Using Kalman Filter to Extract and Test for Common Stochastic Trends

Yoosoon Chang†, Bibo Jiang‡ and Joon Y. Park§

Abstract

This paper considers a state space model with integrated latent variables. The model provides an effective framework to specify, identify and extract common stochastic trends for a set of integrated time series. The model can be readily estimated by the standard Kalman filter, whose asymptotics are fully developed in the paper. In particular, we establish the consistency and asymptotic mixed normality of the maximum likelihood estimator, which validates the use of conventional methods of inference for our model. Moreover, we show that the standard information criteria are consistent and can be used to determine the number of common stochastic trends in our model. Our simulation study clearly demonstrates the relevancy of our asymptotic theory in finite samples. For illustrations, we apply our methodology to analyze common stochastic trends in the fluctuations of macroeconomic aggregates across countries and in the prices of Dow Jones Industrial Average (DJIA) component stocks.

This Version: March 14, 2013
JEL Classification: C22, C51

Key words and phrases: state space model, Kalman filter, common stochastic trends, maximum likelihood estimation, asymptotic theory.

*We are grateful for their helpful comments to the participants at 2012 Symposium on Econometric Theory and Applications, 2009 ZEW Conference on Recent Developments in Macroeconomics, and 2008 New Zealand Econometric Study Group Meeting, and also to the seminar participants at University of Rochester, Hitotsubashi University, Hong Kong University of Science and Technology, Academia Sinica, Indiana University, Yale University, Texas A&M University and Rice University. Chang gratefully acknowledges the financial support from the NSF under Grant SES-0453069/0730152.

†Department of Economics, Indiana University
‡School of Economics, Fudan University
§Department of Economics, Indiana University and Sungkyunkwan University
1. Introduction

The Kalman filter is the basic tool used in the standard state space models, which typically deals with dynamic time series models that involve unobserved variables. The applications of Kalman filter can be found in many fields including economics and finance. The asymptotic behavior of maximum likelihood (ML) estimators based on the filter is well known under regular conditions, i.e., linearity, Gaussianity and stationarity. If linearity is violated, the extended Kalman filter is a standard alternative. Moreover, it is well known that the pseudo-ML estimation performs well when Gaussianity does not hold. To the best of our knowledge, however, no research has been done to investigate the properties of the filter for the case that stationarity is violated. Only recently, Chang, Miller and Park (2009), which will be referred to as CMP hereafter, pioneered in developing a rigorous asymptotic theory for the state space models with one integrated latent variable.

Since CMP allows for only one integrated latent factor, it does not provide any test for the number of distinct latent factors. This would certainly be a critical limitation in practical applications. In many empirical analysis, we see some strong evidence that the common stochastic trends in systems consisting of multiple integrated time series cannot be explained by a single factor. The presence of a single common stochastic trend would imply the presence of as many cointegrating relations as only one less of the total number of integrated time series included in the system. This is highly unlikely, especially when the underlying system is large with multiple integrated time series involved, as is often the case in many practical applications. The reader is referred to, e.g., Kim and Nelson (1999) for various models used in practice and previous empirical researches.

In this paper, we extend CMP to allow for multiple latent factors, and analyze the standard information criteria applied to determine the number of latent factors in our model. Our framework is completely general, except that we require the latent common factors follow random walks in a strict sense. Within this general framework, we establish that the ML estimators of the parameters in the model are consistent and asymptotically mixed normal. The standard inference based on the ML procedure is therefore valid. The convergence rate for the ML estimators is $\sqrt{n}$ as in the standard model. However, we have a faster $n$ rate of convergence for the coefficient of latent common stochastic trends along the cointegration space. This is in parallel to the convergence rates in other types of cointegrated models. Moreover, we show that various information criteria are all asymptotically consistent and can be used to determine the number of common stochastic trends in our model. The information criteria appear to be particularly useful in a large system of integrated time series which shares a relatively small number of common stochastic trends.

The state space modeling with latent integrated factors provides an alternative way of analyzing cointegrated systems. It is in contrast with the cointegrating regressions considered by, for instance, Phillips (1991) and Park (1992), but closely related to the error correction formulation used in Johansen (1988, 1991) and Ahn and Reinsel (1990). They all can be used in modeling a system of cointegrated processes which share common stochastic trends. The state space model, however, is unique and distinguishes itself from other competing models in that it may allow for the common stochastic trends to be modeled as pure random walks. As we show in the paper, the state space model with common stochastic
trends specified as pure random walks is not compatible with a finite order error correction model (ECM) or vector autoregression (VAR). Therefore, the testing procedures that are based on a finite order ECM or VAR are not applicable for the state space models we consider in the paper.

Our simulation clearly demonstrates that the asymptotic theory developed in the paper is useful to better understand the distributions of ML estimators and the performances of information criteria in finite samples. The finite sample distributions of ML estimators are symmetric and well centered as predicted by our asymptotic theories, and consequently the corresponding $t$-ratios are distributed very much like standard normal in finite samples. In determining the number of common stochastic trends, BIC performs truly well even with moderate sample size. Consistent with our asymptotic theory, BIC never over-selects the number of common stochastic trends even in relatively small samples. The under-selection problem also gets improved dramatically as the sample size increases. Indeed, in our simulation with sample size greater or equal to 50, BIC selects the number of common stochastic trends at 100% accuracy.

We provide two empirical illustrations. For the first, we apply our model and methodology to consumption and GDP series from twelve industrialized countries to analyze international consumption risk sharing. If risk sharing is indeed effective, we expect to see less common stochastic trends in consumption series than in GDP series, since some of regional or country specific shocks can be smoothed out by borrowing and lending through international financial markets. In our study, we find three common stochastic trends in consumption series and five in GDP series, which provides some evidence of international consumption risk sharing. For the second illustration, we consider the 30 individual stocks comprising Dow Jones Industrial Average (DJIA) to investigate how many common stochastic trends they have and how their common stochastic trends are related to the DJIA index. Detected are two common stochastic trends, one of which clearly dominates the other. The dominant trend has a realized sample path that is remarkably similar to the historical time series of the DJIA index. We find that the dominant first factor is the most important source of the variation in most of the individual DJ stocks. For some DJ stocks, however, the second dominant factor explains a much larger portion of their variation than the first dominant factor.

The rest of the paper is organized as follows. In Section 2, we introduce our state space model and outline the Kalman filtering technique used to estimate the model. Some preliminary results are also included in this section. Sections 3 and 4 present the main theoretical findings. In Section 3, we establish the consistency and asymptotic mixed normality of the ML estimators. The determination of the number of common stochastic trends is considered in Section 4. In particular, we analyze the standard information criteria applied to select the number of common stochastic trends in our model. Section 5 presents simulation results for the finite sample performance of our estimators and test statistic. Two empirical illustrations follow in Section 6. Here we use our methodology to investigate the common stochastic trends in a system of macroeconomic aggregates from a group of 12 industrialized countries and the 30 stocks comprising the DJIA index. Section 7 concludes the paper, and mathematical proofs are given in Appendix.
2. The Model and Preliminary Results

We consider the state space model given by

\[
\begin{align*}
    y_t &= A_0 x_t + u_t \\
    x_t &= x_{t-1} + v_t
\end{align*}
\]

under the following assumptions:

SSM1: \((y_t)\) is a \(p\)-dimensional observable time series,

SSM2: \((x_t)\) is a \(q\)-dimensional vector of latent variable,

SSM3: \(A_0\) is a \(p \times q\) matrix of unknown parameters of rank \(q\), where \(q \leq p\),

SSM4: \((u_t)\) and \((v_t)\) are \(p\)- and \(q\)-dimensional independent, identically distributed (iid) errors that are normal with mean zero and variance \(\Lambda_0\) and identity matrix \(I_q\), respectively, and independent of each other, and

SSM5: \(x_0\) is independent of \((u_t)\) and \((v_t)\), and assumed to be given.

Our model can be used to extract common stochastic trends in time series \((y_t)\). Notice that latent variable \((x_t)\) is defined as a vector of random walks, and therefore our model provides a natural way to decompose a cointegrated time series into a permanent and transitory components.

The parameter \(A_0\) and the latent common stochastic trends \((x_t)\) are not globally identified in our model. Obviously, the observable time series \((y_t)\) have the same likelihood under joint transformation

\[
A_0 \mapsto A_0 H \quad \text{and} \quad x_t \mapsto H' x_t
\]

for any \(q\)-dimensional orthogonal matrix \(H\). They are identified only up to the equivalence class defined by the transformation in (2). However, both \(A_0\) and \((x_t)\) are locally identified. Indeed, we may easily see that, for any \(q\)-dimensional orthogonal matrix \(H\), \(A_0H\) is not in the neighborhood of any \(p \times q\) matrix \(A_0\) of rank \(q\) defined by the Euclidean or any equivalent norm in the vector space of \(p \times q\) matrices. Of course, \((x_t)\) is identified locally if \(A_0\) is.

To globally identify \(A_0\), we need to introduce further restrictions. The most natural way to achieve global identification of \(A_0\) is to set

\[
A_0' \Lambda_0^{-1} A_0 = D,
\]

where \(D\) is a diagonal matrix. The motivation for this identifying restriction will be discussed in detail later. Once \(\Lambda_0\) is given, we may obtain \(A_0\) satisfying the restriction from any locally identified \(A_0\). Note that, for any locally identified \(A_0\), we may find a diagonal matrix \(D\) and an orthogonal matrix \(H\) such that \(A_0' \Lambda_0^{-1} A_0 = HDH'\), due to the spectral representation of a symmetric matrix. As is well known, the diagonal entries of \(D\) are given by the eigenvalues of \(A_0' \Lambda_0^{-1} A_0\), whose corresponding eigenvectors compose the column vectors of \(H\). It is easy to see that \(A_0H\) with this choice of \(H\) satisfies the diagonality
restriction in (3). Obviously, $A_0H$ is uniquely determined by this procedure as long as no diagonal entries of $D$ are repeated.

For more meaningful interpretations, the identifying restriction in (3) will be imposed in presenting our empirical analysis of macroeconomic aggregates across countries and DJIA component stock prices. However, in developing asymptotic theory of our model and conducting a simulation study to evaluate our testing procedure, we will not introduce any extra restrictions to globally identify $A_0$ and $(x_t)$. Our asymptotic and simulation results for $A_0$ and $(x_t)$ should therefore be interpreted as applying to any locally identified pair of $A_0$ and $(x_t)$ and their estimates obtained in the same local neighborhoods. To ease the exposition of the paper, we first assume that $q$, the dimension of $(x_t)$ and rank of $A_0$, is known, and explain how to extract $(x_t)$ and develop the asymptotic theory for the ML estimation of $A_0$. To determine $q$, we propose to use any of the conventional information criteria, whose asymptotic theory will be developed in a later section.

Throughout the paper, we will mainly look at the simple model given by (1). This is purely for expositional convenience. Our subsequent results extend trivially to a more general class of state space models with measurement equation given by

$$y_t = A_0x_t + \sum_{k=1}^{m} \Pi_k \triangle y_{t-k} + u_t, \quad (4)$$

in place of the one given in (1). The inclusion of the lagged differences of $(y_t)$ in (4) only introduces more parameters associated with the observable stationary components of the model, and would not affect our asymptotic theory in any important manner. In our subsequent development of the theory, we will mention explicitly what modifications are needed to accommodate the general model in (4). In all cases, the necessary modifications are obvious and straightforward.

The model defined in (1) can be estimated by the usual Kalman filter. Let $\mathcal{F}_t$ be the $\sigma$-field generated by $y_1, \ldots, y_t$, and for $z_t = x_t$ or $y_t$, we denote by $z_{t|s}$ the conditional expectation of $z_t$ given $\mathcal{F}_s$ and by $\Omega_{t|s}$ and $\Sigma_{t|s}$ the conditional variances of $x_t$ and $y_t$ given $\mathcal{F}_s$, respectively. The Kalman filter consists of the prediction and updating steps. For any given values of loading matrix $A$ and covariance matrix $\Lambda$, we utilize the relationships in the prediction step

$$x_{t|t-1} = x_{t-1|t-1},$$
$$y_{t|t-1} = Ax_{t|t-1},$$

and

$$\Omega_{t|t-1} = \Omega_{t-1|t-1} + I_q,$$
$$\Sigma_{t|t-1} = A\Omega_{t-1}A' + \Lambda.$$

On the other hand, the updating step relies on the relationships

$$x_{t|t} = x_{t|t-1} + \Omega_{t|t-1}A'\Sigma_{t|t-1}^{-1}(y_t - y_{t|t-1}),$$
$$\Omega_{t|t} = \Omega_{t|t-1} - \Omega_{t|t-1}A'\Sigma_{t|t-1}^{-1}A\Omega_{t|t-1}.$$
The ML estimation method is used in estimating the unknown parameters $A$ and $\Lambda$.

In many cases, the primary goal of using Kalman filter is to calculate a forecast and also the conditional variance of the observed time series $(y_t)$ as a function of previous observations. However, in the case that the value of the unobserved variable is of interest for its own sake, smoothing technique is often used. The smoothed series, which we denote as $(x_t|n)$, is estimated conditionally on the information in the entire sample - not just the information up to time $t$ only, and is accordingly given by $x_t|n = \mathbb{E}(x_t|F_n)$. The following is the key equation for smoothing:

$$x_t|n = x_t|t + \Omega_t|t \Omega_{t+1|t}^{-1}(x_{t+1|n} - x_{t+1|t}).$$

The smoothing procedure works recursively by starting from $t = n - 1$. Starting value $x_{n|n}$ and series $(x_t|t)$, $(x_{t+1|t})$, $(\Omega_t|t)$ and $(\Omega_{t+1|t})$ that are required for the smoothing procedure are obtained in the estimation procedure. The reader is referred to Hamilton (1994) or Kim and Nelson (1999) for more details of this technique. Note that smoothing is implemented after the model parameters are estimated, therefore this procedure has no effect on the parameter estimates.

For any given values of nonsingular $A$ and $\Lambda$, there exist steady state values of $\Omega_{t|t-1}$ and $\Sigma_{t|t-1}$, which we denote by $\Omega$ and $\Sigma$.

**Lemma 2.1** The steady state values $\Omega$ and $\Sigma$ exist and are given by

$$\Omega = \frac{1}{2}(I_q + (I_q + 4(A'\Lambda^{-1}A)^{-1})^{1/2}),$$

$$\Sigma = \frac{1}{2}A(I_q + (I_q + 4(A'\Lambda^{-1}A)^{-1})^{1/2})A' + \Lambda$$

for $p \times q$ matrix $A$ and $p \times p$ matrix $\Lambda$, and $\Omega > I_q$.

If $A$ is identified by the diagonality restriction in (3), then $\Omega$ becomes diagonal. We let

$$D = \text{diag}(\delta_1, \ldots, \delta_q)$$

with $\delta_1 > \cdots > \delta_q$, and write $\Omega$ correspondingly as

$$\Omega = \text{diag}(\omega_1, \ldots, \omega_q)$$

where

$$\omega_i = \frac{1}{2}\left(1 + \sqrt{1 + 4\delta_i^{-1}}\right).$$

Moreover, we set $x_t = (x_{1t}, \ldots, x_{qt})'$ so that the $i$-th factor $(x_{it})$ has the steady state value $\omega_i$ for its conditional variance given $(y_t)$. Note that $\omega_1 < \cdots < \omega_q$, which implies that the $i$-th factor has the $i$-th smallest steady state variance conditional on $(y_t)$. We may therefore interpret the $i$-th factor $(x_{it})$ as the factor having the $i$-th largest correlation with $(y_t)$. Clearly, any factor uncorrelated with $(y_t)$ has a conditional variance diverging to infinity in our framework, since it is specified as a random walk whose unconditional variance increases...
over time without any bound. Of course, the leading factor \((x_t)\) has the largest correlation with \((y_t)\), and therefore, it becomes the most relevant factor.

We let \(A = (a_{ij})\) for \(i = 1, \ldots, p, j = 1, \ldots, q\) and \(\Lambda = (\lambda_{ij})\) for \(i, j = 1, \ldots, p, \) and \(\Sigma = (\sigma_{ij})\) for \(i, j = 1, \ldots, p\). Furthermore, we let \(y_t = (y_{1t}, \ldots, y_{pt})\), so that \((y_{it})\) becomes the \(i\)-th observable time series for \(i = 1, \ldots, p\). Note in particular that \((\sigma_{ii})\) signifies the steady state conditional variance of \((y_{it})\) for \(i = 1, \ldots, p\). Under the identifying restriction in (3), it follows that

\[
\sigma_{ii} = \sum_{j=1}^{q} \omega_j a_{ij}^2 + \lambda_{ii} \tag{5}
\]

for \(i = 1, \ldots, p\), where \((\omega_i)\) is defined above. We define

\[
\pi_{ij} = \frac{\omega_j a_{ij}^2}{\sigma_{ii}} \tag{6}
\]

for \(i = 1, \ldots, p\) and \(j = 1, \ldots, q\). \(\pi_{ij}\) represents the portion of conditional variance of \((y_{it})\) explained by the \(j\)-th common factor.

We will set \(\Omega_{0|0} = \Omega - I_q\) for the rest of the paper, so that \(\Omega_{t|t-1}\) takes its steady state value \(\Omega\) for all \(t \geq 1\). Of course, \(\Sigma_{t|t-1}\) also becomes time invariant and takes its steady state value \(\Sigma\) under this convention.

The following lemma specifies \((x_{t|t-1})\) more explicitly as a function of the observed time series \((y_t)\) and the initial value \(x_0\).

**Lemma 2.2**  We have

\[
x_{t|t-1} = (A'\Lambda^{-1}A)^{-1}A'\Lambda^{-1}y_t - \sum_{k=0}^{t-1} (I_q - \Omega^{-1})^k (A'\Lambda^{-1}A)^{-1}A'\Lambda^{-1}\Delta y_{t-k} + (I_q - \Omega^{-1})^{t-1}x_0
\]

for all \(t \geq 2\).

The result of Lemma 2.2 is given entirely by the prediction and updating steps of the Kalman filter. In particular, it holds even under misspecification of our model in (1).

It follows from Lemma 2.1 that \(\Omega > I_q\), and therefore, \(0 < \Omega^{-1} < I_q\). As a consequence, we have \(0 < I_q - \Omega^{-1} < I_q\), and hence, the magnitude of the term \((I_q - \Omega^{-1})^{t-1}x_0\) geometrically declines as \(t \to \infty\). It implies that the effect of \(x_0\) on \(x_{t|t-1}\) becomes negligible out as \(t \to \infty\), as long as \(x_0\) is fixed and finite a.s. Therefore, we may set

\[
x_0 = 0 \tag{8}
\]

without affecting our asymptotic results.

Let \(\Omega_0\) be the value of \(\Omega\) defined with the true values \(A_0\) and \(\Lambda_0\) of \(A\) and \(\Lambda\). If we denote by \(x_{0|t-1}^0\) the value of \(x_{t|t-1}\) under model (1), we may deduce from Lemma 2.2 and the smoothing procedure that
Proposition 2.3  We have

\[ x_{t|t-1}^0 = x_t + \Omega_0^{-1} \sum_{k=1}^{t-1} (I_q - \Omega_0^{-1})^{k-1}(A_0'\Lambda_0^{-1}A_0)^{-1}A_0'\Lambda_0^{-1}u_{t-k} - \sum_{k=0}^{t-1} (I_q - \Omega_0^{-1})^ku_{t-k} \]

for all \( t \geq 2 \), and

\[ x_{t|n}^0 = x_{t|t}^0 + \sum_{k=1}^{n-t} (I_q - \Omega_0^{-1})^k \Delta x_{t+k|t+k}^0 \]

for all \( t \leq n-1 \).

Proposition 2.3 implies in particular that

\[ x_{t|t-1}^0 - x_t = \Omega_0^{-1}a_{t-1} - b_{t-1}, \]

where

\[ a_{t-1} = \sum_{k=1}^{t-1} (I_q - \Omega_0^{-1})^{k-1}(A_0'\Lambda_0^{-1}A_0)^{-1}A_0'\Lambda_0^{-1}u_{t-k} \quad \text{and} \quad b_{t-1} = \sum_{k=0}^{t-1} (I_q - \Omega_0^{-1})^kv_{t-k}. \]

Under the assumption that \((u_t)\) and \((v_t)\) are iid random sequences, the time series \((a_t)\) and \((b_t)\) become the stationary first-order VAR processes given by

\[ a_t = (I_q - \Omega_0^{-1})a_{t-1} + (A_0'\Lambda_0^{-1}A_0)^{-1}A_0'\Lambda_0^{-1}u_{t}, \]
\[ b_t = (I_q - \Omega_0^{-1})b_{t-1} + v_{t} \]

respectively, since \(0 < I_q - \Omega_0^{-1} < I_q\).

Clearly, every component of \((x_{t|t-1}^0)\) or \((x_{t|n}^0)\) is cointegrated with the corresponding component of \((x_t)\) with unit cointegrating coefficient. The stochastic trends in \((x_t)\) may therefore be identified and represented by those in \((x_{t|t-1}^0)\) or \((x_{t|n}^0)\). It seems worth noting that the results in Proposition 2.3 do not rely on the iid assumption of \((u_t)\) and \((v_t)\). In particular, our results here imply that we may extract common stochastic trends in \((x_t)\) using the predicting and smoothing steps of the Kalman filter, as long as \((u_t)\) and \((v_t)\) are general stationary processes. Apparently, we need to know the true parameter values to obtain \((x_{t|t-1}^0)\) or \((x_{t|n}^0)\). The true parameter values are typically unknown and have to be estimated. In most practical applications, we should therefore use parameter estimates to compute \((x_{t|t-1}^0)\) or \((x_{t|n}^0)\). Clearly, the estimates of \((x_{t|t-1}^0)\) and \((x_{t|n}^0)\) based on consistent parameter estimates are close to \((x_{t|t-1}^0)\) and \((x_{t|n}^0)\).

Once we obtain \((x_{t|t-1}^0)\), we may decompose the time series \((y_t)\) into permanent and transitory (PT) components. If we denote them as \((y_t^p)\) and \((y_t^T)\), respectively, they are given by

\[ y_t^p = A_0x_{t|t-1}^0 \quad \text{and} \quad y_t^T = y_t - A_0x_{t|t-1}^0. \]

The permanent component \((y_t^p)\) is I(1), whereas the transitory component \((y_t^T)\) is I(0). Note that the permanent component \((y_t^p)\) is predictable, while the transitory component \((y_t^T)\) is a martingale difference sequence (mds) and unpredictable.
The Kalman filter has exactly the same prediction and updating steps for the measurement equation (4), if we let

\[ y_{t|t-1} = Ax_{t|t-1} + \sum_{k=1}^{m} \Pi_k \Delta y_{t-k}. \]

in place of \( y_{t|t-1} = Ax_{t|t-1} \). Therefore, it is clear that Lemma 2.1 and Proposition 2.3 hold for this general model without any further modification. Moreover, Lemma 2.2 also continues to be valid if we replace \((y_t)\) with \((y_t - \sum_{k=1}^{m} \Pi_k \Delta y_{t-k})\). The theory of Kalman filter for the general model thus follows immediately.

3. Asymptotics for Maximum Likelihood Estimation

In this section, we consider the maximum likelihood estimation of our model. In particular, we establish the consistency and asymptotic Gaussianity of the maximum likelihood estimator under normality. Because the integrated processes are involved in our model, the usual asymptotic theory for ML estimation of state space models given by, for instance, Caines (1988), does not apply. CMP develops a general asymptotic theory of ML estimation, which allows for the presence of nonstationary time series. They obtain the asymptotics of ML estimators of the parameters in their model, where the number of latent variable is restricted to one. In this paper, we derive the asymptotic properties of the ML estimators of the parameters in the state space model that is allowed to have multiple stochastic latent variables. In developing our asymptotic theory, we will frequently refer to the results obtained previously in CMP.

We let \( \theta \) be a \( \kappa \)-dimensional parameter vector and define

\[ \varepsilon_t = y_t - y_{t|t-1} \]

to be the prediction error with conditional mean zero and variance matrix \( \Sigma \). Under normality, the log-likelihood function of \( y_1, \ldots, y_n \) is given by

\[
\ell_n(\theta) = -\frac{n}{2} \log \det \Sigma - \frac{1}{2} \text{tr} \Sigma^{-1} \sum_{t=1}^{n} \varepsilon_t \varepsilon'_t - \frac{1}{2} \text{tr} \Sigma^{-1} \sum_{t=1}^{n} \varepsilon_t \varepsilon'_t
\]

ignoring the unimportant constant term. Here, \( \Sigma \) and \( (\varepsilon_t) \) are in general given as functions of \( \theta \). Let \( s_n(\theta) \) and \( H_n(\theta) \) be the score vector and Hessian matrix, i.e.,

\[
s_n(\theta) = \frac{\partial \ell_n(\theta)}{\partial \theta} \quad \text{and} \quad H_n(\theta) = \frac{\partial^2 \ell_n(\theta)}{\partial \theta \partial \theta'}.\]

After applying some algebra, we may deduce that

\[
s_n(\theta) = -\frac{n}{2} \frac{\partial (\text{vec} \Sigma')}{\partial \theta} (\text{vec} (\Sigma^{-1} + \frac{1}{2} \frac{\partial (\text{vec} \Sigma')}{\partial \theta} \text{vec} \left( \Sigma^{-1} \sum_{t=1}^{n} \varepsilon_t \varepsilon'_t \Sigma^{-1} \right) - \sum_{t=1}^{n} \frac{\partial \varepsilon'_t}{\partial \theta} \Sigma^{-1} \varepsilon_t),
\]
and

\[
H_n(\theta) = -\frac{n}{2} \left[ I_n \otimes (vec \Sigma^{-1})' \right] \left[ \frac{\partial^2}{\partial \theta \partial \theta'} \otimes (vec \Sigma) \right] \\
+ \frac{1}{2} \left[ I_n \otimes \left( vec \Sigma^{-1} \left( \sum_{t=1}^{n} \varepsilon_t \varepsilon_t' \right) \Sigma^{-1} \right)' \right] \left[ \frac{\partial^2}{\partial \theta \partial \theta'} \otimes (vec \Sigma) \right] \\
+ \frac{n}{2} \frac{\partial (vec \Sigma)'}{\partial \theta} (\Sigma^{-1} \otimes \Sigma^{-1}) \frac{\partial (vec \Sigma)}{\partial \theta'} \\
- \frac{1}{2} \frac{\partial (vec \Sigma)'}{\partial \theta} \left( \Sigma^{-1} \otimes \Sigma^{-1} \left( \sum_{t=1}^{n} \varepsilon_t \varepsilon_t' \right) \Sigma^{-1} + \Sigma^{-1} \left( \sum_{t=1}^{n} \varepsilon_t \varepsilon_t' \right) \Sigma^{-1} \otimes \Sigma^{-1} \right) \frac{\partial (vec \Sigma)}{\partial \theta'} \\
- \sum_{t=1}^{n} \frac{\partial \varepsilon_t'}{\partial \theta} \Sigma^{-1} \frac{\partial \varepsilon_t}{\partial \theta'} \\
+ \frac{\partial (vec \Sigma)'}{\partial \theta} (\Sigma^{-1} \otimes \Sigma^{-1}) \sum_{t=1}^{n} \left( \frac{\partial \varepsilon_t}{\partial \theta'} \otimes \varepsilon_t \right) + \sum_{t=1}^{n} \left( \frac{\partial \varepsilon_t'}{\partial \theta} \otimes \varepsilon_t' \right) (\Sigma^{-1} \otimes \Sigma^{-1}) \frac{\partial (vec \Sigma)}{\partial \theta'}
\]

as given in CMP. Here and elsewhere in the paper, \( vec A \) denotes the column vector obtained by stacking the rows of matrix \( A \).

Denote by \( \hat{\theta}_n \) the maximum likelihood estimator of \( \theta \), the true value of which is set at \( \theta_0 \). As in the standard stationary model, the asymptotics of \( \hat{\theta}_n \) in our model can be obtained from the first order Taylor expansion of the score vector, which is given by

\[
s_n(\hat{\theta}_n) = s_n(\theta_0) + H_n(\theta_n)(\hat{\theta}_n - \theta_0), \tag{11}
\]

where \( \theta_n \) lies in the line segment connecting \( \hat{\theta}_n \) and \( \theta_0 \). Assuming that \( \hat{\theta}_n \) is an interior solution, we have \( s_n(\hat{\theta}_n) = 0 \) immediately. Therefore, it is now clear from (11) that we may write

\[
\nu_n T^{-1}(\hat{\theta}_n - \theta_0) = -[\nu_n^{-1} T' H_n(\theta_n) T \nu_n^{-1}]^{-1} [\nu_n^{-1} T' s_n(\theta_0)] 
\tag{12}
\]

for appropriately defined \( \kappa \)-dimensional square matrices \( \nu_n \) and \( T \), which are introduced here respectively for the necessary normalization and rotation.

Upon appropriate choices of the normalization matrix sequence \( \nu_n \) and rotation matrix \( T \), we will show that

ML1: \( \nu_n^{-1} T' s_n(\theta_0) \rightarrow_d N \) as \( n \rightarrow \infty \) for some \( N \),

ML2: \( -\nu_n^{-1} T' H_n(\theta_0) T \nu_n^{-1} \rightarrow_p M > 0 \) a.s. as \( n \rightarrow \infty \) for some \( M \), and

ML3: There exists a sequence of invertible normalization matrices \( \mu_n \) such that \( \mu_n \nu_n^{-1} \rightarrow 0 \) a.s. and

\[
\sup_{\theta \in \Theta_n} \left\| \mu_n^{-1} T' (H_n(\theta) - H_n(\theta_0)) T \mu_n^{-1} \right\| \rightarrow_p 0,
\]

where \( \Theta_n = \{ \theta \| \| \mu_n^{-1} T^{-1}(\theta - \theta_0) \| \leq 1 \} \) is a sequence of shrinking neighborhoods of \( \theta_0 \).
As shown by Park and Phillips (2001) in their study of the nonlinear regression with integrated time series, conditions ML1-ML3 above are sufficient to derive the asymptotics for $\hat{\theta}_n$. In fact, under conditions ML1-ML3, we may deduce from (12) and continuous mapping theorem that

$$\nu_n^* T^{-1}(\hat{\theta}_n - \theta_0) = -[\nu_n^{-1} T^* H_n(\theta_0) T \nu_n^{-1}]^{-1} [\nu_n^{-1} T^* s_n(\theta_0)] + o_p(1) \rightarrow_d M^{-1} N$$

as $n \rightarrow \infty$. In particular, ML3 ensures that $s_n(\hat{\theta}_n) = 0$ with probability approaching to one and

$$\nu_n^{-1} T^* (H_n(\hat{\theta}_n) - H_n(\theta_0)) T \nu_n^{-1} \rightarrow_p 0$$

as $n \rightarrow \infty$. This was shown by Wooldridge (1994) for the asymptotic analysis of extremum estimators in models including nonstationary time series.

To obtain the limit distribution of $s_n(\theta_0)$, we first let $\varepsilon_0^t, (\partial/\partial \theta') \varepsilon_0^t$ and $(\partial/\partial \theta') vec \Sigma_0$ be defined respectively as $\varepsilon_t, (\partial/\partial \theta') \varepsilon_t$ and $(\partial/\partial \theta') vec \Sigma$ evaluated at the true parameter value $\theta_0$ of $\theta$. Then we have

$$s_n(\theta_0) = \frac{1}{2} \frac{\partial (vec \Sigma_0')}{\partial \theta}(\Sigma_0^{-1} \otimes \Sigma_0^{-1}) vec \left[ \sum_{t=1}^{n} (\varepsilon_t^0 \varepsilon_t^0' - \Sigma_0) \right] - \sum_{t=1}^{n} (\frac{\partial \varepsilon_t^0'}{\partial \theta} \Sigma_0^{-1} \varepsilon_t^0).$$

As shown in CMP,

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} (\varepsilon_t^0 \varepsilon_t^0' - \Sigma_0) \rightarrow_d N \left( 0, (I + K)(\Sigma_0 \otimes \Sigma_0) \right)$$

as $n \rightarrow \infty$, where $K$ is the commutation matrix, and

$$\sum_{t=1}^{n} (\varepsilon_t^0 \varepsilon_t^0' - \Sigma_0) \quad \text{and} \quad \sum_{t=1}^{n} \frac{\partial \varepsilon_t^0'}{\partial \theta} \Sigma_0^{-1} \varepsilon_t^0 \quad \text{are asymptotically independent. (16)}$$

Note in particular that

$$\varepsilon_t^0 = y_t - y_{t-1}^0 = A_0(x_t - x_{t-1}^0) + u_t,$$

and as a consequence $(\varepsilon_t^0, \mathcal{F}_t)$ is a martingale difference sequence and $(\partial/\partial \theta') \varepsilon_t^0$ is a predictable sequence with respect to the filtration $(\mathcal{F}_t)$.

If our model were stationary, the limit distribution would therefore be easily derivable from (15), (16) and

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \varepsilon_t^0'}{\partial \theta} \Sigma_0^{-1} \varepsilon_t^0 \rightarrow_d N \left( 0, \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^{n} \frac{\partial \varepsilon_t^0'}{\partial \theta} \Sigma_0^{-1} \frac{\partial \varepsilon_t^0}{\partial \theta'} \right),$$

which can be readily obtained by employing the standard martingale CLT. Of course, asymptotics in (17) does not hold for our nonstationary model with integrated latent variables. As we will show below in Lemma 3.1, the multivariate process $(\partial/\partial \theta') \varepsilon_t^0$ is given by a mixture of stationary and nonstationary processes. Our subsequent asymptotic analysis...
will therefore be focused on solving the complexity caused by this mixture of stationarity and nonstationarity.

Now we look at our model more specifically. The parameter $\theta$ for our model is given by

$$\theta = ((vecA)'', v(\Lambda)''),$$

(18)

with the true value $\theta_0 = ((vecA_0)'', v(\Lambda_0)'').$ Here and elsewhere in the paper, $v(\Lambda)$ denotes the subvector of $vec\Lambda$ with all subdiagonal elements of $\Lambda$ eliminated. Therefore, $v(\Lambda)$ vectorizes only the nonredundant elements of $\Lambda$. We may relate $vec(\Lambda)$ and $v(\Lambda)$ by $Dv(\Lambda) = vec(\Lambda)$, where $D$ is the duplication matrix. See, e.g., Magnus and Neudecker (1988, pp.48-49). The dimension of $\theta$ is given by

$$\kappa = pq + p(p+1)/2,$$

since in particular there are only $p(p+1)/2$ number of nonredundant elements in $\Lambda$.

For our model (1), we may easily deduce from Lemma 2.2 and Proposition 2.3 that

**Lemma 3.1** We have

$$\frac{\partial \varepsilon_t^0}{\partial vecA} = -(I_p - \Lambda_0^{-1}A_0(A_0'\Lambda_0^{-1}A_0)^{-1}A_0') \otimes x_t + a_t(u,v) \quad \text{and} \quad \frac{\partial \varepsilon_t^0}{\partial vecA} = b_t(u,v),$$

where $a_t(u,v)$ and $b_t(u,v)$ are stationary linear processes driven by $(u_t)$ and $(v_t)$.

According to Lemma 3.1,

$$\frac{\partial \varepsilon_t^0}{\partial \theta} = \left( \frac{\partial \varepsilon_t^0}{\partial (vecA)'}, \frac{\partial \varepsilon_t^0}{\partial v(\Lambda)'} \right)'$$

is a matrix time series consisting of a mixture of integrated and stationary processes since $a_t(u,v)$ and $b_t(u,v)$ are stationary linear processes driven by $(u_t)$ and $(v_t)$. Notice that

$$P = I_p - \Lambda_0^{-1}A_0(A_0'\Lambda_0^{-1}A_0)^{-1}A_0'$$

(19)

is a $(p-q)$-dimensional (non-orthogonal) projection on the space orthogonal to $A_0$ along $\Lambda_0^{-1}A_0$. Naturally, we have $A_0'P = 0$. Consequently, $(A_0 \otimes I_q)'$ annihilates the common stochastic trends in $(\partial \varepsilon_t^0/\partial vecA)$, and therefore $((A_0 \otimes I_q)'(\partial \varepsilon_t^0/\partial vecA))$ becomes stationary. Unlike $(\partial \varepsilon_t^0/\partial vecA)$, it is rather clear from Lemma 3.1 that $(\partial \varepsilon_t^0/\partial vecA)$ is entirely stationary.

In order to effectively deal with the singularity of the matrix $P$ in (19), we follow CMP and introduce a necessary rotation. Let $B_0$ be an $p \times (p-q)$ matrix satisfying the conditions

$$B_0'\Lambda_0^{-1}A_0 = 0 \quad \text{and} \quad B_0'\Lambda_0^{-1}B_0 = I_{p-q}.$$ 

(20)

Note that if rank($A_0$) = $q = p$, such $B_0$ does not exist. In the following discussion we will focus on the case where $q < p$. It is easy to deduce that

$$P = I_p - \Lambda_0^{-1}A_0(A_0'\Lambda_0^{-1}A_0)^{-1}A_0' = \Lambda_0^{-1}B_0B_0',$$

(21)

since $P$ is a projection matrix such that $A_0'P = P\Lambda_0^{-1}A_0 = 0.$
Now the $\kappa$-dimensional rotation matrix $T$ is defined as
\begin{equation}
T = (T_N, T_S),
\end{equation}
where $T_N$ and $T_S$ are matrices of dimensions $\kappa \times \kappa_1$ and $\kappa \times \kappa_2$ with $\kappa_1 = (p - q)q$ and $\kappa_2 = q^2 + p(p + 1)/2$, which are given by
\begin{equation}
T_N = \begin{pmatrix} B_0 \otimes I_q \\ 0 \end{pmatrix} \quad \text{and} \quad T_S = \begin{pmatrix} A_0 (A_0'^{-1} A_0)^{-1/2} \otimes I_q \\ 0 \\ I_{p(p+1)/2} \end{pmatrix},
\end{equation}
respectively. It follows immediately from Lemma 3.1, (20) and (21) that
\begin{equation}
\kappa \rightarrow \infty
\end{equation}
for $r \in \mathbb{R}$, where $U, V, W$ are independent of $n$, and $U(t, V(t)) \rightarrow_d (U, V, W)$ as $n \rightarrow \infty$, where $U$, $V$, and $W$ are (possibly degenerate) Brownian motions such that $V$ and $W$ are independent of $U$, and $\int_0^1 V(r) \Sigma_0^{-1} V(r')' dr$ is of full rank a.s.

To derive the main asymptotic results for the ML estimator $\hat{\theta}_n$ of $\theta$, we need to establish two lemmas, which will be presented in sequel. They are straightforward extensions of Lemmas 3.3 and 3.4 in CMP.

**Lemma 3.2** If we let
\begin{equation}
(U_n(r), V_n(r), W_n(r)) = \left( \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} \Sigma_0^{-1} \xi_t, \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} \Delta T N \frac{\partial \xi_t}{\partial \theta}, \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} T S \frac{\partial \xi_t}{\partial \theta} \Sigma_0^{-1} \xi_t \right)
\end{equation}
for $r \in [0, 1]$, then it follows that
\begin{equation}
(U_n(r), V_n(r), W_n(r)) \rightarrow_d (U, V, W)
\end{equation}
as $n \rightarrow \infty$. We may readily establish from Lemma 3.2 the joint asymptotics of
\begin{equation}
\frac{1}{n} T_N \sum_{t=1}^{n} \frac{\partial \xi_t}{\partial \theta} \Sigma_0^{-1} \xi_t \rightarrow_d \int_0^1 V(r) dU(r),
\end{equation}
and
\[
\frac{1}{\sqrt{n}} T_S' \sum_{t=1}^{n} \frac{\partial \varepsilon_t^0}{\partial \theta} \Sigma_0^{-1} \varepsilon_t^0 \to_d W, \tag{27}
\]
where we denote \(W(1)\) simply as \(W\). This convention will be made for the rest of the paper. Because of the independence of \(V\) and \(U\), the limiting distribution in (26) is mixed normal. On the other hand, the independence of \(W\) and \(U\) renders the two limit distributions in (26) and (27) to be independent. Clearly, we have
\[
W = d \mathcal{N}(0, \text{var}(W)),
\]
where
\[
\text{var}(W) = \lim_{n \to \infty} T_S' \left( \frac{1}{n} \sum_{t=1}^{n} \frac{\partial \varepsilon_t^0}{\partial \theta} \Sigma_0^{-1} \frac{\partial \varepsilon_t^0}{\partial \theta} \right) T_S.
\]
Moreover,

**Lemma 3.3** If we define
\[
Z_n = \frac{1}{2} T_S' \frac{\partial (\text{vec} \Sigma_0)}{\partial \theta} (\Sigma_0^{-1} \otimes \Sigma_0^{-1}) \text{vec} \left[ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (\varepsilon_t^0 \varepsilon_t^0 - \Sigma_0) \right],
\]
then it follows that \(Z_n \to_d Z\), where \(Z = d \mathcal{N}(0, \text{var}(Z))\) with
\[
\text{var}(Z) = \frac{1}{2} T_S' \left[ \frac{\partial (\text{vec} \Sigma_0)}{\partial \theta} (\Sigma_0^{-1} \otimes \Sigma_0^{-1}) \frac{\partial (\text{vec} \Sigma_0)}{\partial \theta'} \right] T_S,
\]
and \(Z\) is also independent of \(U\), \(V\) and \(W\) introduced in Lemma 3.2.

Now we are ready to derive the limit distribution for the ML estimator \(\hat{\theta}_n\) of \(\theta_0\) defined in (18). They are given by (13) with the rotation matrix \(T\) in (22) and the sequence of normalization matrix
\[
\nu_n = \text{diag}(nI_{e_1}, \sqrt{n}I_{e_2}),
\]
as we state below as a theorem.

**Theorem 3.4** All three conditions in ML1-ML3 are satisfied for our model. In particular, ML1 and ML2 hold, respectively, with
\[
N = \left( - \int_0^1 V(r) dU(r) \right) \frac{Z - W}{Z - W}
\]
and
\[
M = \left( \begin{array}{cc} \int_0^1 V(r) \Sigma_0^{-1} V(r)'dr & 0 \\ 0 & \text{var}(W) + \text{var}(Z) \end{array} \right)
\]
in notations introduced above.

Theorem 3.4 is completely analogous to Theorem 3.5 in CMP. In particular, Theorem 3.4 shows that the results in Theorem 3.5 of CMP extends well to the multi-dimensional
case, though the proof is much more involved to deal with the multi-dimensionality of the integrated latent variable. As in CMP, we let

\[
Q = -\left( \int_0^1 V(r)\Sigma_0^{-1} V(r)' \right)^{-1} \int_0^1 V(r) dU(r)
\]

and

\[
\begin{pmatrix} R \\ S \end{pmatrix} = -[var(W) + var(Z)]^{-1}(W - Z),
\]

(28)

where \( R \) and \( S \) are \( \kappa^2 \)- and \( p(p + 1)/2 \)-dimensional, respectively. Note that \( Q \) has a mixed normal distribution, whereas \( R \) and \( S \) are jointly normal and independent of \( Q \). Now we may easily deduce from Theorem 3.4 that

\[
\sqrt{n} \left( v(\hat{\Lambda}_n) - v(\Lambda_0) \right) \to_d S,
\]

and

\[
\sqrt{n} \left( (B_0\Lambda_0^{-1} \otimes I_q) \text{vec}\hat{A}_n \right) \to_d Q
\]

\[
\sqrt{n} \left( (A_0\Lambda_0^{-1} A_0)^{-1/2} A_0\Lambda_0^{-1} \otimes I_q \right) (\text{vec}\hat{A}_n - \text{vec}A_0) \to_d R,
\]

(29)

(30)

similarly as in CMP. In particular, it follows immediately from (29) and (30) that

\[
\sqrt{n}(\text{vec}\hat{A}_n - \text{vec}A_0) \to_d \left( A_0(A_0' \Lambda_0^{-1} A_0)^{-1/2} \otimes I_q \right) R,
\]

which has a degenerate normal distribution, if \( q < p \).

From Theorem 3.4 and the subsequent remarks, we know that the ML estimators \( \hat{A}_n \) and \( \hat{\Lambda}_n \) converge at the standard rate \( \sqrt{n} \), and have normal limit distributions. However, in the case where \( q < p \) the limit distribution of \( \hat{\Lambda}_n \) is degenerate. In the direction of \( B_0\Lambda_0^{-1} \), it has a rate of convergence \( n \) and a mixed normal limit distribution. The normal and mixed normal asymptotic distributions of ML estimators validate the conventional inference for hypothesis testing in such state space models where multiple integrated latent variables are included.

As discussed in CMP, the asymptotic results for the ML estimators for our model also hold, at least qualitatively, for more general models, such as the type of the models including lagged terms in measurement equations. Even for the case where time series consists not only stochastic integrated trends, but also deterministic linear time trend, after some proper rotation of the time series, see, e.g., Park (1992), our asymptotic theories are applicable for the rotated time series. The rotation simply separates out the component dominated by a deterministic linear time trend and the component represented as a purely stochastic integrated process.
4. Determination of Number of Common Trends

In the asymptotic analysis of the ML estimator for our model defined in (1), we assume that the number of common stochastic trends in \((y_t)\) is known to be \(q\). This of course is equivalent to assuming that the number of cointegrating relationships in the \(p\)-dimensional time series \((y_t)\) is known to be \(p-q\). From our analysis in the previous section, we may indeed readily deduce that

\[
B_0' \Lambda_0^{-1} y_t = B_0' \Lambda_0^{-1} u_t \quad \text{and} \quad \text{var}(B_0' \Lambda_0^{-1} u_t) = I_{p-q}.
\]

It is therefore clearly seen that \(\Lambda_0^{-1} B_0\) is the matrix of \(p-q\) cointegrating vectors, which yield cointegrating errors with identity covariance matrix. However, the number of common stochastic trends or the cointegrating relationships is typically unknown in empirical studies. In this section, we will develop a test based on a conventional information criterion for testing the number of common stochastic trends, and explain how we may use the test to determine the dimensionality of the latent integrated processes in our model.

Needless to say, testing for the number of common stochastic trends is equivalent to testing for the number of cointegrating relationships. Therefore, at least conceptually, we may use the existing test such as Johansen (1998, 1991) to determine \(p-q\) or \(q\), i.e., the number of cointegrating vectors or the number of common stochastic trends. However, using the methods based on a finite order VAR or ECM as Johansen’s approach has two important shortcomings in our context. First, as we will show subsequently, our model cannot be represented as any finite order vector autoregression or error correction model. Any finite order VAR or ECM is therefore inconsistent with our model. Second, our model is potentially more useful for a large system of time series which share a few common stochastic trends. For such systems, VAR or ECM formulations often become too flexible, allowing too many parameters. In particular, it is impossible to use long VAR’s or ECM’s, trying to fit an infinite order VAR or ECM.

Proposition 4.1 We have

\[
\triangle y_t = -B_0(\Lambda_0^{-1} B_0)' y_{t-1} - \sum_{k=1}^{t-1} C_k \triangle y_{t-k} + \epsilon_t^0,
\]

where \(C_k = A_0(I_q - \Omega_0^{-1})^k (A_0 \Lambda_0^{-1} A_0)^{-1} A_0' \Lambda_0^{-1} \).

Proposition 4.1 makes clear the difference between our model and the conventional ECM. From (31), we may immediately see that \((y_t)\) is generated as VAR(\(\infty\)), which in particular implies that our model is not representable as a finite-order VAR. Moreover, we have rank deficiencies in the short-run coefficients \((C_k)\), as well as in error correction term \(B_0(\Lambda_0^{-1} B_0)'\). Note that \((C_k)\) are of rank \(q\) and \((\Lambda_0^{-1} B_0)' C_k = 0\) for all \(k = 1, 2, \ldots\). In the conventional ECM, there is no such rank restriction imposed on the short-run coefficients. As a consequence, Johansen’s approach, based on finite order ECM’s, is not applicable in our model. This is also true for the general measurement equation (4). Indeed, it is
easy to see that Proposition 4.1 continues to hold in this case only with \( (y_t) \) replaced by \( (y_t - \sum_{k=1}^{m} \Pi_k \Delta y_{t-k}) \). Clearly, in order to test for the number of common stochastic trends in our framework, a new testing method is needed.

After establishing the asymptotic properties of the ML estimator for our model, we are now ready to introduce the selection procedure for the number of common stochastic trends. From now on, we use \( r \) to denote the number of fitted common stochastic trends, whereas \( q \) signifies the number of true common stochastic trends. Accordingly, we let \( \hat{\Sigma}_n(r) \) be the covariance matrix of the forecast error evaluated at the ML estimator from a model with \( r \) dimensional latent factor \( (x_t) \), i.e.

\[
\hat{\Sigma}_n(r) = \frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t|t-1)(y_t - \hat{y}_t|t-1)'.
\]

We set \( \hat{\Sigma}_n(0) = (1/n) \sum_{t=1}^{n} y_t y_t' \). The criterion used to evaluate the number of common stochastic trends takes the form

\[
\text{IC}(r) = \log |\hat{\Sigma}_n(r)| + \frac{c_n K}{n},
\]

where \( K = (pr + p(p+1)/2) \) is the number of unknown parameters and \( c_n \) stands for the penalty coefficient satisfying the conditions specified below. We suggest that the dimension of the latent factor \( (x_t) \) be selected such that

\[
\hat{q}_n = \arg\min_{0 \leq r \leq r_{\text{max}}} \text{IC}(r),
\]

where \( r_{\text{max}} \) is the maximum number of common stochastic trends we set a priori. Whenever possible, the value for \( r_{\text{max}} \) may be chosen based on existing theories or general recognition. Of course, we may simply set \( r_{\text{max}} = p \) if no information available to pin down the range of the number of common stochastic trends.

**Theorem 4.2** Let \( (c_n) \) be a sequence of numbers such that (a) \( c_n \to c > 0 \) and (b) \( c_n/(n \log n) \to 0 \) as \( n \to \infty \). Then we have

\[
\mathbb{P} \{ \hat{q}_n = q \} \to 1
\]

as \( n \to \infty \).

In case \( r < q \), the forecast error \( (y_t - y_t|t-1) \) necessarily has some integrated components, since the stochastic trends in \( (y_t) \) cannot be fully represented by any \( r \)-dimensional random walks. Therefore, \( \hat{\Sigma}_n(r) \) diverges up to infinity in probability at a rate of \( n \), from which it follows immediately that \( \log |\hat{\Sigma}_n(r)| \) increases at a \( \log n \) rate as \( n \to \infty \). Consequently, for any \( (c_n) \) in (33) increasing at any rate slower than \( n \log n \), we would pick at least \( q \) stochastic trends asymptotically. On the other hand, as shown in Chang, Jiang and Park (2013), the log-likelihood value does not change in the limit if we increase \( r \) beyond its true value \( q \). Therefore, we would not choose any \( r > q \) stochastic trends asymptotically as long as \( (c_n) \) in (33) becomes strictly positive in the limit. Of course, in actual applications of
our procedure, the log-likelihood values do not remain to be exactly the same for any \( r > q \), due to computational errors in finite samples. According to Chang, Jiang and Park (2013), the magnitude of finite sample computational errors increases as the dimension of the fitted model increases, and decreases as the sample size increases.

Theorem 4.2 implies that all commonly used information criteria are consistent if they are applied to determine the number of common trends in our model. BIC (Schwarz 1978), Hannan and Quinn (1979) information criterion and AIC (Akaike 1973) respectively set \( c_n = \log n \), \( c_n = 2 \log \log n \) and \( c_n = 2 \). It is clear that the choice of \((c_n)\) in all of them satisfies the condition that we require in Theorem 4.2. Note in particular that AIC, as well as other information criteria, becomes consistent. Typically, the log-likelihood values decrease nontrivially even for overparametrized models, and AIC becomes inconsistent if it is used in the usual model selection problem. Moreover, in the conventional situation, the log-likelihood values of under-parametrized models are an \( n \)-order, instead of \( n \log n \)-order as in our case, of magnitude smaller than those of correctly specified model, and therefore, all information criteria are expected to select under-parametrized models in our model relatively less frequently. In our subsequent simulations and empirical illustrations, we focus on BIC since it appears to be most widely used. The use of other information criteria yields largely the same results.

5. Simulations

In this section, we provide two sets of simulations, one to illustrate the finite sample properties of the ML estimators and the other to demonstrate those of BIC procedure to determine the number of common stochastic trends.

To investigate the finite sample properties of the ML estimators, we specify a model with three observable time series sharing two common trends. In order to satisfy Assumption SSM4, the error terms \((u_t)\) and \((v_t)\) are generated as

\[
\begin{align*}
    u_t &= \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} + \varepsilon_{3t} \\ \varepsilon_{2t} \end{pmatrix} \quad \text{and} \quad v_t = \begin{pmatrix} \varepsilon_{4t} \\ \varepsilon_{5t} \end{pmatrix},
\end{align*}
\]

where \((\varepsilon_{1t}), \ldots, (\varepsilon_{5t})\) are independent and randomly drawn from \(N(0,1)\). The independence between \((u_t)\) and \((v_t)\) therefore follows and the covariance matrix of \((v_t)\) is identity matrix. The covariance matrix \(\Lambda_0\) of \((u_t)\) can be easily derived as well. We present \(\Lambda_0\) and an arbitrarily selected loading matrix \(A\) as

\[
A = \begin{pmatrix} 1 & 2 \\ 2 & 0 \\ 3 & 2 \end{pmatrix} \quad \text{and} \quad \Lambda_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 2 \\ 0 & 2 & 4 \end{pmatrix}.
\]

Following the global identification condition specified in (3), \(A_0 = AH\) is the true value of the globally identified loading matrix, with \(H\) being defined through the spectral representation of \(A'\Lambda_0^{-1}A\), s.t. \(A'\Lambda_0^{-1}A = HDH'\), where \(D\) is the diagonal matrix of its eigenvalues and the columns of \(H\) are the corresponding eigenvectors. The initial value of the state variable
Figure 1: Densities of ML estimators and $t$-ratios, $n=500$

In the simulations, the samples of size 500 are drawn 2000 times to estimate the ML estimators. The $t$-statistics based on these estimators are also derived. In estimating $\Lambda_0$, we estimate its cholesky triangle instead. In this way, we only need to estimate the nonredundant parameters and at the same time may ensure the estimated covariance matrix to be positive definite. To choose the matrix $B_0$ in the rotation matrix $T$ given in (22), we first regress a randomly picked 3 dimensional vector on $\Lambda_0^{-1}A_0$, and then normalize the residual, such that the normalized residual $e$ satisfies the condition $e'\Lambda_0^{-1}e = 1$. We take $e$ as our $B_0$. It is clear that $B_0$ satisfies the conditions specified in (20). The simulation results are summarized in Figure 1. The distributions of the ML estimators are centered at their true values.

The finite sample behavior of the ML estimators are as expected. The distributions of $\hat{A}_n$ and $\hat{\Lambda}_n$ are symmetric and well centered as predicted by their asymptotic theories. The bottom left panel of Figure 1 presents the estimated densities of rotated $\hat{A}_n$. The solid curves represent the densities of $\hat{A}_n$ in the direction of $B_0^t\Lambda_0^{-1}$ and the rest of the curves represent those in the directions orthogonal to $B_0^t\Lambda_0^{-1}$. As expected, the solid curves are steeper since the rotated $\hat{A}_n$ converges at a faster rate in the directions defined by the cointegrating space. The estimated densities of $t$-ratios are presented in the bottom right panel.
Table 1: BIC performance, \( q=2 \)

<table>
<thead>
<tr>
<th>( p )</th>
<th>( n )</th>
<th>( \hat{q}_n = 1 )</th>
<th>( \hat{q}_n = 2 )</th>
<th>( \hat{q}_n \geq 3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>30</td>
<td>11</td>
<td>989</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>0</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>30</td>
<td>0</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>0</td>
<td>1000</td>
<td>0</td>
</tr>
</tbody>
</table>

panel of Figure 1. In total, there are 13 curves appearing in the figure: 12 of them are the estimated densities of the \( t \)-ratios constructed from the ML estimators of the unknown parameters and one represents the standard normal density. The 13 curves do look very close to each other, which corroborates our asymptotic theory.

In the second set of simulations, we investigate the finite sample performance of BIC in selecting the number of stochastic common trends. We explore two systems, one with 5 observables and the other with 20 observables, i.e., with \( p = 5 \) and \( p = 20 \), respectively. Both \( (u_t) \) and \( (v_t) \) are randomly and independently drawn from \( N(0, 1) \). In particular, \( (v_t) \) is two dimensional, i.e., both systems contain two common stochastic trends, and hence the true number of trends \( q \) is set at \( q = 2 \) for both systems. To evaluate accuracy of our test, we estimate the model (1) with the number of common trends \( r \) varying from \( r = 1 \) through \( r = 5 \), and calculate the corresponding BIC values for each case. The one that generates the smallest BIC value is selected as the estimate \( \hat{q}_n \) for the true number of common stochastic trends \( q \). The simulation is repeated 1000 times for samples of sizes 30 and 50, and the frequency of \( \hat{q}_n \) for each \( (p, n) \) combination is reported in Table 1.

Table 1 shows that BIC works very well in finite samples. Only in the case of 5 observables with sample size 30, BIC under-selects the number of common stochastic trends 11 out of 1000 times, but never over-selects. However, the under-selection problem disappears quickly as the sample size increases to 50. Indeed, BIC selects the true number of common stochastic trends with 100% accuracy in all other cases we consider here. We may therefore conclude that BIC performs well even in relatively small samples. More specifically, for a system with a fixed number of observables, BIC performs better as the sample size gets larger. This is well expected since BIC is consistent as our theory predicts. On the other hand, given sample size and true number of common stochastic trends, the increase of cross-sectional units can also help improve the performance of BIC. This is not surprising either since more observations provide more information and therefore generate stronger signal about the common trends.
Table 2: BIC values

<table>
<thead>
<tr>
<th>Number of Trends ($r$)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>BICs (GDP)</td>
<td>52.182</td>
<td>51.029</td>
<td>50.184</td>
<td>49.612</td>
<td>49.028</td>
<td>49.034</td>
<td>49.427</td>
</tr>
<tr>
<td>BICs (Consumption)</td>
<td>56.062</td>
<td>55.323</td>
<td>54.611</td>
<td>54.927</td>
<td>55.162</td>
<td>55.468</td>
<td>55.925</td>
</tr>
</tbody>
</table>

6. Empirical Illustrations

**International Consumption Risk Sharing** Since Backus, Kehoe and Kydland (1992) documented the now famous consumption correlation puzzle, the literature on international consumption risk sharing has grown tremendously. Based on macroeconomic theory, in an open economy, a country’s consumption decision is not constrained by its own production since agents can smooth consumption by borrowing and lending in international markets. If agents worldwide share the same utility function, consumption across countries should be highly correlated because of consumption risk sharing. In particular, the consumption correlation across countries should be substantially larger than the output correlation. However, most empirical studies, e.g. Backus et al. (1992) among others, find the opposite in data, i.e., outputs are shown to be generally more highly correlated across countries than consumptions, even though international financial markets have become increasingly integrated since the beginning of the 1980s. In this section, we revisit the issue about international consumption risk sharing using the new approach, which studies the number of independent stochastic trends present in both output and consumption series from a group of countries. Our empirical study is based on the conjecture that if risk sharing is indeed effective, regional or country-specific shocks are likely to be smoothed out by borrowing and lending through international markets. As a result, we expect that the number of independent trends present in the consumption system should be smaller than that present in the output system.

Following Beckus et al. (1992), we focus on major industrialized countries. Our data include twelve countries: Australia, Canada, France, Japan, South Africa, Switzerland, United Kingdom, United States, Austria, Finland, Italy and Germany. All the data are quarterly, and taken from the International Financial Statistics (IFS) database, covering the period 1980 Q1 - 2011 Q2. As in Backus et al. (1992), we use private final consumption expenditure as consumption series and GDP as output. With the exception of Austria and Finland, the data are seasonally adjusted. We seasonally adjust the data for these two countries using the X-12 method. The real GDP and real consumption series are constructed as the corresponding nominal series divided by GDP deflators which are also seasonally adjusted. Due to the unification of the West and East Germanies, the real GDP and real consumption series for Germany show an abrupt jump from 1990 Q4 to 1991 Q1, and stay high afterwards. Because our model does not allow for structural breaks,
we re-scale the German data from 1991 Q1 to 2011 Q2 to make them consistent with the pre-unification West German data by the ratios of their values in 1990 Q4 to those in 1991 Q1, assuming that the ratio of economic scale of the West Germany to that of the East Germany stays constant. To remove the ‘growth trend’, we detrend the data using the HP filter with the filter parameter $\lambda = 10000$, which is larger than 1600, used most commonly in empirical macroeconomic studies for quarterly data. This is to remove only ‘growth trend’ and leave all other stochastic trends in the series, as is necessary for our long-run analysis. Our empirical results are fairly robust with respect to the choice of $\lambda$, which we discuss more later.

To allow for short-run dynamics, we include two year lags of the differences of the observables in the model, and hence the general model specified in (4) is used in this application. The model is estimated under the assumption that there exist at least one and possibly up to seven stochastic trends for both GDP and consumption systems. BIC values are calculated for each model. The results are reported in Table 2 and the minimums are highlighted in bold. According to Table 2, there are three common stochastic trends in the consumption system, and five in the GDP system. The fact that the consumption system contains less number of stochastic trends provides some evidence of international consumption risk sharing.

To examine the sensitivity of the selected number of trends to the choice of the HP filter
parameter $\lambda$ in the detrending procedure, we fit our model for a wide range of values of $\lambda$ and observe how our empirical results change accordingly. Overall, our empirical results are strongly robust for the choice of $\lambda$ unless it is too small or too big. For $\lambda$ too small, the GDP and consumption series become over-detrended, and consequently most of their stochastic trends disappear, rendering our test to under-select the number of stochastic trends. For $\lambda$ too big, on the other hand, the outcome is similar to detrend the data with a linear trend, leaving all nonlinear growth trends in the data. Since different countries tend to have different growth rates and patterns, it is of no surprise that more trends would be detected in both systems when the data are under-detrended. In sum, as long as $\lambda$ is within a reasonable range, the number of trends detected for both GDP and consumption systems are the same. In particular, we confirm that the detected numbers of trends for both systems do not change for $\lambda$ in the fairly wide range 8000 – 12000.

To further analyze the extracted common stochastic trends, we conduct variance decomposition analysis for each country. Because our model deals with nonstationary systems and unconditional variance of nonstationary time series diverges as sample size increases, we consider conditional variances instead. The conditional variance decomposition is conducted based on (5) and (6). The portion of conditional variance of $(y_{it})$ explained by each factor can be used to measure the relative importance of different factors to each series. Consequently, we calculate the variance shares attributable to each common factor for each country, and as a summary, report the maximum, first quartile and median of
conditional variance shares of each factor in Table 3 and Table 4, respectively for the GDP and consumption series.

Table 3 indicates that Factor 1 and Factor 2 affect GDP of most countries in an important manner. More specifically, Factor 1 accounts for at least 22.46% GDP variation for 1/4 countries, and 18.41% for half countries. Factor 2, although not as important, also explains at least 14.35% GDP variation for 1/4 countries and 11.28% for half countries. Since a majority of countries are affected by Factor 1 and 2, it is reasonable to consider them as global factors or global shocks. Factor 3, 4 and 5, although important to one or two countries in the system since the maximum shares of conditional variance are reasonably high, are far from influential to other countries. This is clear since the first quartiles of these three factors are all below 4%. Corresponding to global factors, we consider these three factors as regional or country specific. As is well recognized, consumption smoothing through international financial markets are more effective in response to regional or country specific shocks than global shocks. If international consumption risk sharing is indeed effective, we should expect to see at most two trends in the consumption system since all the regional and country specific shocks are smoothed out. However, in the consumption system, besides two dominant global factors, we also detect Factor 3, which is regional since the first quartile is only 6.26% in Table 4. According to this result, we conclude that some but not all regional or country specific shocks are smoothed out.

It is worth noting that conditional variance decomposition analysis can also be important
to an individual country, since it provides useful information on the main source of economic fluctuations, which is critically relevant for policy makers. As discussed in Kose et al. (2003), if a significant fraction of domestic output fluctuations is due to a factor which affects most countries worldwide in an important manner, policies targeting external balances by running trade surpluses or deficits to stabilize sudden movements in economic activity might be ineffective.

With the extracted trends, we decompose each series into permanent and transitory components, and present them separately in the upper and lower panels of Figure 2 and Figure 3, respectively for the GDP and consumption series. As discussed below equation (9), the permanent components are nonstationary and predictable, while the transitory components are martingale difference sequences and therefore unpredictable. As one may easily notice, the permanent components of the GDP and consumption series move together quite closely, indicating that consumption and output are still quite relevant to each other. This in turn implies that agents seem to adjust their consumption decisions according to those predictable changes in output. In sum, although we do find some evidence for international consumption risk sharing, the degree of consumption risk sharing in real world is much weaker than what is predicted by the theory, which argues that consumptions should not be constrained by the contemporaneous output levels.

Common Stochastic Trends in Dow Jones Industrial Average Stocks Chang, Miller and Park (2009) extract a common stochastic trend embedded in the 30 individual stock price series that comprise the Dow Jones Industrial Average (DJIA) by using a one-factor nonstationary state space model. They find that the time evolution of the extracted trend is similar to the sample path of the DJIA. However, the assumption that the 30 individual series are driven by only one stochastic trend is rather restrictive particularly since the component stocks represent 10 sectors, which are likely to be affected by multiple independent shocks. In this application, we relax this assumption and use our BIC procedure to test for the number of trends embedded in the 30 stock price series and analyze the relative importance of each of the extracted common stochastic trends to both the individual component series and the DJIA index.

We use the same data set as in Chang, Miller and Park (2009), which employs the daily closing prices from Yahoo Finance. The data are adjusted by taking into account stock splits and dividends using the methodology developed by the Center for Research in Security Prices. The data cover the period from December 2, 1999 to April 7, 2004, which is the longest recent stretch during which the companies comprising the DJIA did not change. Over this sample period, we have 1,092 daily observations. As is well recognized, the stock price volatility varies over time and shows strong persistence. Since our model does not allow for heteroscedasticity over time, we obtain the data at a monthly equivalent frequency from the daily observations using the volatility time instead of the usual chronological time to remove the persistent and time varying volatility present in the DJIA index. Readers are referred to Park (2010) for the details of the time change methodology. All the data are then log transformed and detrended using a linear trend.

In this application, we estimate the model (1) with the number of common trends varying
Table 5: BIC values

<table>
<thead>
<tr>
<th>Number of Trends (r):</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>BICs</td>
<td>111.38</td>
<td><strong>110.61</strong></td>
<td>111.58</td>
<td>111.80</td>
<td>112.42</td>
<td>114.04</td>
<td>115.10</td>
</tr>
</tbody>
</table>

Figure 4: Scree plot for $\hat{A}'\hat{\Lambda}\hat{A}$

from $r = 1$ to $r = 7$. The resulting BIC values are reported in Table 5 and the minimum is highlighted in bold. According to Table 5, the BIC value reaches its minimum at $r = 2$ and hence $\hat{q}_n = 2$, which indicates that instead of one, there are two stochastic trends driving the 30 DJ stock series. Since the rank of $A_0$ determines the number of trends, we present the scree plot for $\hat{A}'\hat{\Lambda}^{-1}\hat{A}$ from the model with $r = 7$ in Figure 4 as an additional check. As usual, we sort the eigenvalues in descending order along the x-axis and the y-axis represents their value. The scree plot clearly confirms the existence of two common trends.

To analyze the relative importance of the two extracted trends in explaining the stock price variation, we calculate the share of the conditional variance of each component stock attributable to each trend and report the maximum, first quartile and median of the variance shares of the 30 DJ stocks in Table 6. Table 6 shows that Factor 1 is dominant since it explains a substantial portion of the total variation for most stocks. Factor 2 is relatively less important, but not negligible since it explains almost 60% of the variation in the stock which has the maximum variance share and at least 6.98% of total stock price variation for 1/4 component stocks.

The sample paths of the DJIA index and the extracted trends are plotted in Figure 5. By comparing the DJIA in the upper panel and the first trend in the lower panel, one can easily see that the time evolution of the first trend is very similar to that of the DJIA. To further investigate the association between the DJIA and the first trend, we regress the DJIA on the first trend only. The fitted DJIA is plotted on the top panel of Figure 5, which clearly shows that the sample path of the fitted DJIA is very similar to that
Table 6: Conditional variance shares for DJ stocks

<table>
<thead>
<tr>
<th></th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum</td>
<td>72.42%</td>
<td>59.59%</td>
</tr>
<tr>
<td>First Quartile</td>
<td>41.18%</td>
<td>6.98%</td>
</tr>
<tr>
<td>Median</td>
<td>26.34%</td>
<td>1.87%</td>
</tr>
</tbody>
</table>

Figure 5: DJIA and extracted common stochastic trends

of the DJIA, confirming again that the first trend captures most of the variation in the DJIA index. To analyze the marginal contribution of the second trend on the DJIA, we regress the DJIA on both trends. The estimated coefficients are 4.04(0.56) for the first trend, and −0.66(0.28) for the second trend, where the numbers in the parentheses are the corresponding standard deviations. Obviously, the second trend is barely significant both statistically and economically in explaining the variation of the DJIA. This, on the other hand, indicates that the DJIA and the second trend are not much related.

Similar to what we do for the component price series, we conduct conditional variance decomposition for the DJIA, and find that the first trend explains about 34% of the variation in the DJIA, whereas the second trend explains only 1%. This result is consistent to that of the simple regression analysis.

Based on the previous analysis, we conclude that the dominant trend is closely related to the DJIA, and thus we may say that the DJIA represents the variation of the dominant
trend fairly well. However, the DJIA fails to capture the variation of the second trend which is important to some of the component stocks. As a result, the DJIA does not represent the trending movements of these stocks. If only one index shall be chosen to represent the long-run movement of the component stocks, the DJIA seems to be a good choice since it nicely captures the variation of the dominant trend. However, if one is interested in studying the long-run movement of all the component stocks, both trends should be taken into account, since the DJIA alone does not fully catch the trending movements of the stocks. Moreover, studying the individual component stocks which are mostly affected by the second trend might be of particular interest to the investors who have intention to further diversify risks since the second trend is almost uncorrelated with the DJIA.

7. Conclusion

In this paper, we consider a state-space model with multiple integrated latent factors. The model provides a new framework, within which we may effectively specify and analyze common stochastic trends in a cointegrated system as latent factors. The standard Kalman filter is used to estimate the model and to extract the common stochastic trends. We establish the consistency and asymptotic normality of the ML estimators of the model parameters, and therefore validate the conventional method of inference based on ML estimators for this class of models. In particular, the ML estimator for the factor loading coefficient matrix has a mixed rate of convergence. It converges at \( n \) rate in the direction of cointegration, while the overall convergence rate is \( \sqrt{n} \) as in the standard stationary model. Its asymptotic distribution is therefore degenerate if normalized with the conventional \( \sqrt{n} \) rate.

In order to determine the number of common stochastic trends, or equivalently the number of cointegrating relationships, we propose to use the standard information criteria. The existing methods relying on the ECM such as Johansen’s test are not applicable for our model, since it cannot be represented as a finite order VAR. Various information criteria can be readily implemented as a part of our estimation procedure for the model. Moreover, they diverge whenever the model has more or less number of common stochastic trends than are present, and therefore, they are all consistent. At least asymptotically, the probability of picking up the correct number of common stochastic trends is unity. Our method is particularly appropriate to deal with a large dimensional system sharing a few common stochastic trends. The simulation reported in the paper shows that information criteria work extremely well in finite samples. For the empirical illustrations, we apply our model to both macro aggregates and stock prices. In one application, we analyze international consumption risk sharing behavior by using the extended nonstationary state space model. We find some evidence for consumption risk sharing, but only to an extent far less than what is predicted by macro theory. In our application on stock prices, we find two common stochastic trends embedded in the 30 DJ stocks. One trend is closely related to the DJIA index, but the other is not, which provides a room for further investment risk diversification.
References


Appendix: Mathematical Proofs

Proof of Lemma 2.1  According to the prediction and updating steps, we have
\[ \Omega_{t+1|t} - I_q = \Omega_{t|t-1} - \Omega_{t|t-1} A' (A \Omega_{t|t-1} A' + \Lambda)^{-1} A \Omega_{t|t-1}. \]  

(34)

Moreover, since
\[ (A \Omega_{t|t-1} A' + \Lambda)^{-1} = \Lambda^{-1} - \Lambda^{-1} A (I + \Omega_{t|t-1} A' \Lambda^{-1} A)^{-1} \Omega_{t|t-1} A' \Lambda^{-1}, \]

it follows that
\[ \Omega_{t|t-1} A' (A \Omega_{t|t-1} A' + \Lambda)^{-1} A \Omega_{t|t-1} \]
\[ = \left( \Omega_{t|t-1} A' \Lambda^{-1} A - \Omega_{t|t-1} A' \Lambda^{-1} A (I + \Omega_{t|t-1} A' \Lambda^{-1} A)^{-1} \Omega_{t|t-1} A' \Lambda^{-1} A \right) \Omega_{t|t-1} \]
\[ = \Omega_{t|t-1} A' \Lambda^{-1} A (I + \Omega_{t|t-1} A' \Lambda^{-1} A)^{-1} \Omega_{t|t-1} \]
\[ = \left( I_q - (I_q + \Omega_{t|t-1} A' \Lambda^{-1} A)^{-1} \right) \Omega_{t|t-1}. \]  

(35)

Therefore, we may deduce from (34) and (35) that
\[ \Omega_{t+1|t} = I_q + (I_q + \Omega_{t|t-1} A' \Lambda^{-1} A)^{-1} \Omega_{t|t-1}, \] 

(36)

which implies in particular that \( (\Omega_{t|t-1}) \), \( t \geq 1 \), is given as a function of nonnegative symmetric matrix \( A' \Lambda^{-1} A \).

Let \( \lambda > 0 \) be an eigenvalue of \( A' \Lambda^{-1} A \), and denote by \( (\omega_t) \) the corresponding eigenvalue of \( (\Omega_{t|t-1}) \). Then we have
\[ \omega_{t+1} = 1 + \frac{\omega_t}{1 + \lambda \omega_t}. \]

If we define \( f : \mathbb{R}_+ \to \mathbb{R} \) by
\[ f(x) = 1 + \frac{x}{1 + \lambda x}, \]

we may easily see that \( f(0) = 1 \), \( f(\infty) = 1 + 1/\lambda \) and
\[ 0 < \frac{d}{dx} f(x) = \frac{1}{(1 + \lambda x)^2} < 1. \]
As a result, \( f \) takes value 1 at the origin and increases monotonically to \( 1 + 1/\lambda \) as \( x \to \infty \), and intersects with the 45 degree line at

\[
x = \frac{1}{2} \left( 1 + \sqrt{1 + \frac{4}{\lambda}} \right).
\]

We may therefore deduce that \((\omega_t)\) converges to \((1 + \sqrt{1 + 4/\lambda})/2\) as \( t \to \infty \), which implies that \((\Omega_{t|t-1})\) converges to the steady state value

\[
\Omega = \frac{1}{2} \left( I_q + [I_q + 4(A'A^{-1}A)^{-1}]^{1/2} \right)
\]

as \( t \to \infty \). It also follows that \((\Sigma_{t|t-1})\), \(\Sigma_{t|t-1} = A\Omega_{t|t-1}A' + \Lambda\), converges to the corresponding steady state value \(\Sigma = A\Omega A' + \Lambda\) accordingly. This was to be shown.

**Proof of Lemma 2.2** From the prediction and updating steps of the Kalman filter, we have

\[
x_{t+1|t} = x_{t|t-1} + \Omega A'\Sigma^{-1}(y_t - y_{t|t-1}) \\
= x_{t|t-1} + \Omega A'\Sigma^{-1}(y_t - Ax_{t|t-1}) \\
= (I_q - \Omega A'\Sigma^{-1}A)x_{t|t-1} + \Omega A'\Sigma^{-1}y_t
\]

with the steady state values \(\Omega\) and \(\Sigma\). However, it follows from (34) that

\[
\Omega A'\Sigma^{-1}A\Omega = I_q,
\]

i.e.,

\[
\Omega A'\Sigma^{-1}A = \Omega^{-1}.
\]

We may also deduce that

\[
\Sigma^{-1} = (A\Omega A' + \Lambda)^{-1} = \Lambda^{-1} - \Lambda^{-1}A(I_q + \Omega A'\Lambda^{-1}A)^{-1}\Omega A'\Lambda^{-1},
\]

which yields

\[
\Omega A'\Sigma^{-1}A = \Omega A'\Lambda^{-1}A - \Omega A'\Lambda^{-1}A(I_q + \Omega A'\Lambda^{-1}A)^{-1}\Omega A'\Lambda^{-1}A \\
= \Omega A'\Lambda^{-1}A \left[ I_q - (I_q + \Omega A'\Lambda^{-1}A)^{-1}\Omega A'\Lambda^{-1}A \right].
\]

Therefore, it follows from (38) and (40) that

\[
I_q - (I_q + \Omega A'\Lambda^{-1}A)^{-1}\Omega A'\Lambda^{-1}A = (\Omega A'\Lambda^{-1}A)^{-1}\Omega^{-1}.
\]

Furthermore, we have

\[
\Sigma^{-1}A = \Lambda^{-1}A - \Lambda^{-1}A(I_q + \Omega A'\Lambda^{-1}A)^{-1}\Omega A'\Lambda^{-1}A \\
= \Lambda^{-1}A \left[ I_q - (I_q + \Omega A'\Lambda^{-1}A)^{-1}\Omega A'\Lambda^{-1}A \right] \\
= \Lambda^{-1}A (\Omega A'\Lambda^{-1}A)^{-1}\Omega^{-1}
\]
and
\[ \Omega A' \Sigma^{-1} = \Omega [\Lambda^{-1} A (\Omega A' A^{-1} A)^{-1} \Omega^{-1}]' = \Omega^{-1} (A' \Lambda^{-1} A)^{-1} A' \Lambda^{-1}, \] (42)
due to (39) and (41).

Now we have from (37), (38) and (42) that
\[ x_{t+1|t} = (I_q - \Omega^{-1}) x_{t|t-1} + \Omega^{-1} (A' \Lambda^{-1} A)^{-1} A' \Lambda^{-1} y_t, \]
and consequently,
\[ x_{t|t-1} = \sum_{k=1}^{t-1} (I_q - \Omega^{-1})^{k-1} \Omega^{-1} (A' \Lambda^{-1} A)^{-1} A' \Lambda^{-1} y_{t-k} + (I_q - \Omega^{-1})^{t-1} x_{1|0}. \] (43)

Moreover,
\[ \sum_{k=1}^{t-1} (I_q - \Omega^{-1})^{k-1} \Omega^{-1} (A' \Lambda^{-1} A)^{-1} A' \Lambda^{-1} y_{t-k} \]
\[ = \sum_{k=1}^{t-1} (I_q - \Omega^{-1})^{k-1} [I_q - (I_q - \Omega^{-1})] (A' \Lambda^{-1} A)^{-1} A' \Lambda^{-1} y_{t-k} \]
\[ = (A' \Lambda^{-1} A)^{-1} A' \Lambda^{-1} y_t - \sum_{k=0}^{t-2} (I_q - \Omega^{-1})^k (A' \Lambda^{-1} A)^{-1} A' \Lambda^{-1} \Delta y_{t-k} \]
\[ - (I_q - \Omega^{-1})^{t-1} (A' \Lambda^{-1} A)^{-1} A' \Lambda^{-1} y_1 \]
\[ = (A' \Lambda^{-1} A)^{-1} A' \Lambda^{-1} y_t - \sum_{k=0}^{t-1} (I_q - \Omega^{-1})^k (A' \Lambda^{-1} A)^{-1} A' \Lambda^{-1} \Delta y_{t-k}. \] (44)

The stated result now follows directly from (43) and (44). Note that \( x_{1|0} = x_{0|0} = x_0 \) and \( y_0 = 0 \). The proof is therefore complete. \( \Box \)

**Proof of Proposition 2.3** For the proof of Proposition 2.3, the readers are referred to the proof of Proposition 2.4 in CMP for the details. In order to tailor it to our model, we only need to replace \( \omega \) with \( \Omega_0 \) and \( 1/\omega_0 \) with \( \Omega_0^{-1} \). Now let us look at the proof of Proposition 2.3. It follows from Lemma 2.2 that
\[ x_{t|t-1}^0 = (A_0' \Lambda_0^{-1} A_0)^{-1} A_0' \Lambda_0^{-1} (A_0 x_t + u_t) \]
\[ - \sum_{k=0}^{t-1} (I_q - \Omega_0^{-1})^k (A_0' \Lambda_0^{-1} A_0)^{-1} A_0' \Lambda_0^{-1} (A_0 v_{t-k} + (u_{t-k} - u_{t-k-1})) \]
\[ = x_t + (A_0' \Lambda_0^{-1} A_0)^{-1} A_0' \Lambda_0^{-1} u_t - \sum_{k=0}^{t-1} (I_q - \Omega_0^{-1})^k (A_0' \Lambda_0^{-1} A_0)^{-1} A_0' \Lambda_0^{-1} (u_{t-k} - u_{t-k-1}) \]
\[ - \sum_{k=0}^{t-1} (I_q - \Omega_0^{-1})^k v_{t-k}. \] (45)
Moreover, we may easily deduce that

\[ \sum_{k=0}^{t-1} (I_q - \Omega_0^{-1})^k (A'_0 \Lambda_0^{-1} A_0)^{-1} A'_0 \Lambda_0^{-1} (u_{t-k} - u_{t-k-1}) \]

\[ = (A'_0 \Lambda_0^{-1} A_0)^{-1} A'_0 \Lambda_0^{-1} u_t - \Omega_0^{-1} \sum_{k=1}^{t-1} (I_q - \Omega_0^{-1})^k (A'_0 \Lambda_0^{-1} A_0)^{-1} A'_0 \Lambda_0^{-1} u_{t-k}. \]

(46)

The stated result now follows immediately from (45) and (46).

\[ \square \]

**Proof of Lemma 3.1** In the proof, we use the generic notation \((w_t)\) to signify any stationary linear process driven by \((u_t)\) and \((v_t)\). In particular, the definition of \((w_t)\) may be different from line to line. It follows from Lemma 2.2 that

\[ x_{t|t-1} = (A' \Lambda^{-1} A)^{-1} A' \Lambda^{-1} y_t + w_t, \]

(47)

under our convention here. We define the commutation matrix \(K_{ab}\) by

\[ K_{ab} \text{vec} A = \text{vec} A' \]

(48)

for \(a \times b\) matrix \(A\). Recall that we define \(\text{vec}\) to be the operator stacking rows, not the columns, of a matrix. Therefore, if we let \(\overline{\text{vec}}\) be the operator stacking columns of a matrix, and let \(\overline{K}_{ab}\) be the commutation matrix such that \(\overline{K}_{ab} \overline{\text{vec}} A = \overline{\text{vec}} A'\), then we have \(K_{ab} = \overline{K}_{ba}\). The readers are referred to Magnus and Neudecker (1988) for more on the commutation matrix.

Since

\[ \varepsilon_t = y_t - y_{t|t-1} = y_t - A x_{t|t-1} \]

and

\[ \text{vec} A x_{t|t-1} = (I_p \otimes x'_{t|t-1}) \text{vec} A, \]

we may easily deduce that

\[ \frac{\partial \varepsilon_t}{\partial (\text{vec} A)} = -A \frac{\partial x_{t|t-1}}{\partial (\text{vec} A)} - I_p \otimes x'_{t|t-1}, \]

(49)

and

\[ \frac{\partial \varepsilon_t}{\partial (\text{vec} \Lambda)} = -A \frac{\partial x_{t|t-1}}{\partial (\text{vec} \Lambda)}'. \]

(50)

The partial derivatives of \(\varepsilon_t\) with respect to \(\text{vec} A\) and \(\text{vec} \Lambda\) may therefore be easily obtained from (49) and (50), once we find the partial derivatives of \(x_{t|t-1}\) with respect to \(\text{vec} A\) and \(\text{vec} \Lambda\). This is what we will focus on in the rest of the proof.
Firstly, in order to get the partial derivative of $x_{t|t-1}$ with respect to $vecA$, we assume $\Lambda$ to be fixed. It follows from (47) that

$$dx_{t|t-1} = -(A'\Lambda^{-1}A)^{-1}(dA'\Lambda^{-1}A + A'\Lambda^{-1}dA)(A'\Lambda^{-1}A)^{-1}A'\Lambda^{-1}yt$$

$$+ (A'\Lambda^{-1}A)^{-1}dA'\Lambda^{-1}yt + wt$$

$$= -(A'\Lambda^{-1}A)^{-1}dA'(\Lambda^{-1}A(A'\Lambda^{-1}A)^{-1}A'\Lambda^{-1}yt)$$

$$+ (A'\Lambda^{-1}A)^{-1}A'\Lambda^{-1}dA((A'\Lambda^{-1}A)^{-1}A'\Lambda^{-1}yt)$$

$$+ (A'\Lambda^{-1}A)^{-1}dA'(A'\Lambda^{-1}yt) + wt,$$

and after applying $vec$, we get

$$dx_{t|t-1} = -\left[(A'\Lambda^{-1}A)^{-1} \otimes y'_t\Lambda^{-1}A(A'\Lambda^{-1}A)^{-1}A'\Lambda^{-1}\right] vecA'$$

$$- \left[(A'\Lambda^{-1}A)^{-1}A'\Lambda^{-1} \otimes y'_t\Lambda^{-1}A(A'\Lambda^{-1}A)^{-1}\right] vecA$$

$$+ \left[(A'\Lambda^{-1}A)^{-1} \otimes y'_t\Lambda^{-1}\right] vecA' + wt$$

$$= -\left[(A'\Lambda^{-1}A)^{-1} \otimes y'_t\Lambda^{-1}A(A'\Lambda^{-1}A)^{-1}A'\Lambda^{-1}\right] K_{pq}vecA$$

$$- \left[(A'\Lambda^{-1}A)^{-1}A'\Lambda^{-1} \otimes y'_t\Lambda^{-1}A(A'\Lambda^{-1}A)^{-1}\right] vecA$$

$$+ \left[(A'\Lambda^{-1}A)^{-1} \otimes y'_t\Lambda^{-1}\right] K_{pq}vecA + wt.$$ 

Consequently, we have

$$\frac{\partial x_{t|t-1}}{\partial (vecA)} = -\left[(A'\Lambda^{-1}A)^{-1} \otimes y'_t\Lambda^{-1}A(A'\Lambda^{-1}A)^{-1}A'\Lambda^{-1}\right] K_{pq}$$

$$- (A'\Lambda^{-1}A)^{-1}A'\Lambda^{-1} \otimes y'_t\Lambda^{-1}A(A'\Lambda^{-1}A)^{-1}$$

$$+ \left[(A'\Lambda^{-1}A)^{-1} \otimes y'_t\Lambda^{-1}\right] K_{pq} + wt$$

$$= -y'_t\Lambda^{-1}A(A'\Lambda^{-1}A)^{-1}A'\Lambda^{-1} \otimes (A'\Lambda^{-1}A)^{-1}$$

$$- (A'\Lambda^{-1}A)^{-1}A'\Lambda^{-1} \otimes y'_t\Lambda^{-1}A(A'\Lambda^{-1}A)^{-1}$$

$$+ y'_t\Lambda^{-1} \otimes (A'\Lambda^{-1}A)^{-1} + wt.$$ 

(51)

It follows immediately from (51) that

$$\frac{\partial x_{t|t-1}}{\partial vecA} = -(\Lambda_{0}^{-1}A_0x_t + wt) \otimes (A_0'\Lambda_0^{-1}A_0)^{-1} - \Lambda_{0}^{-1}A_0(A_0'\Lambda_0^{-1}A_0)^{-1} \otimes (x_t + wt)$$

$$+ (\Lambda_{0}^{-1}A_0x_t + wt) \otimes (A_0'\Lambda_0^{-1}A_0)^{-1}$$

$$= -\Lambda_{0}^{-1}A_0(A_0'\Lambda_0^{-1}A_0)^{-1} \otimes x_t + wt.$$ 

(52)

From (49) and (52), we can easily deduce that

$$\frac{\partial \xi^0_t}{\partial vecA} = -\frac{\partial x_{t|t-1}}{\partial vecA} \Lambda_0^t \otimes x_{t|t-1}^0$$

$$= \Lambda_0^{-1}A_0(A_0'\Lambda_0^{-1}A_0)^{-1}A_0^t \otimes x_t - I_p \otimes x_{t|t-1}^0$$

$$= -[I_p - \Lambda_0^{-1}A_0(A_0'\Lambda_0^{-1}A_0)^{-1}A_0^t] \otimes x_t + wt,$$
as was to be shown.

Secondly, we consider the partial derivative of \( x_{t|t-1} \) with respect to \( \text{vec}\Lambda \). Similarly, we assume \( A \) to be fixed, and then derive

\[
\begin{align*}
\frac{\partial x_{t|t-1}}{\partial (\text{vec}\Lambda)} &= (A'\Lambda^{-1}A)^{-1}A'\Lambda^{-1} (x_t A'\Lambda^{-1} A^{-1} A'\Lambda^{-1} - I_p) + w_t \\
&= (A'\Lambda^{-1}A)^{-1} A'\Lambda^{-1} \otimes y_t A'\Lambda^{-1} A^{-1} A'\Lambda^{-1} - I_p + w_t \\
&= (A'\Lambda^{-1}A)^{-1} A'\Lambda^{-1} \otimes (x_t A_0 + u_t)A^{-1} A'\Lambda^{-1} - I_p + w_t
\end{align*}
\]

from which it is obvious that

\[
\frac{\partial x_{t|t-1}}{\partial \text{vec}\Lambda} = w_t. \tag{53}
\]

It is now straightforward to have

\[
\frac{\partial \varepsilon_t}{\partial \text{vec}\Lambda} = w_t, \tag{54}
\]

due to (50) and (53). The proof is therefore complete.

\[\square\]

**Proof of Lemma 3.2**  It follows immediately from (23) that

\[
V_n(r) = -B_0' \otimes \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} v_t + o_p(1).
\]

Moreover, due to (24), \( T_n' (\partial \varepsilon_t') / \partial \theta \) is a stationary linear process and \( \mathcal{F}_{t-1} \)-measurable. Consequently, \( W_n \) is a partial sum process of the martingale difference sequence \( T_n' (\partial \varepsilon_t' / \partial \theta) \Sigma_0^{-1} \varepsilon_t \).

The stated results can therefore be readily deduced from the invariance principle for the martingale difference sequence.

\[\square\]

**Proof of Lemma 3.3**  The stated result can be deduced just as in the proof of Lemma 3.4 in CMP.

\[\square\]
Proof of Theorem 3.4  The proof will be done in three steps, each of which will establish ML1, ML2 and ML3. As in CMP, we use the following notational convention in the proof:

(a) \( (w_t) \) denotes a linear process driven by \( (u_s)_{s=1}^T \) and \( (v_s)_{s=1}^T \) that has geometrically decaying coefficients, and

(b) \( (\bar{w}_t) \) is such a process that is \( \mathcal{F}_t \)-measurable.

The notations \( (w_t) \) and \( (\bar{w}_t) \) are generic and signify any processes satisfying the conditions specified above. In general, \( (w_t) \) and \( (\bar{w}_t) \) appearing in different lines may represent different processes.

First Step  ML1 holds with \( N \) given in the theorem, as shown in the proof of Theorem 3.5 in CMP. In particular, we have

\[
\frac{1}{n} T'_N s_n(\theta_0) = \frac{1}{2\sqrt{n}} T'_N \frac{\partial (\text{vec} \Sigma_0)}{\partial \theta} (\Sigma_0^{-1} \otimes \Sigma_0^{-1}) \text{vec} \left[ \frac{1}{\sqrt{n}} \sum_{t=1}^n (\varepsilon_0^t \varepsilon_0'^t - \Sigma_0) \right] \\
- \frac{1}{n} \sum_{t=1}^n T'_N \frac{\partial \Sigma_0^{-1}}{\partial \theta} \Sigma_0^{-1} \varepsilon_t \\
= -\frac{1}{n} \sum_{t=1}^n T'_N \frac{\partial \Sigma_0^{-1}}{\partial \theta} \Sigma_0^{-1} \varepsilon_t + O_p(n^{-1/2}) \\
\to_d - \int_0^1 V(r) \, dU(r)
\]

and

\[
\frac{1}{\sqrt{n}} T'_S s_n(\theta_0) = \frac{1}{2} T'_S \frac{\partial (\text{vec} \Sigma_0)}{\partial \theta} (\Sigma_0^{-1} \otimes \Sigma_0^{-1}) \text{vec} \left[ \frac{1}{\sqrt{n}} \sum_{t=1}^n (\varepsilon_0^t \varepsilon_0'^t - \Sigma_0) \right] \\
- \frac{1}{\sqrt{n}} \sum_{t=1}^n T'_S \frac{\partial \Sigma_0^{-1}}{\partial \theta} \Sigma_0^{-1} \varepsilon_t \\
\to_d Z - W
\]

as \( n \to \infty \).

Second Step  Now we establish ML2. Similar to the proof of Theorem 3.5 in CMP, we can easily deduce that

\[
\frac{1}{n^2} T'_N H_n(\theta_0) T_N \to_p - \int_0^1 V(r) \Sigma_0^{-1} V(r)' \, dr \\
(55)
\]

and

\[
\frac{1}{n^{3/2}} T'_N H_n(\theta_0) T_S = O_p(n^{-1/2}) \\
(56)
\]
as \( n \to \infty \), which are in particular due to
\[
\sum_{t=1}^{n} \left( I \otimes \varepsilon_{t}^{0} \Sigma_{0}^{-1} \right) \left( \frac{\partial^{2}}{\partial \theta \partial \theta'} \otimes \varepsilon_{t}^{0} \right) = O_p(n)
\]
\[
\frac{\partial (vec \Sigma_{0})'}{(\Sigma_{0}^{-1} \otimes \Sigma_{0}^{-1})} \sum_{t=1}^{n} \left( \frac{\partial \varepsilon_{t}^{0}}{\partial \theta'} \otimes \varepsilon_{t}^{0} \right) = O_p(n)
\]
\[
\sum_{t=1}^{n} \left( \frac{\partial \varepsilon_{t}^{0}}{\partial \theta'} \otimes \varepsilon_{t}^{0} \right) (\Sigma_{0}^{-1} \otimes \Sigma_{0}^{-1}) \frac{\partial (vec \Sigma_{0})}{\partial \theta'} = O_p(n)
\]
for large \( n \). Note that \( \Sigma \) is the steady state value and only depends on \( A \) and \( \Lambda \).

Hence, ML2 will be established if we can further show that
\[
\frac{1}{n} T_{S}^{T} H_{n}(\theta_{0}) T_{S} \rightarrow_{p} -[\text{var}(W) + \text{var}(Z)].
\]  
(57)

Notice that
\[
\frac{1}{n} T_{S}^{T} H_{n}(\theta_{0}) T_{S} = A_{n} + B_{n} + C_{n} + (D_{n} + D_{n}') + o_{p}(1),
\]
where
\[
A_{n} = -\frac{1}{2} T_{S} \left[ \frac{\partial (vec \Sigma_{0})'}{(\Sigma_{0}^{-1} \otimes \Sigma_{0}^{-1})} \frac{\partial (vec \Sigma_{0})}{\partial \theta'} \right] T_{S} + o_{p}(1)
\]
\[
B_{n} = -\frac{1}{n} \sum_{t=1}^{n} T_{S} \left( \frac{\partial \varepsilon_{t}^{0} \Sigma_{0}^{-1}}{\partial \theta} \right) T_{S}
\]
\[
C_{n} = -\frac{1}{n} \sum_{t=1}^{n} T_{S} \left[ (I \otimes \varepsilon_{t}^{0} \Sigma_{0}^{-1}) \left( \frac{\partial^{2}}{\partial \theta \partial \theta'} \otimes \varepsilon_{t}^{0} \right) \right] T_{S}
\]
\[
D_{n} = T_{S} \left[ \frac{\partial (vec \Sigma_{0})'}{(\Sigma_{0}^{-1} \otimes \Sigma_{0}^{-1})} \frac{1}{n} \sum_{t=1}^{n} \left( \frac{\partial \varepsilon_{t}^{0}}{\partial \theta'} \otimes \varepsilon_{t}^{0} \right) \right] T_{S}.
\]

As shown in CMP,
\[
A_{n} = -\text{var}(Z) + o_{p}(1)
\]
\[
B_{n} = -\text{var}(W) + o_{p}(1)
\]
\[
D_{n} = O_{p}(n^{-1/2})
\]
for large \( n \). Therefore, it suffices to show that
\[
C_{n} = \begin{pmatrix} C_{n}(A, A) & C_{n}(A, \Lambda) \\ C_{n}(\Lambda, A) & C_{n}(\Lambda, \Lambda) \end{pmatrix} = O_{p}(n^{-1/2})
\]  
(58)

to deduce (57). Note that we have from (51)
\[
I_{q} \otimes x_{t|t-1} + (A' \otimes I_{q}) \frac{\partial x_{t|t-1}}{\partial vec A} = \overline{w}_{t-1},
\]  
(59)
which will be used below to establish (58).

First, we prove

$$C_n(A, A) = O_p(n^{-1/2}).$$

(60)

It follows from (49) that

$$\text{vec} \frac{\partial \varepsilon'_t}{\partial \text{vec}A} = \text{vec}(I_p \otimes x_{t|t-1}) - \text{vec} \frac{\partial x'_{t|t-1}}{\partial \text{vec}A} A'.$$

(61)

Because

$$\text{vec}(I_p \otimes x_{t|t-1}) = (I_p \otimes K_{pq})[(\text{vec} I_p) \otimes x_{t|t-1}]$$

and

$$\frac{\partial x'_{t|t-1}}{\partial \text{vec}A} A' = (I_{pq} \otimes A) \frac{\partial x'_{t|t-1}}{\partial \text{vec}A},$$

we derive that

$$\frac{\partial}{\partial (\text{vec}A)} \frac{\partial \varepsilon'_t}{\partial \text{vec}A} = -(I_p \otimes K_{pq}) \left[(\text{vec} I_p) \otimes \frac{\partial x'_{t|t-1}}{\partial (\text{vec}A)}\right] - \left(\frac{\partial x'_{t|t-1}}{\partial \text{vec}A} \otimes I_p\right) K_{pq} - (I_{pq} \otimes A) \frac{\partial}{\partial (\text{vec}A)} \frac{\partial x'_{t|t-1}}{\partial \text{vec}A}$$

(62)

In what follows, we will show

$$(A'_0 \otimes I_q) \left(I_{pq} \otimes \varepsilon_t^0 \Sigma_{0}^{-1}\right) \left(\frac{\partial}{\partial (\text{vec}A)} \frac{\partial \varepsilon'_t}{\partial \text{vec}A}\right) (A_0 \otimes I_q) = \bar{w}_{t-1} \varepsilon'_t,$$

(63)

from which (60) follows immediately. Given (62), it is equivalent to showing

$$(A'_0 \otimes I_q) \left(I_{pq} \otimes \varepsilon_t^0 \Sigma_{0}^{-1}\right) (I_p \otimes K_{pq}) \left[(\text{vec} I_p) \otimes \frac{\partial x'_{t|t-1}}{\partial (\text{vec}A)}\right] (A_0 \otimes I_q)$$

$$+ (A'_0 \otimes I_q) \left(I_{pq} \otimes \varepsilon_t^0 \Sigma_{0}^{-1}\right) \left(\frac{\partial x'_{t|t-1}}{\partial \text{vec}A} \otimes I_p\right) K_{pq} (A_0 \otimes I_q)$$

$$+ (A'_0 \otimes I_q) \left(I_{pq} \otimes \varepsilon_t^0 \Sigma_{0}^{-1}\right) (I_{pq} \otimes A_0) \left(\frac{\partial}{\partial (\text{vec}A)} \frac{\partial x'_{t|t-1}}{\partial \text{vec}A}\right) (A_0 \otimes I_q)$$

$$= \bar{w}_{t-1} \varepsilon'_t$$

(64)
To establish (64), we will analyze the three terms on the left hand side of equation (64) separately. For the first term in (64), we have

\[
(A_0' \otimes I_q) \left( I_{pq} \otimes \varepsilon^0_t \Sigma^{-1}_0 \right) (I_p \otimes K_{pq}) \left[ \left( vec I_p \right) \otimes \frac{\partial x^0_{t|t-1}}{\partial (vec A)} \right] (A_0 \otimes I_q)
\]

\[
= (A_0' \otimes I_q \otimes \varepsilon^0_t \Sigma^{-1}_0) (I_p \otimes K_{pq}) \left[ \left( vec I_p \right) \otimes \frac{\partial x^0_{t|t-1}}{\partial (vec A)} \right] (A_0 \otimes I_q)
\]

\[
= A_0' \otimes (I_q \otimes \varepsilon^0_t \Sigma^{-1}_0) \left( vec I_p \right) \otimes \frac{\partial x^0_{t|t-1}}{\partial (vec A)} (A_0 \otimes I_q)
\]

\[
= (A_0' \otimes \varepsilon^0_t \Sigma^{-1}_0 \otimes I_q) \left[ \left( vec I_p \right) \otimes \frac{\partial x^0_{t|t-1}}{\partial (vec A)} \right] (A_0 \otimes I_q)
\]

\[
= A_0' \Sigma^{-1}_0 \varepsilon^0_t \otimes I_q \otimes x'_t + \varpi_{t-1} \varepsilon^0_t. \tag{65}
\]

For the second term in (64), we may deduce that

\[
(A_0' \otimes I_q) \left( I_{pq} \otimes \varepsilon^0_t \Sigma^{-1}_0 \right) \left( \frac{\partial x^0_{t|t-1}}{\partial (vec A)} \otimes I_q \right) K_{pq} (A_0 \otimes I_q)
\]

\[
= (A_0' \otimes I_q \otimes \varepsilon^0_t \Sigma^{-1}_0) \left( \frac{\partial x^0_{t|t-1}}{\partial (vec A)} \otimes I_q \right) (I_q \otimes A_0) K_{pq}
\]

\[
= (A_0' \otimes I_q \otimes \varepsilon^0_t \Sigma^{-1}_0) \left( \frac{\partial x^0_{t|t-1}}{\partial (vec A)} \otimes A_0 \right) K_{qq}
\]

\[
= \left[ \left( A_0' \otimes I_q \right) \frac{\partial x^0_{t|t-1}}{\partial (vec A)} \otimes \varepsilon^0_t \Sigma^{-1}_0 A_0 \right] K_{qq}
\]

\[
= \varepsilon^0_t \Sigma^{-1}_0 A_0 \otimes \left[ \left( A_0' \otimes I_q \right) \frac{\partial x^0_{t|t-1}}{\partial (vec A)} \right]
\]

\[
= \varepsilon^0_t \Sigma^{-1}_0 A_0 \otimes I_q \otimes x'_t + \varpi_{t-1} \varepsilon^0_t. \tag{66}
\]

The third term in (64) can be written as

\[
(A_0' \otimes I_q) \left( I_{pq} \otimes \varepsilon^0_t \Sigma^{-1}_0 \right) \left( I_{pq} \otimes A_0 \right) \left( \frac{\partial}{\partial (vec A)} vec \frac{\partial x^0_{t|t-1}}{\partial (vec A)} \right) (A_0 \otimes I_q)
\]

\[
= (A_0' \otimes I_q \otimes \varepsilon^0_t \Sigma^{-1}_0) \left( I_{pq} \otimes A_0 \right) \left( \frac{\partial}{\partial (vec A)} vec \frac{\partial x^0_{t|t-1}}{\partial (vec A)} \right) (A_0 \otimes I_q)
\]

\[
= (A_0' \otimes I_q \otimes \varepsilon^0_t \Sigma^{-1}_0 A_0) \left( \frac{\partial}{\partial (vec A)} vec \frac{\partial x^0_{t|t-1}}{\partial (vec A)} \right) (A_0 \otimes I_q), \tag{67}
\]
Note that after taking derivative with respect to $\text{vec}A$ on equation (59), we get
\[
(I_q \otimes K_{qq}) \left[ (\text{vec}I_q) \otimes \frac{\partial x^0_{t|t-1}}{\partial (\text{vec}A)^T} \right] + (A' \otimes I_q \otimes I_q) \left( \frac{\partial}{\partial (\text{vec}A)^T} \text{vec} \frac{\partial x^0_{t|t-1}}{\partial \text{vec}A} \right) \\
+ \left( I_q \otimes I_q \otimes \frac{\partial x^0_{t|t-1}}{\partial (\text{vec}A)^T} \right) (I_q \otimes K_{pq} \otimes I_q) [K_{pq} \otimes (\text{vec}I_q)] = \bar{w}_{t-1}, \tag{68}
\]
which due in particular to
\[
\text{vec}(I_q \otimes x_{t|t-1}) = (I_q \otimes K_{qq}) \left[ (\text{vec}I_q) \otimes x_{t|t-1} \right]
\]
and
\[
\text{vec}(A' \otimes I_q) \frac{\partial x^0_{t|t-1}}{\partial \text{vec}A} = (A' \otimes I_q \otimes I_q) \text{vec} \frac{\partial x^0_{t|t-1}}{\partial \text{vec}A} \\
= \left( I_q \otimes I_q \otimes \frac{\partial x^0_{t|t-1}}{\partial (\text{vec}A)^T} \right) \text{vec}(A' \otimes I_q),
\]
where
\[
\text{vec}(A' \otimes I_q) = (I_q \otimes K_{pq} \otimes I_q) [(\text{vec}A') \otimes (\text{vec}I_q)] \\
= (I_q \otimes K_{pq} \otimes I_q) [(K_{pq} \text{vec}A) \otimes (\text{vec}I_q)].
\]
See, e.g., Magnus and Neudecker (1988) for the rules in matrix algebra used here.

Now we pre- and post-multiply all three terms in (68) by $I_q \otimes I_q \otimes \varepsilon^0_t \Sigma^{-1}_0 A_0$ and $A_0 \otimes I_q$, separately. The first term after multiplication is
\[
(I_q \otimes I_q \otimes \varepsilon^0_t \Sigma^{-1}_0 A_0) (I_q \otimes K_{qq}) \left[ (\text{vec}I_q) \otimes \frac{\partial x^0_{t|t-1}}{\partial (\text{vec}A)^T} \right] (A_0 \otimes I_q) \\
= (I_q \otimes \varepsilon^0_t \Sigma^{-1}_0 A_0 \otimes I_q) \left( (\text{vec}I_q) \otimes \left[ \frac{\partial x^0_{t|t-1}}{\partial (\text{vec}A)^T} (A_0 \otimes I_q) \right] \right) \\
= A_0 \Sigma^{-1}_0 \varepsilon^0_t \otimes I_q \otimes x^0_t + \bar{w}_{t-1} \varepsilon^0_t, \tag{69}
\]
and the third term after multiplication is
\[
(I_q \otimes I_q \otimes \epsilon_t^{0} \Sigma_0^{-1} A_0) \left( I_q \otimes I_q \otimes \frac{\partial x_{t|t-1}^0}{\partial (\text{vec} A)^t} \right) (I_q \otimes K_{pq} \otimes I_q) [K_{pq}(A_0 \otimes I_q) \otimes (\text{vec} I_q)] (A_0 \otimes I_q)
\]
\[
= \left[ I_q \otimes I_q \otimes \epsilon_t^{0} \Sigma_0^{-1} A_0 (A_0^0 \Sigma_0^{-1} A_0)^{-1} A_0^0 \Sigma_0^{-1} \otimes I_q \otimes x_t' \right] (I_q \otimes K_{pq} \otimes I_q) [K_{pq}(A_0 \otimes I_q) \otimes (\text{vec} I_q)] + \bar{w}_{t-1} \epsilon_t^0
\]
\[
= \left[ \epsilon_t^{0} \Sigma_0^{-1} A_0 (A_0^0 \Sigma_0^{-1} A_0)^{-1} A_0^0 \Sigma_0^{-1} \otimes I_q \right] (A_0 \otimes I_q) \otimes x_t + \bar{w}_{t-1} \epsilon_t^0.
\]
It then follows from (68), (69) and (70) that the second term after multiplication is
\[
(A_0^0 \otimes I_q \otimes \epsilon_t^{0} \Sigma_0^{-1} A_0) \left( \frac{\partial}{\partial (\text{vec} A)^t} \right) \left( \frac{\partial x_{t|t-1}^0}{\partial \text{vec} A} \right) (A_0 \otimes I_q)
\]
\[
= -A_0^0 \Sigma_0^{-1} \epsilon_t^0 \otimes I_q \otimes x_t' - \epsilon_t^{0} \Sigma_0^{-1} A_0 \otimes I_q \otimes x_t + \bar{w}_{t-1} \epsilon_t^0,
\]
(71) (64) now follows immediately from (65), (66) (67) and (71). (60) is therefore established.
Second, we prove that
\[
C_n(A, \Lambda) = O_p(n^{-1/2}).
\]
From (61), we can easily derive
\[
\text{vec} \frac{\partial \epsilon_t'}{\partial \text{vec} A} = - (I_p \otimes K_{pq})[(\text{vec} I_p) \otimes x_{t|t-1}] - (I_{pq} \otimes A) \text{vec} \frac{\partial x_{t|t-1}^0}{\partial \text{vec} A},
\]
which gives
\[
\frac{\partial}{\partial (\text{vec} A)^t} \frac{\partial \epsilon_t'}{\partial \text{vec} A} = - (I_p \otimes K_{pq}) \left[ (\text{vec} I_p) \otimes \frac{\partial x_{t|t-1}^0}{\partial (\text{vec} A)^t} \right] - (I_{pq} \otimes A) \frac{\partial}{\partial (\text{vec} A)^t} \frac{\partial x_{t|t-1}^0}{\partial \text{vec} A}.
\]
We will now show that
\[
(A_0^0 \otimes I_q) (I_{pq} \otimes \epsilon_t^{0} \Sigma_0^{-1}) \left[ \frac{\partial}{\partial (\text{vec} A)^t} \frac{\partial \epsilon_t'}{\partial \text{vec} A} \right] \lambda = \bar{w}_{t-1} \epsilon_t^0
\]
for any \(p^2\)-dimensional vector \(\lambda\). Clearly, (72) can be deduced immediately from (74).
We may easily derive that

\[
(A_0' \otimes I_q) \left( I_{pq} \otimes \varepsilon_t^{0 \Sigma_0^{-1}} \right) (I_p \otimes K_{pq}) \left( (\text{vec} I_p) \otimes \frac{\partial x_0^{t|t-1}}{\partial (\text{vec} \Lambda)^{t}} \right) \lambda \\
= (A_0' \otimes I_q) \left( I_{pq} \otimes \varepsilon_t^{0 \Sigma_0^{-1}} \right) (I_p \otimes K_{pq}) \left( (\text{vec} I_p) \otimes \frac{\partial x_0^{0|t-1}}{\partial (\text{vec} A)^{t}} \right) \lambda \\
= (A_0' \otimes I_q) \left( I_{pq} \otimes \varepsilon_t^{0 \Sigma_0^{-1}} \otimes I_q \right) (I_p \otimes K_{pq}) \left( (\text{vec} I_p) \otimes \frac{\partial x_0^{0|t-1}}{\partial (\text{vec} A)^{t}} \right) \lambda \\
= A_0' \Sigma_0^{-1} \varepsilon_t^{0} \otimes \frac{\partial x_0^{0|t-1}}{\partial (\text{vec} A)^{t}} \lambda \\
= \overline{w}_{t-1} \varepsilon_t^{0}, \tag{75}
\]

where the last step is due to (53). Consequently, in order to establish (74), it suffices to show that

\[
(A_0' \otimes I_q) \left( I_{pq} \otimes \varepsilon_t^{0 \Sigma_0^{-1}} \right) (I_p \otimes A_0) \frac{\partial}{\partial (\text{vec} \Lambda)^{t}} \text{vec} \frac{\partial x_0^{0|t-1}}{\partial (\text{vec} A)^{t}} \lambda \\
= (A_0' \otimes I_q) \left( I_{pq} \otimes \varepsilon_t^{0 \Sigma_0^{-1}} \right) (I_p \otimes A_0) \frac{\partial}{\partial (\text{vec} \Lambda)^{t}} \text{vec} \frac{\partial x_0^{0|t-1}}{\partial (\text{vec} A)^{t}} \lambda \\
= (A_0' \otimes I_q) \left( I_{pq} \otimes \varepsilon_t^{0 \Sigma_0^{-1}} \otimes I_q \right) (I_p \otimes A_0) \frac{\partial}{\partial (\text{vec} \Lambda)^{t}} \text{vec} \frac{\partial x_0^{0|t-1}}{\partial (\text{vec} A)^{t}} \lambda \\
= \overline{w}_{t-1} \varepsilon_t^{0}, \tag{76}
\]

To establish (76), we use equation (59). After taking \( \text{vec} \) on (59), we get

\[
(I_q \otimes K_{qq})[(\text{vec} I_q) \otimes x_{t|t-1}] + (A' \otimes I_q \otimes I_q) \text{vec} \frac{\partial x_{t|t-1}'}{\partial (\text{vec} A)^{t}} = \overline{w}_{t-1},
\]

from which it follows that

\[
(I_q \otimes K_{qq}) \left( \frac{\partial x_{t|t-1}}{\partial (\text{vec} \Lambda)^{t}} \right) + (A' \otimes I_q \otimes I_q) \frac{\partial}{\partial (\text{vec} \Lambda)^{t}} \text{vec} \frac{\partial x_{t|t-1}'}{\partial (\text{vec} A)^{t}} = \overline{w}_{t-1}. \tag{77}
\]

We evaluate (77) at the true values of \( A \) and \( \Lambda \), and then pre- and post-multiply both sides by

\[
I_q \otimes I_q \otimes \varepsilon_t^{0 \Sigma_0^{-1}} A_0 \quad \text{and} \quad \lambda,
\]

respectively, to get

\[
A_0' \Sigma_0^{-1} \varepsilon_t^{0} \otimes \frac{\partial x_0^{0|t-1}}{\partial (\text{vec} A)^{t}} \lambda + (A_0' \otimes I_q \otimes \varepsilon_t^{0 \Sigma_0^{-1}} A_0) \frac{\partial}{\partial (\text{vec} \Lambda)^{t}} \text{vec} \frac{\partial x_0^{0|t-1}}{\partial (\text{vec} A)^{t}} \lambda = \overline{w}_{t-1} \varepsilon_t^{0}, \tag{78}
\]
noting that

\[(I_q \otimes I_q \otimes \varepsilon_t^0 \Sigma_0^{-1} A_0)(I_q \otimes K_{qq}) \left[ (\text{vec} I_q) \otimes \frac{\partial x_0^{t-1}}{\partial (\text{vec} \Lambda)} \right] \lambda \]

\[= (I_q \otimes \varepsilon_t^0 \Sigma_0^{-1} A_0 \otimes I_q) \left[ (\text{vec} I_q) \otimes \frac{\partial x_0^{t-1}}{\partial (\text{vec} \Lambda)} \right] \]

\[= A'_0 \Sigma_0^{-1} \varepsilon_t^0 \left[ \frac{\partial x_0^{t-1}}{\partial (\text{vec} \Lambda)} \right] \lambda \].

(76) then follows immediately from (53) and (78). From (75) and (76), (74) is established.

The proof for (72) is now complete.

The proof for \(C_n(\Lambda, \Lambda)\) is straightforward, as in CMP. We therefore have established (58), and the proof for the second step is complete.

**Third Step** To establish ML3, as in CMP, we let

\[\mu_n = \nu_n^{1-\delta}\]

for some \(\delta > 0\) small, and let \(\theta \in \Theta_n\) be arbitrarily chosen. Since

\[ (B'_0 \Lambda_0^{-1} \otimes I_k) (\text{vec} A - \text{vec} A_0) = O(n^{-1+\delta}) \]

\[ (A'_0 \Lambda_0^{-1} A_0)^{-1/2} A'_0 \Lambda_0^{-1} \otimes I_k) (\text{vec} A - \text{vec} A_0) = O(n^{-1/2+\delta}) \]

\[\text{vec} A - \text{vec} A_0 = O(n^{-1/2+\delta}),\]

we have

\[\text{vec} A = \text{vec} A_0 + O_p(n^{-1/2+\delta}) \]

\[\text{vec} \Lambda = \text{vec} \Lambda_0 + O_p(n^{-1/2+\delta}).\] (79) (80)

We will show that

\[\frac{1}{n^{2(1-\delta)}} T_N' \sum_{t=1}^{n} \left( \frac{\partial \varepsilon_t'}{\partial \theta} - \frac{\partial \varepsilon_0'}{\partial \theta} \right) \Sigma_0^{-1} \left( \frac{\partial \varepsilon_t'}{\partial \theta'} - \frac{\partial \varepsilon_0'}{\partial \theta'} \right) T_N \rightarrow_p 0 \] (81)

\[\frac{1}{n^{2(1-\delta)}} T_N' \left[ \sum_{t=1}^{n} \left( \frac{\partial \varepsilon_t'}{\partial \theta} - \frac{\partial \varepsilon_0'}{\partial \theta} \right) \Sigma_0^{-1} \left( \frac{\partial \varepsilon_t'}{\partial \theta'} - \frac{\partial \varepsilon_0'}{\partial \theta'} \right) \right] \rightarrow_p 0 \] (82)

\[\frac{1}{n^{1-\delta}} \sum_{t=1}^{n} T_S' \left[ (I \otimes \varepsilon_t^0 \Sigma_0^{-1}) \left( \frac{\partial^2}{\partial \theta \partial \theta'} \right) \otimes (\varepsilon_t^0 - \varepsilon_0^0) \right] T_S \rightarrow_p 0 \] (83)

\[\frac{1}{n^{1-\delta}} T_S' \left[ \frac{\partial (\text{vec} \Sigma_0')}{\partial \theta} \left( \Sigma_0^{-1} \otimes \Sigma_0^{-1} \right) \sum_{t=1}^{n} \left( \frac{\partial \varepsilon_0^0}{\partial \theta'} \right) \otimes (\varepsilon_t^0 - \varepsilon_0^0) \right] T_S \rightarrow_p 0 \] (84)
and

\[
\frac{1}{n^{1-\delta}} T_S \left[ \sum_{t=1}^{n} \left( \frac{\partial \varepsilon_t^l}{\partial \theta} - \frac{\partial \varepsilon_t^0}{\partial \theta} \right) \left( \Sigma_{0}^{-1} \frac{\partial \varepsilon_t^0}{\partial \theta'} \right) \right] T_S \rightarrow_p 0 \quad (85)
\]

\[
T_S' \left[ \frac{\partial (\text{vec} \Sigma_0)' (\Sigma_0^{-1} \otimes \Sigma_0^{-1})}{\partial \theta} \frac{1}{n^{1-\delta}} \sum_{t=1}^{n} \left( \frac{\partial \varepsilon_t^l}{\partial \theta} - \frac{\partial \varepsilon_t^0}{\partial \theta} \right) \right] \left( \Sigma_{0}^{-1} \frac{\partial \varepsilon_t^0}{\partial \theta'} \right) T_S \rightarrow_p 0 \quad (86)
\]

\[
\frac{1}{n^{1-\delta}} T_S' \left[ \sum_{t=1}^{n} \left( \frac{\partial \varepsilon_t^l}{\partial \theta} - \frac{\partial \varepsilon_t^0}{\partial \theta} \right) \Sigma_{0}^{-1} \left( \frac{\partial \varepsilon_t^0}{\partial \theta'} \right) \right] T_S \rightarrow_p 0 \quad (87)
\]

\[
\frac{1}{n^{1-\delta}} \sum_{t=1}^{n} T_S' \left[ (I \otimes (\varepsilon_t^l - \varepsilon_t^0)) \Sigma_0^{-1} \right] \left( \frac{\partial^2 (\varepsilon_t^0}{\partial \theta \partial \theta'} \right) T_S \rightarrow_p 0 \quad (88)
\]

\[
\frac{1}{n^{1-\delta}} \sum_{t=1}^{n} T_S' \left[ (I \otimes (\varepsilon_t^l - \varepsilon_t^0)) \Sigma_0^{-1} \right] \left( \frac{\partial^2 \varepsilon_t^0}{\partial \theta \partial \theta'} \right) T_S \rightarrow_p 0 \quad (89)
\]

for all \( A \) and \( \Lambda \) satisfying (79) and (80) to establish ML3. In what follows, we use the generic notation \( \Delta(n^\kappa d_t) \) to denote the terms which include \( n^\kappa \) (or a lower order) times \( (d_t) \), where \( (d_t) \) can be stationary or nonstationary. Clearly, we have

\[
\varepsilon_t - \varepsilon_t^0 \frac{\partial \varepsilon_t^l}{\partial \theta} - \frac{\partial \varepsilon_t^0}{\partial \theta}, \quad \frac{\partial^2 \varepsilon_t^0}{\partial \theta \partial \theta'} \otimes (\varepsilon_t - \varepsilon_t^0) = \Delta(n^{-1/2+\delta} x_t) + w_t, \quad (91)
\]

since both \( A = A_0 + O(n^{-1/2+\delta}) \) and \( \Lambda = \Lambda_0 + O(n^{-1/2+\delta}) \). The results in (81)-(84) follow immediately from (91). In (84), note that

\[
\sum_{t=1}^{n} \left( \frac{\partial \varepsilon_t^0}{\partial \theta} \otimes (\varepsilon_t - \varepsilon_t^0) \right) T_S = \sum_{t=1}^{n} \left( \frac{\partial \varepsilon_t^0}{\partial \theta} T_S \otimes (\varepsilon_t - \varepsilon_t^0) \right) \quad \text{and} \quad \frac{\partial \varepsilon_t^0}{\partial \theta} T_S = w_t.
\]

The proofs for the results in (85)-(90) are more involved. In doing that, we need to show that

\[
T_S' \left( \frac{\partial \varepsilon_t^l}{\partial \theta} - \frac{\partial \varepsilon_t^0}{\partial \theta} \right) = \Delta(n^{-1/2+\delta} w_t) + \Delta(n^{-1+2\delta} d_t), \quad (92)
\]

which is equivalent to showing the following two equations:

\[
(A'_0 \otimes I_k) \left( \frac{\partial \varepsilon_t^l}{\partial \text{vec} A} - \frac{\partial \varepsilon_t^0}{\partial \text{vec} A} \right) = \Delta(n^{-1/2+\delta} w_t) + \Delta(n^{-1+2\delta} d_t), \quad (93)
\]

\[
\frac{\partial \varepsilon_t^l}{\partial \text{vec} A} - \frac{\partial \varepsilon_t^0}{\partial \text{vec} A} = \Delta(n^{-1/2+\delta} w_t) + \Delta(n^{-1+2\delta} d_t). \quad (94)
\]
From
\[
vec\left(\frac{\partial \varepsilon'_t}{\partial vec\Lambda} - \frac{\partial \varepsilon'_0}{\partial vec\Lambda}\right) = \frac{\partial}{\partial (vec\Lambda)' vec\varepsilon'_t} \frac{\partial \varepsilon'_t}{\partial vec\Lambda} (vec\Lambda - vec\Lambda_0) + \frac{\partial}{\partial (vec\Lambda)' vec\varepsilon'_0} \frac{\partial \varepsilon'_0}{\partial vec\Lambda} (vec\Lambda - vec\Lambda_0) + \Delta(n^{-1+2\delta} w_t),
\]
the result in (94) follows immediately from Lemma 3.1. In order to show (93), note that the left hand side of equation (93) is a matrix with elements
\[
A'_{i0} \left( \frac{\partial \varepsilon_{ij}}{\partial A_k} - \frac{\partial \varepsilon_{ij}^0}{\partial A_k} \right),
\]
where \(A_i\) and \(A_k\) represent the \(i\)th and \(k\)th columns of \(A\), respectively, \(A_{i0}\) is the true value of \(A_i\), and \(\varepsilon_{ij}\) is the \(j\)th element of \(\varepsilon_t\), for \(i, k = 1, 2, \ldots, q\) and \(j = 1, 2, \ldots, p\). We will follow the same notations in the rest of the proof. It is rather clear that showing (93) is equivalent to showing the following
\[
A'_{i0} \left( \frac{\partial \varepsilon_{ij}}{\partial A_k} - \frac{\partial \varepsilon_{ij}^0}{\partial A_k} \right) = A'_{i0} \left( \frac{\partial \varepsilon_{ij}^0}{\partial A_k} (A_1 - A_{10}) + \cdots + \frac{\partial \varepsilon_{ij}^0}{\partial A_k} (A_q - A_{q0}) \right) + \Delta(n^{-1+2\delta} d_t)
\]
\[
= \Delta(n^{-1/2+\delta} w_t) + \Delta(n^{-1+2\delta} d_t). \tag{95}
\]
Since \((A'_0 \otimes I_q') \frac{\partial \varepsilon'_t}{\partial vecA} = w_t\) due to Lemma 3.1, we have
\[
A'_{i0} \frac{\partial \varepsilon_{ij}^0}{\partial A_k} = w_t, \tag{96}
\]
which implies
\[
A'_{i0} \frac{\partial \varepsilon_{ij}^0}{\partial A_k} A'_l = w_t, \quad \text{for } l \neq i \tag{97}
\]
\[
A'_{i0} \frac{\partial \varepsilon_{ij}^0}{\partial A_k} A'_l + \frac{\partial \varepsilon_{ij}^0}{\partial A'_k} = w_t \tag{98}
\]
\[
A'_{i0} \frac{\partial \varepsilon_{ij}}{\partial vecA} = w_t. \tag{99}
\]
Due to (96) and (98), we have
\[
A'_{i0} \frac{\partial \varepsilon_{ij}^0}{\partial A_k} (A_i - A_{i0}) = \left( w_t - \frac{\partial \varepsilon_{ij}^0}{\partial A'_k} \right) (A_i - A_{i0}) = \Delta(-n^{1/2+\delta} w_t). \tag{100}
\]
The result in (95) follows immediately from those in (97), (99) and (100). Consequently, the results in (85)-(88) are established due to (91) and (92). Note that in (88),

\[
\left( \frac{\partial \varepsilon_t}{\partial \theta'} - \frac{\partial \varepsilon_0}{\partial \theta'} \right) \otimes \varepsilon_0^t \bigg| T_S = \left( \frac{\partial \varepsilon_t}{\partial \theta'} - \frac{\partial \varepsilon_0}{\partial \theta'} \right) T_S \otimes \varepsilon_0^t.
\]

Furthermore, the result in (92) implies that

\[
\left( \frac{\partial}{\partial \theta \partial \theta'} \bigg| T_S \otimes \varepsilon_0^t \right) = w_t,
\]

which proves the required results in (89) and (90). We therefore finish the third step and the proof for Theorem 3.4 is now complete. □

**Proof of Proposition 4.1**  The proof is trivial and therefore omitted. □

**Proof of Theorem 4.2**  We consider two cases, \( r > q \) and \( r < q \), separately below.

**Case 1: \( r > q \)**  All our asymptotics in the paper are not applicable for this case, since the true value \( A_0 \) of \( A \) has deficiency in column rank. The asymptotics for this degenerated case are developed in Chang, Jiang and Park (2013). In particular, they show that we have

\[
\log |\hat{\Sigma}_n(q)| - \log |\hat{\Sigma}_n(r)| = o_p(n^{-1})
\]

in this case. However, since \( c_n \to c > 0 \), an increase in the penalty term is asymptotically of order bigger than a decrease in the value of the leading term as \( r \) goes beyond \( q \). Therefore, we have \( \hat{q}_n \leq q \) with probability one asymptotically.

**Case 2: \( r < q \)**  In this case, the model is misspecified and the ML estimators \( \hat{A}_n \) and \( \hat{\Lambda}_n \) are no longer consistent. We have

\[
\hat{x}_{t|t-1} = (\hat{A}_n' \hat{\Lambda}_n^{-1} \hat{A}_n)^{-1} \hat{A}_n' \hat{\Lambda}_n^{-1} y_t - \sum_{k=0}^{t-1} (I_q - \hat{\Omega}_n^{-1})^k (\hat{A}_n' \hat{\Lambda}_n^{-1} \hat{A}_n)^{-1} \hat{A}_n' \hat{\Lambda}_n^{-1} \triangle y_{t-k} + (I_q - \hat{\Omega}_n^{-1})^{t-1} x_0,
\]

due to Lemma 2.2. It then follows that

\[
\hat{\varepsilon}_t = y_t - \hat{y}_{t|t-1} = y_t - \hat{A}_n \hat{x}_{t|t-1} = \left( I - \hat{\Lambda}_n (\hat{A}_n' \hat{\Lambda}_n^{-1} \hat{A}_n)^{-1} \hat{A}_n' \hat{\Lambda}_n^{-1} \right) A_0 x_t + \hat{\omega}_t,
\]

where we have \( (1/n) \sum_{t=1}^n \hat{\omega}_t \hat{\omega}_t' = O_p(1) \). Note that

\[
I - \hat{\Lambda}_n (\hat{A}_n' \hat{\Lambda}_n^{-1} \hat{A}_n)^{-1} \hat{A}_n' \hat{\Lambda}_n^{-1}
\]
is a (non-orthogonal) projection on the range $\mathcal{R}(\hat{B}_n)$ of $\hat{B}_n$, where $\hat{B}_n$ is a $p \times (p - r)$ matrix of full column rank defined by $\hat{A}_n'\hat{\Lambda}_n^{-1}\hat{B}_n = 0$, along the range $\mathcal{R}(\hat{A}_n)$ of $p \times r$ matrix $\hat{A}_n$ of full column rank.

Since $A_0$ is $p \times q$ matrix of full column rank and $r < q$, it is clear that the matrix

$$
\left( I - \hat{A}_n(\hat{A}_n'\hat{\Lambda}_n^{-1}\hat{A}_n)^{-1}\hat{A}_n'\hat{\Lambda}_n^{-1} \right) A_0
$$

is of rank at least $q - r > 0$ for all $n$, and we may deduce from (101) that

$$
\hat{\Sigma}_n(r) = \frac{1}{n} \sum_{t=1}^{n} \hat{\varepsilon}_t\hat{\varepsilon}_t' = O_p(n).
$$

On the other hand, we have $\hat{\Sigma}_n(q) = O_p(1)$. Therefore, it follows that

$$
\log |\hat{\Sigma}_n(r)| - \log |\hat{\Sigma}_n(q)| = O_p(\log(n))
$$

for all $r < q$. Consequently, for any choice of $(c_n)$ such that $c_n/(n \log n) \to 0$, an increase in the leading term completely dominates in the limit the corresponding decrease in the penalty term of (33) as $r$ deviates from $q$ and takes a value below $q$. This implies that $\hat{q}_n \geq q$ with probability one asymptotically. \qed