = y_{t+h-1} - Proj(y_{t+h-1}|y_{t+h-2}, y_{t+h-3}, \ldots)$, which implies that $\{\varepsilon_{t+h-1}, y_{t+h-1}, y_{t+h-2}, y_{t+h-3}, \ldots\}$ are linearly dependent. This implies that y_{t+h-1} can be dropped from $Proj(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_t, y_{t+h-1}, y_{t+h-2}, \ldots)$ since it contains redundant information. Therefore,

$$Proj(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_t, y_{t+h-1}, y_{t+h-2}, \ldots) = Proj(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_t, y_{t+h-2}, y_{t+h-3}, \ldots).$$

Similarly, $\varepsilon_{t+h-2} = y_{t+h-2} - Proj(y_{t+h-2}|y_{t+h-3}, y_{t+h-4}, \ldots)$, which implies that $\{\varepsilon_{t+h-2}, y_{t+h-2}, y_{t+h-3}, y_{t+h-4}, \ldots\}$ are linearly dependent. This implies that y_{t+h-2} can be dropped from $Proj(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_t, y_{t+h-2}, y_{t+h-3}, \ldots)$ since it contains redundant information. Therefore,

$$Proj(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_t, y_{t+h-2}, y_{t+h-3}, \ldots) = Proj(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_t, y_{t+h-3}, y_{t+h-4}, \ldots).$$

This process is repeated until y_t is being dropped due to linear dependence yielding

$$Proj(y_{t+h}|\varepsilon_{t+h-1},\ldots,\varepsilon_t,y_t,y_{t-1},\ldots) = Proj(y_{t+h}|\varepsilon_{t+h-1},\ldots,\varepsilon_t,y_{t-1},y_{t-2},\ldots).$$

Therefore, if the data are stationary and the horizon 0 LP residuals are uncorrelated,

$$Proj(y_{t+h}|\varepsilon_{t+h-1},\ldots,\varepsilon_t,y_{t-1},y_{t-2},\ldots) = Proj(y_{t+h}|\varepsilon_{t+h-1},\ldots,\varepsilon_t,y_{t+h-1},y_{t+h-2},\ldots).$$

Since conditional independence is satisfied it follows that

$$[y_{t+h} - Proj(y_{t+h}|\varepsilon_{t+h-1},\ldots,\varepsilon_t,y_{t-1},y_{t-2},\ldots)] \perp [y_{t+h-i} - Proj(y_{t+h-i}|\varepsilon_{t+h-i-1},\ldots,\varepsilon_{t-i},y_{t-i-1},y_{t-i-2},\ldots)] \forall i \geq 1,$$

where $\perp$ is the orthogonal symbol. □

Therefore, if the data are stationary and the residuals $\{\varepsilon_t\}$ are uncorrelated, autocorrelation can be eliminated in the horizon h LP by including $\{\varepsilon_t,\varepsilon_{t+1},\ldots,\varepsilon_{t+h-1}\}$ in the conditioning set. Of course, if the true model requires only finitely many lags in the LP specification, then the proof above applies to that case as well, since the longer lags will all have coefficients of zero in population.

Theorem 1. ***The autocorrelation process of the horizon h LP residuals ($e_{t+h}^{(h)}$) is known.***

Proof. We know from the Wold representation that $\varepsilon_t \perp y_{t-1}, y_{t-2},\ldots$, hence $\varepsilon_t \perp \varepsilon_s \; for \; t \neq s$. Recall that the infinite lag horizon h LP is

$$y_{t+h} = B_1^{(h+1)} y_{t-1} + B_2^{(h+1)} y_{t-2} + \ldots + e_{t+h}^{(h)} = Proj(y_{t+h}|y_{t-1},y_{t-2},\ldots) + e_{t+h}^{(h)}. \tag{2}$$

By Proposition 1, including $\{\varepsilon_t,\varepsilon_{t+1},\ldots,\varepsilon_{t+h-1}\}$ in the conditioning set eliminates autocorrelation, so the horizon h LP can be rewritten as

$$y_{t+h} = Proj(y_{t+h}|\varepsilon_{t+h-1},\ldots,\varepsilon_t,y_{t-1},y_{t-2},\ldots) + u_{t+h}^{(h)}, \tag{3}$$

where $u_{t+h}^{(h)} = e_{t+h}^{(h)} - Proj(y_{t+h}|\varepsilon_{t+h-1},\ldots,\varepsilon_t) = e_{t+h}^{(h)} - Proj(y_{t+h}|\varepsilon_{t+h-1}) - \ldots - Proj(y_{t+h}|\varepsilon_t)$. The $Proj$ can be broken up additively because $\{\varepsilon_t,\ldots,\varepsilon_{t+h-1}\}$ are orthogonal to each other and to $\{y_{t-1},y_{t-2},\ldots\}$. By Proposition 1, $u_{t+h}^{(h)}$ is not autocorrelated. By the Wold representation we know that

$$Proj(y_{t+h}|\varepsilon_t) = \Theta_h \varepsilon_t. \tag{4}$$

This implies, the horizon h LP can be written as

$$y_{t+h} = B_1^{(h+1)} y_{t-1} + B_2^{(h+1)} y_{t-2} + \ldots + \Theta_h \varepsilon_t + \ldots + \Theta_1 \varepsilon_{t+h-1} + u_{t+h}^{(h)}, \tag{5}$$

which implies

$$e_{t+h}^{(h)} = \Theta_h \varepsilon_t + \ldots + \Theta_1 \varepsilon_{t+h-1} + u_{t+h}^{(h)}.$$

As a result, the autocorrelation process of $e_{t+h}^{(h)}$ is known. Using the same linear dependence arguments as in Proposition 1, it can be shown that

$$Proj(y_{t+h} | \varepsilon_{t+h-1}, \ldots, \varepsilon_t, y_{t-1}, y_{t-2}, \ldots) = Proj(y_{t+h} | y_{t+h-1}, y_{t+h-2}, \ldots),$$

which implies that

$$u_{t+h}^{(h)} = \varepsilon_{t+h},$$

in population. □

Thus in population, the error process is a VMA(h) even if the true model is not a VAR. In population

$$B_1^{(h)} = \Theta_h,$$

which implies

$$e_{t+h}^{(h)} = B_1^{(h)} \varepsilon_t + \ldots + B_1^{(1)} \varepsilon_{t+h-1} + \varepsilon_{t+h}.$$

2.3 LP GLS and Its Properties

Since $e_{t+h}^{(h)}$ can be written as

$$e_{t+h}^{(h)} = B_1^{(h)} \varepsilon_t + \ldots + B_1^{(1)} \varepsilon_{t+h-1} + u_{t+h}^{(h)}, \tag{6}$$

GLS can be used to eliminate autocorrelation in LP while avoiding increasing the number of parameters by including $\{\varepsilon_t, \varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$ in the horizon h conditioning set. To understand how, I'll first explain what happens when $\{\varepsilon_t, \varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$ is included in the conditioning set. Just like it is impossible to estimate a VAR(∞) in practice, one cannot estimate LP with infinite lags since there is insufficient data. In practice truncated LP are used where the lags are truncated at k. The proofs of consistency and asymptotic normality discuss the rate at which k needs to grow with the sample size to ensure consistent estimation of the impulse responses. In practice, k, needs to be large enough that the estimated residuals from the horizon 0 LP are

uncorrelated, which is what will be assumed for now. From Theorem 1 we know the horizon h LP is

$$y_{t+h} = B_1^{(h+1)} y_{t-1} + \ldots + B_k^{(h+1)} y_{t-k} + B_1^{(h)} \varepsilon_t + \ldots + B_1^{(1)} \varepsilon_{t+h-1} + u_{t+h}^{(h)}. \tag{7}$$

Due to $\{\varepsilon_t, \varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$ being unobserved, the estimates $\{\hat{\varepsilon}_t, \hat{\varepsilon}_{t+1}, \ldots, \hat{\varepsilon}_{t+h-1}\}$ from the horizon 0 LP must be used instead. Estimates of the impulse responses are still consistent (see appendix for proof), however, even if the sample size is large, inference on the parameters will underrepresent uncertainty because $\{\hat{\varepsilon}_t, \hat{\varepsilon}_{t+1}, \ldots, \hat{\varepsilon}_{t+h-1}\}$, are generated regressors (Pagan, 1984). In order to do valid inference, one must take into account that the generated regressors were estimated.[15]

For now, I will ignore the additional uncertainty from the generated regressors $\{\hat{\varepsilon}_t, \hat{\varepsilon}_{t+1}, \ldots, \hat{\varepsilon}_{t+h-1}\}$. Including $\{\hat{\varepsilon}_t, \hat{\varepsilon}_{t+1}, \ldots, \hat{\varepsilon}_{t+h-1}\}$ in the conditioning set increases the number of parameters in each equation in the system by $h \times r$. If consistent estimates of $\{\hat{B}_1^{(h)}, \hat{B}_1^{(h-1)}, \ldots, \hat{B}_1^{(1)}\}$ are obtained in previous horizons, one can do a Feasible GLS (FGLS) transformation. Let $\widetilde{y}_{t+h}^{(h)} = y_{t+h} - \hat{B}_1^{(h)} \hat{\varepsilon}_t - \ldots - \hat{B}_1^{(1)} \hat{\varepsilon}_{t+h-1}$. Then one can estimate horizon h via the following equation

$$\widetilde{y}_{t+h}^{(h)} = B_1^{(h+1)} y_{t-1} + \ldots + B_k^{(h+1)} y_{t-k} + \tilde{u}_{t+h}^{(h)}. \tag{8}$$

$\widetilde{y}_{t+h}^{(h)}$ is just a GLS transformation that eliminates the autocorrelation problem in LP without having to sacrifice degrees of freedom and $\tilde{u}_{t+h}^{(h)}$ is the error term corresponding to this GLS transformation. If the impulse responses are estimated consistently, then by the continuous mapping theorem, $\widetilde{y}_{t+h}^{(h)}$ converges in probability to the true GLS transformation $y_{t+h}^{(h)} = y_{t+h} - B_1^{(h)} \varepsilon_t - \ldots - B_1^{(1)} \varepsilon_{t+h-1}$ asymptotically. For clarification LP can be estimated sequentially horizon by horizon as follows. First estimate the horizon 0 LP

$$y_t = B_1^{(1)} y_{t-1} + \ldots + B_k^{(1)} y_{t-k} + u_t^{(0)},$$

and due to the horizon 0 LP being a VAR $\varepsilon_t = u_t^{(0)}$. $\hat{B}_1^{(1)}$ and $\hat{\varepsilon}_t$ are estimates of $B_1^{(1)}$ and ε_t respectively . Horizon 1 can be estimated as

$$\widetilde{y}_{t+1}^{(1)} = B_1^{(2)} y_{t-1} + \ldots + B_k^{(2)} y_{t-k} + \tilde{u}_{t+1}^{(1)},$$

where $\tilde{y}_{t+1}^{(1)} = y_{t+1} - \hat{B}_1^{(1)} \hat{\varepsilon}_t$, and $\hat{B}_1^{(2)}$ is the GLS estimate of $B_1^{(2)}$. Horizon 2 can be estimated as

$$\tilde{y}_{t+2}^{(2)} = B_1^{(3)} y_{t-1} + \ldots + B_k^{(3)} y_{t-k} + \tilde{u}_{t+2}^{(2)},$$

[15] In the proof of asymptotic normality of the limiting distribution, it can be seen that the impact of the generated regressors does not disappear asymptotically.

where $\widetilde{y}_{t+2}^{(2)} = y_{t+2} - \hat{B}_1^{(2)}\hat{\varepsilon}_t - \hat{B}_1^{(1)}\hat{\varepsilon}_{t+1}$, and $\hat{B}_1^{(3)}$ is the GLS estimate of $B_1^{(3)}$. Horizon 3 can be estimated as

$$\tilde{y}_{t+3}^{(3)} = B_1^{(4)} y_{t-1} + \ldots + B_k^{(4)} y_{t-k} + \tilde{u}_{t+3}^{(3)},$$

where $\widetilde{y}_{t+3}^{(3)} = y_{t+3} - \hat{B}_1^{(3)}\hat{\varepsilon}_t - \hat{B}_1^{(2)}\hat{\varepsilon}_{t+1} - \hat{B}_1^{(1)}\hat{\varepsilon}_{t+2}$, and so on.

The LP GLS estimator has the following three properties:

Theorem 2. *Under the assumptions stated in Lewis and Reinsel (1985), the LP GLS estimator is consistent. In particular*

$$\hat{B}_1^{(h)} \xrightarrow{p} \Theta_h.$$

Theorem 3. *Under the assumptions stated in Lewis and Reinsel (1985), the limiting distribution of the LP GLS estimates are asymptotically normal.*

Theorem 4. *Under the assumptions stated in Lewis and Reinsel (1985), the limiting distribution of the LP GLS estimates are asymptotically more efficient than the limiting distribution of the LP OLS estimates.*

Under the assumptions stated in Lewis and Reinsel (1985), Jordà and Kozicki (2011) show the consistency and asymptotically normality of $\{B_1^{(h+1)}, B_2^{(h+1)}, \ldots, B_k^{(h+1)}\}$ when estimated via OLS.

Remark. The assumptions are general enough to include most stationary invertible VARMA models. Jordà and Kozicki (2011) proof is an extension of Lewis and Reinsel (1985), who show consistency and asymptotic normality of the VAR(∞). The conditions in Lewis and Reinsel (1985) state the rate at which the lag length, k, needs to grow in order for the estimates to be consistent and asymptotically normal.

Proof. See appendix. The explicit assumptions and proofs are in section A.4 of the appendix. □

As noted earlier, the parameters used in the GLS correction are not known, and their uncertainty should be taken into account in order to do valid inference. To take into account the uncertainty in the generated regressors, frequentist can use bootstrapping, multi-step estimation (Murphy and Topel, 1985), or joint estimation (Newey and McFadden, 1994). Bayesian's can marginalize uncertainty via Monte Carlo integration. Estimation for both frequentist and Bayesian methods will be discussed in the next section.

3 LP GLS Estimation

3.1 Frequentist Estimation via Bootstrapping

For frequentist estimation, LP GLS can be implemented using a circular block bootstrap scheme (Politis and Romano, 1994). Bootstrap samples are first created using the circular block scheme, then for each bootstrap sample, FGLS estimates of the LP horizons are constructed. To illustrate, first one must decide on the number of bootstrap draws, J, the maximum number of of impulse response horizons to be estimated, H, and the number of consecutive blocks, L. There are no good rules of thumb for choosing L in general, so I follow Berkowitz et al. (1999) and set $L = T^{1/3}$. To construct the bootstrap data sets, the original data, $\{y_1, y_2, y_{t-2}, \ldots, y_T\}$, is wrapped around in a circle so that y_1 follows y_T. By construction, the horizon h LP depends on the $\{y_{t+h}, y_{t-1}, y_{t-2}, \ldots, y_{t-k}\}$ tuple. Since LP will be estimated using FGLS and the transformation must be done using the same data, one must first construct all possible $\{y_{t+H}, \ldots, y_{t-k}\}$ tuples. Then to preserve correlation in the data blocks of L consecutive tuples are drawn at random and concatenated to generate a bootstrap sample. Then for each bootstrap sample, the impulse responses $\{B_1^{(h+1)}, B_2^{(h+1)}, \ldots, B_k^{(h+1)}\}$ are estimated for each horizon using the FGLS estimation described in the previous section. This is done for each of the J bootstrap draws. To clarify,

Algorithm 1: Block Bootstrapping Without Bias Adjustment

1: for each bootstrap replication $j = 1, \ldots, J$
2: draw blocks of L consecutive $\{y_{t+H}, y_{t-1}, y_{t-2}, \ldots, y_{t-k}\}$ tuples to generate a bootstrap sample.
3: estimate $\{B_1^{(h+1)}, B_2^{(h+1)}, \ldots, B_k^{(h+1)}\}$ for each horizon via the FGLS procedure.
4: end.

Denoting $\{B_1^{(h+1),<j>}, B_2^{(h+1),<j>}, \ldots, B_k^{(h+1),<j>}\}$ as jth bootstrap replication for the impulse responses, 95% confidence intervals can then be constructed by taking the 2.5% and 97.5% quantiles of the parameter(s) of interest. The bootstrap can also be implemented with bias adjustment. The bias adjustment of the LP GLS bootstrap follows the general procedure of Efron and Tibshirani (1993), and bias adjustment is implemented for each horizon. The bias of any parameter, for example $B_1^{(h)}$, can be calculated via $bias = J^{-1}(\sum_{j=1}^{J} B_1^{(h),<j>}) - \hat{B}_1^{(h)}$. The bias adjusted bootstrap replications are then $B_1^{(h),<j>,BA} = B_1^{(h),<j>} - bias$. To summarize how to block bootstrap with bias adjustment,

Algorithm 2: Block Bootstrapping With Bias Adjustment

1: for each bootstrap replication $j = 1, \ldots, J$
2: draw blocks of L consecutive $\{y_{t+H}, \ldots, y_{t-k}\}$ tuples to generate a bootstrap sample.
3: end
4: for each LP horizon $h = 0, \ldots, H-1$
5: for each bootstrap replication $j = 1, \ldots, J$
6: estimate $\{B_1^{(h+1)}, B_2^{(h+1)}, \ldots, B_k^{(h+1)}\}$ via the FGLS procedure.
7: end
8: calculate the bias and bias adjust the bootstrap estimates as in Efron and Tibshirani (1993).
9: end

The Monte Carlo simulations analyzing the finite sample properties implements the bias adjustment version of the bootstrap.

3.2 Bayesian Estimation

3.2.1 The Likelihood

Despite LP growing in popularity, there has not been a fully Bayesian treatment, which is probably due to the belief that Newey-West standard errors must be used because the autocorrelation process is unknown. Since LP can be estimated using GLS, it is now possible to give a fully Bayesian treatment of LP. Due to LP being standard linear regressions, one just needs to be able to set up the likelihood and elicit a prior. The default prior used in this paper is the conjugate normal inverse-Wishart prior. Conjugate priors need not be used. LP are linear regressions, so any prior that can be used with a linear regression can be used with Bayesian LP.

LP at horizon 0 are equivalent to VARs. To estimate LPs at horizon 0 just estimate

$$y_t = B_1^{(1)} y_{t-1} + B_2^{(1)} y_{t-2} + \ldots + B_k^{(1)} y_{t-k} + u_t^{(0)},$$

as one would a standard Bayesian VAR. Define $\beta^{(0)} \equiv vec(\left[B_1^{(1)}, B_2^{(1)}, \ldots, B_k^{(1)}\right]')$, $X_t^{'(0)} \equiv I_n \otimes [y_{t-1}, y_{t-2}, \ldots, y_{t-k}]'$, then

$$y_t = X_t^{'(0)} \beta^{(0)} + u_t^{(0)},$$

where $u_t^{(0)} \sim N(0, \Sigma_u^{(0)})$. Assume a conditional normal inverse-Wishart prior for $p(\beta^{(0)}, \Sigma_u^{(0)})$. That is

$$p(\beta^{(0)} | \Sigma_u^{(0)}) \sim N(\underline{b}, \Sigma_u^{(0)} \otimes \underline{\Omega}),$$

$$p(\Sigma_u^{(0)}) \sim IW(\underline{n}, \underline{\Psi}),$$

where $\underline{b}$, $\underline{\Omega}$, $\underline{\Psi}$, and $\underline{n}$ are prior hyperparameters. The posterior is also conditional normal inverse-Wishart

$$p(\beta^{(0)}|\Sigma_u^{(0)}, y_{1:T}) \sim N(\overline{b}, \Sigma_u^{(0)} \otimes \overline{\Omega}),$$

$$p(\Sigma_u^{(0)}|y_{1:T}) \sim IW(\overline{n}, \overline{\Psi}),$$

where $\overline{b}$, $\overline{\Omega}$, $\overline{\Psi}$, and $\overline{n}$ are posterior hyperparameters whose formulas are well known and can be found in the appendix. After estimating horizon 0, one can obtain J posterior draws of residuals $\{\varepsilon_{k+1}, \ldots, \varepsilon_T\}$ using the fact that $\varepsilon_t^{<j>} = y_t - X_t^{'(0)}\beta^{(0),<j>}$, where $\beta^{(0),<j>}$ is the jth posterior draw of $\beta^{(0)}$. Now posterior draws of $y_{t+1}^{(1)}$ can be constructed via $\tilde{y}_{t+1}^{(1),<j>} = y_{t+1} - B_1^{(1),<j>}\varepsilon_t^{<j>}$. To understand why posterior draws for $y_{t+1}^{(1)}$ are needed, note that in GLS, one uses parameter estimates in the transformation and treat the transformation as known. The transformation, however, does not take into account uncertainty in the parameters used for the transformation, so to properly take into account uncertainty, one must marginalize out uncertainty in the transformation.

For each J, define $y_{t+1}^{(1)} \equiv \tilde{y}_{t+1}^{(1),<j>}$ and $B_1^{(1)} \equiv B_1^{(1),<j>}$, which means for each J we treat the GLS transformation as known. The horizon 1 LP is

$$y_{t+1}^{(1)} = B_1^{(2)} y_{t-1} + B_2^{(2)} y_{t-2} + \ldots + B_k^{(2)} y_{t-k} + u_{t+1}^{(1)},$$

where $u_{t+1}^{(1)} \sim N(0, \Sigma_u^{(1)})$. Define $\beta^{(1)} \equiv vec(\left[B_1^{(2)}, B_2^{(2)}, \ldots, B_k^{(2)}\right]')$ and $X_t^{'(1)} \equiv I_n \otimes [y_{t-1}, y_{t-2}, \ldots, y_{t-k}]'$. Then the horizon 1 LP can be rewritten as

$$y_{t+1}^{(1)} = X_t^{'(1)}\beta^{(1)} + u_{t+1}^{(1)}.$$

Again, assume a conditional normal inverse Wishart prior for $p(\beta^{(1)}, \Sigma_u^{(1)})$

$$p(\beta^{(1)}|\Sigma_u^{(1)}) \sim N(\underline{b}^{(1)}, \Sigma_u^{(1)} \otimes \underline{\Omega}^{(1)}),$$

$$p(\Sigma_u^{(1)}) \sim IW(\underline{n}^{(1)}, \underline{\Psi}^{(1)}).$$

The posterior is conditional normal inverse-Wishart. That is

$$p(\beta^{(1)}|\Sigma_u^{(1)}, y_{1:T}) \sim N(\overline{b}^{(1)}, \Sigma_u^{(1)} \otimes \overline{\Omega}^{(1)}),$$

$$p(\Sigma_u^{(1)}|y_{1:T}) \sim IW(\overline{n}^{(1)}, \overline{\Psi}^{(1)}).$$

One Monte Carlo draw is obtained from the conditional posterior for each J, which marginalizes out uncertainty in the GLS transformation.

This is done at each horizon in the LP. Before estimation of horizon h, one can obtain posterior draws of $y_{t+h}^{(h)}$ via $\tilde{y}_{t+h}^{(h),<j>} = y_{t+h} - B_1^{(h),<j>}\varepsilon_t^{<j>} - \ldots - B_1^{(1),<j>}\varepsilon_{t+h-1}^{<j>}$. For each J, define $y_{t+h}^{(h)} \equiv \tilde{y}_{t+h}^{(h),<j>}$ and $B_1^{(1)} \equiv B_1^{(1),<j>}, \ldots, B_1^{(h)} \equiv B_1^{(h),<j>}$. The horizon h LP is

$$y_{t+h}^{(h)} = B_1^{(h+1)} y_{t-1} + B_2^{(h+1)} y_{t-2} + \ldots + B_k^{(h+1)} y_{t-k} + u_{t+h}^{(h)},$$

where $u_{t+h}^{(h)} \sim N(0, \Sigma_u^{(h)})$. Define $\beta^{(h)} \equiv vec(\left[B_1^{(h+1)}, B_2^{(h+1)}, \ldots, B_k^{(h+1)}\right]')$, $X_t^{'(h)} \equiv I_n \otimes [y_{t-1}, y_{t-2}, \ldots, y_{t-k}]'$. The horizon h LP can be rewritten as

$$y_{t+h}^{(h)} = X_t^{'(h)} \beta^{(h)} + u_{t+h}^{(h)}.$$

Again, assume a conditional normal inverse gamma prior for $p(\beta^{(h)}, \Sigma_u^{(h)})$

$$p(\beta^{(h)}|\Sigma_u^{(h)}) \sim N(\underline{b}^{(h)}, \Sigma_u^{(h)} \otimes \underline{\Omega}^{(h)}),$$

$$p(\Sigma_u^{(h)}) \sim IW(\underline{n}^{(h)}, \underline{\Psi}^{(h)}).$$

The posterior is conditional normal inverse-Wishart

$$p(\beta^{(h)}|\Sigma_u^{(h)}, y_{1:T}) \sim N(\overline{b}^{(h)}, \Sigma_u^{(h)} \otimes \overline{\Omega}^{(h)}),$$

$$p(\Sigma_u^{(h)}|y_{1:T}) \sim IW(\overline{n}^{(h)}, \overline{\Psi}^{(h)}).$$

One Monte Carlo draw is obtained from the conditional posteriors for each J, which marginalized out uncertainty in the GLS transformation. To summarize,

Algorithm 3: Bayesian LP

1: Estimate the Bayesian VAR/horizon 0 LP.
2: Generate J posterior draws for $\{B_1^{(1)}, B_2^{(1)}, \ldots, B_k^{(1)}\}$
3: for each LP horizon $h = 1, \ldots, H-1$
4: for each posterior draw $j = 1, \ldots, J$
5: estimate $\{B_1^{(h+1)}, B_2^{(h+1)}, \ldots, B_k^{(h+1)}\}$ via the Bayesian version of the FGLS procedure.
6: end
7: end

3.2.2 Priors

Bayesian LP allow the researcher to incorporate prior information for impulse responses at each horizon. Incorporating prior information has multiple advantages. Prior information can be used to shrink impulse responses at any horizon to prevent overfitting, which is often desirable in forecasting or when the number of parameters is large (Giannone et al., 2015). Economic theory can be incorporated into the prior to inform the shape of the impulse responses (e.g. the impulse response is monotonic or hump shaped) and to discipline the long-run behavior (Giannone et al., 2018). Prior information from economic theory can also be used to smooth impulse responses across horizons, which may be desirable in certain contexts (Barnichon and Brownlees, 2018, Stock and Watson, 2018).

The default prior used in this paper is a conjugate training sample prior. When using a training sample prior in Bayesian LP, the researcher must decide how many horizons they are going to estimated before they choose the size of the training sample. To understand why, assume that the training sample is of size $\underline{T}$. The same training sample must be used for each horizon, so the training sample must be large enough to estimate a training sample prior at each horizon. Recall that when estimating horizon h, h observations will be lost from the original sample, so the training sample for horizon h has $\underline{T} - h$ observations.[16]

As shown in the previous section, the horizon 0 LP

$$y_t = B_1^{(1)} y_{t-1} + B_2^{(1)} y_{t-2} + \ldots + B_k^{(1)} y_{t-k} + u_t^{(0)}.$$

can be recast as

$$y_t = X_t^{'(0)} \beta^{(0)} + u_t^{(0)},$$

where $\beta^{(0)} \equiv vec(\left[B_1^{(1)}, B_2^{(1)}, \ldots, B_k^{(1)}\right]')$, $X_t^{'(0)} \equiv I_n \otimes [y_{t-1}, y_{t-2}, \ldots, y_{t-k}]'$, and $u_t^{(0)} \sim N(0, \Sigma_u^{(0)})$. The conjugate training sample prior for $p(\beta^{(0)}, \Sigma_u^{(0)})$ is

$$p(\beta^{(0)} | \Sigma_u^{(0)}) \sim N(\underline{b}, \Sigma_u^{(0)} \otimes \underline{\Omega}),$$

$$p(\Sigma_u^{(0)}) \sim IW(\underline{n}, \underline{\Psi}).$$

$\underline{n}$ is the prior degrees of freedom, $\underline{b} = \hat{\beta}_{OLS}$ and $\underline{\Psi} = \underline{n}\hat{\Sigma}_{OLS}$, where $\hat{\beta}_{OLS}$ and $\hat{\Sigma}_{OLS}$ are the OLS results from the training sample. $\underline{\Omega} = \frac{T}{\underline{n}}(\underline{X}'\underline{X})^{-1}$ where $\underline{X}$ is the design matrix for the training sample and $\frac{T}{\underline{n}}$ rescales the conditional variance of $\beta^{(0)}$ so the conditional distribution will have the asymptotic variance of the OLS results based on the average of $\underline{n}$ observations.[17] $\underline{n}$, which determines the informativeness of the prior, can be chosen by the researcher or a prior can be placed on $\underline{n}$ and estimated using Griddy Gibbs or sampling

[16] This does not account for the k presample observations that will be treating as deterministic in the VAR(k).

[17] This is in the spirit of the unit information prior (Kass and Wasserman, 1995), but since this is done over a training sample, it does not make double use of the data.

importance resampling. In order for the prior mean of $\Sigma_u^{(0)}$ to be defined, $\underline{n} \geq p+2$. By default, I set $\underline{n} = p+2$ to make the prior weakly informative but still proper. The diagonal of $\underline{\Omega}$ can be taken to prevent collinearity issues if the prior is only based on small training sample (Brodersen et al., 2015). When estimating the training sample prior for horizons 1 and greater, autocorrelation is corrected for in the training sample estimates using the GLS procedure discussed in Section 2.2.[18]

Even though the conjugate normal inverse-Wishart training sample prior is the only prior presented, many priors can be used with Bayesian LP. The priors need not be conjugate. LP are linear regressions, so any prior that can be used with a linear regression can be used with Bayesian LP. Again, Bayesian LP allow the researcher to incorporate prior information for impulse responses at each horizon. Prior information can be used to shrink impulse responses at any horizon to prevent overfitting. Economic theory can be incorporated into the prior to inform the shape of the impulse responses and to discipline the long-run behavior, which would help smooth impulse responses across horizons and alleviate the sometimes erratic impulse responses estimated from frequentist LP.

3.3 Structural Identification

This subsection briefly discusses structural identification in LP GLS. These techniques can be applied to both the bootstrapped LP and Bayesian LP. For an extensive review of structural identification in VARs and LP see Ramey (2016), and for an extensive treatment of identification in VARs and LP using external instruments see Stock and Watson (2018). Structural identification in Bayesian LP is essentially the same as identification with frequentist LP. Going back to the horizon 0 LP

$$y_t = B_1^{(1)} y_{t-1} + B_2^{(1)} y_{t-2} + \ldots + B_k^{(1)} y_{t-k} + u_t^{(0)},$$

let $u_t^{(0)} = Rs_t$ where s_t is a vector of structural shocks and R is an invertible matrix. If R is known, after estimating $\{B_1^{(1)}, B_1^{(2)}, \ldots, B_1^{(h+1)}\}$, one can construct the structural impulse responses, $\{G^{(1)}, G^{(2)}, \ldots, G^{(h+1)}\}$, via Monte Carlo integration where $G^{(h)} = B_1^{(h)} R$. Typically R is not known, but can be estimated, so Monte Carlo integration can still be applied. An example of R being estimated would be a triangular (recursive) ordering.[19] One would estimated horizon 0 LP, and then apply a recursive ordering to posterior (or bootstrap) draws of $\Sigma_u^{(0)}$ to obtain draws of R, and then draws of $G^{(h)}$ can be constructed via $G^{(h)} = B_1^{(h)} R$.

It is often the case that the researcher may not know all of the identifying restrictions in R or may believe that R is not invertible, but the researcher has an instrument that they believe can trace out impulse

[18] Uncertainty is not marginalized out in the GLS transformation, $\widetilde{y}_{t+h}^{(h)}$, for the training sample.

[19] In the literature a triangular (recursive) ordering is often called a cholesky ordering because people often apply a cholesky decomposition to impose the ordering. It should be noted that the cholesky normalizes the variances of the structural shocks to unity. If one does not want to normalize the structural shocks one can instead use the LDL decomposition to impose recursive the ordering.

responses of interest. The impulse responses of interest can instead be estimated by LP instrumental variable regressions (LP-IV). Stock and Watson (2018) show that in order for LP-IV to be valid, 3 conditions need to be satisfied. Decompose s_t into $s_{1,t}$ and $s_{2,t}$ where $s_{1,t}$ is the structural shock of interest at time t and $s_{2,t}$ represents all other structural shocks at time t. Let z_t be an instrument that the researcher believes can trace out the impulse responses of $s_{1,t}$. The instrument must satisfy the following three conditions

$$E[s_{1,t}z_t] \neq 0,$$

$$E[s_{2,t}z_t] = 0,$$

$$E[s_{t+j}z_t] = 0 \; for \; j \neq 0.$$

The first two conditions are just the standard relevance and exogeneity conditions for instrumental variable regression. The third condition is a lead-lag exogeneity condition, which guarantees that the instrument, z_t, is only identifying the impulse response of the shock $s_{1,t}$. If the third condition is not satisfied, then z_t will amalgamate the impulse responses at different horizons. It may be the case that these conditions are only satisfied after conditioning on suitable control variables (e.g. the lags of a VAR/horizon 0 LP).

Frequentist typically estimate LP-IV via two-stage least squares (2SLS). For example, say I want to estimate the impulse response, $g^{(h)}$, the impact a shock to monetary policy has on output at horizon h. Let output be denoted as $output_t$ and the monetary policy variable mp_t. The frequentists approach is to estimate LP-IV by running

$$output_{t+h} = g^{(h)} mp_t + control\, variables + error_{t+h}^{(h)} \tag{9}$$

via 2SLS and using z_t as an instrument for mp_t. Newey and West (1987) standard errors are used to account for autocorrelation, but as shown section 2, this ignores the increasing variance problem. The increasing variance problem is particularly problematic with LP-IV because the increasing variance can weaken the strength of instrument for $h \geq 1$.[20] Alternatively, the impulse responses of shocks to $s_{1,t}$ can be recovered if z_t is included as an endogenous variable in the system and ordered first (Paul, ming, Plagborg-Møller and Wolf, 2019). Let $\mathring{y}_t = \begin{bmatrix} z_t \\ y_t \end{bmatrix}$ where y_t contains mp_t, $output_t$, and the control variables at time t, then the horizon 0 LP/VAR is

$$\mathring{y}_t = \mathring{B}_1^{(1)} \mathring{y}_{t-1} + \mathring{B}_2^{(1)} \mathring{y}_{t-2} + \ldots + \mathring{B}_k^{(1)} \mathring{y}_{t-k} + \mathring{u}_t^{(0)}.$$

Since z_t is ordered first due to its exogeneity, the residual for the z_t equation, $\mathring{u}_{1,t}^{(0)}$, will be able to trace out the structural impulse responses of interest.[21] Going back to the monetary policy example, the impulse

[20] Whether the strength of the instrument is weakened depends in part on the type of impulse response being estimated. For example if one is estimated a cumulative multiplier directly like in Ramey and Zubairy (2018), the autocorrelation would weaken the strength of the instrument since the first stage of the 2SLS procedure has an increasing variance problem.

[21] Even if the control variables are exogenous to the system, any VARX can be written as a VAR with the exogenous variables ordered

response $g^{(h)}$ can be constructed as the ratio of the impulse response of $output_{t+h}$ to $\mathring{u}_{1,t}^{(0)}$ divided the impulse response of mp_t to $\mathring{u}_{1,t}^{(0)}$. Hence by imbedding z_t as an endogenous variable in the system and ordering it first, one can just estimate equation (2) via their preferred LP GLS method and construct the impulse responses of interest.

4 LP GLS and Relative Efficiency

To give a sense of potential efficiency gains of estimating LP via GLS, I will compare the asymptotic relative efficiency of the LP GLS estimator and the standard LP estimator when the true model is an AR(1). The asymptotic results apply for frequentists and Bayesians estimation due to the Bernstein-Von Mises theorem. Take the simple AR(1) model

$$y_t = ay_{t-1} + \varepsilon_t,$$

where $|a| < 1$ and $a \neq 0$ and ε_t is a white noise error process with $E(\varepsilon_t) = 0$ and $var(\varepsilon_t) = \sigma^2$. This implies that $E(y_t) = 0$ and the $var(y_t) = E(y_t' y_t) = \frac{\sigma^2}{(1-a^2)}$. Define $\{b^{(1)}, b^{(2)}, \ldots, b^{(h+1)}\}$ as the LP impulse responses for the AR(1) model. The limiting distribution of the LP GLS impulse response at horizon h is

$$\sqrt{T}(\hat{b}^{(h)} - a^h) \xrightarrow{d} N(0, [1 + (h^2 - 1)a^{2h-2}](1 - a^2)),$$

(follows from Theorem 4). The limiting distribution of the LP impulse response estimated by OLS with Newey-West standard errors at horizon h is

$$\sqrt{T}(\hat{b}^{(h)} - a^h) \xrightarrow{d} N(0, (1 - a^2)^{-1}[1 + a^2 - \{2h + 1\}a^{2h} + \{2h - 1\}a^{2h+2}]),$$

(Bhansali, 1997). The relative efficiency between the LP GLS and LP impulse responses,

$$\frac{[1 + (h^2 - 1)a^{2h-2}](1 - a^2)^2}{[1 + a^2 - \{2h + 1\}a^{2h} + \{2h - 1\}a^{2h+2}]},$$

determines which specification is more efficient. Note that the relative efficiency not only depends on the persistence, a, but on the horizon as well.

first in a block recursive scheme, therefore estimates from this setup are consistent.

Table 3: Relative Efficiency of LP (GLS) to LP (NW)						
Autocorrelation	Horizons					
Coefficient	3	5	10	20	30	40
$a = .99$	.993	.979	.945	.88	.818	.759
$a = .975$	.983	.948	.864	.713	.580	.464
$a = .95$	.966	.896	.735	.475	.288	.165
$a = .9$	.931	.792	.508	.179	.061	.029
$a = .75$	.827	.53	.195	.123	.123	.123
$a = .5$	.727	.496	.45	.45	.45	.45
$a = .25$	.854	.828	.827	.827	.827	.827
$a = .1$	.971	.97	.97	.97	.97	.97
$a = .01$	1	1	1	1	1	1

The gains from LP GLS can be large but they are not necessarily monotonic. This is because if the persistence is not that high, the impulse responses decay to zero quickly making the variance of the impulse responses small, and the gains from correcting for correcting for autocorrelation are not as large.

The efficiency gains of estimating LP via GLS, do not stop there. It turns out that when the true model is a AR(1) and the system is persistent enough, LP estimated with GLS can be approximately as efficient as the AR(1). Let $\hat{a}$ be the OLS estimate, the OLS estimate of a has the limiting distribution

$$\sqrt{T}(\hat{a} - a) \xrightarrow{d} N(0, 1 - a^2).$$

By the delta method, the horizon h impulse response has the limiting distribution

$$\sqrt{T}(\hat{a}^h - a^h) \xrightarrow{d} N(0, h^2 a^{2h-2}(1 - a^2)).$$

The asymptotic relative efficiency between the AR and LP GLS impulse responses

$$\frac{h^2 a^{2h-2}}{h^2 a^{2h-2} + (1 - a^{2h-2})},$$

determines which specification is more efficient. Since the true model is an AR(1), if the errors are normal, the AR(1) model will be asymptotically more efficient due to the Cramer-Rao lower bound (Bhansali, 1997). Below is a table of the relative efficiency between the AR and LP impulse responses for different values of a.

Table 4: Asymptotic Relative Efficiency of AR to LP (GLS)					
Horizons	5	10	20	30	40
$a = .99$	.997	.998	.999	.999	.999
$a = .975$	.991	.994	.996	.996	.996
$a = .95$	.980	.985	.985	.980	.968
$a = .9$	.95	.946	.881	.667	.302
$a = .75$	.736	.362	.007	0	0
$a = .5$	0	0	0	0	0

If the data is persistent enough, the LP impulse responses have approximately the same variance for horizons relevant in macro. For example, the economics profession has still not determined if GDP has a unit root or not. Assume that GDP is stationary but highly persistent with an AR(1) coefficient of .99. In this case, the AR(1) impulse responses has approximately the same variance for at least the first 40 horizons. Müller (2014) estimates the AR(1) coefficient for unemployment to be approximately .973. This would lead to the AR(1) impulse responses having approximately the same variance for at least the first 40 horizons. Other important macroeconomic variables such as inflation and the 3 month interest rate and most macro aggregates are also highly persistent and would display similar results. It is not until the AR(1) coefficient is .9 that you can see a notable difference over the first 40 horizons, and even then it is not until about 20 or so horizons out.

When the true model is a multivariate VAR things become more complicated. Efficiency still depends on the horizon and persistence, but because persistence can vary across the equations in the system, then for any horizon, LP could be approximately as efficient for some impulse responses and much less efficient for others. To see why, let us return to the VAR(1) model

$$y_t = A_1 y_{t-1} + \varepsilon_t.$$

Take the eigenvalue decomposition of $A_1 = E\Lambda_1 E'$, where $\Lambda_1 = diag(\lambda_1, \ldots, \lambda_k)$ is the diagonal matrix of distinct nonzero eigenvalues and $E = [e_1, \ldots, e_k]$ is the corresponding eigenmatrix and $EE^{-1} = I$ where I is the identity matrix. As a result $A_1^h = E\Lambda_1^h E' = \sum \lambda_1^h e_1 e_1'$. Define $w_t = E^{-1} y_t$ and $\eta_t = E^{-1}\varepsilon_t$. For simplicity assume E is known. This implies the VAR can be transformed into

$$w_t = \Lambda_1 w_{t-1} + \eta_t,$$

which will be called the transformed model. Consequently

$$w_{t+h} = \Lambda_1^{h+1} w_{t-1} + \Lambda_1^h \eta_t + \ldots + \Lambda_1 \eta_{t+h-1} + \eta_{t+h}.$$

Since Λ is diagonal, each equation in the transformed VAR(1) is an AR(1) model. Therefore the results derived earlier in this subsection for the AR(1) model apply.

More generally, it should be noted that, the efficiency gains of impulse responses estimated via LP GLS impulse for a particular horizon depends on the relative efficiency of the eigenvalues, and how much an eigenvalue contributes to the variance of an impulse response. So if A_1 contains different eigenvalues, the eigenmatrices and the correlation among eigenvalues would determine how much the variance of an eigenvalue contributes to the variance of an impulse responses in the untransformed model and hence determine the relative efficiency of LP GLS impulse response to the VAR impulse responses. Essentially, the efficiency gains of the VAR come from the less persistent components. Depending on how many persistent eigenvalues there are and how much they contribute to the variance of the impulse responses, it is possible for LP GLS to be approximately as efficient as the VAR, when the true model is a VAR. Whether impulse responses of LP would be approximately as efficient would depend on the true data generating process, the persistence of the system, the dependence structure of the variables, and the horizon. In other words, it would be specific to the situation. It follows that the efficiency of the VAR relative to the LP has been overstated in the literature.

5 Monte Carlo Evidence

In this section I present Monte Carlo evidence of the finite sample properties of the bias adjusted LP GLS bootstrap. The properties of LP GLS estimator will analyzed along with the properties of the LP estimated via OLS with Newey-West standard errors (which will be referred to as LP NW), and the bias adjusted VAR bootstrap (Kilian, 1998). Bayesian LP will not be included in the Monte Carlo exercise because Bayesian methods do not carry the same interpretation as frequentist confidence intervals and their coverage cannot be assessed the same way (Rubin, 1984, Hoff, 2009).

The Monte Carlo simulations will deal with AR(1) models since it is easy to isolate the persistence and, as shown in the previous subsection, the results will be informative toward VARs generally. The population model is

$$y_t = ay_{t-1} + \varepsilon_t,$$

where $a \in \{.99, .975, .95, .9, .75, .5\}$ and $\varepsilon_t \sim N(0,1)$ and the sample size $T \in \{250\}$. The different values of a represent a range of eigenvalues encountered in macro. The sample size of 250 is representative of a quarterly data set dating back to 1960. Even though the most prominent macro variables such as GDP, inflation, and unemployment date back to at least 1948, many do not date back that far. The comprehensive McCracken and Ng (2016) data set goes back to 1959, so a sample size of 250 seems reasonable.

Simulations are conducted 1,000 times for each combination of a and T. For each simulation, I estimated

the model for each desired horizon using all three methods and then check if the 95% confidence intervals contain the true impulse response. I then calculate the probability that the 95% confidence interval contains the true impulse response over the Monte Carlo simulations which gives me the coverage of the different methods. For each simulation draw, I also save the length of the 95% interval for the the different methods for each horizon. The lengths are then averaged over each Monte Carlo simulation for each method and horizon to get the respective average length of the 95% confidence intervals for each method and horizon. For the bias adjusted LP GLS and VAR bootstraps, I generate 1,000 bootstrap draws. I set the lag truncation parameter for the Newey-West standard errors to be $h-1$, when estimating the horizon h LP. Note that for correctly specified VARs, this is the true lag truncation parameter. 15 horizons are analyzed, which would be representative of analyzing four years of impulse responses for quarterly data.

Figures 1 and 2 displays the coverage and average length respectively. In general the bias-adjusted LP bootstrap has good finite sample properties. It tends to have coverage at or near the nominal level, but coverage can decline somewhat at longer horizons, dropping as low as 87%, if the autocorrelation coefficient is very persistent. This decline can be remedied to a degree by increasing the number of bootstraps. It is also the case that for the most persistent autocorrelation coefficients, coverage tends to be closer to 90% than 95%. The bias adjusted VAR bootstrap has coverage at or near the nominal level at all horizons. Consistent with the theoretical prediction in the previous section, the relative efficiency of the LP relative to the VAR depends on the persistence, with high persistence levels tending to have similar average lengths, which is consistent with asymptotic relative efficiency results in the previous section.

The LP NW estimator's performance can be much worse than the other two estimators. Coverage can drop drastically at higher horizons, below 60%, and if the data is persistent enough, coverage can be quite below the nominal level even at shorter horizons. The lack of coverage is due not only to the small sample bias but to Newey-West standard errors underestimating uncertainty. Note that the LP NW estimator tends to have shorter length than the bias adjusted LP GLS and VAR bootstraps. To "test" if this is due to Newey-West standard errors underestimating uncertainty, I estimate the "true" Monte Carlo variance of the different methods. That is, for each simulation I generated the AR(1) model and estimated the point estimates for each horizon using LP GLS, LP via OLS, and VAR via OLS. There is no bias correction here because the point is to show that even without bias correction, using GLS would lead to efficiency gains. To construct the VAR impulse responses, the OLS estimate was just raised to it's respective power. I saved the point estimates for each horizon for each simulation, and then I calculated the 95% quintiles (95% Monte Carlo confidence interval) for across the saved simulation estimates. This give me an approximation of the "true" 95% confidence intervals for these methods for the specific model and sample size. Figure 3 displays the average length for the "true" 95% confidence intervals. The efficiency gains of using GLS are pretty clear and because the LP GLS and the VAR OLS do not use bias correction in this Monte Carlo, it follows that LP NW

standard errors is underestimating uncertainty, and the lack of coverage is not solely due to the small sample bias. Newey-West standard errors underestimating uncertainty is a common problem when the process is persistent (Müller, 2014). It is also important to note that Newey-West standard errors are underestimating uncertainty, even though the true lag truncation parameter is being used.

Even though the bias adjusted LP bootstrap displays good finite sample properties, the Monte Carlo analysis of the "true" confidence intervals also indicate there is some efficiency loss from using the block bootstrap. That is, the block bootstrap is not as efficient as it could be. The efficiency loss is probably due to the chosen block length, particularly the choice of consecutive blocks $L = T^{1/3}$. Additional Monte Carlos also show that changing the block length or implementing the stationary bootstrap can also improve the slight decline in coverage for the highly persistent processes, but there can also be a loss in efficiency. Since the block length involves a bias variance tradeoff with longer block lengths yielding less biased test statistics with larger variances and shorter block lengths yielding the opposite, a rule or a cross validation method such as Hall et al. (1995) needs to be developed.

6 Issues of Nonstationarity

It is also worth reiterating that the GLS procedure presented in section 2 and the consistency and asymptotic normality of the procedure assumes stationarity. Nonstationarity can be caused by unit roots or structural breaks. In the case of unit roots, inference from the frequentist perspective could differ depending on which variables have unit roots and what the parameters of interest are (Sims et al., 1990, Jordà, 2009). Consistency of the results can still hold if the errors have enough moments (Sims et al., 1990, Jordà, 2009), so the procedure can still eliminate autocorrelation, but asymptotic normality of the results could break down, so inference based on the frequentist procedures presented could be invalid if unit roots are present. If unit roots are an issue and the order of integration is known, the data could just be difference to stationarity. However the order of integration is probably not known. One could test for unit roots, but frequentist tests unit roots lack power and can create considerable coverage distortions depending on the conclusion of the test (Pesavento and Rossi, 2006). In the case of Bayesian LP, Bayesian methods do not need to do anything special to take into account "explosive" nonstationarity behavior (e.g. unit roots) (Sims et al., 1990, Del Negro and Schorfheide, 2011), so estimation and inference involving Bayesian LP could proceed as usual.[22]

When nonstationarity is caused by structural breaks, both the frequentist and Bayesian methods presented will break down if they do not properly take into account change(s) in the parameters. Stationarity guarantees that the model has a linear time-invariant VMA representation. If the data are not stationary and structural breaks are the cause, then the procedure may not eliminate autocorrelation. To understand why

[22]There is a lively debate about how to construct priors for Vector Error Correction models (Del Negro and Schorfheide, 2011).

it matters if structural breaks are present, note that if the data are not stationary, it is possible for the estimated horizon 0 LP residuals to be uncorrelated since the VAR can still produce reasonable one-step ahead forecasts when the model is misspecified (Jordà, 2005). A "Wold representation" exists for nonstationary data, but the impulse responses for this VMA representation are allowed to be time dependent (Granger and Newbold, 1977, Priestley, 1988).[23] Assuming there is no deterministic component, any time series process can be written as

$$y_t = \varepsilon_t + \sum_{i=1}^{\infty} \Theta_{i,t}\varepsilon_{t-i},$$

where $\Theta_{i,t}$ is now indexed by the horizon and time period and $var(\varepsilon_t) = \Sigma_{\varepsilon,t}$. Using recursive substitution, the time dependent Wold representation can be written as a time dependent VAR or a time dependent LP.[24] It can be shown that a time dependent version of Theorem 1 exists. The horizon h time dependent LP is

$$y_{t+h} = B_{1,t}^{(h+1)} y_{t-1} + B_{2,t}^{(h+1)} y_{t-2} + \ldots + e_{t+h}^{(h)}, \tag{10}$$

where

$$e_{t+h}^{(h)} = \Theta_{h,t}\varepsilon_t + \ldots + \Theta_{1,t}\varepsilon_{t+h-1} + \varepsilon_{t+h}$$

$$B_{1,t}^{(h)} = \Theta_{h,t}.$$

If impulse responses are time dependent at higher horizons, but a time invariant version of LP GLS is applied, autocorrelation may not be eliminated at these horizons because the time-invariant LP are misspecified. In other words, if the data are nonstationary and the nonstationarity is caused by structural breaks, the time invariant version of LP GLS may not eliminate autocorrelation in the residuals since the estimates of the impulse responses may not be consistent. In this sense, LP GLS is a type of general misspecification test, because if one had estimated LP using OLS and Newey-West standard errors, the autocorrelation in the residuals would not hint toward potential misspecification since the residuals are inherently autocorrelated.

As noted in Granger and Newbold (1977), macro data encountered in practice are unlikely to be stationary, implying that the Wold representation may be time dependent. If the impulse responses of the Wold representation are time dependent, since time-varying parameter models can approximate any form of nonlinearity (Granger, 2008), a time varying version of LP GLS may be applied. The time-varying parameter version of the above GLS procedure presented in section 2 will be able to eliminate autocorrelation as long as the parameter changes are not so violent that a time-varying parameter model cannot track them. All else equal, the more adaptive the time-varying parameter model

[23] Nonstationarity in economics typically refers to explosive behavior (e.g. unit roots), but nonstationarity is more general and refers to a distribution that does not have a constant mean and/or variance over time (e.g. threshold models or models with stochastic volatility). Depending on the true model, differencing may not make the data stationarity (Leybourne et al., 1996, Priestley, 1988).

[24] The lag lengths can be infinite. Obviously in practice, a finite lag length would be chosen.

will be able to track changes and the better the approximation.[25] Time-varying parameter LP are presented in the next section. If the nature of the time dependence is known, that is, the researcher knows when the structural breaks occur or the nature of the time variation (i.e. regime switching models for expansions and recessions), then that specific time dependent model can be applied to the LP GLS procedure.

7 Time-Varying Parameter LP

As noted in the introduction, a researcher may be interested in allowing for time-varying parameters. Stock and Watson (1996) and Ang and Bekaert (2002) show many macroeconomic and financial time series exhibit parameter instability. It is also commonplace for regressions with macroeconomic time series to display heteroskedasticity of unknown form (Stock and Watson, 2007), and in order to do valid inference, the heteroskedasticity must be taken into account. Parameter instability can occur for many reasons such as policy changes, technological evolution, changing economic conditions, etc. If parameter instability is not appropriately taken into account, it can lead to invalid inference, poor out of sample forecasting, and incorrect policy evaluation. Moreover, time-varying parameter models can approximate any non-linear model (non-linear in the variables and/or the parameters), which makes them more robust to model misspecification (Granger, 2008).

As mentioned in the previous section, for any time series process, there exists a time dependent Wold representation

$$y_t = \varepsilon_t + \sum_{i=1}^{\infty} \Theta_{i,t}\varepsilon_{t-i},$$

where $\Theta_{i,t}$ is now indexed by the horizon and time period and $var(\varepsilon_t) = \Sigma_{\varepsilon,t}$. Using recursive substitution, the time dependent Wold representation can be written as a time dependent VAR or a time dependent LP. It can be shown that a time dependent version of Theorem 1 exists. The horizon h time dependent LP is

$$y_{t+h} = B_{1,t}^{(h+1)} y_{t-1} + B_{2,t}^{(h+1)} y_{t-2} + \ldots + B_{k,t}^{(h+1)} y_{t-k} + e_{t+h}^{(h)}, \tag{11}$$

where

$$e_{t+h}^{(h)} = \Theta_{h,t}\varepsilon_t + \ldots + \Theta_{1,t}\varepsilon_{t+h-1} + \varepsilon_{t+h}$$

$$B_{1,t}^{(h)} = \Theta_{h,t}.$$

Just like the time invariant case, k can be infinite in population but will be truncated to a finite value in finite samples. Similarly to the time-invariant transformation, one can do a GLS transformation $\widetilde{y}_{t+h}^{(h)} =$

[25] Baumeister and Peersman (2012) show via Monte Carlo simulations that time-varying parameter models are able to capture discrete breaks in a satisfactory manner should they occur.

$y_{t+h} - \hat{B}_{1,t}^{(h)}\hat{\varepsilon}_t - \ldots - \hat{B}_{1,t}^{(1)}\hat{\varepsilon}_{t+h-1}$. Then one can estimate horizon h via the following equation

$$\widetilde{y}_{t+h}^{(h)} = B_{1,t}^{(h+1)} y_{t-1} + B_{2,t}^{(h+1)} y_{t-2} + \ldots + B_{k,t}^{(h+1)} y_{t-k} + \tilde{u}_{t+h}^{(h)}. \tag{12}$$

Estimation is carried out in the same way as in the time-invariant case, except the models are being estimated with time-varying parameters.

Just like a static LP model can be more robust to model misspecification than a static VAR, a time-varying parameter LP model can be more robust to model misspecification than a time-varying parameter VAR. If the true model is time varying, then the misspecification of the VAR can extend to the time variation as well. Due to the iterative nature of the VAR, misspecification in time variation would be compounded in the construction of the impulse responses alongside other misspecifications in the VAR. Time-varying parameter LP, however, allow for the amount and nature of time variation to change across horizons. Since time-varying parameter models can also approximate any non-linear model, time-varying parameter LP can do a to better job capture the time variation in the impulse responses at each horizon.

There are several ways to estimate time-varying parameter models. Bayesian methods are the primary methods used to estimate time-varying parameter models, and because autocorrelation is explicitly corrected for in Bayesian LP, it is straightforward to apply time-varying parameters to Bayesian LP. For the rest of this section, I will describe a computationally convenient way to estimate time-varying parameter models. This procedure is based off of Prado and West (2010). Let

$$y_t = X_t'\beta_t + v_t,$$

$$\beta_t = \beta_{t-1} + w_t,$$

where y_t is a $r \times 1$ vector β_t is the $p \times 1$ state vector at time t, X_t is a $p \times r$ vector of regressors at time t, v_t is a $r \times 1$ vector observation noise with $v_t \sim N(0, \Sigma_t)$, w_t is the state evolution noise with $w_t \sim N(0, \Sigma_t \otimes W_t)$, and v_s and w_t are independent and mutually independent $\forall s, t$. Notice that the variance of v_t is allowed to be time-varying. Stochastic volatility (time-varying variance) is modeled as a beta-Bartlett Wishart random walk. Define D_{t-1} is the amount of information known at time $t-1$. The beta-Bartlett Wishart random walk is defined using the following $t-1$ to time t update

$$p(\Sigma_{t-1}|D_{t-1}) \sim IW(n_{t-1}, \Psi_{t-1})$$

and

$$p(\Sigma_t|D_{t-1}) \sim IW(\theta n_{t-1}, b_t\Psi_{t-1}),$$

where θ is a discount factor for stochastic volatility and $b_t = (\theta n_{t-1} + k - 1)/(n_{t-1} + k - 1)$.[26] The models are estimated using discount factors and the Forward Filter Backward Sampler (FFBS) algorithm, and details about the estimation procedure can be found in the Appendix.[27] Because discount factors and conjugate priors are used, MCMC is not needed. This is crucial for three reasons. First, if the number or parameters is even moderately large, time-varying parameter models such as Cogley and Sargent (2005), Primiceri (2005) become computationally demanding to estimate if not infeasible (Koop and Korobilis, 2013). Second, LP are estimated horizon by horizon in a sequential fashion which can make procedures such as Cogley and Sargent (2005), Primiceri (2005) impractical. Third, in order to do model comparison or hypothesis testing, it is often necessary to calculate the marginal likelihood, which is no trivial task for models estimated using MCMC. In recent years discount factors have been used in the as a solution to when the procedures of Cogley and Sargent (2005), Primiceri (2005) are burdensome (Koop and Korobilis, 2013, Koop et al., 2018). This is not to suggest that time-varying parameter procedures such as Cogley and Sargent (2005), Primiceri (2005) or other cannot be used, just that depending on the goal of the analysis and the computational power available to the researcher, these procedures may not be practical.[28]

Discount factors (also known as forgetting factors) are a natural framework for allowing and controlling for time variation in regression coefficients and the variance and are a core part of the Bayesian forecasting literature (West and Harrison, 1997, Prado and West, 2010). Discount factors lie in the interval $(0, 1]$. If a discount factor, say $\theta = .99$ is used, then from period $t \rightarrow t + 1$, $\frac{1}{\theta} - 1 \approx 1\%$ of information known at time t is discounted or forgotten in the Kalman filtration process.[29] And if $\theta = .99$, observations from 20 periods ago receive approximately 80% as much weight as this period's observation. The loss of information over time allows more recent data to have a larger impact on the parameter value and is the crux for controlling for time variation in the parameters. The discount factors are estimated using Griddy Gibbs. Including the the discount factor as a parameter to be estimated takes into account uncertainty in the hyperparameters and is a natural way to safeguard against overfitting (Giannone et al., 2015).

Due to the number of parameters being estimated, the priors for time-varying parameter models are quite important (Koop and Korobilis, 2009), otherwise parameter estimates may be imprecise if the sample size is not large. Like Cogley and Sargent (2005), Primiceri (2005), a training sample prior can be used. The prior is the same as the one presented earlier in section 3.2.[30]

[26] The model uses different discount factors for the regression coefficients and stochastic volatility.

[27] See West and Harrison (1997), Prado and West (2010) for derivations and more details about time-varying parameter methods using discount factors.

[28] If time-varying parameter procedures such as Cogley and Sargent (2005), Primiceri (2005) are used, it is recommended that the MCMC be implemented using the more computationally efficient precision sampler in Chan and Jeliazkov (2009).

[29] A discount factor of .99 has properties similar to what Cogley and Sargent (2005) call their "business as usual" prior, and it can be shown that the choice of prior shrinkage coefficient in Cogley and Sargent (2005) allows for variation in the regression coefficients roughly similar to that allowed for by a regression coefficient discount factor of .99 (Koop and Korobilis, 2013).

[30] It should be noted that non-informative priors (such as reference priors) cannot be used in Bayesian model comparison due to Bartlett's paradox. If a training sample is not available, other priors can be used. See Koop and Korobilis (2009), Koop (2017) for a review.

9 Concluding Remarks

I show that LP can be estimated with GLS. Estimating LP with GLS has three major implications. First, LP GLS can be substantially more efficient and less biased than estimation by OLS with Newey-West standard errors. Moreover, if the data are persistent and the true model is a VAR, it can be shown that impulse responses from LP can be approximately as efficient as impulse responses from VARs. Whether or not the LP is approximately as efficient depends on the persistence of the system, the horizon, and the dependence structure of the system. All else equal, the more persistent the system, the more likely LP impulse responses will be approximately as efficient for horizons typically relevant in practice. Given that most macro data are nonstationary or nearly nonstationary, even if the true model is a VAR, the efficiency of the VAR relative to the LP has been overstated in the literature.

Second, because autocorrelation process can be modeled explicitly, it is possible to give a fully Bayesian treatment of LP. That is, LP can be estimated using fully Bayesian or frequentist methods. Bayesian LP have many advantages over frequentist LP and/or Bayesian VARs such as allowing the researcher to incorporate prior information for impulse responses at each horizon. Prior information can be used to shrink impulse responses at any horizon to prevent overfitting. Economic theory can be incorporated into the prior to inform the shape of the impulse responses (e.g. the impulse response is monotonic or hump shaped) and to discipline the long-run behavior. Bayesian methods do not need to do anything special to take into account nonstationarity.

Third, since autocorrelation process can be modeled explicitly, it is now possible to estimate time-varying parameter LP. Bayesian LP can easily be adapted to handle time-varying parameter models, but one does not have to use Bayesian methods. Time-varying parameter LP can take into account structural instability in the regression coefficients and/or the covariance matrix, and since time-varying parameter models can approximate any form of non-linearity, makes them more robust to model misspecification (Granger, 2008).

The results in this paper have many potential extensions for both frequentist and Bayesian analysis. It would be useful for frequentist to have a data dependent rule or cross validation method for the optimal block length when using block bootstrapping for LP. It may be useful to extend some of the big data, sparsity, and variable selection methods used for VARs to LP.[31] It may also be useful to extend LP GLS to a non-linear (in the variables) or non-parametric setting. Even though time-varying parameter models can approximate any non-linear model (non-linear in the variables and/or the parameters), the approximation is for the conditional mean, so if the true model is non-linear in the variables, estimation of the linear (in the variables) time-invariant or time-varying parameter LP GLS would lead to inconsistent estimates of the true impulse responses. One potential solution would be to extend polynomial LP, which are motivated by non-linear

[31] See Koop and Korobilis (2009), Koop (2017) for a review.

version of the Wold representation (see Jordà (2005) section 3 for more details). If one is does not want to make assumptions about the functional form or the model, the second potential solution would be to extend nonparametric LP. Lastly, since LP are direct multistep forecasts, the results in this paper have the potential to improve the forecast accuracy of direct multistep forecasts.

References

Ang, A. and G. Bekaert (2002). Regime switches in interest rates. *Journal of Business and Economic Statistics 20*(2), 163–182.

Barnichon, R. and C. Brownlees (2018). Impulse response estimation by smooth local projections. *Working Paper*.

Baumeister, C. and G. Peersman (2012). The role of time-varying price elasticities in accounting for volatility changes in the crude oil market. *Journal of Applied Econometrics 28*(7), 1087–1109.

Berkowitz, J., I. Birgean, and L. Kilian (1999). On the finite-sample accuracy of nonparametric resampling algorithms foreconomic time series. In *Advances in Econometrics: Applying Kernel and Nonparametric Estimationto Economic Topics*, Volume 14, pp. 77–107.

Bhansali, R. J. (1997). Direct autoregressive predictors for multistep prediction: Order selection and performance relative to the plug in predictors. *Statistica Sinica 7*, 425–449.

Brodersen, K. H., F. Gallusser, J. Koehler, N. Remy, and S. L. Scott (2015). Inferring causal impact using bayesian structural time series models. *Annals of Applied Statistics 9*, 247–274.

Chan, J. C. and I. Jeliazkov (2009). Efficient simulation and integrated likelihood estimation in state space models. *International Journal of Mathematical Modelling and Numerical Optimisation 1*(1), 101–120.

Cogley, T. and T. J. Sargent (2005). Drifts and volatilities: Monetary policies and outcomes in the post-ww ii us. *Review of Economic Dynamics. 8*(2), 262–302.

Del Negro, M. and F. Schorfheide (2011). *Bayesian Macroeconometrics*, pp. 293–389. The Oxford Handbook of Bayesian Econometrics.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. CHAPMAN HALL/CRC.

Giannone, D., M. Lenza, and G. E. Primiceri (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics 97*(2), 412–435.

Giannone, D., M. Lenza, and G. E. Primiceri (2018). Priors for the long run. *Journal of the American Statistical Association Forthcoming*.

Granger, C. and P. Newbold (1977). *Forecasting Economic Time Series*.

Granger, C. W. (2008). Non-linear models: Where do we go next - time varying parameter models? *Studies in Nonlinear Dynamics Econometrics 12*(3), Online.

Greene, W. H. (2012). *Econometric Analysis* (7th ed.). Prentice Hall.

Gruber, L. F. and M. West (2016). Gpu-accelerated bayesian learning and forecasting in simultaneous graphical dynamiclinear models. *Bayesian Analysis 11*(1), 125–149.

Hall, P., J. L. Horowitz, and B.-Y. Jing (1995). On blocking rules for the bootstrap with dependent data. *Biometrika 82*(3), 561–574.

Hamilton, J. (1994). *Time Series Analysis*.

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

Hoff, P. (2009). *A First Course in Bayesian Statistical Methods* (1 ed.). Springer Texts in Statistics. Springer-Verlag New York.

Hoff, P. and J. Wakefield (2013). Bayesian sandwich posteriors for pseudo-true parameters. *Journal of Statistical Planning and Inference 143*(10), 1638–1642.

Jordà, Ò. (2005). Estimation and inference of impulse responses by local projections. *American Economic Review 95*(1), 161–182.

Jordà, Ò. (2009). Simultaneous confidence regions for impulse responses. *Review of Economics and Statistics 91*(3), 629–647.

Jordà, Ò. and S. Kozicki (2011). Estimation and inference by the method of projection minimum distance: An application to the new keynesian hybrid phillips curve. *International Economic Review 52*(2), 461–487.

Kass, R. and L. Wasserman (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association 90*(431), 928–934.

Kilian, L. (1998). Small-sample confidence intervals for impulse response functions. *Review of Economics and Statistics 80*(2), 218–230.

Kilian, L. and Y. J. Kim (2011). How reliable are local projection estimators of impulse responses? *Review of Economics and Statistics 93*(4), 1460–1466.

Koop, G. (2017). Bayesian methods for empirical macroeconomics with big data. *Review of Economic Analysis 9*(1), 33–56.

Koop, G. and D. Korobilis (2009). Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends in Econometrics 3*(4), 267–358.

Koop, G. and D. Korobilis (2013). Large time-varying parameter vars. *Journal of Econometrics 177*(2), 185–198.

Koop, G., D. Korobilis, and D. Pettenuzzo (2018). Bayesian compressed vector autoregressions. *Journal of Econometrics Forthcoming*.

Lazarus, E., D. J. Lewis, J. H. Stock, and M. W. Watson (2018). Har inference: Recommendations for practice. *Journal of Business and Economic Statistics 36*(4), 541–559.

Lewis, R. and G. C. Reinsel (1985). Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis 16*(3), 393–411.

Leybourne, S. J., B. P. M. McCabe, and A. R. Tremayne (1996). Can economic time series be differenced to stationarity? *Journal of Business and Economic Statistics 14*(4), 435–446.

Lopes, H. F., A. R. B. Moreira, and A. M. Schmidt (1999). Hyperparameter estimationin forecast models. *Computational Statistics Data Analysis 29*, 387–410.

McCracken, M. and S. Ng (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business and Economic Statistics 34*(4), 574–589.

Miranda-Agrippino, S. and G. Ricco (2018). The transmission of monetary policy shocks. *Working Paper*.

Müller, U. K. (2014). Hac corrections for strongly autocorrelated time series. *Journal of Business and Economic Statistics 32*(3), 311–322.

Murphy, K. M. and R. H. Topel (1985). Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics 3*(4), 370–379.

Nakamura, E. and J. Steinsson (2018). Identification in macroeconomics. *Journal of Economic Perspectives 32*(3), 59–86.

Newey, W. K. and D. McFadden (1994). *Large sample estimation and hypothesis testing*, Volume 4, Chapter 36, pp. 2111–2245. Amsterdam: Elsevier.

Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*(3), 703–708.

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review 25*(1), 221–247.

Paul, P. (Forthcoming). The time-varying effect of monetary policy on asset prices. *Review of Economics and Statistics*.

Pesavento, E. and B. Rossi (2006). Small-sample confidence intervals for multivariate impulse response functions at long horizons. *Journal of Applied Econometrics 21*(8), 1135–1155.

Plagborg-Møller, M. and C. K. Wolf (2019). Local projections and vars estimate the same impulse responses. *Working Paper*.

Politis, D. N. and J. P. Romano (1994). The stationary bootstrap. *Journal of the American Statistical Association 89*(428), 1303–1313.

Pope, A. L. (1990). Biases of estimators in multivariate non-gaussian autoregressions. *Jour-nal of Time Series Analysis 11*(3), 249–258.

Prado, R. and M. West (2010). *Time Series: Modeling, Computation, and Inference*. Chapman Hall/CRC Press Taylor and Francis Group.

Priestley, M. B. (1988). *Non-linear and Non-stationary Time Series Analysis*. Academic Press.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *Review of Economics Studies 72*(3), 821–852.

Ramey, V. A. (2016). Macroeconomic shocks and their propagation. In J. B. Taylor and H. Uhlig (Eds.), *Handbook of Macroeconomics*, Volume 2, Chapter 2, pp. 71–162.

Ramey, V. A. and S. Zubairy (2018). Government spending multipliers in good times and in bad: Evidence from u.s. historical data. *Journal of Political Economy 126*(2), 850–901.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics 12*(4), 1151–1172.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica 48*(1), 1–48.

Sims, C. A., J. Stock, and M. Watson (1990). Inference in linear time series models with some unit roots. *Econometrica 58*(1), 113–144.

Sims, C. A. and T. Zha (2006). Were there regime switches in u.s. monetary policy? *American Economic Review 96*(1), 54–81.

Stock, J. and M. Watson (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business and Economic Statistics 14*(1), 11–30.

Stock, J. and M. Watson (2007). *Introduction to Econometrics*. Addison Wesley Longman.

Stock, J. and M. Watson (2018). Identification and estimation of dynamic causal effects in macroeconomics using external instruments. *The Economic Journal 128*(610), 917–948.

West, M. and J. Harrison (1997). *Bayesian Forecasting and Dynamic Models* (2nd ed.). Springer-Verlag.

White, H. (2001). *Asymptotic Theory for Econometricians*. Emerald Group Publishing Limited.

Zhao, Z. Y., M. Xie, and M. West (2016). Dynamic dependence networks: Financial time series forecasting and portfolio decisions. *Applied Stochastic Models in Business and Industry 32*(3), 311–332.

Appendix

A.1 Normal Inverse-Wishart Posterior Equations

Let

$$y_t = B_1^{(1)} y_{t-1} + B_2^{(1)} y_{t-2} + \ldots + B_k^{(1)} y_{t-k} + u_t^{(0)},$$

as one would a standard Bayesian VAR. Define $\beta^{(0)} \equiv vec(\left[B_1^{(1)}, B_2^{(1)}, \ldots, B_k^{(1)}\right]')$, $X_t^{'(0)} \equiv I_n \otimes [y_{t-1}, y_{t-2}, \ldots, y_{t-k}]'$, then

$$y_t = X_t^{'(0)} \beta^{(0)} + u_t^{(0)},$$

where $u_t^{(0)} \sim N(0, \Sigma_u^{(0)})$. Assume a conditional normal inverse-Wishart prior for $p(\beta^{(0)}, \Sigma_u^{(0)})$. That is

$$p(\beta^{(0)} | \Sigma_u^{(0)}) \sim N(\underline{b}, \Sigma_u^{(0)} \otimes \underline{\Omega}),$$

$$p(\Sigma_u^{(0)}) \sim IW(\underline{n}, \underline{\Psi}),$$

where $\underline{b}$, $\underline{\Omega}$, $\underline{\Psi}$, and $\underline{n}$ are prior hyperparameters. Define $y \equiv [y'_{k+1}, \ldots, y'_T]'$ and $X \equiv [X_{k+1}^{(0)}, \ldots, X_T^{(0)}]'$, The posterior is also conditional normal inverse-Wishart. That is

$$p(\beta | \Sigma, y_{1:T}) \sim N(\overline{b}, \Sigma \otimes \overline{\Omega}),$$

$$p(\Sigma | y_{1:T}) \sim IW(\overline{n}, \overline{\Psi}),$$

where

$$\overline{\Omega} = (X'X + \underline{\Omega}^{-1})^{-1},$$

$$\hat{A} = (X'X)^{-1} X'y,$$

$$\overline{B} = \overline{\Omega}[\underline{\Omega}^{-1} \underline{B} + X'X\hat{A}],$$

$$\overline{b} = vec(\overline{B}),$$

$$\underline{b} = vec(\underline{B}),$$

$$S = (y - X\hat{A})'(y - X\hat{A}),$$

$$\overline{\Psi} = S + \underline{\Psi} + \hat{A}X'X\hat{A} + \underline{B\Omega}^{-1}\underline{B} - \overline{B}'(X'X + \underline{\Omega}^{-1})\overline{B},$$

$$\overline{n} = \underline{n} + T - k.$$

A.2 Forward Filter Backward Sampler (FFBS)

Forward Filtering

More detail about the algorithm can be found in Prado and West (2010). Recall that a TVP model can be characterized as follows:

$$y_t = X_t'\beta_t + v_t,$$

$$\beta_t = \beta_{t-1} + w_t,$$

where y_t is a $r \times 1$ vector β_t is the $p \times 1$ state vector at time t, X_t is a $p \times r$ vector of regressors at time t, ϵ_t is a $r \times 1$ vector observation noise with $v_t \sim N(0, \Sigma_t)$, w_t is the state evolution noise with $w_t \sim N(0, \Sigma_t \otimes W_t)$, and v_s and w_t are independent and mutually independent $\forall s, t$. Notice that the variance of v_t is allowed to be time-varying. Stochastic volatility (time-varying variance) is modeled as a beta-Bartlett Wishart random walk. Stochastic volatility is modeled as a beta-Bartlett Wishart random walk which is defined as following $t-1$ to time t update

$$p(\Sigma_{t-1}|D_{t-1}) \sim IW(n_{t-1}, \Psi_{t-1})$$

then

$$p(\Sigma_t|D_{t-1}) \sim IW(\theta n_{t-1}, b_t \Psi_{t-1})$$

where θ is a discount factor for stochastic volatility and $b_t = (\theta n_{t-1} + k - 1)/(n_{t-1} + k - 1)$. Let D_0 represents initial prior information and the current information set represented by $D_t = \{D_{t-1}, y_t\}$. The estimates of a standard TVP DLM can be obtained as follows. First recall that for a VAR(k) $X_t \equiv I_n \otimes [y_{t-1}^{'}, \ldots, y_{t-k}^{'}]$. Imagine we have the posterior distributions of β_t and v_t at time $t-1$. The posteriors are:

$$\beta_{t-1}|\Sigma_{t-1}, D_{t-1} \sim N(m_{t-1}, \Sigma_{t-1} \otimes C_{t-1}),$$

$$\Sigma_{t-1}|D_{t-1} \sim IW(n_{t-1}, \Psi_{t-1}),$$

where

$$M_t = M_{t-1} + A_t \epsilon_t^{'},$$

$$m_t = vec(M_t),$$

$$C_t = R_t - A_t A_t' q_t,$$

$$A_t = R_t X_t / q_t,$$

$$R_t = C_{t-1} + W_t = C_{t-1}/\delta.$$

$$n_t = \theta n_{t-1} + 1,$$

$$\Psi_t = \Psi_{t-1} + \epsilon_t \epsilon_t^{'}/q_t,$$

$$\epsilon_t = y_t - f_t,$$

$$f_t = X_t' M_{t-1},$$

$$q_t = X_t' R_t X_t' + 1,$$

where δ is the discount factor for the regression coefficients.The volatility evolves from the Σ_{t-1} posterior to the prior of Σ_t according to

$$p(\Sigma_t|D_{t-1}) \sim IW(\theta n_{t-1}, \theta\Psi_{t-1})$$

State evolves from the β_{t-1} prior to the β_t posterior as follows:

$$\beta_t|\Sigma_t, D_{t-1} \sim N(m_{t-1}, \Sigma_t \otimes R_t),$$

$\beta_t|D_{t-1}$ and $\Sigma_t|D_{t-1}$ are now the priors for β_t and Σ_t respectively. This leads to the following one-step-ahead predictive of y_t:

$$y_t|D_{t-1} \sim T_{\theta n_{t-1}}(f_t, q_t \frac{\Psi_{t-1}}{n_{t-1}}),$$

where The posterior for $\beta_t|D_t$ and $\Sigma_t|D_t$ can be now be calculated.

Backward Sampling

Initialize at T draw

$$\Sigma_T|D_T \sim IW(n_t, \Psi_t),$$

$$\beta_T|\Sigma_T, D_T \sim N(M_T, \Sigma_T \otimes C_T).$$

For t – 1 to 1

$$\Sigma_t^{-1} = \theta\Sigma_{t+1}^{-1} + \gamma_t,$$

where

$$\gamma_t^{-1} \sim IW((1-\theta)n_t, \Psi_t),$$

and

$$\beta_t = m_t + \delta(\beta_{t+1} - m_t) + N(0, \Sigma_t \otimes C_t^*),$$

where

$$C_t^* = C_t - \delta^2 R_{t+1}.$$

A.3 Choosing Lag Length and Estimating Discount Factors

The optimal the lag length is chosen by maximizing the joint log likelihood functions defined in terms of the predictive densities

$$\log[p(y_{1:T}|D_0, \delta, \theta, lag\,length)] = \sum_{t=1}^{T} \log[p(y_t|D_{t-1}, \delta, \theta, lag\,length)],$$

where

$$p(y_t|D_{t-1}, \delta, \theta, lag\,length),$$

is the one step ahead predictive density, δ is the discount factor that controls for time variation in the regression coefficients, and D_{t-1} is the amount of information known at time $t-1$.[32] Maximizing the joint log likelihood functions is equivalent to maximizing the marginal likelihood. If each model is assumed to have the same prior probability, it is also equivalent to choosing the model with the highest posterior probability. Let $M_1, M_2, \ldots, M_I$ denote I models of the same structure that only differ in their lag lengths. The posterior probability for model i can be calculated by:

$$p(M_i|y_{1:T}, D_0) = \frac{p(M_i)p(y_{1:T}|D_0, M_i)}{\sum_{j=1}^{I} p(M_j)p(y_{1:T}|D_0, M_j)}.$$

Assuming all models have equal prior probability ($p(M_i) = I^{-1}\ \forall i$):

$$p(M_i|y_{1:T}, D_0) = \frac{p(y_{1:T}|D_0, M_i)}{\sum_{j=1}^{I} p(y_{1:T}|D_0, M_j)}.$$

Then conditional on the optimal lag length, the posterior distributions for the regression coefficients and the variance are model averaged over the grid of discount factors in order to take into account the uncertainty in the discount factors. Model averaging over the grid of discount factors is equivalent to placing a uniform prior on the discount factors and estimating them using Griddy Gibbs. Ideally one would use sampling importance resampling (see (Lopes et al., 1999)), but this is computationally impractical.

The regression coefficients' discount factor is estimated over a default grid of [.7, 1] where the grid is partitioned by .01. The stochastic volatility discount factor is also chosen over a default grid of [.7, 1] where the grid is partitioned by .01. The initial grid size and partition are chosen because they cover fairly rapid

[32] X_t is suppressed in the marginal likelihood for clarity.

parameter changes to no parameter change and should cover most situations (**?**) .[33] It is important to note that if posterior distribution of the discount factors pile up at the bottom of the grid, the grid must be lowered. For example let us say that the median regression coefficient discount factor is .95, but the median variance discount factor is .7. The grid for the variance discount factor must be lowered (e.g., to [.6, 1]). The reason for this is because the true discount factor for the variance may be .62 and the regression coefficient discount factor 1, but because the grid initially only searched over [.7, 1], it may be optimal for the regression coefficients to allow for time variation in order to compensate for the bound on the amount of stochastic volatility. Theoretically, one could allow just for regression coefficient instability or only for stochastic volatility. One would just have to restrict the discount factor not of interest to be equal to 1 and then search the grid for the other discount factor. This is not recommended because the restriction may exaggerate the results of the test. For example, if the true model has stochastic volatility and the test is restricted not to allow for stochastic volatility, it may be optimal for the time-varying parameter model to exaggerate the amount of time variation in the regression coefficients in order to compensate for the restriction.[34]

Depending on the situation more flexible time-varying parameter models may be needed. It is possible to allow subsets of regression coefficients to have different discount factors. To do so one, would use block discounting (Prado and West, 2010). However, it should be noted that as the number of discount factors becomes large, the computational demands increase exponentially because a grid must be searched for each discount factor. It is also possible to change discount factors over the sample period (Koop and Korobilis, 2013). Using cholesky style decoupling and recoupling (Zhao et al., 2016) or simultaneous graphical dynamic linear models(Gruber and West, 2016), it is also possible to allow each equation in a system to have different discount factors.

A.4 Proofs of Consistency, Asymptotic Normality, and Efficiency of LP GLS

Preliminaries and Assumptions

Let y_t be an $r \times 1$ vector with Wold representation given by

$$y_t = \varepsilon_t + \sum_{h=1}^{\infty} \Theta_h \varepsilon_{t-h}$$

[33] Depending on the context, these grid values may not be appropriate and can be adjusted accordingly. If desired, one can also conduct a sensitivity analysis with the size of the grid partitions.

[34] A similar argument is made by Sims and Zha (2006) on an earlier version of Cogley and Sargent (2005) that did not allow for stochastic volatility in their time-varying parameter model.

where ε_t is i.i.d. with $E(\varepsilon_t) = 0$ and $E(\varepsilon_t\varepsilon_t') = \Sigma_\varepsilon$ and the Θ_h satisfy $\sum_{h=0}^{\infty} \parallel \Theta_h \parallel < \infty$ where $\parallel \Theta_h \parallel^2 = tr(\Theta_h'\Theta_h)$ with $\Theta_0 = I_r$. Further, assume $det\{\Theta(z)\} \neq 0$ for $|z| \leq 1$ where $\Theta(z) = \sum_{h=0}^{\infty} \Theta_h z^h$ so that process can be written as an infinite order VAR representation

$$y_t = \sum_{j=1}^{\infty} A_j y_{t-j} + \varepsilon_t$$

with $\sum_{j=1}^{\infty} \parallel A_j \parallel < \infty$ and $A(z) = \Theta(z)^{-1}$. By recursive substitution

$$y_{t+h} = B_1^{(h)} y_t + B_2^{(h)} y_{t-1} + \ldots + \varepsilon_{t+h} + \Theta_1 \varepsilon_{t+h-1} + \ldots + \Theta_{h-1}\varepsilon_{t+1},$$

where $B_1^{(h)} = \Theta_h$, $B_j^{(h)} = \Theta_{h-1}A_j + B_{j+1}^{(h-1)}$ for $h \geq 1$ and with $B_{j+1}^{(0)} = 0$; $\Theta_0 = I_r$ with $j \geq 1$. The horizon h LP consists of estimating Θ_h from a least squares estimate of A_1^h with truncated regression

$$y_{t+h} = B_1^{(h)} y_t + \ldots + B_k^{(h)} y_{t-k+1} + e_{k,t+h}^{(h)},$$

where

$$e_{k,t+h}^{(h)} = \sum_{j=k+1}^{\infty} B_j^{(h)} y_{t-j+1} + \varepsilon_{t+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{t+h-l}.$$

For standard LP

$\hat{B}(k,h,OLS) = (\hat{B}_1^{(h)}, \ldots, \hat{B}_k^{(h)}) = \hat{\Gamma}_{1-k,h}'\hat{\Gamma}_k^{-1}$

$\hat{\Gamma}_{1-k,h} = (T-k-H)^{-1} \sum_{t=k}^{T-h} X_{t,k} y_{t+h}'$

$\hat{\Gamma}_k = (T-k-H)^{-1} \sum_{t=k}^{T-H} X_{t,k} X_{t,k}'$

$X_{t,k} = (y_t', y_{t-1,}', \ldots, y_{t-k+1}')'$

$$\hat{B}(k,h,OLS) - B(k,h) = \{(T-k-H)^{-1} \sum_{t=k}^{T-H} (\sum_{j=k+1}^{\infty} B_j^{(h)} y_{t-j+1} + \varepsilon_{t+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{t+h-l}) X_{t,k}'\} \hat{\Gamma}_k^{-1}$$

Define

$$U_{1T} = \{(T-k-H)^{-1} \sum_{t=k}^{T-H} (\sum_{j=k+1}^{\infty} B_j^{(h)} y_{t-j+1}) X_{t,k}'\}$$

$$U_{2T} = \{(T-k-H)^{-1} \sum_{t=k}^{T-H} \varepsilon_{t+h} X_{t,k}'\}$$

$$U_{3T} = \{(T-k-H)^{-1} \sum_{t=k}^{T-H} (\sum_{l=1}^{h-1} \Theta_l \varepsilon_{t+h-l}) X_{t,k}'\}$$

Proof of Consistency for LP OLS Correction

Assumption 2. *Let y_t satisfy the Wold representation as presented above. Assume that in addition,*

$$(i)\ E|\varepsilon_{it}\varepsilon_{jt}\varepsilon_{kt}\varepsilon_{lt}| < \infty$$

for $1 \leq i, j, k, l \leq n$.

(ii) k satisfies

$$\frac{k^2}{T} \to 0;\ T, k \to \infty$$

(iii) k satisfies

$$k^{1/2} \sum_{j=k+1}^{\infty} \| A_j \| \to 0\ T, k \to \infty.$$

These assumptions were used to show consistency of the VAR(∞) (Lewis and Reinsel, 1985) and the LP(∞) (Jordà and Kozicki, 2011).

Proposition 2. *Assume assumption 2 holds, then*

$$\| \hat{B}(k, h, OLS) - B(k, h) \| \xrightarrow{p} 0.$$

Proof. Lewis and Reinsel (1985) establish that $\| \hat{\Gamma}_k^{-1} \|_1$ is bounded in probability, so consistency in standard LP consists of showing that $\| U_{1T} \|, \| U_{2T} \|$, and $\| U_{3T} \|$ converge in probability to 0. This was shown in Jordà and Kozicki (2011). However, their proof showing $\| U_{3T} \|$ converging to 0 is incorrect. It is incorrect because $(\sum_{l=1}^{h-1} \Theta_l \varepsilon_{t+h-l}) X'_{t,k}$ is assumed to be independent across time. It is not. A correct proof is

$$U_{3T} = \{(T-k-H)^{-1} \sum_{t=k}^{T-H} (\sum_{l=1}^{h-1} \Theta_l \varepsilon_{t+h-l}) X'_{t,k}\}$$

$$\| U_{3T} \|^2 = \| \{(T-k-H)^{-1} \sum_{t=k}^{T-H} (\sum_{l=1}^{h-1} \Theta_l \varepsilon_{t+h-l}) X'_{t,k}\} \|^2$$

$$\| U_{3T} \|^2 = (T-k-H)^{-2} trace\{[\sum_{n=k}^{T-H} (\sum_{l=1}^{h-1} \Theta_l \varepsilon_{n+h-l}) X'_{n,k}]'[\sum_{m=k}^{T-H} (\sum_{l=1}^{h-1} \Theta_l \varepsilon_{m+h-l}) X'_{m,k}]\}$$

$$\| U_{3T} \|^2 = (T-k-H)^{-2} trace\{\sum_{m=k}^{T-H} \sum_{n=k}^{T-H} (\sum_{l=1}^{h-1} \Theta_l \varepsilon_{n+h-l})' (\sum_{l=1}^{h-1} \Theta_l \varepsilon_{m+h-l}) X'_{m,k} X_{n,k}\}$$

by the cyclic property of traces.

$$E \| U_{3T} \|^2 = (T-k-H)^{-2} trace \sum_{m=k}^{T-H} \sum_{n=k}^{T-H} E\{(\sum_{l=1}^{h-1} \Theta_l \varepsilon_{n+h-l})' (\sum_{l=1}^{h-1} \Theta_l \varepsilon_{m+h-l}) X'_{m,k} X_{n,k}\}.$$

For $|n-m| > h-1$

$$E\{(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{n+h-l})'(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{m+h-l})X'_{m,k}X_{n,k}\} = E\{(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{n+h-l})'\}E\{(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{m+h-l})X'_{m,k}X_{n,k}\} = 0$$

by independence. So

$$E \parallel U_{3T} \parallel^2 = (T-k-H)^{-2} trace \sum_{m=k}^{T-H}\sum_{|n-m|<h} E\{(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{n+h-l})'(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{m+h-l})X'_{m,k}X_{n,k}\}.$$

Note that

$$| E\{(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{n+h-l})'(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{m+h-l})X'_{m,k}X_{n,k}\} | \leq (E[\{(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{n+h-l})'(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{m+h-l})\}^2])^{1/2}(E[\{X'_{m,k}X_{n,k}\}^2])^{1/2}$$

by Cauchy-Schwarz inequality. And

$$X'_{m,k}X_{n,k} = y'_m y_n + y'_{m-1}y_{n-1} + \ldots + y'_{m-k+1}y_{n-k+1}$$

$$(X'_{m,k}X_{n,k})^2 = (y'_m y_n + y'_{m-1}y_{n-1} + \ldots + y'_{m-k+1}y_{n-k+1})^2$$

$$E[(X'_{m,k}X_{n,k})^2] = O_p(k^2)$$

and

$$| E([(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{n+h-l})'(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{m+h-l})]^2) | = constant$$

due to the finite fourth moments of ε and $\sum_{h=0}^{\infty} \parallel \Theta_h \parallel < \infty$. Consequently for $|n-m| \leq h-1$,

$$trace\{(E[\{(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{n+h-l})'(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{m+h-l})\}^2])^{1/2}(E[\{X'_{m,k}X_{n,k}\}^2])^{1/2}\} = O_p(k).$$

This implies there exists some finite constant M such that

$$E \parallel U_{3T} \parallel^2 = (T-k-H)^{-2} trace \sum_{m=k}^{T-H}\sum_{|n-m|<h} E\{(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{n+h-l})'(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{m+h-l})X'_{m,k}X_{n,k}\} \leq (T-k-H)^{-2}(T-k-H)(kh)M$$

$$E \parallel U_{3T} \parallel^2 \leq (T-k-H)^{-1}k \times constant \xrightarrow{p} 0.$$

$$\Longrightarrow \parallel U_{3T} \parallel \xrightarrow{p} 0$$

That completes the correction that shows that LP OLS is consistent. □

Proof of Consistency for LP GLS

Theorem 2. *Assume assumption 2 holds, then for LP GLS*

$$\| \hat{B}(k,h,GLS) - B(k,h) \| \xrightarrow{p} 0.$$

Proof. To show consistency in LP GLS, there is an additional term

$$U_{4T} = \{(T-k-H)^{-1} \sum_{t=k}^{T-H} (\sum_{l=1}^{h-1} \hat{\Theta}_l \hat{\varepsilon}_{t+h-l}) X'_{t,k}\}$$

that must be taken into account. To see why note that the horizon h LP GLS is

$$y_{t+h} - \sum_{l=1}^{h-1} \hat{\Theta}_j \hat{\varepsilon}_{t+h-l} = B_1^{(h)} y_t + \ldots + B_k^{(h)} y_{t-k+1} + \widetilde{u}_{k,t+h}^{(h)},$$

where

$$\widetilde{u}_{k,t+h}^{(h)} = \sum_{j=k+1}^{\infty} B_j^{(h)} y_{t-j+1} + \varepsilon_{t+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{t+h-l} - \sum_{l=1}^{h-1} \hat{\Theta}_j \hat{\varepsilon}_{t+h-l}.$$

and $\hat{\Theta}_h = \hat{B}_1^{(h)}$. To show consistency of LP GLS it suffices to show that $\| U_{4T} \| \xrightarrow{p} 0$ because for LP GLS

$$\| \hat{B}(k,h) - B(k,h) \| = \| U_{1T}\hat{\Gamma}_k^{-1} + U_{2T}\hat{\Gamma}_k^{-1} + U_{3T}\hat{\Gamma}_k^{-1} - U_{4T}\hat{\Gamma}_k^{-1} \|$$

$$\leq \| U_{1T} \| \| \hat{\Gamma}_k^{-1} \|_1 + \| U_{2T} \| \| \hat{\Gamma}_k^{-1} \|_1 + \| U_{3T} \| \| \hat{\Gamma}_k^{-1} \|_1 - \| U_{4T} \| \| \hat{\Gamma}_k^{-1} \|_1 .$$

Lewis and Reinsel (1985) establish that $\| \hat{\Gamma}_k^{-1} \|_1$ is bounded in probability. Jordà and Kozicki (2011) show $\| U_{1T} \|$ and $\| U_{2T} \|$ converges in probability to 0, and Proposition 2 shows $\| U_{3T} \|$ converges in probability to 0. The proof showing $\| U_{4T} \| \xrightarrow{p} 0$ will be a proof by induction. Assume the consistency for the previous $h-1$ horizons has been proven. Hence $\| \hat{\Theta}_l \| \xrightarrow{p} \| \Theta_l \| < \infty$ for $1 \leq l \leq h-1$. Note

$$\hat{\varepsilon}_t = \varepsilon_t + (\sum_{j=1}^{\infty} A_j y_{t-j}) - (\sum_{i=1}^{k} \hat{A}_i y_{t-i}).$$

Therefore

$$U_{4T} = \{(T-k-H)^{-1} \sum_{t=k}^{T-H} (\sum_{l=1}^{h-1} \hat{\Theta}_l \hat{\varepsilon}_{t+h-l}) X'_{t,k}\}$$

$$= \{(T-k-H)^{-1} \sum_{t=k}^{T-H} (\sum_{l=1}^{h-1} \hat{\Theta}_l (\varepsilon_{t+h-l} + (\sum_{j=1}^{\infty} A_j y_{t+h-l-j}) - (\sum_{i=1}^{k} \hat{A}_i y_{t+h-l-i}))) X'_{t,k}\}$$

$$= \sum_{l=1}^{h-1} \hat{\Theta}_l \{(T-k-H)^{-1} \sum_{t=k}^{T-H} ((\varepsilon_{t+h-l} + (\sum_{j=1}^{\infty} A_j y_{t+h-l-j}) - (\sum_{i=1}^{k} \hat{A}_i y_{t+h-l-i}))) X'_{t,k}\}.$$

It was shown earlier that

$$\| U_{3T} \| = \| \sum_{l=1}^{h-1} \Theta_l \{(T-k-H)^{-1} \sum_{t=k}^{T-H} \varepsilon_{t+h-l} X'_{t,k}\} \| \xrightarrow{p} 0.$$

Since $h-1$ is finite and $\| \hat{\Theta}_l \| \xrightarrow{p} \| \Theta_l \| < \infty$

$$\| \sum_{l=1}^{h-1} \hat{\Theta}_l \{(T-k-H)^{-1} \sum_{t=k}^{T-H} \varepsilon_{t+h-l} X'_{t,k}\} \| \leq \sum_{l=1}^{h-1} \underbrace{\| \hat{\Theta}_l \|}_{bounded} \underbrace{\| \{(T-k-H)^{-1} \sum_{t=k}^{T-H} \varepsilon_{t+h-l} X'_{t,k}\} \|}_{plim=0} \xrightarrow{p} 0.$$

To show $\| U_{4T} \| \xrightarrow{p} 0$ it suffices to show that

$$\| \sum_{l=1}^{h-1} \hat{\Theta}_l \{(T-k-H)^{-1} \sum_{t=k}^{T-H} (\sum_{j=1}^{\infty} A_j y_{t+h-l-j} - \sum_{i=1}^{k} \hat{A}_i y_{t+h-l-i}) X'_{t,k}\} \| \xrightarrow{p} 0.$$

Owing to $h-1$ in finite and $\| \hat{\Theta}_l \| \xrightarrow{p} \| \Theta_l \| < \infty$, this simplifies to showing

$$\| \{(T-k-H)^{-1} \sum_{t=k}^{T-H} (\sum_{j=1}^{\infty} A_j y_{t+h-l-j} - \sum_{i=1}^{k} \hat{A}_i y_{t+h-l-i}) X'_{t,k}\} \|$$

$$= \| \{(T-k-H)^{-1} \sum_{t=k}^{T-H} ((\sum_{j=k+1}^{\infty} A_j y_{t+h-l-j}) - (\hat{B}(k,1) - B(k,1)) X_{t+h-l-1,k}) X'_{t,k}\} \|$$

$$= \| \{(T-k-H)^{-1} \sum_{t=k}^{T-H} ((\sum_{j=k+1}^{\infty} A_j y_{t+h-l-j})) X'_{t,k}\} - \{(T-k-H)^{-1} \sum_{t=k}^{T-H} ((\hat{B}(k,1) - B(k,1)) X_{t+h-l-1,k}) X'_{t,k}\} \| \xrightarrow{p} 0.$$

Jordà and Kozicki (2011) and Lewis and Reinsel (1985) already showed

$$\| \{(T-k-H)^{-1} \sum_{t=k}^{T-H} ((\sum_{j=k+1}^{\infty} A_j y_{t+h-l-j})) X'_{t,k}\} \| \xrightarrow{p} 0.$$

Now all that is left to show is

$$\| \{(T-k-H)^{-1} \sum_{t=k}^{T-H} ((\hat{B}(k,1) - B(k,1)) X_{t+h-l-1,k}) X'_{t,k}\} \| \xrightarrow{p} 0.$$

Note that $(\hat{B}(k,1) - B(k,1)$ does not depend on the t subscript so it can be factored out. That is,

$$\| \{(T-k-H)^{-1} \sum_{t=k}^{T-H} ((\hat{B}(k,1) - B(k,1)) X_{t+h-l-1,k}) X'_{t,k}\} \|$$

$$= \| \{(\hat{B}(k,1) - B(k,1))(T-k-H)^{-1} \sum_{t=k}^{T-H} X_{t+h-l-1,k} X'_{t,k}\} \|$$

$$\leq \underbrace{\| \{(\hat{B}(k,1) - B(k,1))\} \|}_{plim=0} \underbrace{\| (T-k-H)^{-1} \sum_{t=k}^{T-H} X_{t+h-l-1,k} X'_{t,k}\} \|_1}_{bounded} \xrightarrow{p} 0.$$

Since this is a proof by induction, it was assumed that the first $h-1$ horizons are consistent, so the first term converges in probability. The second term is bounded due to $\| \hat{\Gamma}_k \|_1 = \| (T-k-H)^{-1} \sum_{t=k}^{T-H} X_{t,k} X'_{t,k} \|_1$ being bounded and since the autocovariances are absolutely summable. It follows that

$$\| \hat{\Theta}_l \{(T-k-H)^{-1} \sum_{t=k}^{T-H} ((\varepsilon_{t+h-l} + (\sum_{j=1}^{\infty} A_j y_{t+h-l-j}) - (\sum_{i=1}^{k} \hat{A}_i y_{t+h-l-i}))) X'_{t,k}\} \| \xrightarrow{p} 0$$

for each $1 \leq l \leq h-1$. Therefore, $\| U_{4T} \| \xrightarrow{p} 0$ since the sum of a finite number of terms that each converge to zero also converges to 0. That is

$$\| U_{4T} \| = \| \sum_{l=1}^{h-1} \hat{\Theta}_l \{(T-k-H)^{-1} \sum_{t=k}^{T-H} ((\varepsilon_{t+h-l} + (\sum_{j=1}^{\infty} A_j y_{t+h-l-j}) - (\sum_{i=1}^{k} \hat{A}_i y_{t+h-l-i}))) X'_{t,k}\} \| \xrightarrow{p} 0.$$

To complete the proof by induction, note that the horizon 0 LP is a VAR, and the consistency results for the VAR were proved in Lewis and Reinsel (1985), so the first step in the induction process was proved. □

Proof of Asymptotic Normality for LP OLS Correction

Assumption 3. *Let y_t satisfy the Wold representation as presented in the preliminary section. Assume that in addition,*

$$(i)\, E|\varepsilon_{it}\varepsilon_{jt}\varepsilon_{kt}\varepsilon_{lt}| < \infty$$

for $1 \leq i, j, k, l \leq r$.

(ii) k satisfies

$$\frac{k^3}{T} \to 0;\ T, k \to \infty$$

(iii) k satisfies

$$T^{1/2} \sum_{j=k+1}^{\infty} \| A_j \| \to 0\ T, k \to \infty.$$

Proposition 3. *Assume assumption 3 holds, then for LP OLS*

$$\sqrt{T-k-H} vec[\hat{B}(k,h,OLS) - B(k,h)] \xrightarrow{d} N(0, \Omega_h).$$

where

$$\Omega_h = \sum_{t=0}^{\infty}\sum_{s=0}^{\infty} cov[vec\{(\varepsilon_{t+h} + \sum_{l=1}^{h-1}\Theta_l\varepsilon_{t+h-l})X'_{t,k}\Gamma_k^{-1}\}, vec\{(\varepsilon_{s+h} + \sum_{l=1}^{h-1}\Theta_l\varepsilon_{s+h-l})X'_{s,k}\Gamma_k^{-1}\}'].$$

These assumptions were used to show asymptotic normality of the VAR(∞) (Lewis and Reinsel, 1985) and the LP(∞) (Jordà and Kozicki, 2011). It turns out Jordà and Kozicki (2011) use the incorrect Central Limit Theorem. Jordà and Kozicki (2011) proof follows the same argument as Lewis and Reinsel (1985). Lewis and Reinsel (1985) use a martingale CLT to prove asymptotic normality. This is possible because in the case of a VAR since

$$vec\{(T-k-H)^{-1/2}\sum_{t=k}^{T-H}(\varepsilon_{t+1}X'_{t,k})\Gamma_k^{-1}\}$$

is a martingale, because ε_{t+1} and $X'_{t,k}$ are independent of each other, and ε_{t+1} is an i.i.d. and is therefore uncorrelated over time. In order to use the martingale CLT theorem for standard LP

$$vec\{(T-k-H)^{-1/2}\sum_{t=k}^{T-H}(\varepsilon_{t+h} + \sum_{l=1}^{h-1}\Theta_l\varepsilon_{t+h-l})X'_{t,k}\Gamma_k^{-1}\}$$

would need to be a martingale. But it is not a martingale. Even though $(\varepsilon_{t+h} + \sum_{l=1}^{h-1}\Theta_l\varepsilon_{t+h-l})$ is independent of $X'_{t,k}$, the process is not a martingale because the error term $(\varepsilon_{t+h} + \sum_{l=1}^{h-1}\Theta_l\varepsilon_{t+h-l})$ is correlated across h horizons and $X'_{t,k}$ is correlated for potentially infinite horizons. Instead of using the Martingale Central Limit Theorem, the Gordin Central Limit Theorem should have been used. Given that the ε_t are i.i.d. and strongly stationary, $(\varepsilon_{t+h} + \sum_{l=1}^{h-1}\Theta_l\varepsilon_{t+h-l})$ are strongly stationary and ergodic. Due to the assumptions placed on y_t, $X'_{t,k}$ is strongly stationary and ergodic. Hence

$$\{vec\{(\varepsilon_{t+h} + \sum_{l=1}^{h-1}\Theta_l\varepsilon_{t+h-l})X'_{t,k}\Gamma_k^{-1}\}\}_{t=-\infty}^{t=\infty}$$

is strongly stationary and ergodic (Hayashi, 2000, White, 2001). The Gordin CLT states that if a time series process is strongly stationary and ergodic and satisfies the following three conditions:

1. Asymptotic uncorrelatedness
2. Summability of autocovariances
3. Asymptotic negligibility of innovations,

then it is asymptotically normal (Greene, 2012). The corrected proof of standard LP can be shown as follows.

Proof. To show asymptotic uncorrelatedness need to show that

$$lim_{j\to\infty}E[\{(\varepsilon_{t+h} + \sum_{l=1}^{h-1}\Theta_l\varepsilon_{t+h-l})X'_{t,k}\}\Gamma_k^{-1}\Big|\{(\varepsilon_{t+h-j} + \sum_{l=1}^{h-1}\Theta_l\varepsilon_{t+h-l-j})X'_{t-j,k}\}\Gamma_k^{-1}] = 0,$$

where $E[\cdot|\cdot]$ is the conditional expectation. Asymptotic uncorrelatedness is trivially satisfied because when j is greater than $h-1$, the process is independent since $(\varepsilon_{t+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{t+h-l})$ would be independent of $(\varepsilon_{t+h-j} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{t+h-l-j})$.

To show Summability of autocovariances, need to show

$$lim_{T\to\infty} var((T-k-H)^{-1/2} vec\{ \sum_{t=k}^{T-H} (\varepsilon_{t+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{t+h-l})) X'_{t,k} \Gamma_k^{-1}\})$$

is finite and constant. Define

$$s_T = (T-k-H)^{-1/2} vec\{ \sum_{t=k}^{T-H} (\varepsilon_{t+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{t+h-l})) X'_{t,k} \Gamma_k^{-1}\}.$$

Note that

$$vec\{(\varepsilon_{t+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{t+h-l}) X'_{t,k} \Gamma_k^{-1}\} = (\Gamma_k^{-1} X_{t,k} \otimes I_r) vec((\varepsilon_{t+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{t+h-l}))$$

so

$$var(s_T) = (T-k-H) \sum_{m=k}^{T-H} \sum_{n=k}^{T-H} E[(\Gamma_k^{-1} X_{m,k} \otimes I_r) vec((\varepsilon_{m+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{m+h-l})) vec((\varepsilon_{n+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{n+h-l}))' (X'_{n,k} \Gamma_k'^{-1} \otimes I_r)]$$

for $|n-m| > h-1$ the most future $(\varepsilon_{t+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{t+h-l})$ in the couple is independent of everything else. Therefore

$$var(s_T) = (T-k-H)^{-1} \sum_{m=k}^{T-H} \sum_{|n-m|<h} E[(\Gamma_k^{-1} X_{m,k} \otimes I_r) vec((\varepsilon_{m+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{m+h-l})) vec((\varepsilon_{n+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{n+h-l}))' (X'_{n,k} \Gamma_k'^{-1} \otimes I_r)].$$

If one conditions on information known up to time n ($\mathcal{F}_n$ will denote the time n information set)

$$E[(\Gamma_k^{-1} X_{m,k} \otimes I_r) vec((\varepsilon_{m+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{m+h-l})) vec((\varepsilon_{n+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{n+h-l}))' (X'_{n,k} \Gamma_k'^{-1} \otimes I_r) \mid \mathcal{F}_n]$$

$$= [(\Gamma_k^{-1} X_{m,k} \otimes I_r) \Sigma_{e^{(h)},(m-n)} (X'_{n,k} \Gamma_k'^{-1} \otimes I_r)]$$

where

$$\Sigma_{e^{(h)},(m-n)} = E[(\varepsilon_{m+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{m+h-l})(\varepsilon'_{n+h} + \sum_{l=1}^{h-1} \varepsilon'_{n+h-l} \Theta'_l) \mid \mathcal{F}_n]$$

$$= E[(\varepsilon_{m+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{m+h-l})(\varepsilon'_{n+h} + \sum_{l=1}^{h-1} \varepsilon'_{n+h-l} \Theta'_l)]$$

which is constant and finite for all m and n due to the finite fourth moments of ε and $\sum_{h=0}^{\infty} \| \Theta_h \| < \infty$.

$$[(\Gamma_k^{-1} X_{m,k} \otimes I_r) \Sigma_{e^{(h)},(m-n)} (X'_{n,k} \Gamma_k'^{-1} \otimes I_r)]$$

$$= [(\Gamma_k^{-1} X_{m,k} \otimes I_r)(1 \otimes \Sigma_{e^{(h)},(m-n)})(X'_{n,k}\Gamma_k'^{-1} \otimes I_r)]$$

$$= [(\Gamma_k^{-1} X_{m,k} \otimes \Sigma_{e^{(h)},(m-n)})(X'_{n,k}\Gamma_k'^{-1} \otimes I_r)]$$

$$= [\Gamma_k^{-1} X_{m,k} X'_{n,k}\Gamma_k'^{-1} \otimes \Sigma_{e,m,n}]$$

and

$$E[\Gamma_k^{-1} X_{m,k} X'_{n,k}\Gamma_k'^{-1} \otimes \Sigma_{e^{(h)},(m-n)}] = \Gamma_k^{-1}\Gamma_{(m-n),k}\Gamma_k'^{-1} \otimes \Sigma_{e^{(h)},(m-n)}$$

where $E(X_{m,k}X'_{n,k}) = \Gamma_{(m-n),k}$. Due to

$$E[\Gamma_k^{-1} X_{m,k} X'_{n,k}\Gamma_k'^{-1} \otimes \Sigma_{e^{(h)},(m-n)}] = \Gamma_k^{-1}\Gamma_{(m-n),k}\Gamma_k'^{-1} \otimes \Sigma_{e^{(h)},(m-n)}$$

being constant

$$var(s_T) = (T-k-H)^{-1} \sum_{m=k}^{T-H} \sum_{|n-m|<h} E[(\Gamma_k^{-1} X_{m,k} \otimes I_r) vec((\varepsilon_{m+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{m+h-l}))vec((\varepsilon_{n+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{n+h-l}))'(X'_{n,k}\Gamma_k'^{-1} \otimes I_r)]$$

$$= \sum_{|n-m|<h} E[(\Gamma_k^{-1} X_{m,k} \otimes I_r) vec((\varepsilon_{m+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{m+h-l}))vec((\varepsilon_{n+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{n+h-l}))'(X'_{n,k}\Gamma_k'^{-1} \otimes I_r)]$$

$$= \sum_{|n-m|<h} \Gamma_k^{-1}\Gamma_{(m-n),k}\Gamma_k'^{-1} \otimes \Sigma_{e^{(h)},(m-n)}$$

which is finite for finite h.

To show the Asymptotic negligibility of innovations, note that for $k > h-1$, the innovation is zero (this point ends up not mattering). Since $(\Gamma_k^{-1} X_{t,k} \otimes I_r) vec((\varepsilon_{t+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{t+h-l}))$ is second order stationary (it has mean zero and it has been shown that the autocovariances are finite and constant at all horizons), then there exists a Wold VMA representation. This Wold representation can be written as a stationary VAR (∞). If I write the VAR (∞) as a VAR(1),

$$Z_t = AZ_{t-1} + e_t$$

so

$$r_{t0} = e_t$$

$$r_{t1} = Ae_t$$

$$r_{t2} = A^2 e_t$$

$$\vdots$$

Because the VAR is stationary, the impact of an innovation decays over time, and asymptotic negligibility

trivially follows. □

Proof of Asymptotic Normality LP GLS

For LP OLS

$$\sqrt{T-k-H}[\hat{B}(k,h,OLS) - B(k,h)] = \sqrt{T-k-H}[U_{1T}\hat{\Gamma}_k^{-1} + U_{2T}\hat{\Gamma}_k^{-1} + U_{3T}\hat{\Gamma}_k^{-1}]$$

For LP GLS

$$\sqrt{T-k-H}[\hat{B}(k,h,GLS) - B(k,h)] = \sqrt{T-k-H}[U_{1T}\hat{\Gamma}_k^{-1} + U_{2T}\hat{\Gamma}_k^{-1} + U_{3T}\hat{\Gamma}_k^{-1} - U_{4T}\hat{\Gamma}_k^{-1}]$$

where again

$$U_{4T} = \{(T-k-H)^{-1}\sum_{t=k}^{T-H}(\sum_{l=1}^{h-1}\hat{\Theta}_l\hat{\varepsilon}_{t+h-l})X'_{t,k}\}$$

Theorem 3. *If assumption 3 holds, then for LP GLS*

$$\sqrt{T-k-H}vec[\hat{B}(k,h,GLS) - B(k,h)] \xrightarrow{d} N(0,\Omega_h^{GLS}),$$

where

$$\Omega_h^{GLS} = var(\aleph) + var(\Upsilon) + cov(\aleph,\Upsilon') + cov(\Upsilon,\aleph')$$

$$\aleph = (T-k-H)^{-1/2}vec[\sum_{t=k}^{T-H}\varepsilon_{t+h}X'_{t,k}\Gamma_k^{-1}]$$

$$\Upsilon = vec[\sum_{l=1}^{h-1}\Theta_l\sqrt{T-k-H}(\hat{B}(k,1) - B(k,1))\Gamma_{(h-l-1),k}\Gamma_k^{-1}]$$

Proof. To show that

$$\sqrt{T-k-H}[\hat{B}(k,h,GLS) - B(k,h)] = \sqrt{T-k-H}[U_{1T}\hat{\Gamma}_k^{-1} + U_{2T}\hat{\Gamma}_k^{-1} + U_{3T}\hat{\Gamma}_k^{-1} - U_{4T}\hat{\Gamma}_k^{-1}]$$

is normally distributed, it will first help to simplify the expression by showing

$$\sqrt{T-k-H}[\hat{B}(k,h,GLS) - B(k,h)] \xrightarrow{p} \sqrt{T-k-H}[U_{1T}\Gamma_k^{-1} + U_{2T}\Gamma_k^{-1} + U_{3T}\Gamma_k^{-1} - U_{4T}\Gamma_k^{-1}]$$

This can be done by showing that

$$\| \sqrt{T-k-H}[U_{1T} + U_{2T} + U_{3T} - U_{4T}](\hat{\Gamma}_k^{-1} - \Gamma_k^{-1}) \| \xrightarrow{p} 0.$$

Jordà and Kozicki (2011) already showed that

$$\| \sqrt{T-k-H}[U_{1T}+U_{2T}+U_{3T}](\hat{\Gamma}_k^{-1}-\Gamma_k^{-1}) \| \xrightarrow{p} 0.$$

So I just need to show

$$\| \sqrt{T-k-H}U_{4T}(\hat{\Gamma}_k^{-1}-\Gamma_k^{-1}) \| \xrightarrow{p} 0.$$

To simplify the expression into something more manageable, I'll begin by simplifying

$$\sqrt{T-k-H}[U_{4T}](\hat{\Gamma}_k^{-1}-\Gamma_k^{-1}).$$

Let

$$\hat{\varepsilon}_t = \varepsilon_t + (\sum_{j=1}^{\infty} A_j y_{t-j}) - (\sum_{i=1}^{k} \hat{A}_i y_{t-i}),$$

then

$$\sqrt{T-k-H}U_{4T}(\hat{\Gamma}_k^{-1}-\Gamma_k^{-1}) = \{(T-k-H)^{-1/2} \sum_{t=k}^{T-H} (\sum_{l=1}^{h-1} \hat{\Theta}_l \hat{\varepsilon}_{t+h-l}) X'_{t,k}\}(\hat{\Gamma}_k^{-1}-\Gamma_k^{-1}).$$

$$= \{(T-k-H)^{-1/2} \sum_{t=k}^{T-H} (\sum_{l=1}^{h-1} \hat{\Theta}_l (\varepsilon_{t+h-l} + (\sum_{j=1}^{\infty} A_j y_{t+h-l-j}) - (\sum_{i=1}^{k} \hat{A}_i y_{t+h-l-i}))) X'_{t,k}\}(\hat{\Gamma}_k^{-1}-\Gamma_k^{-1}).$$

$$= \{(T-k-H)^{-1/2} \sum_{t=k}^{T-H} (\sum_{l=1}^{h-1} \hat{\Theta}_l (\varepsilon_{t+h-l} + (\sum_{j=k+1}^{\infty} A_j y_{t+h-l-j}) - (\hat{B}(k,1)-B(k,1))X_{t+h-l-1,k})) X'_{t,k}\}(\hat{\Gamma}_k^{-1}-\Gamma_k^{-1}).$$

$$= \{(T-k-H)^{-1/2} \sum_{l=1}^{h-1} \hat{\Theta}_l \sum_{t=k}^{T-H} (\varepsilon_{t+h-l} + (\sum_{j=k+1}^{\infty} A_j y_{t+h-l-j}) - (\hat{B}(k,1)-B(k,1))X_{t+h-l-1,k})) X'_{t,k}\}(\hat{\Gamma}_k^{-1}-\Gamma_k^{-1}).$$

So

$$\| \sqrt{T-k-H}U_{4T}(\hat{\Gamma}_k^{-1}-\Gamma_k^{-1}) \|$$

$$= \| \{(T-k-H)^{-1/2} \sum_{l=1}^{h-1} \hat{\Theta}_l \sum_{t=k}^{T-H} (\varepsilon_{t+h-l} + (\sum_{j=k+1}^{\infty} A_j y_{t+h-l-j}) - (\hat{B}(k,1)-B(k,1))X_{t+h-l-1,k})) X'_{t,k}\}(\hat{\Gamma}_k^{-1}-\Gamma_k^{-1}) \|$$

$$\leq \| \{(T-k-H)^{-1/2} \sum_{l=1}^{h-1} \hat{\Theta}_l \sum_{t=k}^{T-H} (\varepsilon_{t+h-l} + (\sum_{j=k+1}^{\infty} A_j y_{t+h-l-j}) - (\hat{B}(k,1)-B(k,1))X_{t+h-l-1,k})) X'_{t,k}\} \|$$

$$\times \| (\hat{\Gamma}_k^{-1}-\Gamma_k^{-1}) \|_1$$

$$\leq \sum_{l=1}^{h-1} \| \hat{\Theta}_l \| \left(\| \{(T-k-H)^{-1/2} \sum_{t=k}^{T-H} (\varepsilon_{t+h-l} + (\sum_{j=k+1}^{\infty} A_j y_{t+h-l-j}) - (\hat{B}(k,1)-B(k,1))X_{t+h-l-1,k})) X'_{t,k}\} \| \right)$$

$$\times \{ \| (\hat{\Gamma}_k^{-1}-\Gamma_k^{-1}) \|_1 \}$$

$$= \sum_{l=1}^{h-1} \| \hat{\Theta}_l \| \left(\| \{[k(T-k-H)]^{-1/2} \sum_{t=k}^{T-H} (\varepsilon_{t+h-l} + (\sum_{j=k+1}^{\infty} A_j y_{t+h-l-j}) - (\hat{B}(k,1)-B(k,1))X_{t+h-l-1,k})) X'_{t,k}\} \| \right)$$

$$\times\{k^{1/2} \parallel (\hat{\Gamma}_k^{-1} - \Gamma_k^{-1}) \parallel_1\}$$

It has already been shown that

$$\parallel \hat{\Theta}_l \parallel \xrightarrow{p} \parallel \Theta_l \parallel < \infty,$$

for each $1 \leq l \leq h-1$. And we know from Lewis and Reinsel (1985) that $k^{1/2} \parallel (\hat{\Gamma}_k^{-1} - \Gamma_k^{-1}) \parallel_1 \xrightarrow{p} 0$. Since $h-1$ is finite, to show

$$\parallel \sqrt{T-k-H} U_{4T} (\hat{\Gamma}_k^{-1} - \Gamma_k^{-1}) \parallel \xrightarrow{p} 0,$$

I just need to show that

$$\Big(\parallel \{[k(T-k-H)]^{-1/2} \sum_{t=k}^{T-H} (\varepsilon_{t+h-l} + (\sum_{j=k+1}^{\infty} A_j y_{t+h-l-j}) - (\hat{B}(k,1) - B(k,1)) X_{t+h-l-1,k})) X'_{t,k}\} \parallel \Big)$$

is bounded for each $1 \leq l \leq h-1$.

$$\Big(\parallel \{[k(T-k-H)]^{-1/2} \sum_{t=k}^{T-H} (\varepsilon_{t+h-l} + (\sum_{j=k+1}^{\infty} A_j y_{t+h-l-j}) - (\hat{B}(k,1) - B(k,1)) X_{t+h-l-1,k})) X'_{t,k}\} \parallel \Big)$$

$$\leq \parallel [k(T-k-H)]^{-1/2} \sum_{t=k}^{T-H} \varepsilon_{t+h-l} X'_{t,k} \parallel +$$

$$\parallel [k(T-k-H)]^{-1/2} \sum_{t=k}^{T-H} (\sum_{j=k+1}^{\infty} A_j y_{t+h-l-j}) X'_{t,k} \parallel - \parallel [k(T-k-H)]^{-1/2} \sum_{t=k}^{T-H} (\hat{B}(k,1) - B(k,1)) X_{t+h-l-1,k} X'_{t,k} \parallel .$$

The first term is bounded since it was shown in the proof of consistency that

$$\parallel (T-k-H)^{-1} \sum_{t=k}^{T-H} \varepsilon_{t+h-l} X'_{t,k} \parallel = O_p(\Big(\frac{k}{T-k-H}\Big)^{1/2}).$$

Jordà and Kozicki (2011) show that the second term converges in probability to 0. For the final term note that

$$\parallel [k(T-k-H)]^{-1/2} \sum_{t=k}^{T-H} (\hat{B}(k,1) - B(k,1)) X_{t+h-l-1,k} X'_{t,k} \parallel$$

$$\leq \underbrace{\Big(\frac{T-k-H}{k}\Big)^{1/2} \parallel (\hat{B}(k,1) - B(k,1)) \parallel}_{bounded} \underbrace{\parallel (T-k-H)^{-1} \sum_{t=k}^{T-H} X_{t+h-l-1,k} X'_{t,k} \parallel_1}_{bounded} .$$

Consequently

$$\parallel \sqrt{T-k-H} U_{4T} (\hat{\Gamma}_k^{-1} - \Gamma_k^{-1}) \parallel \xrightarrow{p} 0,$$

and this completes the proof showing

$$\sqrt{T-k-H}[\hat{B}(k,h,GLS)-B(k,h)] \xrightarrow{p} \sqrt{T-k-H}[U_{1T}\Gamma_k^{-1}+U_{2T}\Gamma_k^{-1}+U_{3T}\Gamma_k^{-1}-U_{4T}\Gamma_k^{-1}].$$

From Jordà and Kozicki (2011) we know that

$$\| \sqrt{T-k-H}U_{1T}\Gamma_k^{-1} \| \xrightarrow{p} 0.$$

As a result

$$\sqrt{T-k-H}[\hat{B}(k,h,GLS)-B(k,h)] \xrightarrow{p} \sqrt{T-k-H}[U_{2T}\Gamma_k^{-1}+U_{3T}\Gamma_k^{-1}-U_{4T}\Gamma_k^{-1}].$$

Therefore

$$\sqrt{T-k-H}[\hat{B}(k,h,GLS)-B(k,h)] \xrightarrow{p} (T-k-H)^{-1/2}(\sum_{t=k}^{T-H}\varepsilon_{t+h}X_{t,k}')\Gamma_k^{-1}+(T-k-H)^{-1/2}\sum_{t=k}^{T-H}(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{t+h-l})X_{t,k}'\Gamma_k^{-1}$$

$$-(T-k-H)^{-1/2}\sum_{t=k}^{T-H}(\sum_{l=1}^{h-1}\hat{\Theta}_l\hat{\varepsilon}_{t+h-l})X_{t,k}'\Gamma_k^{-1}.$$

Since

$$\hat{\varepsilon}_t = \varepsilon_t + (\sum_{j=1}^{\infty}A_j y_{t-j}) - (\sum_{i=1}^{k}\hat{A}_i y_{t-i})$$

$$= \varepsilon_t + (\sum_{j=k+1}^{\infty}A_j y_{t-j}) - (\hat{B}(k,1)-B(k,1))X_{t-1,k}),$$

it can be shown that

$$(T-k-H)^{-1/2}\sum_{t=k}^{T-H}(\sum_{l=1}^{h-1}\hat{\Theta}_l\hat{\varepsilon}_{t+h-l})X_{t,k}'\Gamma_k^{-1} \xrightarrow{p} (T-k-H)^{-1/2}\sum_{t=k}^{T-H}(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{t+h-l})X_{t,k}'\Gamma_k^{-1}$$

$$-\sum_{l=1}^{h-1}\Theta_l\sqrt{T-k-H}(\hat{B}(k,1)-B(k,1))\Gamma_{(h-l-1),k}\Gamma_k^{-1},$$

(the proof is omitted for brevity). Therefore

$$\sqrt{T-k-H}[\hat{B}(k,h,GLS)-B(k,h)] \xrightarrow{p} (T-k-H)^{-1/2}(\sum_{t=k}^{T-H}\varepsilon_{t+h}X_{t,k}')\Gamma_k^{-1}+\sum_{l=1}^{h-1}\Theta_l\sqrt{T-k-H}(\hat{B}(k,1)-B(k,1))\Gamma_{(h-l-1),k}\Gamma_k^{-1}.$$

Note that

$$vec[(T-k-H)^{-1/2}\{\sum_{t=k}^{T-H}\varepsilon_{t+h}X_{t,k}'\}\Gamma_k^{-1}+\sum_{l=1}^{h-1}\Theta_l\sqrt{T-k-H}(\hat{B}(k,1)-B(k,1))\Gamma_{(h-l-1),k}\Gamma_k^{-1}]$$

$$= vec[(T-k-H)^{-1/2}\{\sum_{t=k}^{T-H}\varepsilon_{t+h}X'_{t,k}\}\Gamma_k^{-1}] + vec[\sum_{l=1}^{h-1}\Theta_l\sqrt{T-k-H}(\hat{B}(k,1)-B(k,1))\Gamma_{(h-l-1),k}\Gamma_k^{-1}]$$

$$= \{I_{kr}\otimes I_r\}vec[(T-k-H)^{-1/2}\{\sum_{t=k}^{T-H}\varepsilon_{t+h}X'_{t,k}\}\Gamma_k^{-1}] + \Big(\sum_{l=1}^{h-1}\{\Gamma_k^{-1}\Gamma'_{(h-l-1),k}\otimes\Theta_l\}\Big)vec[\sqrt{T-k-H}(\hat{B}(k,1)-B(k,1))]$$

To show that

$$\sqrt{T-k-H}vec[\hat{B}(k,h,GLS)-B(k,h)] \xrightarrow{d} N(0,\Omega_h^{GLS}),$$

it suffices to show that the joint distribution of

$$\begin{bmatrix} \{I_{kr}\otimes I_r\}vec[(T-k-H)^{-1/2}\{\sum_{t=k}^{T-H}\varepsilon_{t+h}X'_{t,k}\}\Gamma_k^{-1}] \\ \Big(\sum_{l=1}^{h-1}\{\Gamma_k^{-1}\Gamma'_{(h-l-1),k}\otimes\Theta_l\}\Big)vec[\sqrt{T-k-H}(\hat{B}(k,1)-B(k,1))] \end{bmatrix}$$

converge to a normal distribution.

$$\begin{bmatrix} \{I_{kr}\otimes I_r\}vec[(T-k-H)^{-1/2}\{\sum_{t=k}^{T-H}\varepsilon_{t+h}X'_{t,k}\}\Gamma_k^{-1}] \\ \Big(\sum_{l=1}^{h-1}\{\Gamma_k^{-1}\Gamma'_{(h-l-1),k}\otimes\Theta_l\}\Big)vec[\sqrt{T-k-H}(\hat{B}(k,1)-B(k,1))] \end{bmatrix}$$

$$= \begin{bmatrix} \{I_{kr}\otimes I_r\} & 0 \\ 0 & \Big(\sum_{l=1}^{h-1}\{\Gamma_k^{-1}\Gamma'_{(h-l-1),k}\otimes\Theta_l\}\Big) \end{bmatrix} \begin{bmatrix} vec[(T-k-H)^{-1/2}\{\sum_{t=k}^{T-H}\varepsilon_{t+h}X'_{t,k}\}\Gamma_k^{-1}] \\ vec[\sqrt{T-k-H}(\hat{B}(k,1)-B(k,1))] \end{bmatrix}$$

Define

$$s_T = \begin{bmatrix} \{I_{kr}\otimes I_r\} & 0 \\ 0 & \Big(\sum_{l=1}^{h-1}\{\Gamma_k^{-1}\Gamma'_{(h-l-1),k}\otimes\Theta_l\}\Big) \end{bmatrix} \begin{bmatrix} vec[(T-k-H)^{-1/2}\{\sum_{t=k}^{T-H}\varepsilon_{t+h}X'_{t,k}\}\Gamma_k^{-1}] \\ vec[(T-k-H)^{-1/2}\{\sum_{t=k}^{T-H}\varepsilon_{t+1}X'_{t,k}\}\Gamma_k^{-1}] \end{bmatrix}.$$

To show asymptotic normality of s_T, the Gordin's CLT will be used. Using similar reasoning as for standard LP,

$$\Big\{\begin{bmatrix} \{I_{kr}\otimes I_r\} & 0 \\ 0 & \Big(\sum_{l=1}^{h-1}\{\Gamma_k^{-1}\Gamma'_{(h-l-1),k}\otimes\Theta_l\}\Big) \end{bmatrix} \begin{bmatrix} vec[\varepsilon_{t+h}X'_{t,k}\Gamma_k^{-1}] \\ vec[\varepsilon_{t+1}X'_{t,k}\Gamma_k^{-1}] \end{bmatrix}\Big\}_{t=-\infty}^{t=\infty}$$

is a strongly stationary and ergodic sequence.

To show asymptotic normality for LP GLS, it needs to be shown that

$$s_T = \begin{bmatrix} \{I_{kr}\otimes I_r\} & 0 \\ 0 & \Big(\sum_{l=1}^{h-1}\{\Gamma_k^{-1}\Gamma'_{(h-l-1),k}\otimes\Theta_l\}\Big) \end{bmatrix} \begin{bmatrix} vec[(T-k-H)^{-1/2}\{\sum_{t=k}^{T-H}\varepsilon_{t+h}X'_{t,k}\}\Gamma_k^{-1}] \\ vec[(T-k-H)^{-1/2}\{\sum_{t=k}^{T-H}\varepsilon_{t+1}X'_{t,k}\}\Gamma_k^{-1}] \end{bmatrix}$$

is normally distributed. Since s_T is a strongly stationary and ergodic sequence, all that is left is to show the following conditions are satisfied:

1. Asymptotic uncorrelatedness

2. Summability of autocovariances

3. Asymptotic negligibility of innovations.

Asymptotic uncorrelatedness follows along the same lines as the standard LP and is omitted for brevity.

To show Summability of autocovariances, must show that

$$lim_{T\to\infty} var(s_T)$$

is finite and constant. Note that

$$\begin{bmatrix} \{I_{kr}\otimes I_r\} & 0 \\ 0 & \Big(\sum_{l=1}^{h-1}\{\Gamma_k^{-1}\Gamma'_{(h-l-1),k}\otimes\Theta_l\}\Big) \end{bmatrix}\begin{bmatrix} vec[\varepsilon_{t+h}X'_{t,k}\Gamma_k^{-1}] \\ vec[\varepsilon_{t+1}X'_{t,k}\Gamma_k^{-1}] \end{bmatrix}$$

$$= \begin{bmatrix} \{I_{kr}\otimes I_r\} & 0 \\ 0 & \Big(\sum_{l=1}^{h-1}\{\Gamma_k^{-1}\Gamma'_{(h-l-1),k}\otimes\Theta_l\}\Big) \end{bmatrix}\begin{bmatrix} (\Gamma_k^{-1}X_{t,k}\otimes I_r)vec[\varepsilon_{t+h}] \\ (\Gamma_k^{-1}X_{t,k}\otimes I_r)vec[\varepsilon_{t+1}] \end{bmatrix}$$

$$= \begin{bmatrix} \{I_{kr}\otimes I_r\} & 0 \\ 0 & \Big(\sum_{l=1}^{h-1}\{\Gamma_k^{-1}\Gamma'_{(h-l-1),k}\otimes\Theta_l\}\Big) \end{bmatrix}\begin{bmatrix} (\Gamma_k^{-1}X_{t,k}\otimes I_r)\varepsilon_{t+h} \\ (\Gamma_k^{-1}X_{t,k}\otimes I_r)\varepsilon_{t+1} \end{bmatrix}$$

$$= \begin{bmatrix} (\Gamma_k^{-1}X_{t,k}\otimes I_r)\varepsilon_{t+h} \\ \Big(\sum_{l=1}^{h-1}\{\Gamma_k^{-1}\Gamma'_{(h-l-1),k}\otimes\Theta_l\}\Big)(\Gamma_k^{-1}X_{t,k}\otimes I_r)\varepsilon_{t+1} \end{bmatrix}$$

The autocovariances for lag $m-n$ is

$$E[\begin{bmatrix} (\Gamma_k^{-1}X_{m,k}\otimes I_r)\varepsilon_{m+h} \\ l_k(\Gamma_k^{-1}X_{m,k}\otimes I_r)\varepsilon_{m+1} \end{bmatrix}\begin{bmatrix} (\Gamma_k^{-1}X_{n,k}\otimes I_r)\varepsilon_{n+h} \\ l_k(\Gamma_k^{-1}X_{n,k}\otimes I_r)\varepsilon_{n+1} \end{bmatrix}']$$

$$= E[\begin{bmatrix} (\Gamma_k^{-1}X_{m,k}\otimes I_r)\varepsilon_{m+h} \\ l_k(\Gamma_k^{-1}X_{m,k}\otimes I_r)\varepsilon_{m+1} \end{bmatrix}\begin{bmatrix} \varepsilon'_{n+h}(\Gamma_k^{-1}X_{n,k}\otimes I_r)' & \varepsilon'_{n+1}(\Gamma_k^{-1}X_{n,k}\otimes I_r)'l'_k \end{bmatrix}]$$

$$= E\begin{bmatrix} (\Gamma_k^{-1}X_{m,k}\otimes I_r)\varepsilon_{m+h}\varepsilon'_{n+h}(\Gamma_k^{-1}X_{n,k}\otimes I_r)' & (\Gamma_k^{-1}X_{m,k}\otimes I_r)\varepsilon_{m+h}\varepsilon'_{n+1}(\Gamma_k^{-1}X_{n,k}\otimes I_r)'l'_k \\ l_k(\Gamma_k^{-1}X_{m,k}\otimes I_r)\varepsilon_{m+1}\varepsilon'_{n+h}(\Gamma_k^{-1}X_{n,k}\otimes I_r)' & l_k(\Gamma_k^{-1}X_{m,k}\otimes I_r)\varepsilon_{m+1}\varepsilon'_{n+1}(\Gamma_k^{-1}X_{n,k}\otimes I_r)'l'_k \end{bmatrix}$$

where $l_k = \Big(\sum_{l=1}^{h-1}\{\Gamma_k^{-1}\Gamma'_{(h-l-1),k}\otimes\Theta_l\}\Big)$. Now taking the conditional expectation based on the time n information set, $\mathcal{F}_n$,

$$= E\Big(\begin{bmatrix} (\Gamma_k^{-1}X_{m,k}\otimes I_r)\varepsilon_{m+h}\varepsilon'_{n+h}(\Gamma_k^{-1}X_{n,k}\otimes I_r)' & (\Gamma_k^{-1}X_{m,k}\otimes I_r)\varepsilon_{m+h}\varepsilon'_{n+1}(\Gamma_k^{-1}X_{n,k}\otimes I_r)'l'_k \\ l_k(\Gamma_k^{-1}X_{m,k}\otimes I_r)\varepsilon_{m+1}\varepsilon'_{n+h}(\Gamma_k^{-1}X_{n,k}\otimes I_r)' & l_k(\Gamma_k^{-1}X_{m,k}\otimes I_r)\varepsilon_{m+1}\varepsilon'_{n+1}(\Gamma_k^{-1}X_{n,k}\otimes I_r)'l'_k \end{bmatrix} | \mathcal{F}_n\Big)$$

$$= \Big(\begin{bmatrix} (\Gamma_k^{-1} X_{m,k} \otimes I_r)\Sigma_{\varepsilon,(m-n)}(\Gamma_k^{-1} X_{n,k} \otimes I_r)' & (\Gamma_k^{-1} X_{m,k} \otimes I_r)\Sigma_{\varepsilon,(m+1-n-h)}(\Gamma_k^{-1} X_{n,k} \otimes I_r)' l_k' \\ l_k(\Gamma_k^{-1} X_{m,k} \otimes I_r)\Sigma_{\varepsilon,(m+1-n-h)}(\Gamma_k^{-1} X_{n,k} \otimes I_r)' & l_k(\Gamma_k^{-1} X_{m,k} \otimes I_r)\Sigma_{\varepsilon,(m-n)}(\Gamma_k^{-1} X_{n,k} \otimes I_r)' l_k' \end{bmatrix} \Big)$$

$$= \Big(\begin{bmatrix} (\Gamma_k^{-1} X_{m,k} \otimes I_r)(1 \otimes \Sigma_{\varepsilon,(m-n)})(\Gamma_k^{-1} X_{n,k} \otimes I_r)' & (\Gamma_k^{-1} X_{m,k} \otimes I_r)(1 \otimes \Sigma_{\varepsilon,(m+1-n-h)})(\Gamma_k^{-1} X_{n,k} \otimes I_r)' l_k' \\ l_k(\Gamma_k^{-1} X_{m,k} \otimes I_r)(1 \otimes \Sigma_{\varepsilon,(m+1-n-h)})(\Gamma_k^{-1} X_{n,k} \otimes I_r)' & l_k(\Gamma_k^{-1} X_{m,k} \otimes I_r)(1 \otimes \Sigma_{\varepsilon,(m-n)})(\Gamma_k^{-1} X_{n,k} \otimes I_r)' l_k' \end{bmatrix} \Big)$$

$$= \Big(\begin{bmatrix} (\Gamma_k^{-1} X_{m,k} \otimes \Sigma_{\varepsilon,(m-n)})(\Gamma_k^{-1} X_{n,k} \otimes I_r)' & (\Gamma_k^{-1} X_{m,k} \otimes \Sigma_{m+h,n+1})(\Gamma_k^{-1} X_{n,k} \otimes I_r)' l_k' \\ l_k(\Gamma_k^{-1} X_{m,k} \otimes \Sigma_{m+1,n+h})(\Gamma_k^{-1} X_{n,k} \otimes I_r)' & l_k(\Gamma_k^{-1} X_{m,k} \otimes \Sigma_{\varepsilon,(m-n)})(\Gamma_k^{-1} X_{n,k} \otimes I_r)' l_k' \end{bmatrix} \Big)$$

$$= \Big(\begin{bmatrix} (\Gamma_k^{-1} X_{m,k} X_{n,k}' \Gamma_k^{-1} \otimes \Sigma_{\varepsilon,(m-n)}) & (\Gamma_k^{-1} X_{m,k} X_{n,k}' \Gamma_k^{-1} \otimes \Sigma_{\varepsilon,(m+1-n-h)}) l_k' \\ l_k(\Gamma_k^{-1} X_{m,k} X_{n,k}' \Gamma_k^{-1} \otimes \Sigma_{\varepsilon,(m+1-n-h)}) & l_k(\Gamma_k^{-1} X_{m,k} X_{n,k}' \Gamma_k^{-1} \otimes \Sigma_{\varepsilon,(m-n)}) l_k' \end{bmatrix} \Big)$$

It follows that

$$E\Big(\begin{bmatrix} (\Gamma_k^{-1} X_{m,k} X_{n,k}' \Gamma_k^{-1} \otimes \Sigma_{\varepsilon,(m-n)}) & (\Gamma_k^{-1} X_{m,k} X_{n,k}' \Gamma_k^{-1} \otimes \Sigma_{\varepsilon,(m+1-n-h)}) l_k' \\ l_k(\Gamma_k^{-1} X_{m,k} X_{n,k}' \Gamma_k^{-1} \otimes \Sigma_{\varepsilon,(m+1-n-h)}) & l_k(\Gamma_k^{-1} X_{m,k} X_{n,k}' \Gamma_k^{-1} \otimes \Sigma_{\varepsilon,(m-n)}) l_k' \end{bmatrix} \Big)$$

$$= \Big(\begin{bmatrix} (\Gamma_k^{-1} \Gamma_{(m-n),k} \Gamma_k^{-1} \otimes \Sigma_{\varepsilon,(m-n)}) & (\Gamma_k^{-1} \Gamma_{(m-n),k} \Gamma_k^{-1} \otimes \Sigma_{\varepsilon,(m+1-n-h)}) l_k' \\ l_k(\Gamma_k^{-1} \Gamma_{(m-n),k} \Gamma_k^{-1} \otimes \Sigma_{\varepsilon,(m+1-n-h)}) & l_k(\Gamma_k^{-1} \Gamma_{(m-n),k} \Gamma_k^{-1} \otimes \Sigma_{\varepsilon,(m-n)}) l_k' \end{bmatrix} \Big)$$

which is finite since $\Gamma_k^{-1}, \Gamma_{(m-n),k}, \Sigma_{\varepsilon,(m-n)}, \Sigma_{\varepsilon,(m+1-n-h)}, and\ l_k$ are bounded in probability. Therefore, the autocovariances of

$$\begin{bmatrix} (\Gamma_k^{-1} X_{t,k} \otimes I_r)\varepsilon_{t+h} \\ l_k(\Gamma_k^{-1} X_{t,k} \otimes I_r)\varepsilon_{t+1} \end{bmatrix}$$

are finite at all leads and lags.

For notational brevity let

$$q_m = \begin{bmatrix} (\Gamma_k^{-1} X_{m,k} \otimes I_r)\varepsilon_{m+h} \\ l_k(\Gamma_k^{-1} X_{m,k} \otimes I_r)\varepsilon_{m+1} \end{bmatrix}.$$

Note that

$$var(s_T) = (T-k-H)^{-1} \sum_{m=k}^{T-H} \sum_{n=k}^{T-H} E[q_m q_n'].$$

For $|n-m| \geq h$, $E[q_m q_n'] = 0$ due to independence so

$$var(s_T) = (T-k-H)^{-1} \sum_{m=k}^{T-H} \sum_{|n-m|<h} E[q_m q_n'].$$

Since the expectations are constant and $\frac{h}{T} \to 0$,

$$lim_{T\to\infty} var(s_T) = \sum_{|n-m|<h} E[q_m q_n'].$$

which is finite. This completes the proof of summability of autocovariances.

It was shown in the proof of summability of autocovariances that q_t is stationary. Hence asymptotic negligibility of innovations follows along the same lines as the proof of asymptotic normality for standard LP, so it is omitted for brevity. □

Proof of Asymptotic Efficiency LP GLS Relative to LP OLS

Theorem 4. *Under Assumption 3,*

$$var\{\sqrt{T-k-H}vec[\hat{B}(k,h,GLS) - B(k,h)]\} - var\{\sqrt{T-k-H}vec[\hat{B}(k,h,OLS) - B(k,h)]\} \leq 0$$

in the negative semi-definite sense. That is, the GLS estimator is more efficient than the OLS estimator.

Proof. The Wold representation can be inverted into an infinite order VAR representation

$$y_t = \sum_{j=1}^{\infty} A_j y_{t-j} + \varepsilon_t.$$

Any VAR(p) (including a VAR(∞)) can be written as a companion VAR(1). Denote this VAR(1) as

$$Y_t = AY_{t-1} + Z_t.$$

Take the eigenvalue decomposition of $A = E\Lambda E^{-1}$, where Λ is the diagonal matrix of distinct nonzero eigenvalues and E is the corresponding eigenmatrix and $EE^{-1} = I$ where I is the identity matrix. As a result $A^h = E\Lambda^h E^{-1}$. Define $W_t = E^{-1}Y_t$ and $\eta_t = E^{-1}Z_t$. This implies the VAR can be transformed into

$$W_t = \Lambda W_{t-1} + \eta_t.$$

Consequently

$$W_{t+h} = \Lambda^{h+1}W_{t-1} + \Lambda^h\eta_t + \ldots + \Lambda\eta_{t+h-1} + \eta_{t+h}.$$

Theorems 2 and 3 establish the consistency and asymptotic normality of LP OLS and LP GLS. If I can show the limiting distribution of GLS estimator is more efficient than the limiting distribution of OLS estimator for a stationary VAR(1) model at every horizon, it follows that the LP GLS estimator is asymptotically more

efficient than the LP OLS estimator, since the mapping function from the LP estimates to the Wold coefficients is continuous and differentiable. Define

$$\sqrt{T-k-H}q = \sqrt{T-k-H}[\hat{B}(k,h,OLS) - \hat{B}(k,h,GLS)].$$

Note that

$$lim_{T\to\infty} var[\sqrt{T-H}vec\{\hat{B}(k,h,OLS) - B(k,h)\}]$$

$$= lim_{T\to\infty}\{var[\sqrt{T-H}vec\{\hat{B}(k,h,GLS) - B(k,h)\} + \sqrt{T-H}vec\{q\}]$$

$$= lim_{T\to\infty}\{var[\sqrt{T-k-H}vec[\hat{B}(k,h,GLS) - B(k,h)] + var[\sqrt{T-H}vec(q)]$$

$$+cov[\sqrt{T-k-H}[vec\{\hat{B}(k,h,GLS)-B(k,h)\}, \sqrt{T}vec\{q\}]+cov[\sqrt{T-k-H}[vec\{\hat{B}(k,h,GLS)-B(k,h)\}, \sqrt{T}vec\{q\}]'\}.$$

To show that LP GLS is more efficient, it suffices to show that

$$lim_{T\to\infty} cov[\sqrt{T-k-H}vec\{q\}, \sqrt{T-k-H}vec\{\hat{B}(k,h,GLS) - B(k,h)\}] \geq 0,$$

in the positive semi-definite sense. Note that

$$\sqrt{T-k-H}[\hat{B}(k,h,GLS)-B(k,h)] \xrightarrow{p} (T-k-H)^{-1/2}(\sum_{t=k}^{T-H}\varepsilon_{t+h}X'_{t,k})\Gamma_k^{-1}+\sum_{l=1}^{h-1}\Theta_l\sqrt{T-k-H}(\hat{B}(k,1)-B(k,1))\Gamma_{(h-l-1),k}\Gamma_k^{-1},$$

$$\sqrt{T-k-H}q \xrightarrow{p} \sqrt{T-k-H}U_{4T}\Gamma_k^{-1},$$

$$\sqrt{T-k-H}U_{4T}\Gamma_k^{-1} \xrightarrow{p} (T-k-H)^{-1/2}\sum_{t=k}^{T-H}(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{t+h-l})X'_{t,k}\Gamma_k^{-1}-\sum_{l=1}^{h-1}\Theta_l\sqrt{T-k-H}(\hat{B}(k,1)-B(k,1))\Gamma_{(h-l-1),k}\Gamma_k^{-1}.$$

So

$$lim_{T\to\infty} cov[\sqrt{T-k-H}vec\{U_{4T}\Gamma_k^{-1}\}, \sqrt{T-k-H}vec\{\hat{B}(k,h,GLS) - B(k,h)\}]$$

$$= (T-k-H)^{-1}\sum_{m=k}^{T-H}\sum_{n=k}^{T-H}E[(\Gamma_k^{-1}X_{m,k}\otimes I_r)(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{m+h-l})\varepsilon'_{n+h}(\Gamma_k^{-1}X_{n,k}\otimes I_r)']$$

$$+(T-k-H)^{-1}\sum_{m=k}^{T-H}\sum_{n=k}^{T-H}E[(\Gamma_k^{-1}X_{m,k}\otimes I_r)(\sum_{l=1}^{h-1}\Theta_l\varepsilon_{m+h-l})\varepsilon'_{n+1}(\Gamma_k^{-1}X_{n,k}\otimes I_r)'\Big(\sum_{l=1}^{h-1}\{\Gamma_k^{-1}\Gamma'_{(h-l-1),k}\otimes\Theta_l\}\Big)']$$

$$-(T-k-H)^{-1}\sum_{m=k}^{T-H}\sum_{n=k}^{T-H}E[\Big(\sum_{l=1}^{h-1}\{\Gamma_k^{-1}\Gamma'_{(h-l-1),k}\otimes\Theta_l\}\Big)(\Gamma_k^{-1}X_{m,k}\otimes I_r)\varepsilon_{m+1}\varepsilon'_{n+h}(\Gamma_k^{-1}X_{n,k}\otimes I_r)']$$

$$-(T-k-H)^{-1}\sum_{m=k}^{T-H}\sum_{n=k}^{T-H}E[\Big(\sum_{l=1}^{h-1}\{\Gamma_k^{-1}\Gamma'_{(h-l-1),k}\otimes\Theta_l\}\Big)(\Gamma_k^{-1}X_{m,k}\otimes I_r)\varepsilon_{m+1}\varepsilon'_{n+1}(\Gamma_k^{-1}X_{n,k}\otimes I_r)'\Big(\sum_{l=1}^{h-1}\{\Gamma_k^{-1}\Gamma'_{(h-l-1),k}\otimes\Theta_l\}\Big)'].$$

By independence and since the expectations are finite

$$= \sum_{l=1}^{h-1} \Big((\Gamma_k^{-1} \Gamma'_{k,(-l)} \Gamma_k^{-1} \otimes \Theta_l \Sigma_\varepsilon) \Big)$$

$$+ \sum_{l=1}^{h-1} \Big((\Gamma_k^{-1} \Gamma'_{k,(h-l-1)} \Gamma_k^{-1} \otimes \Theta_l \Sigma_\varepsilon) \Big) \Big(\sum_{l=1}^{h-1} \{ \Gamma_k^{-1} \Gamma'_{(h-l-1),k} \otimes \Theta_l \} \Big)'$$

$$- [\Big(\sum_{l=1}^{h-1} \{ \Gamma_k^{-1} \Gamma'_{(h-l-1),k} \otimes \Theta_l \} \Big) (\Gamma_k^{-1} \Gamma'_{(h-1),k} \Gamma_k^{-1} \otimes \Sigma_\varepsilon)]$$

$$- [\Big(\sum_{l=1}^{h-1} \{ \Gamma_k^{-1} \Gamma'_{(h-l-1),k} \otimes \Theta_l \} \Big) (\Gamma_k^{-1} \otimes \Sigma_\varepsilon) \Big(\sum_{l=1}^{h-1} \{ \Gamma_k^{-1} \Gamma'_{(h-l-1),k} \otimes \Theta_l \} \Big)'].$$

Second and fourth lines cancel so

$$lim_{T \to \infty} cov[\sqrt{T-k-H} vec\{ U_{4T} \Gamma_k^{-1} \}, \sqrt{T-k-H} vec\{ \hat{B}(k,h,GLS) - B(k,h) \}]$$

$$= \sum_{l=1}^{h-1} \Big((\Gamma_k^{-1} \Gamma'_{k,(-l)} \Gamma_k^{-1} \otimes \Theta_l \Sigma_\varepsilon) \Big)$$

$$- [\Big(\sum_{l=1}^{h-1} \{ \Gamma_k^{-1} \Gamma'_{(h-l-1),k} \otimes \Theta_l \} \Big) (\Gamma_k^{-1} \Gamma'_{(h-1),k} \Gamma_k^{-1} \otimes \Sigma_\varepsilon)].$$

In the case where the true model can be written as a VAR(1) which has been diagonalized then

$$lim_{T \to \infty} cov[\sqrt{T-k-H} vec\{ U_{4T} \Gamma_k^{-1} \}, \sqrt{T-k-H} vec\{ \hat{B}(k,h,GLS) - B(k,h) \}]$$

$$= \sum_{l=1}^{h-1} \Big((\Gamma_w^{-1} \Gamma_w \Lambda'^l \Gamma^{-1} \otimes \Lambda^l \Sigma_\eta) \Big) - [\Big(\sum_{l=1}^{h-1} \{ \Gamma_w^{-1} \Gamma'_w \Lambda'^{h-l-1} \otimes \Lambda^l \} \Big) (\Gamma_w^{-1} \Gamma'_w \Lambda'^{h-1} \Gamma_w^{-1} \otimes \Sigma_\eta)]$$

$$= \sum_{l=1}^{h-1} \Big((\Lambda'^l \Gamma_w^{-1} \otimes \Lambda^l \Sigma_\eta) \Big) - [\Big(\sum_{l=1}^{h-1} \{ \Lambda'^{h-l-1} \otimes \Lambda^l \} \Big) (\Lambda'^{h-1} \Gamma_w^{-1} \otimes \Sigma_\eta)].$$

$$= \sum_{l=1}^{h-1} \Big((\Lambda'^l \Gamma_w^{-1} \otimes \Lambda^l \Sigma_\eta) \Big) - [\Big(\sum_{l=1}^{h-1} \{ \Lambda^{2h-l-2} \Gamma_w^{-1} \otimes \Lambda^l \Sigma_\eta \} \Big)]$$

$$= \sum_{l=1}^{h-1} \Big((\Lambda'^l - \Lambda^{2h-l-2}) \Gamma_w^{-1} \otimes \Lambda^l \Sigma_\eta) \Big)$$

where $E(W_t W_t') = \Gamma_w$, and since the model is a VAR(1), $E(W_t W_{t-j}') = \Lambda^j \Gamma_w$. Note that the dimensions of the parameters have been suppressed for simplicity. Premultiply corresponding terms in the sum by identity matrix

$$(\Lambda^l \otimes \Lambda^{-l}) = I$$

yields

$$\sum_{l=1}^{h-1}\left((\Lambda^{2l}-\Lambda^{2h-2})\Gamma_w^{-1}\otimes\Sigma_\eta\right)$$

$$=\sum_{l=1}^{h-1}(\Lambda^{2l}\Gamma_w^{-1}\otimes\Sigma_\eta)-(h-1)(\Lambda^{2(h-1)}\Gamma_w^{-1}\otimes\Sigma_\eta)$$

which is positive definite since

$$\sum_{l=1}^{h-1}(\Lambda^{2l-2(h-1)}\Lambda^{2(h-1)}\Gamma_w^{-1}\otimes\Sigma_\eta)-(h-1)(\Lambda^{2(h-1)}\Gamma_w^{-1}\otimes\Sigma_\eta)\geq 0$$

$$\sum_{l=1}^{h-1}(\Lambda^{-2(h-l-1)}\Lambda^{2(h-1)}\Gamma_w^{-1}\otimes\Sigma_\eta)-(h-1)(\Lambda^{2(h-1)}\Gamma_w^{-1}\otimes\Sigma_\eta)\geq 0$$

$$\sum_{l=1}^{h-1}(\Lambda^{-2(h-l-1)}\Lambda^{2(h-1)}\Gamma_w^{-1}\otimes\Sigma_\eta)[(h-1)(\Lambda^{2(h-1)}\Gamma_w^{-1}\otimes\Sigma_\eta)]^{-1}-I\geq 0$$

$$(h-1)^{-1}\sum_{l=1}^{h-1}(\Lambda^{-2(h-l-1)}\otimes I)-I\geq 0\ for\ h=2,3,\ldots$$

since Λ is diagonal matrix where all of diagonal elements are less than one is absolute value and $l+1\leq h$. Therefore GLS is more efficient since

$$lim_{T\to\infty}var[\sqrt{T-k-H}[\hat{B}(k,h,OLS)-B(k,h)]$$

$$=lim_{T\to\infty}\{\underbrace{var[\sqrt{T-k-H}[\hat{B}(k,h,GLS)-B(k,h)]}_{pos}+\underbrace{var[\sqrt{T-H}q]}_{pos}+\underbrace{cov[\sqrt{T-k-H}[\hat{B}(k,h,GLS)-B(k,h),\sqrt{T}q]}_{pos-semi}$$

$$+\underbrace{cov[\sqrt{T-k-H}[\hat{B}(k,h,GLS)-B(k,h),\sqrt{T}q]'}_{pos-semi}\}.$$

□

A.5 Figures

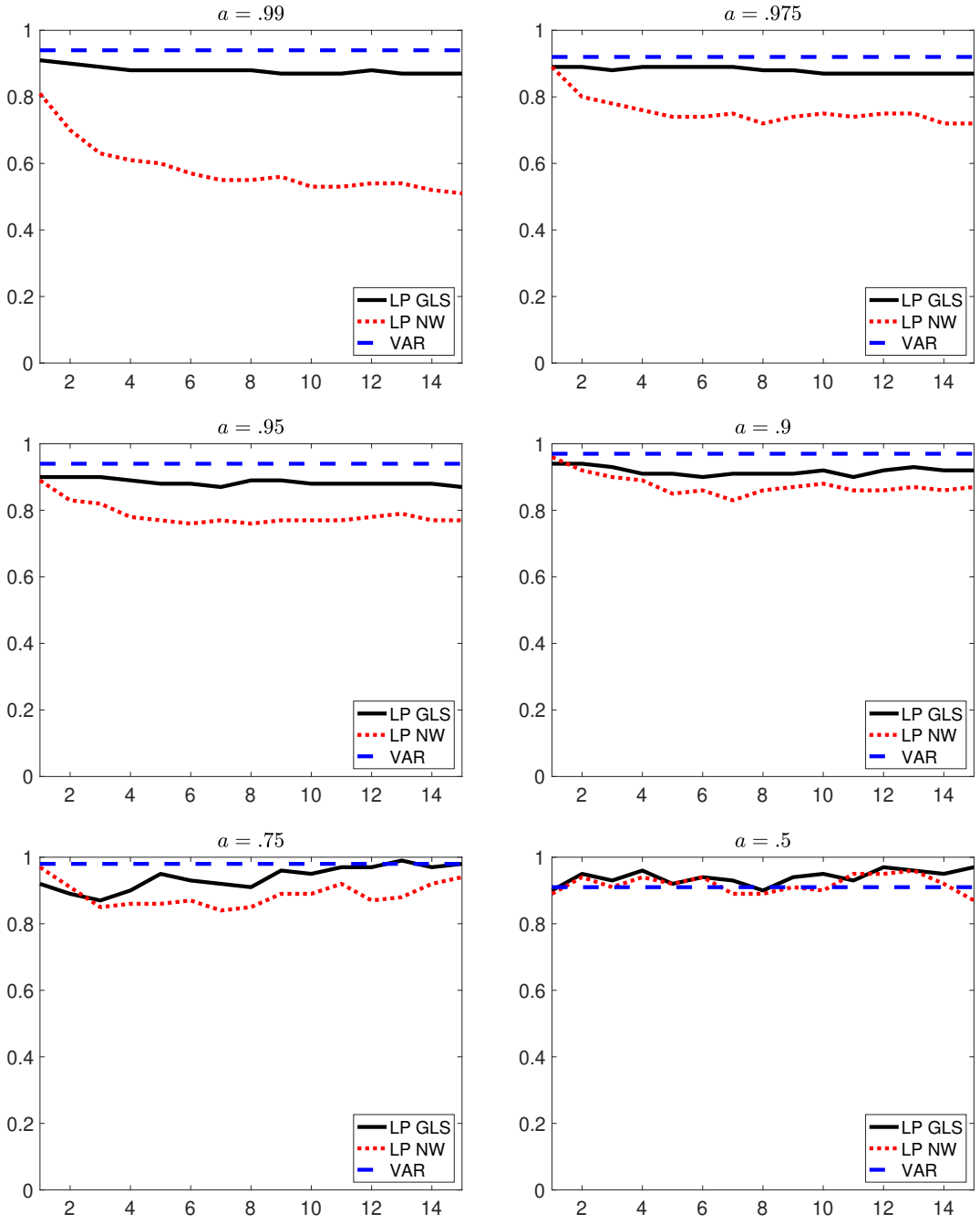

Figure 1: Coverage Rates for 95% Confidence Intervals

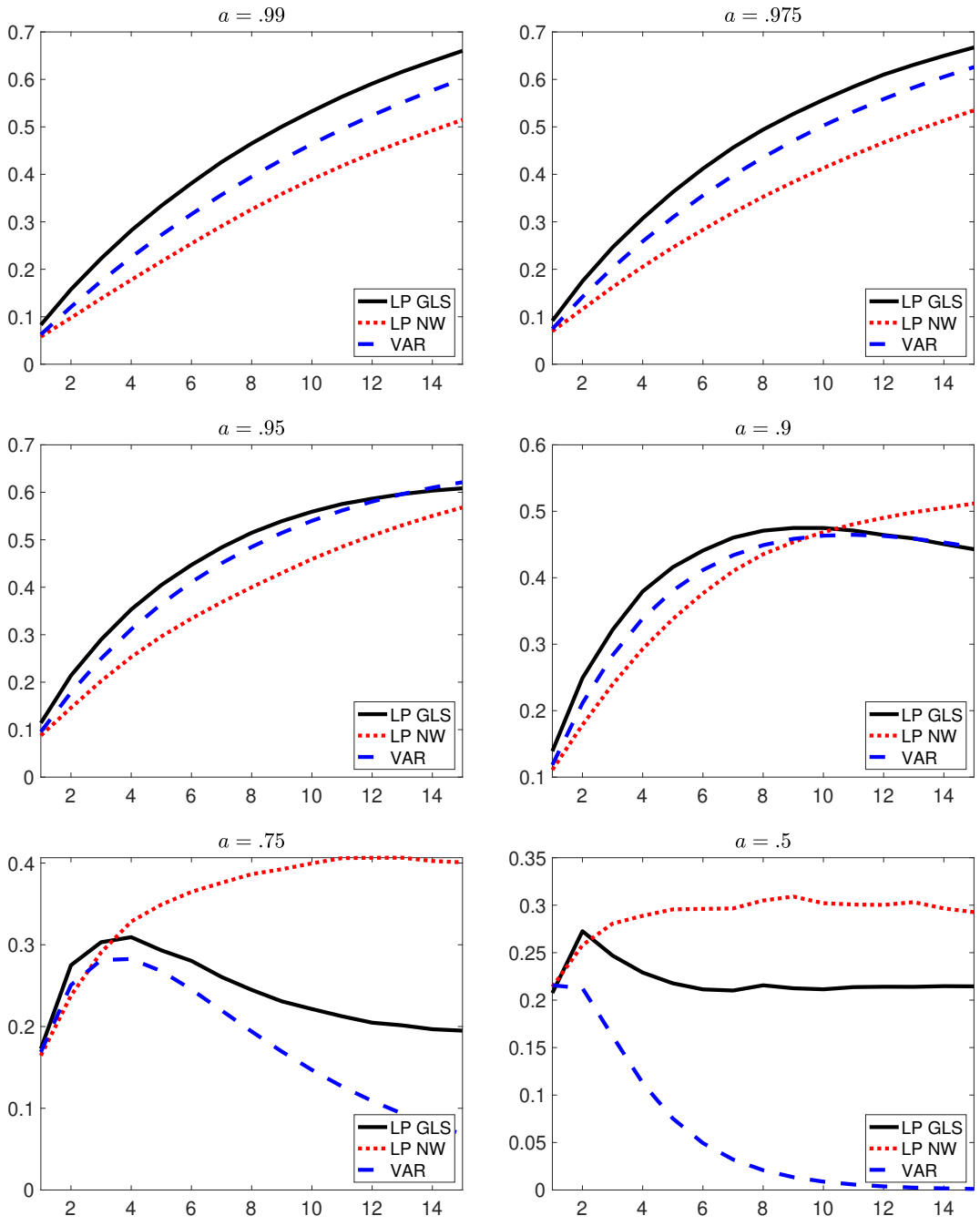

Figure 2: Average Length for 95% Confidence Intervals

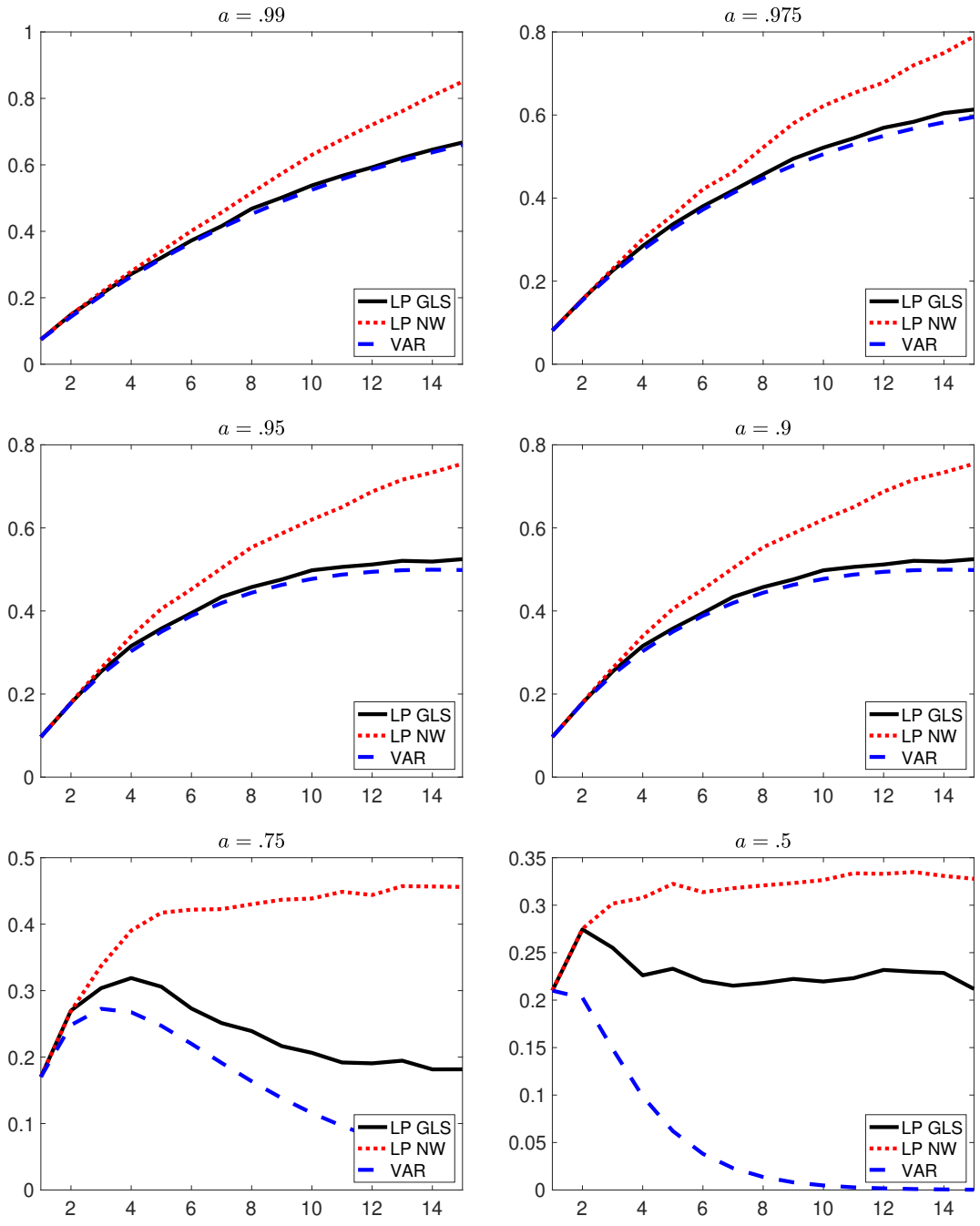

Figure 3: Monte Carlo Simulation of "True" Length for 95% Confidence Intervals