Fitting Dynamically Misspecified Models: An Optimal Transportation Approach

Jean-Jacques Forneron[*] Zhongjun Qu[†]

July 18, 2025

Abstract

This paper considers filtering, parameter estimation, and testing for potentially dynamically misspecified state-space models. When dynamics are misspecified, filtered values of state variables often do not satisfy model restrictions, making them hard to interpret, and parameter estimates may fail to characterize the dynamics of filtered variables. To address this, a sequential optimal transportation approach is used to generate a model-consistent sample by mapping observations from a flexible reduced-form to the structural conditional distribution iteratively. Filtered series from the generated sample are model-consistent. Specializing to linear processes, a closed-form Optimal Transport Filtering algorithm is derived. Minimizing the discrepancy between generated and actual observations defines an Optimal Transport Estimator. Its large sample properties are derived. A specification test determines if the model can reproduce the sample path, or if the discrepancy is statistically significant. Empirical applications to trend-cycle decomposition, DSGE models, and affine term structure models illustrate the methodology and the results.

JEL Classification: C11, C12, C13, C32, C36.
Keywords: Semiparametric estimation, Model Evaluation.

[*] Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215 USA. Email: jjmf@bu.edu, Website: http://jjforneron.com

[†] Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215 USA. Email: qu@bu.edu.

We would like to thank Alfred Galichon, Marc Henry, Ivana Komunjer, seminar participants at Penn State, VTSS, the SED 2024, IAAE 2025 conferences, the 2025 Manchester Econometrics Workshop, and the BC/BU Greenline workshop for useful comments and suggestions.

1 Introduction

Structural estimation is routinely used to evaluate Economic theories and conduct counterfactual analyses with non-experimental data. As noted by Domowitz and White (1982), to make the analysis tractable - either analytically or numerically - some simplification is needed so that the model merely approximates the actual, more complex, data-generating process. Though misspecified, the model still provides tractable insights about causal mechanisms that can be used for policy evaluation and conduct counterfactual experiments. This paper is specifically interested in multivariate models of the form:

$$y_t = g(z_t, v_t; \theta), \quad z_t = h(z_{t-1}, v_t; \theta), \tag{1}$$

where y_t are observed variables such as output or inflation, z_t are unobserved variables such as productivity, and v_t are structural innovations. The functions g and h are known, or solved numerically, up to parameters of interest θ. This is known as a state-space, or hidden Markov model. Examples include DSGE and structural asset pricing models. It is common to fit the model using a filtering algorithm that recovers the latent variables z_t – the Kalman or particle filter – and then maximize the likelihood, or sample Bayesian posterior draws.

This paper shows that several issues can arise when the dynamics in g or h are misspecified. First, for a given value θ, filtered variables may not satisfy the model restrictions given by g and h. For instance, the same average of a filtered series can be substantially non-zero even though the model describes a mean-zero process. This can make inferences on policy-relevant variables difficult, e.g., output gap or natural rate of interest, since their interpretation is model-dependent. Similarly, filtered structural shock processes can be cross-correlated even though the model specify them as independent, a well documented phenomenon in the DSGE literature. Second, likelihood estimates $\hat{\theta}_n$ might be hard to interpret since these coefficients may not characterize the dynamics of the filtered variables: the serial correlation of a Kalman filtered shock can differ substantially from its model-implied value. These two points are illustrated using a medium-scale DSGE model. Third, the likelihood is not defined when there are fewer structural shocks than observables. This limits the potential for model-based dimension reduction where a few structural shocks are used to summarize the comovement of financial or economic variables. This is also illustrated in the applications.

This paper considers an optimal transportation approach to fitting dynamics models in (1) with a mean-squared criterion. Fitting here refers to filtering latent variables and estimating the parameters of interest. The basic idea is to first flexibly approximate the true dynamics of the data; using a reduced-form model. Then, for a given θ, a new sample is recursively constructed for which (1) holds. At each time iteration, the procedure maps the observations to model-consistent data by transporting from the reduced-form to the model-predicted conditional distribution, i.e. the conditional mean-squared error between the sample and the model-consistent data is minimized. Finding the least-squares difference between the original and the new sample produces optimal transport estimates for the parameters of interest. A by-product is an optimal transport filtered series for z_t. The new data, filtered values, and estimates preserve model dynamics by construction, and are thus internally valid. Note that, unlike with i.i.d. data, the dependence structure here requires a different approach to implementing the transport. This is reflected in the use of a flexible reduced-form model and the iterative nature use of a *conditional* transport, where each step builds on the previous ones and is performed as many times as the sample size.

Although there has been much progress in the computation of non-linear filters and numerical optimal transportation, the generic approach described above can be computationally demanding for larger models. Specializing to linear processes, a plugin rule for the optimal transport map is derived leading to closed-form expressions. The true dynamics of the data are approximated using a sieve approach through a vector autoregression of increasing order. The resulting algorithm has closed-form, is easy to implement, and numerically inexpensive. The associated estimator is semiparametric, as only the linear dynamics are specified. The closed-form plugin map extends to a class of non-linear models, though not as general as (1). Because the transport map reduces to the identity map under correct model specification, the framework encompasses correctly specified structural models as a special case.

For stationary linear processes, we derive the large sample frequentist results for the optimal transport estimator $\hat{\theta}_n$ which minimizes the mean-squared difference between original and model-consistent samples. Under standard regularity conditions, the estimates are shown to be consistent and asymptotically normal at a $\sqrt{n}$-rate. An expression for the asymptotic variance is derived under correct specification and misspecification. A specification test based on the mean-squared discrepancy between the sample paths is proposed and studied. The method and results cover a large class of models with an infinite moving average representation which includes linear state-space models. While the results are confined to frequentist estimation, it can also be of interest to extend the framework and consider quasi-Bayesian posterior sampling and inference. This goes beyond the scope of this paper.

In Machine Learning, the sample Wasserstein distance between distributions is a popular tool for data analyses by optimal transportation. It is generally intractable, non-smooth, suffers from a curse of dimensionality, and is computationally demanding for estimation. This can limit its appeal for estimating models with a moderate or large number of observables and parameters. In the scalar case, the minimum Wasserstein distance estimator is fully parametric but has non-standard limiting distribution, which complicates inference. Entropic regularization is a popular way to circumvent some limitations of the Wasserstein distance, but it introduces bias. In contrast, the setting here is semiparametric – only first and second order moments of the data and model are involved. The auxiliary model provides these moments for the data. We show that this allows for a computationally trivial closed-form solution to the transport problem, even for medium-scale DSGE models. The closed-form map is smooth, making estimation regular with $\sqrt{n}$-asymptotically Gaussian estimates.

Three empirical applications illustrate the methodology and the issues discussed above on well-known models. First, a small-scale DSGE model from Lubik and Schorfheide (2004) is estimated. The specification for inflation is rejected at the 5% significance level, which corroborates previous findings. Second, the medium-scale Smets and Wouters (2007) model is estimated. Unlike previous studies, the fit for consumption is rejected as it does not match its volatility and persistence. Further, the Kalman filtered series display irregularities consistent with misspecification, exacerbated by persistence in the data. Finally, an affine term structure model based on Hamilton and Wu (2012) illustrates the dimension-reduction aspect: 3 factors explain most of the variation of 6 yields ranging from 1 month to 5 years.

2 A Motivating Example: Trend-Cycle Decomposition

The following illustrates how misspecification can affect filtered variables and illustrates the methodology introduced in this paper. Trend-cycle decompositions are commonly used to date business cycles for many countries. One approach is to model the logarithm of real GDP using an unobserved component model as in Watson (1986):

$$\underbrace{y_t = \tau_t + c_t}_{\text{log(GDP)}}, \text{ where } \underbrace{\tau_t = \mu + \tau_{t-1} + \eta_t}_{\text{trend component}} \text{ and } \underbrace{c_t = \rho_1 c_{t-1} + \rho_2 c_{t-2} + e_t}_{\text{cycle component}},$$

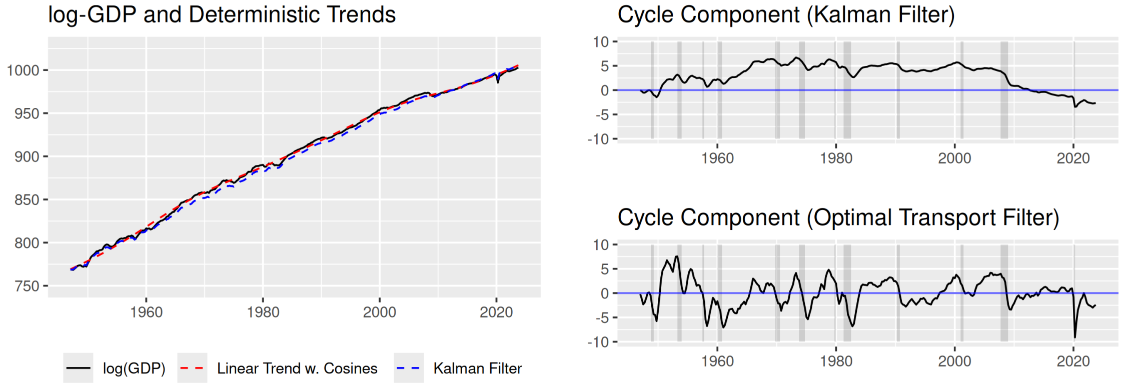
where $(\eta_t, e_t) \overset{iid}{\sim} \mathcal{N}(0, \text{diag}(\sigma_\eta^2, \sigma_e^2))$. The trend component τ_t is modeled as a random walk with drift and the cycle component c_t as a stationary AR(2) process with mean zero. The Kalman filter[1] (KF) can be used to compute from observed y_t: estimates of the trend and cycle components, the likelihood, and ultimately estimate the parameters $\theta = (\mu, \rho_1, \rho_2, \sigma_\eta, \sigma_e)$.

[1] See e.g. Harvey (1990, Ch3) for a textbook introduction.

Here the latent variable $z_t = (\tau_t, c_t)$ includes the trend and cycle components. While this DGP is non-stationary, and thus not covered by the parameter estimation results below, it illustrates that the methodology applies to a broader set of filtering problems.

Figure 1 displays U.S. log-GDP between 1947 and 2023 (left panel) and the extracted cycle components (right panels). The left panel indicates that the KF (blue dashed line) does not fully capture changes in the trend growth rate over the three-quarters of a century spanned by the data. In particular, the estimated trend is systematically below the log-GDP between 1965 and 2008. This issue was already raised by Perron and Wada (2009). A flexible trend estimate (red dashed line), described below the figure, better captures the gradual changes in the trend component.[2]

Figure 1: U.S. GDP – 1947Q1-2023Q4 – Trend-Cycle Decomposition



Legend: Red dashed line: linear trend w. cosines $\delta_0 + \delta_1 t + \delta_2 \cos(2\pi t/n) + \delta_3 \cos(2\pi 2t/n) + \delta_4 \cos(2\pi 3t/n) + \delta_5 \cos(2\pi 4t/n)$. Blue dashed line: Kalman Filter estimates of τ_t. Vertical bars: NBER recession dates.

The top right panel reports the KF estimates for the cycle component. Before discussing the estimates, the following gives a brief overview of the issue. Filtering algorithms, including the KF, infer the unobserved variables by applying Bayes' update recursively. Given current beliefs at time $t-1$ about the unobserved z_{t-1}, the model is used to predict the current z_t and y_t. The observed outcome y_t is then used to update beliefs about the current z_t, using Bayes' rule. When the model is dynamically misspecified, the prediction for (z_t, y_t) is biased.[3] Beliefs are updated from biased prediction errors, and are themselves biased. The next prediction builds on biased beliefs and a biased prediction rule. As a result, misspecification bias propagates over time and accumulates.

[2] The number of cosine terms is chosen to have the shortest periodicity at 20 years, which is larger than the business cycle frequency, i.e. 1.5 to 8 years.

[3] The KF remains valid when the dynamics are correctly specified but the errors are non-Gaussian (Anderson and Moore, 1979, Ch5.2).

In light of the discussion above, notice that the KF estimated trend is systematically below GDP throughout the period 1965-2008 (left panel). Likewise, the filtered cycle component (top right panel) is systematically positive during this period. This would suggest that the economy is characterized by a continuous expansion over this forty-year period. Clearly, the cycle estimates are contaminated by misspecification in the trend component.

Finally, the bottom right panel shows the optimal transport filter (OTF) values, computed using the flexible trend (dashed blue line) as a basis for the reduced form model. See Appendix G for details. No terms were added to handle the Covid recession. Unlike KF estimates, it marks turning points (vertical bars) with good accuracy and visually appears to be stationary with zero mean. To give more quantitative evidence, the Augmented Dickey-Fuller test for a unit root has p-values 0.81 and 0.01 for KF and OTF, respectively, rejecting the null of a unit root only for the latter at the usual significance levels. The KF fits a nonstationary cycle component.

3 Optimal Transport Filtering and Estimation

This section first introduces a general algorithm to perform optimal transportation with state-space models, then provides closed-form recursions for linear state-space models, and describes the optimal transport estimator.

Setup: For the remainder of the paper, $\tilde{y}_t$ denotes the observations collected by the researcher and y_t denotes data generated from (1). The data generating process for $\tilde{y}_t$ is given by the conditional distribution $\tilde{p}(\tilde{y}_t|\tilde{y}_{t-1}, \tilde{y}_{t-2}, \dots) := \tilde{p}_{t|t-1}(\tilde{y}_t)$, which needs to be estimated. For a parameter value θ, filtering methods can be used to evaluate the model's predictive distribution $p(y_t|y_{t-1}, \dots; \theta) := p_{t|t-1}(y_t; \theta)$, this is not estimated. To reduce notation, the same p refers to the joint and marginal distributions of y and z. The goal here is to recursively construct a new sample $y_1, \dots, y_n$ satisfying (1) which is as close to the data $\tilde{y}_1, \dots, \tilde{y}_n$ as possible, under some metric. Algorithm 1 gives a general approach, alternating between a *filtering step* (Filter), used to compute the predictive distribution $p_{t|t-1}$, and an *optimal transport step* (OT) to compute the new observation y_t.

3.1 Optimal Transport Filter (OTF)

Algorithm 1 describes a generic procedure for model (1). An implementation of the filtering (Filter) and optimal transport (OT) steps can be found in Appendix D, Algorithm 3.

Filtering. For general non-linear state-space models, the filtering steps 3 and 6 in the Algorithm can be carried out numerically by Monte Carlo, using a particle filter algorithm. This produces draws which can be used to compute the OT solution below. For discrete or discretized models with finitely many values, these steps involve matrix operations. See Chopin and Papaspiliopoulos (2020) for a textbook introduction. These steps involve standard operations, except that they use the transported data, not the original observations, as the sample for prediction and updating. This crucial difference ensures model consistency.

Algorithm 1 Optimal Transport Filter

```
procedure OTF
Inputs: 1) Sample: $\tilde{y}_1, \dots, \tilde{y}_n$, predictive distribution $\tilde{p}(\tilde{y}_t|\tilde{y}_{t-1}, \dots)$
        2) Model: conditional distribution $p(y_t, z_t|z_{t-1}; \theta)$, initial beliefs $z_0 \sim p_{0|0}(z_0)$
Outputs: 1) Mapped data $y_1, \dots, y_n$, 2) Filtered states $z_{t|t} \sim p(z_t|y_t, \dots, y_1)$.
    for $t \in \{1, \dots, n\}$ do
        Predict: Using the model, compute                                  ▷ (Filter)
            $p_{t|t-1}(y_t, z_t) = \int p(y_t, z_t|z_{t-1}) p_{t-1|t-1}(z_{t-1}) dz_{t-1}$
        Transport: Find a joint distribution $\pi_{t|t-1}$ which solves       ▷ (OT)
            $$\min_{\pi \in \Pi_{t|t-1}} \mathbb{E}_\pi(\|y_t - \tilde{y}_t\|^2)$$
            where $\Pi_{t|t-1} = \{\pi(y_t, \tilde{y}_t) \text{ s.t. } \int \pi(y_t, \tilde{y}_t) d\tilde{y}_t = p_{t|t-1}(y_t), \int \pi(y_t, \tilde{y}_t) dy_t = \tilde{p}_{t|t-1}(\tilde{y}_t)\}$
        Update:                                                            ▷ (Filter)
            $$p_{t|t}(z_t) \propto p_{t|t-1}(z_t, y_t) / \pi_{t|t-1}(y_t|\tilde{y}_t)$$
            using the distributions $p_{t|t-1}$ from step 3 and $\pi_{t|t-1}$ from step 4.
    end for
end procedure
```

Optimal Transport. Step 4 in the Algorithm involves the Wasserstein distance (Villani, 2009, Ch6), here between the conditional distributions $p_{t|t-1}$ and $\tilde{p}_{t|t-1}$. The main appeal of OT in this setting is the construction of a joint distribution $\pi_{t|t-1}$, also known as coupling, which minimizes the mean squared loss $\mathbb{E}_{\pi_{t|t-1}}(\|y_t - \tilde{y}_t\|^2)$ between realizations from the true DGP and the model. In certain cases, the solution $\pi_{t|t-1}$ uniquely maps observations $\tilde{y}_t$ to a single value y_t. In particular, when $p_{t|t-1}$ and $\tilde{p}_{t|t-1}$ have finite second moment and $\tilde{p}_{t|t-1}$ is absolutely continuous with respect to the Lebesgue measure, then the coupling is unique and defines a deterministic map $T_{t|t-1} : \tilde{y}_t \to T_{t|t-1}(\tilde{y}_t) := y_t$ (Villani, 2009, Th9.4). These conditions are met in the applications considered in the paper. If the mapping is not unique, one can report the barycentric projection, as in Algorithm 3, Appendix D.

In the absence of model misspecification, the transport map in Step 4 is the identity map; the resulting y_t are the same as the original $\tilde{y}_t$. Algorithm 1 reduces to standard filtering operations. When the model is dynamically misspecified, the conditional distributions of the model $p_{t|t-1}$ differ from the conditional distribution of the observations $\tilde{p}_{t|t-1}$. Standard filtering algorithms ignore this discrepancy and feed the data into mismatched dynamics, resulting in filtered values of states that do not obey the model's restrictions. This leads to model inconsistencies. By using OT prior to the filtering steps, the Algorithm generates a new sample that is close to the observations while satisfying model restrictions. Subsequently, all filtering operations are based on this new sample. The filtered values are guaranteed to be model-consistent. As a byproduct, the differences between the new sample and the data sample can be compared to gain insights into the model's fit and potential misspecification.

In general, step 4 is not closed-form, and numerical methods are required. One approach is to take B draws from each distribution and solve step 4 between the two empirical distributions. This can be solved exactly by linear programming methods, or approximately using Sinkhorn's algorithm (Peyré and Cuturi, 2019, Ch3.1,4.2). Since this step has to be done n times, the computational cost is crucial for implementation. For filtering only, Sinkhorn's algorithm is sufficiently fast that, when combined with particle filter recursions, Algorithm 1 runs in a little over 1 minute for the model in Section 2 – using a combination of R and C++ codes. The computational cost is prohibitive for estimating larger models. The following shows how to implement Algorithm 1 in closed-form for linear state-space models.

3.1.1 Linear State-Space Models

The following specializes to linear state-space models of the form:

$$y_t = \mu(\theta) + A(\theta)z_t + B(\theta)v_t, \quad z_t = C(\theta)z_{t-1} + D(\theta)v_t, \tag{2}$$

The dependence on the parameters θ will be omitted in the Algorithm below to simplify notation. Specification (2) sets a particular linear structure in (1). The number of structural shocks $v_t \sim (0, I)$ can be greater, equal, or less than the number of observed outcomes y_t. Model (2) includes linearized DSGE models and affine term structure models as special cases.

Turning to the data, under mild conditions (given below), $\tilde{y}_t$ admits an infinite-order vector autoregressive (VAR) representation. A natural auxiliary model, to compute the predictive distribution for $\tilde{y}_t$, is a finite-order VAR(k), a sieve approximation of the VAR(∞):

$$\tilde{y}_t = \tilde{\mu} + \sum_{j=1}^{k} \Psi_j [\tilde{y}_{t-j} - \tilde{\mu}] + e_t. \tag{3}$$

This is a reduced-form VAR, and e_t may differ in dimension from the structural shocks. The VAR coefficients are estimated using ordinary least-squares, setting $\tilde{y}_0 = \tilde{y}_{-1} = \cdots = \bar{y}_n$, the sample mean of $\tilde{y}_t$. This has negligible effect on the estimates but allows to compute residuals $\hat{e}_t$ for all $t = 1, \dots, n$. In practice, the number of lags k should be sufficiently large so that no significant residual autocorrelation remains. See Kuersteiner (2005), and references therein, for automated lag-length selection procedures.

Algorithm 2 describes the Optimal Transport Filter for linear state-space models. It involves only matrix operations and can be readily applied to models where the Kalman Filter (KF) is used. It combines time-invariant KF iterations with an optimal transport (OT) map P. It adjusts the innovations $\hat{e}_t$, whose sample variance is $\tilde{\Sigma}_{nk} = \hat{\mathrm{var}}(\tilde{y}_t|\tilde{y}_{t-1}, \dots)$, to match the variance $\Sigma(\theta) = \mathrm{var}(y_t|y_{t-1}, \dots)$ implied by model (2). The predictive distributions are summarized by $\nu_{t|t} = \mathbb{E}(z_t|y_t, \dots, y_1)$, $\nu_{t|t-1} = \mathbb{E}(z_t|y_{t-1}, \dots, y_1)$, $\mu_{t|t-1} = \mathbb{E}(y_t|y_{t-1}, \dots, y_1)$, and $V = \mathrm{var}(z_t|y_t, \dots)$; $\tilde{\Sigma}_{nk}^{1/2}$ and $\tilde{\Sigma}_{nk}^{-1/2}$ are the matrix square root of $\tilde{\Sigma}_{nk}$ and its inverse.

Algorithm 2 Optimal Transport Filter: Linear State-Space Models

```
procedure OTF
    Inputs: 1) Sample: $\tilde{y}_1, \dots, \tilde{y}_n$, residuals $\hat{e}_1, \dots, \hat{e}_n$
            2) Model: coefficients $\mu, A, B, C, D$. Initial beliefs $z_0 \sim (\nu_{0|0}, V)$
    Outputs: 1) Mapped data $y_1, \dots, y_n$, 2) Filtered states $z_{t|t} \sim (\nu_{t|t}, V)$
    for $t \in \{1, \dots, n\}$ do
        Predict: $\nu_{t|t-1} = C\nu_{t-1|t-1}$, $\mu_{t|t-1} = \mu + A\nu_{t|t-1}$        ▷ (KF)
        Transport: $y_t = \mu_{t|t-1} + P\hat{e}_t$        ▷ (OT)
            where $P = \tilde{\Sigma}_{nk}^{-1/2}[\tilde{\Sigma}_{nk}^{1/2}\Sigma\tilde{\Sigma}_{nk}^{1/2}]^{1/2}\tilde{\Sigma}_{nk}^{-1/2}$ (Transport Map)
            and $\Sigma = \mathrm{var}_{t|t-1}(y_t)$, $\tilde{\Sigma}_{nk} = \widehat{\mathrm{var}}_{t|t-1}(\tilde{y}_t)$ (Innovation Variance)
        Update: $\nu_{t|t} = \nu_{t|t-1} + KP\hat{e}_t$        ▷ (KF)
            where $K = \overline{V}A'\Sigma^{\dagger}$ (Kalman gain)
            and $\overline{V} = \mathrm{var}_{t|t-1}(z_t)$
    end for
end procedure
```

Filtering. The prediction step is a standard KF operation. The matrices V, Σ, and K solve the system of equations (see Anderson and Moore, 1979, Ch4, for further details):

$\bar{V} = CVC' + DD', K = \bar{V}A'\Sigma^{\dagger}, V = (I - KA)\bar{V}, \Sigma = ACVC'A' + (B + AD)(B + AD)'$, where $\Sigma^{\dagger}$ denotes the Moore-Penrose inverse of Σ if it is singular, and otherwise its inverse. Besides V and Σ defined above, the matrix $\overline{V} = \text{var}(z_t|y_{t-1}, \dots)$ measures the one-step-ahead prediction error for z_t. These matrices are standard KF quantities.

Optimal Transport. As in the generic case, OT enables the construction of a joint distribution that minimizes the loss $\mathbb{E}_{\pi_{t|t-1}}(\|y_t - \tilde{y}_t\|^2)$ between the model and the true DGP. Here, the solution to this minimization problem is unique and in closed form, given by $y_t = \mu_{t|t-1} + P\hat{e}_t$, as shown in Step 4 of the Algorithm. This solution maps each observation $\tilde{y}_t$ to a single model-consistent value y_t. Subsequently, all belief updating and filtering operations are based on the new y_t rather than the original $\tilde{y}_t$, ensuring model consistency. In the absence of misspecification, P is the identity matrix and y_t are the same as $\tilde{y}_t$.

For implementation, the key differences with KF are in the transport and update steps. The standard KF update is $\nu_{t|t} = \nu_{t|t-1} + K\tilde{e}_t$ where $\tilde{e}_t = \tilde{y}_t - \mu_{t|t-1}$ are prediction errors computed using model (2). Here, the prediction errors $\hat{e}_t = \tilde{y}_t - \tilde{\mu}_{t|t-1}$ are based on the auxiliary VAR model and are transported using the matrix P to have variance $\Sigma(\theta)$. Enforcing the model-based covariance structure ensures the new data is model-consistent.

The OT literature mainly considers transport plans between parametric and/or sample distributions. Since the setting here is semiparametric, the key idea is to consider transport plans between distributions that are not fully specified. The following explains how the plan is derived and why it is semiparametrically valid. Since both $\tilde{y}_t$ and y_t are linear processes, their predictive distributions, specified up to their second moment, can be written as:

$$y_t|\{y_{t-1}, y_{t-2}, \dots\} \sim (\mu_{t|t-1}, \Sigma), \quad \tilde{y}_t|\{\tilde{y}_{t-1}, \tilde{y}_{t-2}, \dots\} \sim (\tilde{\mu}_{t|t-1}, \tilde{\Sigma}).$$

Their joint distribution $\pi_{t|t-1}$, up to the second moment, takes the form

$$\begin{pmatrix} y_t \\ \tilde{y}_t \end{pmatrix} \Bigg| \begin{pmatrix} y_{t-1}, y_{t-2}, \dots \\ \tilde{y}_{t-1}, \tilde{y}_{t-2}, \dots \end{pmatrix} \sim \left(\begin{pmatrix} \mu_{t|t-1} \\ \tilde{\mu}_{t|t-1} \end{pmatrix}, \begin{pmatrix} \Sigma & C_{t|t-1} \\ C'_{t|t-1} & \tilde{\Sigma} \end{pmatrix} \right),$$

where $C_{t|t-1}$ is the conditional covariance between y_t and $\tilde{y}_t$. Recall that OT minimizes $\mathbb{E}_{\pi_{t|t-1}}(\|y_t - \tilde{y}_t\|^2) = \|\mu_{t|t-1} - \tilde{\mu}_{t|t-1}\|^2 + \text{trace}\left(\Sigma + \tilde{\Sigma}\right) - 2\text{trace}\left(C_{t|t-1}\right)$. In this expression, $\mu_{t|t-1}$ and Σ are computed in the KF recursions, while $\tilde{\mu}_{t|t-1}$ and $\tilde{\Sigma}$ are evaluated with the VAR. Therefore, the only unknown is the covariance matrix $C_{t|t-1}$. There is an additional constraint that $\pi_{t|t-1}$ is a proper distribution, i.e. $C_{t|t-1}$ cannot be arbitrary. This implies that the optimal transportation problem can be written as a semidefinite program:

$$\min_{C} \Big(-2\text{trace}(C) \Big) \text{ subject to } \begin{pmatrix} \Sigma & C \\ C' & \tilde{\Sigma} \end{pmatrix} \geq 0. \tag{4}$$

Any transport map between $\tilde{y}_t$ and y_t with covariance C that solves (4) is optimal. Since the distributions are not fully specified, the map is not uniquely defined. In particular, the linear map which solves the Gaussian case:

$$T : \tilde{y}_t \to \mu_{t|t-1} + P(\tilde{y}_t - \tilde{\mu}_{t|t-1}), \text{ where } P = \tilde{\Sigma}^{-1/2} \left(\tilde{\Sigma}^{1/2} \Sigma \tilde{\Sigma}^{1/2} \right)^{1/2} \tilde{\Sigma}^{-1/2},$$

is optimal and preserves the linearity of the process. These derivations follow from Dowson and Landau (1982), Olkin and Pukelsheim (1982), and Givens and Shortt (1984). See Peyré and Cuturi (2019), Remark 2.31, for additional discussion of the Gaussian case. Algorithm 2 uses the plugin estimates $\tilde{\Sigma}_{nk}$ and $\hat{e}_t$ of $\tilde{\Sigma}$ and $\tilde{y}_t - \tilde{\mu}_{t|t-1}$. Appendix H considers a univariate moving average setting and analytically shows that the transport preserves the impulse responses to shocks at the horizons modeled in the analysis.

Accommodating some non-linearities: The plugin transport map extends to non-linear models of the form:

$$y_t = \mu(x_t; \theta) + \Sigma^{1/2}(x_t; \theta) v_t,$$

where $v_t \sim (0, I)$ and x_t is observed, or can be perfectly inferred, at time $t-1$. The solution to (4) now changes with t: $P_{t|t-1} = \tilde{\Sigma}_{t|t-1}^{-1/2} (\tilde{\Sigma}_{t|t-1}^{1/2} \Sigma(x_t; \theta) \tilde{\Sigma}_{t|t-1}^{1/2})^{1/2} \tilde{\Sigma}_{t|t-1}^{-1/2}$; the map becomes $T_{t|t-1} : \tilde{y}_t \to \mu(x_t; \theta) + P_{t|t-1}(\tilde{y}_t - \tilde{\mu}_{t|t-1})$. The requirement that x_t is observable accommodates (G)ARCH but not stochastic volatility models, for instance. Choices of auxiliary models $\tilde{p}_{t|t-1}$ used for these models in simulation-based estimation are referenced below.

3.2 Optimal Transport Estimation (OTE)

A by-product of the OTF Algorithms 1, 2 is the model-consistent series y_t, which will be referred to as *coupled series*, or *coupling*. The following considers estimating the parameters θ by minimizing the discrepancy between the original sample $\tilde{y}_t$ and its coupling y_t.

The coupling y_t depends on two sets of parameters: the structural coefficients θ, and reduced-form auxiliary parameters ψ_k. For stationary linear processes, a canonical choice for the auxiliary model is the VAR(∞) model. In Algorithm 2, it is approximated by a finite-order VAR(k) with coefficients $\psi_k = (\tilde{\mu}', \mathrm{vech}(\tilde{\Sigma})', \mathrm{vec}(\Psi_1)', \dots, \mathrm{vec}(\Psi_k)')'$ where vec, vech denote the vectorization and half vectorization (Magnus and Neudecker, 2019, Ch2.4). In practice, OTF relies on OLS estimates $\hat{\psi}_{nk} = (\tilde{\mu}_n', \mathrm{vech}(\tilde{\Sigma}_{nk})', \mathrm{vec}(\hat{\Psi}_1)', \dots, \mathrm{vec}(\hat{\Psi}_k)')'$.

For Algorithm 1, the choice of auxiliary model $\tilde{p}$ depends on the particular model (1). This is related to the choice of moments in simulation-based estimation: Gallant and Tauchen (1996) suggest several models, including the SNP estimator of Gallant and Nychka (1987) and a nonparametric ARCH model.[4] Altissimo and Mele (2009), Kristensen and Shin (2012) consider kernel-density estimates when the model is Markovian in the observables.

In either case, the estimation is conducted as follows: given parameters $(\theta, \hat{\psi}_{nk})$, use Algorithm 1 or 2 to generate a series $y_t(\theta; \hat{\psi}_{nk})$ and compute the loss function:

$$Q_n(\theta; \hat{\psi}_{nk}) = \frac{1}{n} \sum_{t=1}^{n} \|y_t(\theta; \hat{\psi}_{nk}) - \tilde{y}_t\|_{W_n}^2,$$

for some symmetric positive definite weighting matrix W_n. The optimal transport estimator (OTE) is the minimizer $\hat{\theta}_n$ of Q_n. For a d-dimensional vector, i.e., $y_t = (y_{t,1}, \dots, y_{t,d})$, setting $W_n = \mathrm{diag}(\mathrm{var}(\tilde{y}_{t,1}), \dots, \mathrm{var}(\tilde{y}_{t,d}))^{-1}$ gives the qualitative interpretation that $\hat{\theta}_n$ maximizes the average R-squared between $\tilde{y}_t$ and its coupling y_t, i.e. $R_j^2 = 1 - [\sum_t (y_{t,j} - \tilde{y}_{t,j})^2] / [\sum_t (\tilde{y}_{n,j} - \tilde{y}_{t,j})^2]$ for $j \in \{1, \dots, d\}$. This choice of W_n was used in all simulated and empirical examples.

For DSGE models, it is common to incorporate prior information. This can be accommodated here by penalization: $Q_n(\theta; \hat{\psi}_{nk}) - \frac{1}{n} \log(\pi(\theta))$, where π is the prior density. Under suitable regularity conditions, the first-order asymptotic properties of $\hat{\theta}_n$ are unchanged.

Interpretation of the loss function. The sample loss function Q_n approximates:

$$Q(\theta; \psi_0) = \min_{\pi \in \Pi(\theta; \psi_0)} \mathbb{E}_\pi [\|y_t - \tilde{y}_t\|_W^2],$$

where $\Pi(\theta; \psi_0)$ is the set of joint distributions with marginals P_θ (given by the structural model, with structural shocks v_t) and $\tilde{P}(\psi_0)$ (given by the DGP, with reduced-form shocks e_t.) For any realized history $(e_t, e_{t-1}, \dots)$, the optimal transport map translates it into a sequence structural shocks *while ensuring model consistency*: $(v_t, v_{t-1}, \dots) = P(\theta; \psi_0)(e_t, e_{t-1}, \dots)$, where $P(\theta; \tilde{\Sigma})$ is the transport map defined above. The loss Q measures the weighted mean squared error between the predictions of the structural model and by the true DGP. Minimizing Q amounts to finding a θ that minimizes the discrepancy between these predictions.

[4] Gallant and Nychka (1987) call $\tilde{p}$ the score generator for the Efficient Method of Moments.

4 Related Literatures

Textbook references on optimal transport (OT) include Villani (2003) for theory, Peyré and Cuturi (2019) for computation, and Galichon (2018) for Economics. In statistics, much of the methodology and theory considers OT between iid samples. Dudley (1969) showed that the empirical Wasserstein distance suffers from a curse of dimensionality, unlike the plug-in approach used here. The literature is much more limited for dependent data. O'Connor et al. (2022, pp8-9) construct couplings between finite state Markov Chains using dynamic programming methods. They do not consider parameter estimation.

Several papers consider parameter estimation using the Wasserstein distance. Bassetti and Regazzini (2006) and Bassetti et al. (2006) study the estimation of location and scale for univariate distributions. They derive consistency and a non-standard limiting distribution for the estimator; Bernton et al. (2019) extend their results. As a minimum-distance estimator, alternatives to OTE include the Simulated Method of Moments, Indirect Inference (Gourieroux and Monfort, 1996), and adversarial estimation using GANs (Kaji et al., 2023). Genevay et al. (2018) discuss the advantages of using OT over classifiers found in GANs. Forneron (2023) considers semi-nonparametric simulation-based estimation, but assumes correctly specified dynamics. These methods do not recover the latent variables which are often an object of interest for policy or prediction. Algorithm 2 is closely related to a goodness-of-fit plot in the Real Business Cycle literature. Plosser (1989, Figures 2-6) and King and Rebelo (1999, Figures 7, 13) compute historical productivity shocks outside the model and use them to simulate a one-shock RBC economy. They plot simulated against real data to show the fit of calibrated models.

Robust filtering also considers model misspecification but aims to recover the true latent variable. The main goal is to reduce sensitivity to local misspecification over a pre-specified neighborhood, see e.g. Sayed (2001) and Shafieezadeh Abadeh et al. (2018). This relates to Hansen and Sargent (2008)'s approach to robustness in Economics. Here, the model can be globally misspecified; the filtered values are computed under model constraints.

Several papers consider estimation and policy analysis with misspecified DGSE models. Del Negro et al. (2007) and Del Negro and Schorfheide (2009) use a DSGE-VAR framework where a hyperparameter penalizes between a reduced form and structural model. Here, the flexible VAR is used to enforce the model structure with the coupling. This ensures the parameters are internally valid, i.e. characterize the dynamics of y_t.

5 Large Sample Properties of OTE

The following derives consistency and asymptotic normality results for a class of linear processes, which includes linear state-space models described by (2):

$$y_t = \mu(\theta) + A(\theta)z_t + B(\theta)v_t, \quad z_t = C(\theta)z_{t-1} + D(\theta)v_t. \tag{2}$$

Notation: The parameters $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$. Norms: for a matrix $A = (a_{ij})$ of size $n \times m$, the baseline norm is $\|A\| = \sqrt{\text{trace}(A'A)}$, the operator norm is $\|A\|_{op} = \sqrt{\lambda_{\max}(A'A)}$, the sup norm is $\|A\|_\infty = \max_{i,j} |a_{ij}|$. Eigenvalues: for a symmetric matrix A of size $n \times n$, $\lambda_j(A)$ denotes the j-th eigenvalue, $1 \leq j \leq n$, in increasing order, $\lambda_{\max}(A) = \lambda_n(A)$ and $\lambda_{\min}(A) = \lambda_1(A)$; $\underline{\lambda} \preceq A$ implies $\lambda_{\min}(A) \geq \underline{\lambda}$ and $A \preceq \overline{\lambda}$ implies $\lambda_{\max}(A) \leq \overline{\lambda}$. For a matrix A of size $n \times m$, the singular values are given by $\sigma_j(A) = \sqrt{\lambda_j(A'A)}$ if $m < n$, or $\sigma_j(A) = \sqrt{\lambda_j(AA')}$ if $n < m$, for $j = 1, \dots, \min(n, m)$; $\sigma_{\min}(A) = \sigma_1(A)$ and $\sigma_{\max}(A) = \sigma_{\min(m,n)}(A) = \|A\|_{op}$.

5.1 Consistency and Asymptotic Normality

The true data-generating process (DGP) and the model are assumed to be left-invertible; i.e. they admit a one-sided infinite vector moving-average (VMA) representation.[5]

Assumption 1. *$\tilde{y}_t$ and y_t admit causal VMA(∞) representations:*

$$\tilde{y}_t = \tilde{\mu} + e_t + \sum_{j=1}^{\infty} \tilde{\Lambda}_j e_{t-j}, \quad y_t(\theta) = \mu(\theta) + \xi_t + \sum_{j=1}^{\infty} \Lambda_j(\theta)\xi_{t-j},$$

for any $\theta \in \Theta$, where e_t and ξ_t are white noise with variance $\tilde{\Sigma}$ and $\Sigma(\theta)$.

The VMA innovations ξ_t need not coincide with, or span, the structural innovations v_t (Fernández-Villaverde et al., 2007). The number of structural innovations can be greater than, equal to, or less than the number of observables. In the latter case, the models are stochastically singular.[6] Our empirical applications cover all three situations.

Using the VMA representation, Algorithm 2 involves the following quantities: $\tilde{\Sigma} = \text{var}(e_t)$, $\Sigma(\theta) = \text{var}(\xi_t)$, $\tilde{\mu}_{t|t-1} = \tilde{\mu} + \sum_{j=1}^{\infty} \tilde{\Lambda}_j e_{t-j}$, and $\mu_{t|t-1} = \mu(\theta) + \sum_{j=1}^{\infty} \Lambda_j(\theta)\xi_{t-j}$. Note

[5] Some models can feature non-invertibility, this is the case with permanent income (Fernández-Villaverde et al., 2007). Algorithm 2 only relies on second-order moments whereas identification and estimation of non-invertible models rely on higher-order cumulants, which is beyond the scope of this paper.

[6] For instance: multivariate RBC models with a single shock to productivity are stochastically singular. See Komunjer and Ng (2011), Qu (2018) for identification and estimation with stochastic singularity.

that $\Sigma(\theta)$ is the same as that in Algorithm 2, where y_t admits a state-space representation. Take the transport map $P(\theta; \tilde{\Sigma})$, computed using $\tilde{\Sigma}$ and $\Sigma(\theta)$, the coupled series y_t is:

$$y_t(\theta; \psi_0) = \mu(\theta) + P(\theta; \tilde{\Sigma})e_t + \sum_{j=1}^{\infty} \Lambda_j(\theta)P(\theta; \tilde{\Sigma})e_{t-j}.$$

The index ψ_0 refers to the innovations and the variance used to compute the coupling. For $\psi = \psi_0$, the true errors e_t and $\tilde{\Sigma}$ are used, as above. For $\psi = \hat{\psi}_{nk}$, the residuals $\hat{e}_t$ and sample variance $\tilde{\Sigma}_{nk}$ are used, with the convention that $\hat{e}_t = 0$ for $t \leq 0$. For $\psi = \psi_k$, the error $e_{t,k}$ and $\tilde{\Sigma}$ are used, where $e_{t,k} = \tilde{y}_t - \tilde{\mu} - \sum_{j=1}^{k} \tilde{\Psi}_j[\tilde{y}_{t-j} - \tilde{\mu}]$ are the VAR(k) errors. The KF steps in Algorithm 2 compute the VMA coefficients $\Lambda_j(\theta)$ using the state-space representation (2). Suppose $\|C(\theta)\|_{op} < 1$, iterate on the KF and OT steps to find:

$$y_t(\theta; \psi_0) = \mu(\theta) + P(\theta; \tilde{\Sigma})e_t + \sum_{j=1}^{\infty} A(\theta)C^j(\theta)K(\theta)P(\theta; \tilde{\Sigma})e_{t-j},$$

where $K(\theta)$ is the Kalman gain. Thus, $\Lambda_0 = I_d$ and $\Lambda_j(\theta) = A(\theta)C^j(\theta)K(\theta)$ for each $j \geq 1$.

Assumption 2. *(i). $\sum_{j=1}^{\infty} j^{1/2}\|\tilde{\Lambda}_j\| < \infty$ and det$\left(\sum_{j=0}^{\infty} \tilde{\Lambda}_j z^j\right) \neq 0$ for all $|z| \leq 1$ with $z \in \mathbb{C}$; (ii). e_t is strictly stationary, $\mathbb{E}_{t-1}(e_t) = 0$, $\mathbb{E}(e_t e_t') = \tilde{\Sigma}$, and $0 < \underline{\lambda} \preceq \tilde{\Sigma} \preceq \overline{\lambda} < \infty$; (iii). for some $r > 4$, $\mathbb{E}(\|e_t\|^{2r}) < \infty$, and e_t is α-mixing with size $-a$, where $a > r/(r-2)$.*

Assumption 2 provides several sufficient conditions for $\tilde{y}_t$ to admit a VAR(∞) representation and to study the OLS estimates (Hannan and Deistler, 2012, Ch7). The mixing conditions are needed to derive near-epoch dependence (NED) properties for $y_t(\theta; \hat{\psi}_{nk})$, its derivatives, and asymptotic results for $\tilde{\Sigma}_{nk}$.[7] Assumption 2 allows for unmodelled dependence in higher-order moments, such as conditional heteroskedasticity (ARCH, GARCH) or stochastic volatility that satisfy a strong-mixing condition.

Assumption 3. *Θ is convex and compact and $\theta \to (\mu(\theta), \Sigma(\theta), \Lambda_1(\theta), \dots)$ is three times continuously differentiable, such that: (i). rank$[\Sigma(\theta)] = r_\Sigma$ for all $\theta \in \Theta$, and $0 \preceq \Sigma(\theta) \preceq \overline{\lambda} < \infty$; (ii). $\sup_{\theta \in \Theta} \|\mu(\theta)\| < \infty$ and $\sum_{j=0}^{\infty} \sup_{\theta \in \Theta} \|\Lambda_j(\theta)\|_{op} < \infty$; (iii). for $s = 1, \dots, 3$ and any $i_1, \dots, i_s \in \{1, \dots, d_\theta\}$, $\sup_{\theta \in \Theta} \|\partial^s_{\theta_{i_1}, \dots, \theta_{i_s}} \mu(\theta)\| < \infty$, $\sup_{\theta \in \Theta} \|\partial^s_{\theta_{i_1}, \dots, \theta_{i_s}} vec[\Sigma(\theta)]\|_\infty < \infty$, and $\sum_{j=0}^{\infty} \sup_{\theta \in \Theta} \|\partial^s_{\theta_{i_1}, \dots, \theta_{i_s}} vec[\Lambda_j(\theta)]\|_\infty < \infty$.*

Assumption 4. *There exists $C \geq 0$, $b \geq 2$, and $\varepsilon > 0$ such that for $s = 1, \dots, 3$ and any $i_1, \dots, i_s \in \{1, \dots, d_\theta\}$: $\sum_{j=m+1}^{\infty} \sup_{\theta \in \Theta} \|\Lambda_j(\theta)\|_{op} \leq Cm^{-(b+\varepsilon)}$, $\sum_{j=m+1}^{\infty} \|\tilde{\Lambda}_j\|_{op} \leq Cm^{-(b+\varepsilon)}$, and $\sum_{j=m+1}^{\infty} \sup_{\theta \in \Theta} \|\partial^s_{\theta_{i_1}, \dots, \theta_{i_s}} vec[\Lambda_j(\theta)]\|_\infty \leq Cm^{-(b+\varepsilon)}$.*

[7] Definitions A1, A2 recall the concepts of strong-mixing and NED. Lemma B1 derives the NED properties.

Assumptions 3 and 4 restrict the dependence of $\tilde{y}_t$, y_t and its derivatives. The constant rank condition is discussed below. The following Lemma gives conditions on the state-space representation (2) for which Assumption 3 holds. Lemma B1, Appendix B further shows that Assumption 4 also holds for any $b \geq 2$ and $\varepsilon > 0$, with an appropriate constant $C > 0$.

Lemma 1 (State-Space Model - VMA representation)**.** *If Θ is convex and compact, and the following conditions hold: (i). rank$[\Sigma(\theta)] = r_\Sigma$ for all $\theta \in \Theta$; (ii). $\Sigma(\cdot)$, $\mu(\cdot)$, $A(\cdot)$, $B(\cdot)$, $C(\cdot)$, and $D(\cdot)$ are three times continuously differentiable with bounded derivatives; (iii). $\inf_{\theta \in \Theta} \inf_{|z| \leq 1, z \in \mathbb{C}} |det(I - C(\theta)z)| > 0$, then Assumption 3 holds.*

The constant rank condition $\text{rank}[\Sigma(\theta)] = r_\Sigma$, which appears in Assumption 4 and Lemma B1, and the full rank condition $0 < \underline{\lambda} \preceq \tilde{\Sigma}$ (Assumption 2) are particularly important for the transport map to be well behaved. Much like the square root of a scalar, $x \rightarrow \sqrt{x}$, the matrix square root $A \rightarrow A^{1/2}$, used in the transport map, is not continuously differentiable at a singular A. The following Lemma derives a new result for the differentiability of $\theta \rightarrow A(\theta)^{1/2}$ when $A(\cdot)$ is singular with a constant rank. The proof involves a constructive local block decomposition which can be used to compute differentials analytically. Unlike an eigenvalue decomposition, the block decomposition is smooth under multiplicity of eigenvalues. This should be of independent interest as the matrix square root appears in a variety of settings.

Lemma 2 (Matrix Square Root, Constant Rank)**.** *Suppose $\Theta \subset \mathbb{R}^{d_\theta}$ is convex and compact and $\theta \rightarrow A(\theta) \geq 0$ is s-times continuously differentiable for some $s \geq 1$. Assume that $A(\theta)$ has constant rank r, where $1 \leq r \leq d = dim(A)$ and $0 < \underline{\lambda} \leq \inf_\theta \lambda_r[A(\theta)] \leq \sup_\theta \lambda_{\max}[A(\theta)] \leq \overline{\lambda} < \infty$. Then: (i). There exists $\delta > 0$, such that for any $\theta_0 \in \Theta$, there exists $M(\theta)$ and $B(\theta)$ that are s-times continuously differentiable on $\mathcal{B}_\delta(\theta_0) = \{\theta \in \Theta, \|\theta - \theta_0\| \leq \delta\}$, such that $0 < \underline{\lambda}_B \preceq B(\theta) \preceq \overline{\lambda}_B < \infty$, $M(\theta)M(\theta)' = I_d$, and $A(\theta) = M(\theta) blockdiag[B(\theta), 0_{m,m}]M(\theta)'$ where $m = d - r$. (ii). For all $\theta_0 \in \Theta$, the square root $A(\theta)^{1/2} = M(\theta) blockdiag[B(\theta)^{1/2}, 0_{m,m}]M(\theta)'$ is s-times continuously differentiable on $\mathcal{B}_\delta(\theta_0)$. (iii). The square root $\theta \rightarrow A(\theta)^{1/2}$ is s-times continuously differentiable on Θ.*

The KF recursions are well defined under stochastic singularity (Anderson and Moore, 1979, p39); however, the likelihood is not defined. If the constant rank condition fails, the transport map becomes non-smooth and the KF steps in Algorithm 2 become sensitive to numerical accuracy and can be unstable (Anderson and Moore, 1979, Ch6.5).

Lemma 3 (Data: VAR(∞) representation, VAR(k) approximation)**.** *Suppose Assumptions 1, 2, and 4 hold. Then $\tilde{y}_t$ admits a VAR(∞) representation: $\tilde{y}_t = \tilde{\mu} + \sum_{j=1}^{\infty} \Psi_j (\tilde{y}_{t-j} - \tilde{\mu}) + e_t$,* *where* $\sum_{j=1}^{\infty} j^{1/2}\|\Psi_j\| < \infty$*,* $\Psi_0 = I_d$*, and det* $\left(\sum_{j=0}^{\infty}\Psi_j z^j\right) \neq 0$ *for any* $|z| \leq 1$*. Further, suppose* $k \to \infty$ *such that* $k^3/n \to 0$ *and* $\sqrt{n}\sum_{j=k+1}^{\infty}\|\Psi_j\| \to 0$*, and let* $\tilde{\Sigma}_{nk} = \frac{1}{n}\sum_{t=1}^{n}\hat{e}_t\hat{e}_t'$ *and* $\tilde{\Sigma}_n = \frac{1}{n}\sum_{t=1}^{n} e_t e_t'$*. Then: (i).* $\max_{j=1,\dots,k}\|\hat{\Psi}_j - \Psi_j\| = O_p(\sqrt{\log(n)/n})$*; (ii).* $\tilde{\Sigma}_{nk} - \tilde{\Sigma}_n = o_p(1/\sqrt{n})$*; (iii).* $\tilde{y}_n - \tilde{\mu} = O_p(n^{-1/2})$ *and* $\tilde{\Sigma}_n - \tilde{\Sigma} = O_p(n^{-1/2})$*.*

Lemma 3 combines several existing results for the auxiliary parameters $\hat{\psi}_{nk}$ from the literature, mainly Lewis and Reinsel (1985) and Hannan and Deistler (2012). The conditions on the order of the VAR order, k, depend on the decay of the VAR coefficients. If model (2) is correctly specified and the conditions for Lemma 1 hold - or if the true model is a finite order stationary VARMA - then $\|\Psi_j\|_{op} = O(\bar{\rho}^j)$ for some $\bar{\rho} \in [0, 1)$ and $\sqrt{n}\sum_{j=k+1}^{\infty}\|\Psi_j\| = o(1)$ as long as $\log(n)/k \to 0$. In these cases, the order k can increase very slowly.

Theorem 1 (Consistency)**.** *Suppose Assumptions 1-4 hold, k satisfies the conditions of Lemma 3,* $W_n \xrightarrow{p} W > 0$*, and* $Q(\theta;\psi_0) = \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left(\|y_t(\theta;\psi_0) - \tilde{y}_t\|_W^2\right)$ *is uniquely minimized at* $\theta = \theta_0$*. If k is such that* $\sqrt{n}\sum_{j=k+1}^{\infty}\sup_{\theta\in\Theta}\|\Lambda_j(\theta)\|_{op} = o(1)$*, then* $\hat{\theta}_n \xrightarrow{p} \theta_0$*.*

Theorem 1 shows that the estimator $\hat{\theta}_n$ is consistent for the minimizer θ_0 of $Q(\cdot;\psi_0)$. Calculations show that $Q(\theta;\psi_0)$ can be represented as:

$$\|\tilde{\mu} - \mu(\theta)\|_W^2 + \sum_{j=0}^{\infty}\text{trace}\left(\tilde{\Sigma}^{1/2}\{\tilde{\Lambda}_j - \Lambda_j(\theta)P(\theta;\tilde{\Sigma})\}'W\{\tilde{\Lambda}_j - \Lambda_j(\theta)P(\theta;\tilde{\Sigma})\}\tilde{\Sigma}^{1/2}\right),$$

where $\tilde{\Lambda}_0 = \Lambda_0(\theta) = I_d$. Thus, θ_0 minimizes a weighted distance between two VMA(∞) representations: one describing the data and the other describing the model.

Theorem 2 (Asymptotic Normality)**.** *Suppose the conditions for Theorem 1 hold with* $\theta_0 \in$ *interior*(Θ)*. Let* $u_{t,k} = y_t(\theta_0;\psi_k) - \tilde{y}_t$*,* $u_t = y_t(\theta_0;\psi_0) - \tilde{y}_t$*,* $G_t(\theta_0;\psi_k) = vec[\partial_\theta y_t(\theta_0;\psi_k)']$ *and*

$$M = \mathbb{E}\left(\partial_\theta y_t(\theta_0;\psi_0)'W\partial_\theta y_t(\theta_0;\psi_0)\right) + \mathbb{E}\left([u_t'W \otimes I]\,\partial_\theta G_t(\theta_0;\psi_0)\right),$$
$$D_{\theta,\psi}(k) = \mathbb{E}\left[\partial_\theta y_t(\theta_0;\psi_k)'W\partial_\psi y_t(\theta_0;\psi_{k0}) + \left[u_{t,k}'W \otimes I\right]\partial_\psi G_t(\theta_0;\psi_{k0})\right].$$

Suppose M is invertible and there exists $\underline{k} \geq 1$ *and* $c_1 > 0$ *such that for all* $k \geq \underline{k}$*:* $0 < c_1 \leq \sigma_{\min}[D_{\theta,\psi}(k)] < \infty$*. Define* $Z_{k,t} = ((\tilde{y}_t - \tilde{\mu})', vec[e_t\tilde{Y}_{t-1,k}'\Gamma_k^{-1}]', vech[e_te_t' - \tilde{\Sigma}]')'$*, with* $\tilde{Y}_{t-1,k} = ((\tilde{y}_{t-1} - \tilde{\mu})', \dots, (\tilde{y}_{t-k} - \tilde{\mu})')'$ *and* $\Gamma_k = \mathbb{E}(\tilde{Y}_{t-1,k}\tilde{Y}_{t-1,k}')$*. Then, the sequence of covariance matrices*

$$V_{n,k} = M^{-1}var\left[\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{\partial_\theta y_t(\theta_0;\psi_k)'Wu_{t,k} + D_{\theta,\psi}(k)Z_{k,t}\}\right]M^{-1},$$

is bounded from above. If, in addition, $V_{n,k}^{-1} = O(1)$, then: $\sqrt{n} V_{n,k}^{-1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I)$.

Theorem 2 establishes the asymptotic normality of the estimates $\hat{\theta}_n$. The invertibility of M and the lower bound c_1 are local identification conditions. The requirement $V_{n,k}^{-1} = O(1)$ is standard for central limit theorems (e.g. White, 2014, Th5.20). The boundedness of $V_{n,k}$ implies a $\sqrt{n}$-rate of convergence for $\hat{\theta}_n - \theta_0$. Note that this rate does not apply to all functionals of $\hat{\psi}_{nk}$; some may converge more slowly (e.g. Lewis and Reinsel, 1985, Th6).

The residual $u_t = y_t(\theta_0; \psi_0) - \tilde{y}_t$ measures the model-data discrepancy. Like in OLS, it reflects a simple decomposition of $\tilde{y}_t$ into fitted values $y_t(\theta_0; \psi_0)$ and residuals u_t. The R^2 introduced earlier measure their relative magnitudes. With $n = \infty$, $R^2 = 1$ indicates correct specification. Formal specification testing is considered in the next subsection.

Computing standard errors. The following describes how to compute standard errors assuming correct specification and allowing for misspecification. For models considered in Section 7, bootstrap inference could be rather computationally cumbersome. The plugin estimates for $M, D_{\theta,\psi}(k)$, etc are shown to be consistent in the proof of Theorem 2.

Under correct specification, $u_t = 0$ and $u_{t,k} = o(n^{-1/2})$ don't contribute to the asymptotic standard errors. First, evaluate $\partial_\theta y_t(\hat{\theta}_n; \hat{\psi}_{nk})$ and $\partial_\psi y_t(\hat{\theta}_n; \hat{\psi}_{nk})$ with finite-differences or by automatic differentiation. Then, compute $\hat{M}_n = \frac{1}{n}\sum_{t=1}^n \partial_\theta y_t(\hat{\theta}_n; \hat{\psi}_{nk})' W_n \partial_\theta y_t(\hat{\theta}_n; \hat{\psi}_{nk})$ and $\hat{D}_{n,\theta,\psi}(k) = \frac{1}{n}\sum_{t=1}^n \partial_\theta y_t(\hat{\theta}_n; \hat{\psi}_{nk})' W_n \partial_\psi y_t(\hat{\theta}_n; \hat{\psi}_{nk})$. Let $L_n(\hat{\psi}_{nk})$ denote the Gaussian quasi-likelihood for the auxiliary VAR(k) model, $H_n(\hat{\psi}_{nk})$ its Hessian, and $\partial_\psi L_t(\hat{\psi}_{nk})$ the score for $\tilde{y}_t$. Take $\hat{S}_{t,k} = \hat{M}_n^{-1} \hat{D}_{n,\theta,\psi}(k) H_n(\hat{\psi}_{nk})^{-1} \partial_\psi L_t(\hat{\psi}_{nk})$. Then, $\hat{V}_{n,k}$ is the HAC estimator for the long-run variance of $\hat{S}_{t,k}$. Standard errors are computed from $\hat{V}_{n,k}/n$ in a standard fashion.

Allowing for misspecification requires estimating several additional terms. Compute $\hat{u}_t = \hat{u}_{t,k} = y_t(\hat{\theta}_n; \hat{\psi}_{nk}) - \tilde{y}_t$, $\partial_\theta \hat{G}_t(\hat{\theta}_n; \hat{\psi}_{nk})$, and $\partial_\psi \hat{G}_t(\hat{\theta}_n; \hat{\psi}_{nk})$ with $\hat{G}_t(\hat{\theta}_n; \hat{\psi}_{nk}) = \text{vec}[\partial_\theta y_t(\hat{\theta}_n; \hat{\psi}_{nk})]$. Then, compute $\hat{M}_n = \frac{1}{n}\sum_{t=1}^n \partial_\theta y_t(\hat{\theta}_n; \hat{\psi}_{nk})' W_n \partial_\theta y_t(\hat{\theta}_n; \hat{\psi}_{nk}) + \frac{1}{n}\sum_{t=1}^n [\hat{u}_t' W_n] \otimes \partial_\theta \hat{G}_t(\hat{\theta}_n; \hat{\psi}_{nk})$ and $\hat{D}_{n,\theta,\psi}(k) = \frac{1}{n}\sum_{t=1}^n \partial_\theta y_t(\hat{\theta}_n; \hat{\psi}_{nk})' W_n \partial_\psi y_t(\hat{\theta}_n; \hat{\psi}_{nk}) + \frac{1}{n}\sum_{t=1}^n [\hat{u}_t' W_n] \otimes \partial_\psi \hat{G}_t(\hat{\theta}_n; \hat{\psi}_{nk})$. Using the same Gaussian quasi-Likelihood terms as above, evaluate $\hat{S}_{t,k} = \hat{M}_n^{-1}\{\partial_\theta y_t(\hat{\theta}_n; \hat{\psi}_{nk})' W_n \hat{u}_t + \hat{D}_{n,\theta,\psi}(k) H_n(\hat{\psi}_{nk})^{-1} \partial_\psi L_t(\hat{\psi}_{nk})\}$. Finally, $\hat{V}_{n,k}$ is the estimator for the long-run variance of $\hat{S}_{t,k}$.

5.2 Specification Testing

The population loss $Q(\theta_0; \psi_0)$ defines a distance between the VMA(∞) representations of $\tilde{y}_t$ and $y_t(\theta; \psi_0)$. When the model is correctly specified in terms of second-order moments, the minimizer θ_0 yields $Q(\theta_0; \psi_0) = 0$. When the model is misspecified, however, the minimum is strictly positive: $Q(\theta_0;\psi_0) > 0$. The following considers a specification test based on the sample analog $Q_n(\hat{\theta}_n;\hat{\psi}_{nk})$ of the optimal transport distance $Q(\theta_0;\psi_0)$.

Assumption 5. *Suppose that: (i). $[k\log(n)]^8/n = o(1)$; (ii). $[\log(n)]^4/k = o(1)$; (iii). $\mathbb{E}(\|e_t\|^{16}) < \infty$; (iv). $\alpha(j) \leq C(1+j)^{-(a+\varepsilon)}$ for $a \geq 6$, $\varepsilon > 0$ and all $j \geq 1$; (v). Assumption 4 holds with $b \geq 6$; (vi). $\|W_n - W\| = O_p(n^{-1/2})$, (vii) $\sqrt{nk}\sum_{j=k+1}^{\infty}\|\Psi_j\|_{op} = o([\log(n)]^{-2})$.*

Assumption 5 is more restrictive than those needed for Theorems 1 and 2. When the model is correctly specified, the loss $Q_n(\hat{\theta}_n;\hat{\psi}_{nk})$ is asymptotically determined by the distance between $\hat{\psi}_{nk}$ and ψ_k. The distributional results below build on a strong approximation result for NED processes with dependence changing with the lag structure, indexed by k.

Theorem 3 (Specification Test)**.** *Suppose the conditions for Theorems 1 and 2 and Assumption 5 hold. If the model is correctly specified, then:*

$$nQ_n(\hat{\theta}_n;\hat{\psi}_{nk}) = n\mathcal{Z}_{n,k}'M_k\mathcal{Z}_{n,k} + o_p(k^{1/2}[\log(n)]^{-2}),$$

where $M_k = \mathbb{E}\left[(\partial_\psi y_t(\theta_0;\psi_{k0}) + \partial_\theta y_t(\theta_0;\psi_k)M^{-1}E_k)'W(\partial_\psi y_t(\theta_0;\psi_{k0}) + \partial_\theta y_t(\theta_0;\psi_k)M^{-1}E_k)\right]$, with $E_k = -\mathbb{E}[\partial_\theta y_t(\theta_0;\psi_{k0})'W\partial_\psi y_t(\theta_0;\psi_{k0})]$, $\mathcal{Z}_{n,k} \sim \mathcal{N}(0, S_{n,k}/n)$, $S_{n,k} = nvar[\overline{Z}_{n,k}]$, and M, $\overline{Z}_{n,k}$ are defined in Theorem 2. If $S_{n,k}$ and M_k are such that trace$(S_{n,k}M_k) \geq O(k)$ and trace$([S_{n,k}M_k]^2) \geq O(k)$, then for any $\alpha \in (0,1)$:

$$\mathbb{P}\left(nQ_n(\hat{\theta}_n;\hat{\psi}_{nk}) > c_{n,k}(1-\alpha)\right) = \alpha + o(1),$$

where $c_{n,k}(1-\alpha)$ is the $1-\alpha$ quantile of $n\mathcal{Z}_{n,k}'M_k\mathcal{Z}_{n,k}$.

Theorem 3 shows that $nQ_n(\hat{\theta}_n;\hat{\psi}_{nk})$ can be approximated by a weighted sum of independent χ_1^2 random variables. The derivatives $\partial_\psi y_t(\theta;\psi_{k0})$ are given in Lemma B1. Lütkepohl (2005, Ch15) provides formulas for $S_{n,k}$ in the homoskedastic case. The conditions trace$(S_{n,k}M_k) \geq O(k)$ and trace$([S_{n,k}M_k]^2) \geq O(k)$ are analogous to the rank conditions needed to ensure the J-test for GMM has a χ_{k-d}^2 distribution, where k is the number of moments and d the number of parameters. The Theorem states that under the null hypothesis of correct specification, the asymptotic size of the test is α. The test is also consistent against distant alternatives, see Lemma F3 in Appendix F. Power against local alternatives depends on the ratio k/n. A detailed analysis of local power is left to future research.

The test in Theorem 3 involves all variables $\tilde{y}_t$ used in the estimation. In certain settings, the researcher might inquire how well the model fits a specific variable $\tilde{y}_{t,j}$, e.g. consumption if the object of interest is welfare. The following Corollary specializes to a single variable, using a selection matrix D_j. The proof is the same as Theorem 3, and it is omitted.

Corollary 1 (Specification Test on a Single Variable)**.** *Suppose the conditions for Theorems 1 and 2 and Assumption 5 hold. Let* $Q_{n,j}(\hat{\theta}_n; \hat{\psi}_{nk}) = \frac{1}{n}\sum_{t=1}^{n} \|y_t - \tilde{y}_t\|^2_{D_j W_n D_j}$ *for* $j \in \{1, \dots, d\}$*, where* $D_j = diag(\mathbb{1}_{j=1}, \dots, \mathbb{1}_{j=d})$*. If the model is correctly specified, then:*

$$nQ_n(\hat{\theta}_n; \hat{\psi}_{nk}) = n\mathcal{Z}_{n,k}' M_{k,j} \mathcal{Z}_{n,k} + o_p(k^{1/2}[\log(n)]^{-2}),$$

where $M_{k,j} = \mathbb{E}\left[(\partial_\psi y_t(\theta_0; \psi_{k0}) + \partial_\theta y_t(\theta_0; \psi_k) M^{-1} E_{k,j})' D_j W D_j (\partial_\psi y_t(\theta_0; \psi_{k0}) + \partial_\theta y_t(\theta_0; \psi_k) M^{-1} E_k)\right]$, *with* $E_{k,j} = -\mathbb{E}[\partial_\theta y_t(\theta_0; \psi_{k0})' D_j W D_j \partial_\psi y_t(\theta_0; \psi_{k0})]$*, and* M *and* $\mathcal{Z}_{n,k}$ *are as defined in Theorem 3. If* $S_{n,k}$ *and* $M_{k,j}$ *are such that* $trace\,(S_{n,k} M_{k,j}) \geq O(k)$ *and* $trace\,([S_{n,k} M_{k,j}]^2) \geq O(k)$*, then for any* $\alpha \in (0, 1)$*:* $\mathbb{P}\left(nQ_{n,j}(\hat{\theta}_n; \hat{\psi}_{nk}) > c_{n,k,j}(1-\alpha)\right) = \alpha + o(1)$*, where* $c_{n,k,j}(1-\alpha)$ *is the* $1-\alpha$ *quantile of* $n\mathcal{Z}_{n,k}' M_{k,j} \mathcal{Z}_{n,k}$*.*

6 Monte Carlo Simulations

The Monte Carlo simulations are based on the Lubik and Schorfheide (2004, LS) model, considered for the first empirical application as well:

$$\begin{aligned} &y_t = E_t y_{t+1} - \tau(r_t - E_t \pi_{t+1}) + g_t, \quad \pi_t = \beta E_t \pi_{t+1} + \kappa(y_t - z_t), \quad g_t = \rho_g g_{t-1} + \varepsilon_{gt} \\ &z_t = \rho_z z_{t-1} + \varepsilon_{zt}, \quad r_t = \rho_r r_{t-1} + (1-\rho_r)\psi_1 \pi_t + (1-\rho_r)\psi_2 (y_t - z_t) + \varepsilon_{rt}, \end{aligned}$$

where y_t, π_t, and r_t are log deviations of output, inflation, and the nominal interest rate from their steady states, respectively. The shocks ε_{rt}, $\varepsilon_{gt}, \varepsilon_{gt}$ are iid Gaussian with mean zero and variances $\sigma_r^2, \sigma_g^2, \sigma_z^2$; ε_{gt}; ε_{zt} are cross-correlated with correlation ρ_{gz}. The observables are log levels of output, inflation, and interest rate (both annualized), which satisfy $Y_t = (0, \pi^*, \pi^* + r^*)' + (y_t, 4\pi_t, 4r_t)'$, where output is detrended, and π^* and r^* are annualized steady-state rates of inflation and real interest rate with $\beta = (1 + r^*/100)^{-1/4}$. The data are generated using the posterior means from Bayesian inference on the full sample with LS's prior.[8] The VAR includes a constant and 4 lags as regressors.

The baseline sample size corresponds to the full sample estimation below with $n = 192$. A larger sample size of $n = 500$ is also considered. The prior from Lubik and Schorfheide (2003) is used to regularize the estimates. Table 1 reports the averages and standard deviations

[8] Table I1, Appendix I, includes a description of the parameters, the bounds imposed on the parameters, and the prior π used to regulate the estimates.

Table 1: LS Model: Average Estimate, Standard Deviation, Rejection Rates

	τ^{-1}	r^*	κ	ψ_1	ψ_2	ρ_r	ρ_g	ρ_z	σ_r	σ_g	σ_z	ρ_{gz}	π^*
TRUE	3.18	1.87	0.50	1.33	0.21	0.76	0.89	0.86	0.26	0.13	0.97	0.80	4.01
	$n = 192$												
MEAN	1.93	1.84	0.39	1.29	0.18	0.73	0.86	0.82	0.26	0.20	1.04	0.60	3.94
STD	0.42	0.33	0.13	0.16	0.02	0.05	0.04	0.05	0.04	0.03	0.17	0.21	0.63
REJ$_c$	0.14	0.07	0.05	0.00	0.00	0.01	0.03	0.04	0.01	0.05	0.03	0.04	0.11
REJ$_r$	0.01	0.07	0.05	0.09	0.00	0.00	0.07	0.08	0.00	0.07	0.03	0.02	0.10
LEN$_c$	5.16	1.20	0.99	2.51	4.14	0.30	0.16	0.19	0.20	0.22	0.68	1.22	2.13
LEN$_r$	7.66	1.25	1.99	6.29	11.21	0.78	0.18	0.24	0.35	0.40	0.95	2.25	2.17
	$n = 500$												
MEAN	2.08	1.87	0.38	1.30	0.18	0.75	0.89	0.84	0.26	0.16	1.03	0.63	4.08
STD	0.45	0.23	0.11	0.11	0.02	0.03	0.02	0.03	0.03	0.02	0.11	0.15	0.44
REJ$_c$	0.25	0.07	0.14	0.00	0.00	0.00	0.01	0.07	0.04	0.04	0.04	0.01	0.10
REJ$_r$	0.04	0.07	0.03	0.00	0.00	0.00	0.02	0.06	0.01	0.06	0.04	0.01	0.10
LEN$_c$	3.59	0.81	0.65	1.44	2.67	0.18	0.10	0.10	0.14	0.12	0.43	0.84	1.57
LEN$_r$	4.50	0.82	0.91	2.84	5.96	0.37	0.10	0.12	0.18	0.16	0.50	1.17	1.58

Legend: 200 Monte Carlo replications. MEAN/STD: average and empirical standard error of estimates. REJ$_c$, REJ$_r$: rejection rates for 5% level t-test. LEN: median length of 95% confidence intervals.

of the estimates, rejection rates using standard errors that assume correct specification and those that allow for misspecification, and the length of resulting 95% confidence intervals. The standard error estimates used to compute the tests and confidence intervals do not account for the prior regularization, which is assumed to be *asymptotically* negligible.

Most estimates are centered at the true value when $n = 192$. A few estimates are somewhat biased towards the prior mode, most notably the risk aversion τ^{-1}, which lies between the true value 3.18 and the prior mode 1.88. The rejection rates are generally close to the 5% level or conservative; significant overrejection is observed only for τ^{-1} when using the non-robust standard errors, driven by prior-induced bias. The robust standard errors tend to be larger, producing lower rejection rates and wider confidence intervals. The estimation precision improves when n is increased to 500, and the other conclusions remain similar. The specification test for all variables has a rejection rate of 0.05 and 0.04 for $n = 192$ and $n = 500$, respectively. For consumption only, the rejection rates are 0.05 and 0.04. Both are close to the nominal level. Additional Monte Carlo simulation for the medium-scale Smets and Wouters (2007) model can be found in Table I4, Appendix I.

7 Empirical Illustrations

Two macroeconomic and one financial applications illustrate different aspects of the OT filter and estimation. The first two revisit a small and medium-scale DSGE model using the same sample period 1960Q1-2007Q4 for both.

7.1 Small New-Keynesian Model

The first empirical application further considers the Lubik and Schorfheide (2003) model. Following LS, there are two specifications: determinacy, with a unique equilibrium, and indeterminacy, where sunspot equilibria exist. The parameters are described in Table 2. Indeterminacy adds 4 sunspot parameters: $M_{r\epsilon}, M_{g\epsilon}$, and $M_{z\epsilon}$ capture the correlation between the 3 shocks and the sunspot shock, σ_ϵ is its standard deviation.

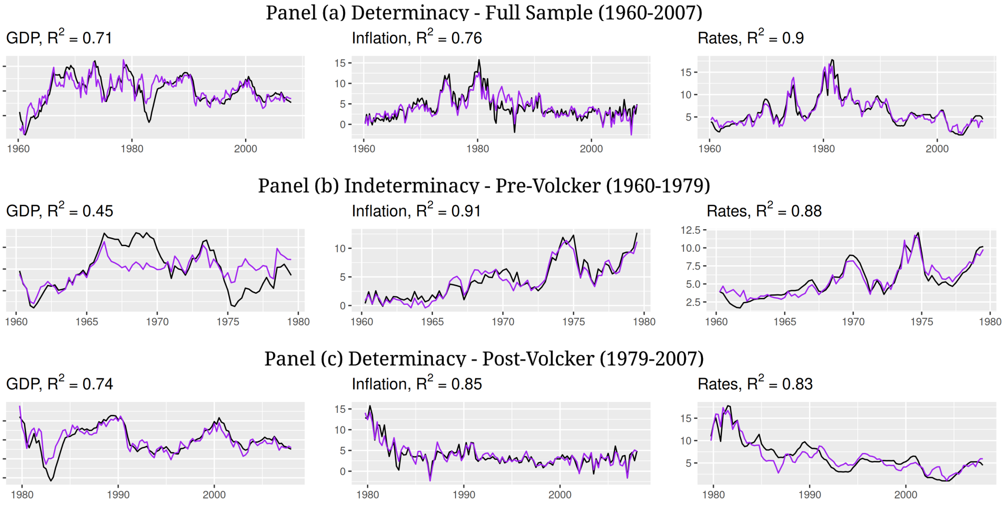
Table 2: LS Model: Parameter Estimates, Specification Test ($k = 4$ lags)

Parameter Estimates		Determinacy (Full Sample)			Indeterminacy (Pre-Volcker)			Determinacy (Post-Volcker)		
θ	Parameter Interpretation	EST	SD$_c$	SD$_r$	EST	SD$_c$	SD$_r$	EST	SD$_c$	SD$_r$
τ^{-1}	risk aversion	2.43	4.12	2.06	1.29	3.96	0.77	1.56	1.29	3.60
r^*	steady state real interest rate	1.87	0.28	0.38	0.99	0.56	0.48	2.49	0.73	1.45
κ	Phillips curve slope	0.31	0.58	0.11	0.54	1.85	0.44	0.30	0.48	4.16
ψ_1	inflation target	1.29	0.18	0.32	0.69	0.11	0.19	1.80	0.67	14.99
ψ_2	output target	0.16	0.72	1.10	0.14	0.67	0.28	0.18	1.17	31.72
ρ_r	interest rate smoothing	0.68	0.06	0.19	0.46	0.18	0.18	0.79	0.08	1.87
ρ_g	exog spending AR	0.89	0.06	0.04	0.74	0.40	0.16	0.92	0.05	0.14
ρ_z	technology shock AR	0.82	0.08	0.04	0.78	0.10	0.12	0.83	0.07	0.24
σ_r	monetary policy shock SD	0.24	0.04	0.06	0.22	0.05	0.08	0.21	0.07	0.73
σ_g	exog spending SD	0.18	0.05	0.05	0.28	1.17	0.21	0.18	0.10	0.43
σ_z	technology shock SD	1.56	0.42	0.25	1.12	0.50	0.34	1.14	0.41	2.38
ρ_{gz}	exog spending-technology cor	0.90	0.27	0.22	0.18	3.19	1.09	0.35	0.68	1.75
$M_{r\epsilon}$	sunspot-monetary coef	–	–	–	0.43	1.82	1.04	–	–	–
$M_{g\epsilon}$	sunspot-exog spending coef	–	–	–	-1.80	5.87	1.66	–	–	–
$M_{z\epsilon}$	sunspot-technology coef	–	–	–	0.63	0.91	0.24	–	–	–
σ_ϵ	sunspot shock SD	–	–	–	0.08	4.41	1.33	–	–	–
π^*	steady state inflation	4.07	0.74	0.73	5.10	1.78	1.55	3.83	0.82	0.79
Specification Test		STAT	10%	5%	STAT	10%	5%	STAT	10%	5%
All		121.2	140.0	189.6	59.4	106.2	169.6	65.9	164.8	235.3
Output		56.0	90.6	123.8	42.8	51.1	86.0	30.1	131.8	188.6
Inflation		45.2	27.1	37.8	7.0	33.9	55.0	16.6	22.2	29.0
Interest Rate		20.0	31.1	42.1	9.6	22.7	36.6	19.2	22.0	30.8

Legend: EST: parameter estimates $\hat{\theta}_n$. SD$_c$: standard errors assuming correct model specification. SD$_r$: misspecification-robust standard errors. STAT: $nQ_n(\hat{\theta}_n; \hat{\psi}_{nk})$. 10%, 5%: critical values for specification test at corresponding significance levels. All: specification test on all variables. Output, Inflation, Interest Rate: specification test on individual variables. $n = 192, 78, 114$ for the full, pre and post-Volcker samples.

The sample is constructed and divided into subsamples as in Clarida et al. (2000): the full sample (1960Q1-2007Q4), the pre-Volcker period (1960Q1-1979Q2), and the post-Volcker period (1979Q3-2007Q4). They are associated with determinate, indeterminate, and determinate policy regimes, respectively. To remain consistent with LS and subsequent analyses,

Figure 2: LS Model: Actual and Fitted Values



Panel (a) Determinacy - Full Sample (1960-2007)

Panel (b) Indeterminacy - Pre-Volcker (1960-1979)

Panel (c) Determinacy - Post-Volcker (1979-2007)

Legend: Black solid line = data, Purple solid line = coupling. $R^2 = [\sum_t(\tilde{y}_t - y_t)^2]/[\sum_t(\tilde{y}_t - \tilde{y}_n)^2]$, computed for each variable.

their log-prior density π is used to penalize the OT loss so that the estimates minimize $nQ_n(\theta; \hat{\psi}_n) - \log \pi(\theta)$. The baseline auxiliary model is a VAR(4), results with $k = 2$ are reported in Appendix I. The weighting matrix W_n is diagonal with the inverse of the variances of the three observables.

Point estimates and standard errors are reported in Table 2. The estimates for the pre-Volcker and post-Volcker periods are in line with those in LS, computed using Bayesian likelihood inference. The steady-state real interest rate r^* is lower in the pre-Volcker sample, and the inflation target coefficient π^* is significantly below 1, suggesting the Fed allowed the real interest rate to decline when facing inflation. This is reversed in the post-Volcker sample. These findings are not new; nevertheless, it is interesting to observe that they are consistent when enforcing model-consistent filtering. The full-sample estimates are broadly similar to the post-Volcker ones. The standard errors that assume correct model specification (SD_c) tend to be wider than those in LS. Those allowing for model misspecification (SD_r) tend to be close to SD_c, but are much wider for the Taylor rule coefficients in the Post-Volcker regime. This likely reflects the weak identification of these coefficients and the small sample size. Finally, the correlation ρ_{gz} hits the upper bound of 0.9, suggesting that the model requires an extreme parameter value to fit the data, reflecting a tension between them.

The specification test, also in Table 2, further investigates tensions between data and model. In the full sample, the test rejects inflation at the 5% significance level. With $k = 2$ the model is rejected overall and for each variable individually (Table I2, Appendix I). The pre and post-Volcker samples do not reject the model. This finding is consistent with LS. None of the variables are individually rejected on the two subsamples.

Moreover, the methods enable us to contrast the actual data with their model-implied values (i.e. the coupling) to obtain an intuitive understanding of their discrepancies. Figure 2 contrasts actual and model-consistent data for GDP, inflation, and interest rate series, respectively, for the full sample and the two subsamples. For GDP, the actual data exhibit a deeper recession and lower inflation rates in the 1980s than those implied by the model. In other words, the model overpredicts the levels of GDP and inflation compared to the data. This finding confirms that this model, with time-invariant parameters, is unable to capture the rich GDP and inflation dynamics present in the data for the full sample period.

7.2 Medium-Scale DSGE Model

The second empirical application considers the Smets and Wouters (2007, SW) model. Table I3, Appendix I, includes parameter interpretations and prior distributions from SW used in our estimation. This model includes 36 free parameters and is estimated on 7 observables: consumption, investment, output and wage growth, hours worked, inflation, and interest rate. There are as many shocks: productivity, exogenous spending, monetary policy, investment-specific technology, price markup, wage markup, and risk premium shocks.

Estimation. To illustrate the scope of OTE, the full model and 3 singular versions are estimated using the same method. Estimates are reported in Tables 3 (for the full model) and I5 (for singular models), Appendix I. Likelihood-based posterior estimates computed with SW's prior are reported in Table 3 as a reference. All specifications rely on a VAR(4) auxiliary model; the inverse of the variances of the observables is used as W_n.

The 3 singular models remove, in order, the risk premium, wage markup, and price markup shocks reducing to 6, 5, and 4 shocks for 7 observables. The choice of shocks to remove follows Qu (2018), to reflect a view that they have a weaker structural interpretation than the remaining 4. It is interesting to examine their impact on the model fit. The standard likelihood approach cannot estimate singular DSGE models because the covariance matrix of the one-step-ahead forecasting errors $\Sigma(\theta)$ is singular. Qu (2018) used a composite likelihood framework and did not formally test the resulting models. OTE handles both singular and nonsingular models within the same framework.

Additionally, note that although SW chose to fit their model to seven variables, it has implications for additional macro variables including the price of capital and capital utiliza-

Table 3: SW Model: Estimates and Standard Errors

θ	Parameter Interpretation	OT Estimate			Posterior		
		EST	SD_c	SD_r	MEAN	5%	95%
ρ_{ga}	Corr.: tech. and exog. spending shocks	0.47	0.27	0.53	0.58	0.45	0.70
μ_w	Wage mark-up shock MA	0.88	0.13	0.20	0.90	0.83	0.95
μ_p	Price mark-up shock MA	0.78	0.25	0.86	0.81	0.63	0.90
α	Share of capital in production	0.24	0.04	0.05	0.23	0.20	0.26
ψ	Elast. of capital utilization adjustment cost	0.44	0.22	0.78	0.49	0.33	0.66
φ	Investment adjustment cost	3.00	1.50	3.42	6.12	4.57	7.87
σ_c	Elast. of Intertemporal substitution	1.01	0.12	0.52	1.50	1.28	1.74
λ	Habit persistence	0.74	0.12	0.38	0.71	0.63	0.78
ϕ_p	Fixed costs in production	1.50	0.26	0.60	1.68	1.55	1.81
ι_w	Wage indexation	0.85	0.30	0.95	0.56	0.33	0.76
ξ_w	Wage stickiness	0.84	0.12	0.04	0.77	0.68	0.85
ι_p	Price indexation	0.27	0.32	0.42	0.25	0.12	0.40
ξ_p	Price stickiness	0.80	0.06	0.12	0.69	0.61	0.77
σ_l	Labor supply elasticity	1.00	2.38	2.92	2.25	1.41	3.21
r_π	Taylor rule: inflation weight	1.80	1.06	0.77	2.03	1.77	2.30
$r_{\Delta y}$	Taylor rule: output gap change weight	0.16	0.05	0.10	0.21	0.17	0.25
r_y	Taylor rule: output gap weight	0.16	0.16	0.08	0.10	0.07	0.14
ρ	Taylor rule: interest rate smoothing	0.90	0.08	0.07	0.82	0.78	0.86
ρ_a	Productivity shock AR	0.94	0.03	0.22	0.97	0.96	0.99
ρ_b	Risk premium shock AR	0.68	0.10	0.19	0.28	0.12	0.49
ρ_g	Exogenous spending shock AR	0.86	0.09	1.22	0.97	0.95	0.98
ρ_i	Investment shock AR	0.47	0.15	0.13	0.70	0.61	0.79
ρ_r	Monetary policy shock AR	0.47	0.28	0.71	0.17	0.07	0.29
ρ_p	Price mark-up shock AR	0.96	0.03	0.04	0.96	0.91	0.99
ρ_w	Wage mark-up shock AR	0.92	0.11	0.21	0.96	0.92	0.98
σ_a	Productivity shock std. dev.	0.32	0.08	0.13	0.46	0.42	0.50
σ_b	Risk premium shock std. dev.	0.11	0.02	0.05	0.23	0.18	0.28
σ_g	Exogenous spending shock std. dev.	0.33	0.04	0.24	0.50	0.46	0.55
σ_i	Investment shock std. dev.	0.33	0.07	0.16	0.41	0.35	0.48
σ_r	Monetary policy shock std. dev.	0.09	0.05	0.07	0.22	0.20	0.25
σ_p	Price mark-up shock std. dev.	0.05	0.04	0.24	0.12	0.09	0.14
σ_w	Wage mark-up shock std. dev.	0.25	0.03	0.11	0.28	0.25	0.32
$\bar{\gamma}$	Trend growth: real GDP, Infl., Wages	0.46	0.01	0.03	0.45	0.41	0.48
r	Discount rate	0.20	0.09	0.23	0.12	0.06	0.21
$\bar{\pi}$	Steady state inflation rate	0.80	0.26	0.25	0.68	0.53	0.85
$\bar{l}$	Steady state hours worked	0.16	0.65	0.77	1.31	-0.09	2.75

Legend: Prior distribution and estimation bounds can be found in Table I3, Appendix I

tion rate. Adding any of these variables to the set of observables will immediately make the model singular. In fact, for all medium-scale DSGE models, nonsingularity arises only because we restrict the estimation to a limited set of macro variables.

For the full model, the OT estimates are similar to the posterior means. In all but two cases, the posterior means fall within the 95% confidence intervals obtained from OT estimates and robust standard errors. For the remaining two cases, OT produces a more persistent risk premium shock process with a lower residual standard deviation. The robust standard errors are almost always greater than those assuming correct model specification.

For singular models, removing the risk premium shock has little effect on the estimates — they are close to the nonsingular case; all confidence intervals overlap with their nonsingular counterparts. When the wage markup shock is also removed, ξ_w (wage stickiness) and ι_w (wage indexation) decrease, while ρ_r (monetary policy shock persistence) increases, though their confidence interval still overlaps due to substantial estimation uncertainty. When further removing the price markup shock, ξ_p (price stickiness) and ι_p (wage stickiness) both drop noticeably. Although not reported here, the effects of these parameter differences on the model can be further assessed by plotting the impulse response functions.

Specification Testing. The specification test does not reject the original model at the 5% significance level, as shown in Table 4. However, the test rejects the model's fit for consumption, even with 7 shocks. Further investigation reveals that the model under-predicts contractions, e.g. in 1974Q4 consumption fell by 2.38% vs. 1.34% for the fitted values. Fitted consumption is less volatile (standard deviation of 0.53 vs. 0.68), and more persistent (autocorrelation of 0.46 vs. 0.18). For singular models, when the risk premium shock is removed, the specification test rejects the full model at the 5% significance level; for individual tests, the results for consumption, wage, and interest rate reject the null hypothesis. When the wage markup shock is removed, the tests on output and labor also reject, implying that 5 out of 7 variables are now rejected. Finally, when the price markup shock is removed, the conclusions remain the same as in the five-shock case, with only two variables—investment and inflation—not rejected by the test at the 5% level.

This is the first attempt to formally test singular DSGE models. The results pinpoint model features that remain compatible or become incompatible with data once shocks are removed. They can be useful tools for researchers to determine which latent processes and mechanisms contribute to the fit of a model within a unified framework.

Filtering the Latent Shock Processes. The OTF produces model-consistent values of latent variables, including shock processes. Using the original SW model, we compare these filtered values with their counterparts produced by the KF, which does not enforce model consistency. The same parameter values (OT estimates) are used to ensure comparability.

Figure 3, panel a) displays the results for the 7 shocks separately. For TFP, the KF yields a puzzling conclusion: the economy was boosted by a positive TFP process from 1960 until about 1980, and then depressed by a negative TFP process between 1980 and 2000. In contrast, the OT filter produces a process with negative values during the early

Table 4: SW Model: Specification Testing With(out) Stochastic Singularity

	7 shocks			6 shocks			5 shocks			4 shocks		
	STAT	10%	5%	STAT	10%	5%	STAT	10%	5%	STAT	10%	5%
All	256.6	320.7	395.0	411.4	282.1	364.5	597.7	291.4	357.0	632.0	288.6	353.3
Cons.	60.4	31.9	36.8	56.5	32.6	39.9	192.9	6.3	8.9	192.7	6.1	8.5
Invest.	25.6	33.2	38.6	34.1	31.9	36.6	45.6	41.0	47.3	45.0	41.1	47.7
Output	34.3	37.0	41.8	34.9	32.4	36.7	54.4	39.8	45.1	57.9	39.2	44.8
Labor	16.3	35.4	47.2	14.7	48.4	66.8	86.9	49.3	63.5	97.6	47.8	60.0
Infl.	52.8	92.5	122.7	75.1	93.5	124.2	67.6	101.1	137.8	69.6	104.1	146.1
Wage	22.6	64.8	80.3	108.3	26.2	33.3	76.4	30.9	37.6	98.3	27.4	33.1
Int. Rate	44.6	47.9	63.3	87.9	33.8	45.2	73.9	49.9	67.3	70.9	48.8	65.6

Legend: All: specification test on all 7 variables (consumption, investment, output, labor, inflation, wage, interest rate). STAT: test statistic for specification test. 5%, 10%: critical values.

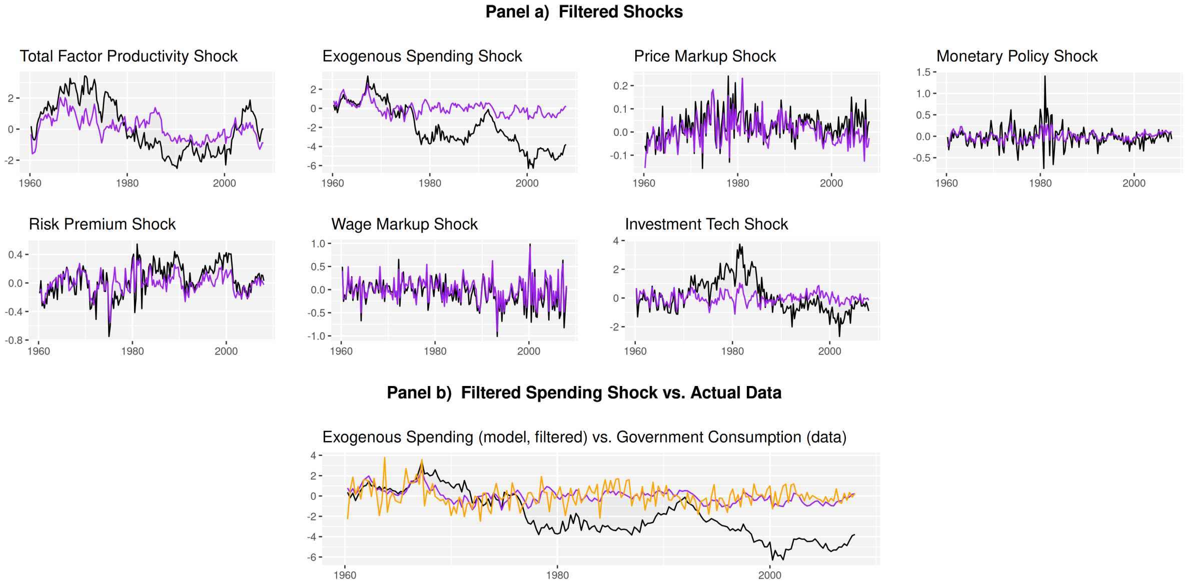
1960s, the mid-1970s recessions, and the slowdown leading up to the 2001 recession. These estimates are clearly more interpretable than the KF values. The investment technology shock shows a similar pattern: the KF yields a mostly positive process from the 1970s to the mid-1980s, which then switches to a mostly negative process. The OT estimates do not have this problem.

For exogeneous spending, KF yields negative values most of the time and exhibits a downward trend, in sharp contrast with the zero-mean assumption. The OTF estimates do not have this problem. For the monetary policy shock, KF produces large oscillating values in the early 1980s, which is puzzling given the monetary tightening that characterizes this period. In contrast, the OTF produces mostly positive values for this period, consistent with this characterization. The last 3 shocks show closer resemblances between KF and OTF.

Table I6, Appendix I, presents the cross and serial-correlations of the 7 shock processes. OT consistently gives values close to the true ones, while KF shows significant discrepancies in several cases; for example, the true first-order serial correlations for investment and monetary policy shocks are both 0.47, but KF yields 0.83 and 0.09. This illustrates that, when using KF, the parameter estimates may not capture the dynamics of the filtered values.

As a validation exercise, Figure 3, panel b), compares the exogenous spending shocks process with government consumption data. They are closely related since exogeneous spending is the difference between output and the sum of consumption, investment and capital utilization (Smets and Wouters, 2007, p588). The correlation between the data and filtered values is 0.36 and 0.05 for OTF and KF, respectively. Although this data was *not used* in the estimation, the OTF captures some of its variation using other series and the model. These results demonstrate that the OT filter, by enforcing model consistency, produces filtered values that obey model assumptions and can be more interpretable in practice.

Figure 3: SW Model: Filtered Shock Processes



Panel a) Filtered Shocks

Panel b) Filtered Spending Shock vs. Actual Data

Legend: Black solid line: Kalman Filter (KF), Purple solid line: Optimal Transport Filter (OTF). Both filters are applied using the same OT estimates found in Table 3, Appendix I. Orange line: Real Government Consumption Expenditures and Gross Investment; Source: FRED (GCEC1).

To wrap up, we have considered this medium-scale DSGE model to illustrate that the proposed methods can be used to estimate singular and nonsingular models, testing their specifications, and produce filtered variables, all within the same framework. The methods' applications are not restricted to macroeconomics; we next consider a financial application.

7.3 Affine Term Structure Model

The third application considers a term structure model where three latent factors explain six yields. Algorithm 2 provides a way to assess the extent to which the unaltered, stochastically singular, structural model fits the empirical data; no measurement errors are introduced.

The specification of the structural model follows Ang and Piazzesi (2003) and Hamilton and Wu (2012). Let F_t denote three latent factors, which follow a Gaussian VAR:

$$F_{t+1} = c + \rho F_t + v_{t+1},$$

with $v_{t+1} \overset{iid}{\sim} \mathcal{N}(0, I)$. Under the assumption of no-arbitrage, the price of a pure discount bond at time t, P_t, is a function of this state vector and a stochastic discount factor:

$$P_t = E_t(P_{t+1}M_{t,t+1}) = \int P_{t+1}(F_{t+1})M_{t,t+1}(F_{t+1})\phi(F_{t+1}; c + \rho F_t, I)dF_{t+1}, \tag{5}$$

where $\phi(\cdot; c+\rho F_t, I)$ represents a multivariate normal density, with mean $c+\rho F_t$ and identity covariance matrix. Affine term structure models assume that the one-period short rate r_t is an affine function of the state vector. They specify the stochastic discount factor $M_{t,t+1}$ to be a function of λ_t, the market prices of risk, which is also an affine function of F_t: $r_t = \delta_0 + \delta_1' F_t$, $M_{t,t+1} = \exp[-r_t - (1/2)\lambda_t'\lambda_t - \lambda_t' v_{t+1}]$, $\lambda_t = \lambda + \Lambda F_t$. As highlighted in Hamilton and Wu (2012), the pricing equation in (5) has an intuitive representation under the risk-neutral measure, under which all assets are discounted by the short-term interest rate: $P_t = \exp(-r_t) \int P_{t+1}(F_{t+1})\phi(F_{t+1}; c^Q + \rho^Q F_t, I) dF_{t+1}$, where $F_{t+1} = c^Q + \rho^Q F_t + v_{t+1}^Q$, with $v_{t+1}^Q \overset{iid}{\sim} \mathcal{N}(0, I)$, $c^Q = c - \lambda$, and $\rho^Q = \rho - \Lambda$.

It is well-documented that parameter normalizations are necessary to identify the model. Following Ang and Piazzesi (2003) and Hamilton and Wu (2012), we set: $c = 0, \delta_1 \geqslant 0$, and ρ^Q lower triangular. Then, the yield y_t^n on an n-period pure-discount bond is given by:

$$\begin{aligned} y_t^n &= a_n + b_n' F_t, \quad b_n = \frac{1}{n}\left(I + \rho^{Q\prime} + ... + \left(\rho^{Q\prime}\right)^{n-1}\right)\delta_1 \\ a_n &= \delta_0 + \frac{1}{n}\left(b_1' + 2b_2' + ... + (n-1)b_{n-1}'\right)c^Q - \frac{1}{2n}\left(b_1'b_1 + 4b_2'b_2 + ... + (n-1)^2 b_{n-1}'b_{n-1}\right). \end{aligned}$$

The parameters to be estimated are $\delta_0, \delta_1, \rho^Q, c^Q$, and ρ. The diagonal elements of ρ^Q are required to be in decreasing order to ensure identification. The auxiliary model is VAR(4). The yields are weighted diagonally by the inverse of their variance. The OT objective Q_n, is penalized by prior distribution which enforces an ordering of the factors.

Table 5: Affine Term Structure Model: Parameter Estimates

	Prior [MEAN, SD]			OT Estimates		
ρ^Q	[0.9,0.2]	[0,0.2]	-	0.999	-	-
	[0,0.2]	[0.8,0.2]	-	0.022	0.963	-
	[0,0.2]	[0,0.2]	[0.6,0.2]	0.018	0.209	0.723
δ_1	[0.01,0.1]	[0.01,0.1]	[0.01,0.1]	0.003	0.023	0.037
δ_0	[0.4,0.2]	-	-	0.441	-	-
c^Q	[0,1]	[0,1]	[0,1]	1.254	-0.026	0.536
ρ	[0.9,0.2]	[0,0.2]	[0,0.2]	0.958	0.012	0.071
	[0,0.2]	[0.8,0.2]	[0,0.2]	0.004	0.908	0.102
	[0,0.2]	[0,0.2]	[0.6,0.2]	0.011	0.133	0.770

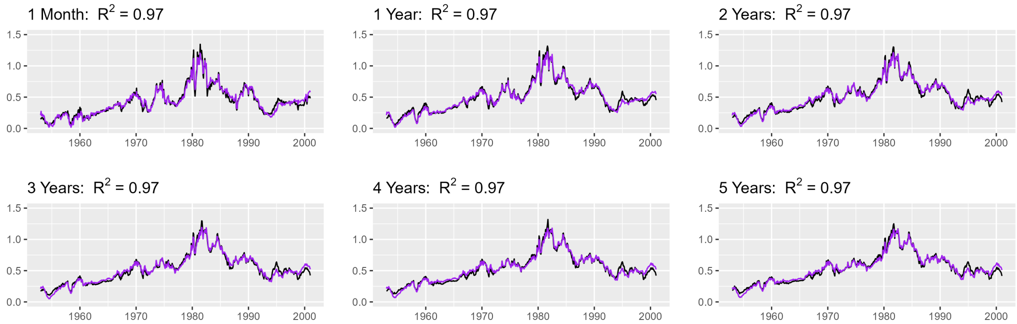
Legend: A weakly informative prior is used to enforce the ordering of the factors.

The estimates in Table 5 are similar to Hamilton and Wu (2012, Table 5).[9] The first factor is very persistent, and the off-diagonal elements of ρ are small in magnitude. The estimate of δ_0 is close to the mean of the short-term rate. The main difference is that the

[9] Hamilton and Wu (2012) use four yields and introduce measurement error in one of the series.

first element of δ_1 in our case is smaller than theirs (their estimate is 0.017). The parameters ρ^Q, δ_1 and c^Q are highly correlated: it is possible to move their values jointly with little effect on Q_n, suggesting weak identification.[10] The prior helps stabilize the estimates.

Figure 4: Affine Term Structure Model: Actual and Model-Consistent Yields



Legend: Actual yields: Black solid line, Coupling: Purple solid line. Fit shown for prior regularized estimates; the fit is virtually identical without prior regularization.

What is intriguing is whether, with three shocks, the model can approximate the dynamics of six observables. Figure 4 compares actual yield data with their couplings. These values track each other closely, with the R^2 equal to 0.97 in all six cases.

With three shocks, the model accounts for 97% of the dynamics in the data for this sample period. The plots suggest no apparent model misspecification for this sample period. Adding measurement errors in this case would be a shortcut to obtaining parameter estimates with a likelihood and cannot reveal additional information regarding the baseline model specification. Our approach allows us to obtain parameter estimates with a graphical method to assess the model fit. Finally, we conjecture that incorporating more information to improve the identification of ρ^Q, δ_1 and c^Q can be beneficial for further improve this model.

8 Conclusion

This paper has introduced a computationally attractive method for filtering and estimating parameters of potentially misspecified dynamic models using dynamic optimal transportation. Empirical applications illustrate how this can be used to visually assess the fit of a model, by comparing the actual and the coupled time-series, and to formally test the model specification over all or some specific variables. Several extensions could be of interest in future research. Deriving a plugin map for general non-linear state-space models in (1) could

[10] For this reason standard errors are not reported. The specification test is not reported either.

be useful if it helps circumvents the curse of dimensionality, as done here. A simple but useful Corollary to Theorem 3 would be to consider an Anderson-Rubin type statistic for models that are potentially weakly identified.

References

Altissimo, F. and A. Mele (2009): "Simulated non-parametric estimation of dynamic models," *The Review of Economic Studies*, 76, 413–450.

Anderson, B. and J. B. Moore (1979): *Optimal filtering.*

Ang, A. and M. Piazzesi (2003): "A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables," *Journal of Monetary economics*, 50, 745–787.

Bassetti, F., A. Bodini, and E. Regazzini (2006): "On minimum Kantorovich distance estimators," *Statistics & probability letters*, 76, 1298–1302.

Bassetti, F. and E. Regazzini (2006): "Asymptotic properties and robustness of minimum dissimilarity estimators of location-scale parameters," *Theory of Probability & Its Applications*, 50, 171–186.

Bernton, E., P. E. Jacob, M. Gerber, and C. P. Robert (2019): "On parameter estimation with the Wasserstein distance," *Information and Inference: A Journal of the IMA*, 8, 657–676.

Buckley, M. J. and G. K. Eagleson (1988): "An approximation to the distribution of quadratic forms in normal random variables," *Australian Journal of Statistics*, 30, 150–159.

Chopin, N. and O. Papaspiliopoulos (2020): *An introduction to sequential Monte Carlo*, vol. 4, Springer.

Clarida, R., J. Gali, and M. Gertler (2000): "Monetary policy rules and macroeconomic stability: evidence and some theory," *The Quarterly journal of economics*, 115, 147–180.

Davidson, J. (2021): *Stochastic Limit Theory: An Introduction for Econometricians*, Oxford University Press.

Del Negro, M. and F. Schorfheide (2009): "Monetary policy analysis with potentially misspecified models," *American Economic Review*, 99, 1415–1450.

Del Negro, M., F. Schorfheide, F. Smets, and R. Wouters (2007): "On the fit of new Keynesian models," *Journal of Business & Economic Statistics*, 25, 123–143.

Domowitz, I. and H. White (1982): "Misspecified models with dependent observations," *Journal of Econometrics*, 20, 35–58.

Dowson, D. and B. Landau (1982): "The Fréchet distance between multivariate normal distributions," *Journal of multivariate analysis*, 12, 450–455.

Dudley, R. M. (1969): "The speed of mean Glivenko-Cantelli convergence," *The Annals of Mathematical Statistics*, 40, 40–50.

Fernández-Villaverde, J., J. F. Rubio-Ramírez, T. J. Sargent, and M. W. Watson (2007): "ABCs (and Ds) of understanding VARs," *American economic review*, 97, 1021–1026.

Forneron, J.-J. (2023): "A Sieve-SMM Estimator for Dynamic Models," *Econometrica*, 91, 943–977.

Galichon, A. (2018): *Optimal transport methods in economics*, Princeton University Press.

Gallant, A. R. and D. W. Nychka (1987): "Semi-nonparametric maximum likelihood estimation," *Econometrica: Journal of the econometric society*, 363–390.

Gallant, A. R. and G. Tauchen (1996): "Which moments to match?" *Econometric theory*, 12, 657–681.

Genevay, A., G. Peyré, and M. Cuturi (2018): "Learning generative models with sinkhorn divergences," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 1608–1617.

Givens, C. R. and R. M. Shortt (1984): "A class of Wasserstein metrics for probability distributions." *Michigan Mathematical Journal*, 31, 231–240.

Gonçalves, S. and L. Kilian (2007): "Asymptotic and bootstrap inference for AR(∞) processes with conditional heteroskedasticity," *Econometric Reviews*, 26, 609–641.

Gourieroux, C. and A. Monfort (1996): *Simulation-based econometric methods*, Oxford university press.

Hamilton, J. D. and J. C. Wu (2012): "Identification and estimation of Gaussian affine term structure models," *Journal of Econometrics*, 168, 315–331.

Hannan, E. J. and M. Deistler (2012): *The statistical theory of linear systems*, SIAM.

Hansen, L. P. and T. J. Sargent (2008): *Robustness*, Princeton university press.

Harvey, A. C. (1990): *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press.

Horn, R. A. and C. R. Johnson (1991): "Topics in matrix analysis," *Cambridge University Presss, Cambridge*, 37, 39.

Jaffard, S. (1990): "Propriétés des matrices "bien localisées" près de leur diagonale et quelques applications," in *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, Elsevier, vol. 7, 461–476.

Kaji, T., E. Manresa, and G. Pouliot (2023): "An adversarial approach to structural estimation," *Econometrica*, 91, 2041–2063.

King, R. G. and S. T. Rebelo (1999): "Resuscitating real business cycles," *Handbook of macroeconomics*, 1, 927–1007.

Komunjer, I. and S. Ng (2011): "Dynamic identification of dynamic stochastic general equilibrium models," *Econometrica*, 79, 1995–2032.

Kristensen, D. and Y. Shin (2012): "Estimation of dynamic models with nonparametric simulated maximum likelihood," *Journal of Econometrics*, 167, 76–94.

Kuersteiner, G. M. (2005): "Automatic inference for infinite order vector autoregressions," *Econometric Theory*, 21, 85–115.

Lee, J. (2010): *Introduction to topological manifolds*, vol. 202, Springer Science & Business Media.

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

Lubik, T. A. and F. Schorfheide (2003): "Computing sunspot equilibria in linear rational expectations models," *Journal of Economic dynamics and control*, 28, 273–285.

——— (2004): "Testing for indeterminacy: An application to US monetary policy," *American Economic Review*, 94, 190–217.

Lütkepohl, H. (2005): *New introduction to multiple time series analysis*, Springer Science & Business Media.

Magnus, J. R. and H. Neudecker (2019): *Matrix differential calculus with applications in statistics and econometrics*, John Wiley & Sons.

O'Connor, K., K. McGoff, and A. B. Nobel (2022): "Optimal transport for stationary Markov chains via policy iteration," *Journal of Machine Learning Research*, 23, 1–52.

Olkin, I. and F. Pukelsheim (1982): "The distance between two random vectors with given dispersion matrices," *Linear Algebra and its Applications*, 48, 257–263.

Peligrad, M. (2002): "Some remarks on coupling of dependent random variables," *Statistics & probability letters*, 60, 201–209.

Perron, P. and T. Wada (2009): "Let's take a break: Trends and cycles in US real GDP," *Journal of monetary Economics*, 56, 749–765.

Peyré, G. and M. Cuturi (2019): "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, 11, 355–607.

Pisier, G. (1983): "Some applications of the metric entropy condition to harmonic analysis," *Banach Spaces, Harmonic Analysis, and Probability Theory*, 123–154.

Plosser, C. I. (1989): "Understanding real business cycles," *Journal of Economic Perspectives*, 3, 51–77.

Pollard, D. (2002): *A user's guide to measure theoretic probability*, 8, Cambridge University Press.

Qu, Z. (2018): "A composite likelihood framework for analyzing singular DSGE models," *Review of Economics and Statistics*, 100, 916–932.

Rio, E. (1999): *Théorie asymptotique des processus aléatoires faiblement dépendants*, vol. 31, Springer Science & Business Media.

Sayed, A. H. (2001): "A framework for state-space estimation with uncertain models," *IEEE Transactions on Automatic Control*, 46, 998–1013.

Shafieezadeh Abadeh, S., V. A. Nguyen, D. Kuhn, and P. M. Mohajerin Esfahani (2018): "Wasserstein distributionally robust Kalman filtering," *Advances in Neural Information Processing Systems*, 31.

Smets, F. and R. Wouters (2007): "Shocks and frictions in US business cycles: A Bayesian DSGE approach," *American economic review*, 97, 586–606.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, vol. 126, Springer New York.

Villani, C. (2003): *Topics in optimal transportation*, vol. 58, American Mathematical Soc.

——— (2009): *Optimal transport: old and new*, vol. 338, Springer.

Watson, M. W. (1986): "Univariate detrending methods with stochastic trends," *Journal of monetary economics*, 18, 49–75.

White, H. (2014): *Asymptotic theory for econometricians*, Academic press.

Wooldridge, J. M. and H. White (1988): "Some invariance principles and central limit theorems for dependent heterogeneous processes," *Econometric theory*, 4, 210–230.

Zaitsev, A. Y. (2013): "The accuracy of strong Gaussian approximation for sums of independent random vectors," *Russian Mathematical Surveys*, 68, 721.

Appendix A Definitions

The following recalls two notions of dependence, further details can be found in Davidson (2021, Ch14,15,18). Take the sequence e_t found in Assumptions 1 and 2. Let $\mathcal{F}_t = \sigma(e_t, e_{t-1}, \dots)$ be the sigma-algebra constructed on $e_t, e_{t-1}, \dots$. Similarly, let $\mathcal{F}_{t-m}^{t+m} = \sigma(e_{t+m}, e_{t+m-1}, \dots, e_{t-m+1}, e_{t-m})$. For two sub sigma-algebras $\mathcal{G}, \mathcal{H}$, their strong-mixing coefficient is defined as $\alpha(\mathcal{G}, \mathcal{H}) = \sup_{G \in \mathcal{G}, G \in \mathcal{G}} |\mathbb{P}(G \cap H) - \mathbb{P}(G)\mathbb{P}(H)|$.

Definition A1 (Strong-mixing)**.** *The strictly stationary sequence e_t is said to be α-mixing, i.e. strong mixing, if the α-mixing coefficients $\alpha_m = \alpha(\mathcal{F}_t, \mathcal{F}_{t-m})$ satisfy $\alpha_m \to 0$ as $m \to \infty$.*

Definition A2 (Near-epoch dependence)**.** *The sequence y_t is said to be near-epoch dependent (NED) in L_p-norm on $\{e_t\}_{t=-\infty}^{+\infty}$, for $p > 0$ if: $(\mathbb{E}[\|y_t - \mathbb{E}(y_t|\mathcal{F}_{t-m}^{t+m})\|^p])^{1/p} \leq d_t \nu_m$, where $\nu_m \to 0$ as $m \to \infty$, d_t is a sequence of positive constants, and $\mathcal{F}_{t-m}^{t+m} = \sigma(e_{t+m}, \dots, e_{t-m})$.*

Appendix B Preliminary Results

Lemma B1 (Dependence)**.** *Suppose Assumptions 1, 2, 3, and 4 hold. Let $y_t(\theta; \psi_{k0}) = \mu(\theta) + \sum_{\ell=0}^{\infty} \Lambda_\ell(\theta) P(\theta; \tilde{\Sigma}) e_{t-\ell}$, computed from $(e_{t-\ell})_{\ell \geq 0}$ instead of $(e_{t-\ell,k})_{t > \ell \geq 0}$ for $y_t(\theta; \psi_k)$, and $\phi_j = vec(\Psi_j)$. Then: (1) For all $\theta \in \Theta$, the sequences $\tilde{y}_t$, $y_t(\theta; \psi_0)$, $y_t(\theta; \psi_k)$, $\partial_\theta y_t(\theta; \psi_0)$, and $\partial_\theta y_t(\theta; \psi_k)$ are NED in L_q-norm of size $-b$ on e_t for some $q \leq 2r$, with r defined in Assumption 2 (iii). (2) For all θ and any $j \geq 1$, $\partial_{\phi_j} y_t(\theta; \psi_{k0})$ is NED in L_q-norm of size $-b$; $\partial_{\tilde{\mu}} y_t(\theta; \psi_{k0})$ is constant and deterministic; and $\partial_{vec(\tilde{\Sigma})} y_t(\theta; \psi_{k0})$ is NED in L_q-norm with size $-b$ if $\tilde{\Sigma}$ is invertible and rank$[\Sigma(\theta)]$ is constant.*

The partial derivatives in the Lemma are expressed as follows and will be used in subsequent proofs: $\partial_{\phi_j} y_t(\theta; \psi_{k0}) = -\sum_{\ell=0}^{\infty} (\tilde{y}_{t-\ell-j-1} - \tilde{\mu})' \otimes [\Lambda_\ell(\theta) P(\theta; \tilde{\Sigma})]$, $\partial_{\tilde{\mu}} y_t(\theta; \psi_{k0}) = I + (\sum_{\ell=0}^{\infty} \Lambda_\ell(\theta) P(\theta; \tilde{\Sigma}))(I_d - \sum_{\ell=1}^{\infty} \Psi_\ell)$, and $\partial_{\mathrm{vec}(\tilde{\Sigma})} y_t(\theta; \psi_{k0}) = \sum_{\ell=0}^{\infty} (e'_{t-\ell} \otimes \Lambda_\ell(\theta)) \partial_{\mathrm{vec}(\tilde{\Sigma})} \mathrm{vec}[P(\theta; \tilde{\Sigma})]$.

Appendix C Proofs for the Main Theorems

Proof of Theorem 1. We have: $Q_n(\theta; \hat{\psi}_{nk}) = \frac{1}{n} \sum_{t=1}^{n} \|y_t(\theta; \hat{\psi}_{nk}) - \tilde{y}_t\|_{W_n}^2$. Assumption 3 implies $Q(\cdot; \psi_0)$ is continous on Θ. Given the identification assumption, it is sufficient to derive uniform equivalence and uniform law of numbers, stated as follows:

$$\sup_{\theta \in \Theta} |Q_n(\theta; \hat{\psi}_{nk}) - Q_n(\theta; \psi_0)| = o_p(1), \quad \sup_{\theta \in \Theta} |Q(\theta; \psi_0) - Q_n(\theta; \psi_0)| = o_p(1).$$

Step 1. Uniform Equivalence: $\sup_{\theta\in\Theta}|Q_n(\theta;\hat{\psi}_{nk}) - Q_n(\theta;\psi_0)| = o_p(1)$**.**

We first establish some properties related to upper bounds and stochastic orders to be used in the proof. By Lemma 3, we have $\sup_{j=1,\ldots,k}\|\hat{\Psi}_j - \Psi_j\| = O_p(\sqrt{\log(n)/n})$, $\|\tilde{\mu}_n - \tilde{\mu}\| = O_p(n^{-1/2})$, and $\|\tilde{\Sigma}_{nk} - \tilde{\Sigma}\| = O_p(n^{-1/2})$. Let $e_{t,k} = \tilde{y}_t - \tilde{\mu} - \sum_{j=1}^{k}\Psi_j(\tilde{y}_{t-j-1} - \tilde{\mu})$, $\hat{e}_t = \tilde{y}_t - \tilde{\mu}_n - \sum_{j=1}^{k}\hat{\Psi}_j(\tilde{y}_{t-j-1} - \tilde{\mu}_n)$. For $t-j-1 \leq 0$, set $\tilde{y}_{t-j-1} - \tilde{\mu} = 0$ and $\tilde{y}_{t-j-1} - \tilde{\mu}_n = 0$. This does not affect the properties of the VAR estimates since $k/n = o(n^{-1/2})$, by assumption.

For $q = 2r$, $\mathbb{E}(\|e_t\|^q) < \infty$ and Assumption 2 imply $\mathbb{E}(\|\tilde{y}_t\|^q) < \infty$. Thus, $\sup_{t=1,\ldots,n}\|\tilde{y}_t\| = O_p(n^{1/q})$ (van der Vaart and Wellner, 1996, Lem2.2.2). Likewise, $\sup_{t=1,\ldots,n}\|\hat{e}_t - e_{t,k}\|^2 \leq (k+1)\max(\|\hat{\Psi}_1 - \Psi_1\|^2, \ldots, \|\hat{\Psi}_k - \Psi_k\|^2, \|\tilde{\mu}_n - \tilde{\mu}\|^2)\sup_{t=1,\ldots,n}\|\tilde{y}_t\|^2 = O_p(k\log(n)n^{2/q-1})$ and $\sup_{t=1,\ldots,n}\|\hat{e}_t - e_{t,k}\| = O_p(\sqrt{k\log(n)}n^{1/q-1/2}) = o_p(1)$, since $q > 8$ and $k = o(n^{1/3})$. Note that $[\mathbb{E}(\|e_{t,k} - e_t\|^q)]^{1/q} \leq [\sum_{j=k+1}^{\infty}\|\Psi_j\|][\mathbb{E}(\|\tilde{y}_t - \tilde{\mu}\|^q)]^{1/q} = O(n^{-1/2})$, for $t \geq k+1$, by assumption. This implies that $\sup_{t=k+1,\ldots,n}\|e_{t,k} - e_t\| = O_p(n^{1/q-1/2}) = o_p(1)$. Also, $\mathbb{E}(\|e_{t,k} - e_t\|^q) < \infty$ implies $\sup_{t=1,\ldots,k}\|e_{t,k} - e_t\| = O_p(k^{1/q})$.[11]

Also, $\tilde{\Sigma}_{nk} = \tilde{\Sigma} + O_p(n^{-1/2})$ implies $0 < \underline{\lambda}/2 \preceq \tilde{\Sigma}_{nk} \preceq 2\overline{\lambda} < \infty$ with probability approaching one. Because $\theta \to \Sigma(\theta)$ has constant rank, $(\theta, \Sigma) \to P(\theta;\Sigma)$ is continuously differentiable with bounded derivative on $\Theta \times \{\Sigma \text{ s.t. } \underline{\lambda}/2 \preceq \Sigma \preceq 2\overline{\lambda}\}$ by Lemma 2. Then, for any θ and any Σ_1 and Σ_2 in this set: $\|P(\theta;\Sigma_1) - P(\theta;\Sigma_2)\| \leq \sup_{\theta\in\Theta,\underline{\lambda}/2\preceq\Sigma\preceq 2\overline{\lambda}}\|\partial_{\text{vec}(\Sigma)}\text{vec}[P(\theta;\Sigma)]\|_\infty\|\Sigma_1 - \Sigma_2\|$. This implies $\sup_{\theta\in\Theta}\|P(\theta;\tilde{\Sigma}_{nk}) - P(\theta;\tilde{\Sigma})\| = O_p(n^{-1/2})$.

Now, we apply these bounds to study $y_t(\theta;\hat{\psi}_{nk}) - y_t(\theta;\psi_0)$. Recall that $y_t(\theta;\hat{\psi}_{nk}) = \mu(\theta) + \sum_{j=0}^{\infty}\Lambda_j(\theta)P(\theta;\tilde{\Sigma}_{nk})\hat{e}_{t-j}$ with $\hat{e}_t = 0$ for $t \leq 0$; $y_t(\theta;\psi_k) = \mu(\theta) + \sum_{j=0}^{\infty}\Lambda_j(\theta)P(\theta;\tilde{\Sigma})e_{t-j,k}$ with $e_{t-j,k} = 0$ for $t \leq 0$; and $y_t(\theta;\psi_0) = \mu(\theta) + \sum_{j=0}^{\infty}\Lambda_j(\theta)P(\theta;\tilde{\Sigma})e_{t-j}$. We have

$$\sup_{\theta\in\Theta}\sup_{t=1,\ldots,n}\|y_t(\theta;\hat{\psi}_{nk}) - y_t(\theta;\psi_0)\|$$

$$\leq \sup_{\theta\in\Theta}\|P(\theta;\tilde{\Sigma}_{nk})\|_{op}\left(\sum_{j=0}^{\infty}\sup_{\theta\in\Theta}\|\Lambda_j(\theta)\|_{op}\right)\sup_{t=1,\ldots,n}\|\hat{e}_t - e_{t,k}\| \tag{A1}$$

$$+ \left(\sum_{j=0}^{\infty}\sup_{\theta\in\Theta}\|\Lambda_j(\theta)\|_{op}\right)\sup_{t=1,\ldots,n}\|e_{t,k}\|\sup_{\theta\in\Theta}\|P(\theta;\tilde{\Sigma}_{nk}) - P(\theta;\tilde{\Sigma})\| \tag{B1}$$

$$+ \sup_{\theta\in\Theta}\|P(\theta;\tilde{\Sigma})\|_{op}\left(\sum_{j=0}^{\infty}\sup_{\theta\in\Theta}\|\Lambda_j(\theta)\|_{op}\right)\sup_{t=1,\ldots,n}\|e_t - e_{t,k}\| \tag{C1}$$

$$+ \sup_{t=1,\ldots,n}\sup_{\theta\in\Theta}\|\sum_{j=k+1}^{\infty}\Lambda_j(\theta;\tilde{\Sigma})P(\theta;\tilde{\Sigma})e_{t-j}\|. \tag{D1}$$

[11] Simply use $[\mathbb{E}(\|e_{t,k} - e_t\|^q)]^{1/q} \leq [\mathbb{E}(\|e_{t,k}\|^q)]^{1/q} + [\mathbb{E}(\|e_t\|^q)]^{1/q}$ and $\sup_{t=1,\ldots,k}\|e_{t,k} - e_t\| \leq O_p(k^{1/q})$ by van der Vaart and Wellner (1996, Lem2.2.2).

From the above and $\sup_{t=1,\dots,n}\|e_{t,k}\| \leq \sup_{t=1,\dots,n}\|e_{t,k}-e_t\| + \sup_{t=1,\dots,n}\|e_t\|$:

$$\begin{aligned} &(A1) \leq O_p(\sqrt{k\log(n)}n^{1/q-1/2}), && (B1) \leq O_p(n^{1/q-1/2}),\\ &(C1) \leq \mathbb{1}_{t\geq k+1}O_p(n^{1/q-1/2}) + \mathbb{1}_{t\leq k}O_p(k^{1/q}), && (D1) \leq O_p(n^{1/q-1/2}), \end{aligned}$$

using $\sqrt{n}\sum_{j=k+1}\sup_{\theta\in\Theta}\|\Lambda_j(\theta)\|_{op}[\mathbb{E}(\|e_t\|^q)]^{1/q} = o(1)$ for the last inequality. Together, these imply: $\sup_{\theta\in\Theta}\sup_{t=k+1,\dots,n}\|y_t(\theta;\hat{\psi}_{nk}) - y_t(\theta;\psi_0)\| \leq O_p(\sqrt{k\log(n)}n^{1/q-1/2})$, which is $o_p(1)$ when $k^3/n = o(1)$ and $q > 8$. Also, $\sup_{\theta\in\Theta}\sup_{t=1,\dots,k}\|y_t(\theta;\hat{\psi}_{nk}) - y_t(\theta;\psi_0)\| \leq O_p(k^{1/q})$.

We apply the above results to evaluate $Q_n(\theta;\hat{\psi}_{nk}) - Q_n(\theta;\psi_0)$. Let $\langle y,\tilde{y}\rangle_{W_n} = y'W_n\tilde{y}$, we have: $Q_n(\theta;\hat{\psi}_{nk}) = Q_n(\theta;\psi_0) + (2/n)\sum_{t=1}^n\langle y_t(\theta;\hat{\psi}_{nk}) - y_t(\theta;\psi_0), \tilde{y}_t - y_t(\theta;\psi_0)\rangle_{W_n} + (2/n)\sum_{t=1}^n\|y_t(\theta;\hat{\psi}_{nk}) - y_t(\theta;\psi_0)\|^2_{W_n}$. By the Cauchy–Schwarz inequality and $\overline{\lambda}_{n,W} = \lambda_{\max}(W_n)$:

$$\begin{aligned} &\sup_{\theta\in\Theta}|Q_n(\theta;\hat{\psi}_{nk}) - Q_n(\theta;\psi_0)|\\ &\leq 2\overline{\lambda}_{n,W}\{\sup_{t=k+1,\dots,n}\sup_{\theta\in\Theta}\|y_t(\theta;\hat{\psi}_{nk}) - y_t(\theta;\psi_0)\| + O_p(k^{1+1/q}/n)\}\sup_{\theta\in\Theta}(1/n)\sum_{t=1}^n\|\tilde{y}_t - y_t(\theta;\psi_0)\|\\ &+ 2\overline{\lambda}_{n,W}\{\sup_{t=1,\dots,n}\sup_{\theta\in\Theta}\|y_t(\theta;\hat{\psi}_{nk}) - y_t(\theta;\psi_0)\|\}^2. \end{aligned}$$

Of the three terms after the inequality: the last term is $o_p(1)$ since $W_n \xrightarrow{p} W$, with W finite, which implies $\overline{\lambda}_{n,W} = O_p(1)$. The second term is $o_p(1)$ if $\sup_{\theta\in\Theta} 1/n\sum_{t=1}^n\|\tilde{y}_t - y_t(\theta;\psi_0)\| = O_p(1)$. This is the case because $\sum_{j=0}^\infty[\|\tilde{\Lambda}_j\|_{op} + \sup_{\theta\in\Theta}\|\Lambda_j(\theta)\|_{op}\sup_{\theta\in\Theta}\|P(\theta;\tilde{\Sigma})\|_{op}]\mathbb{E}(\|e_{t-j}\|) < \infty$. The first term is $o_p(1)$ because $k^{1+1/q}/n = o(k^3/n) = o(1)$, using $\sum_{t=1}^k\sup_{\theta\in\Theta}\|y_t(\theta;\hat{\psi}_{nk}) - y_t(\theta;\psi_0)\| \leq k\sup_{t=1,\dots,k}\sup_{\theta\in\Theta}\|y_t(\theta;\hat{\psi}_{nk}) - y_t(\theta;\psi_0)\| \leq O_p(k^{1+1/q})$. Altogether, this implies the result: $\sup_{\theta\in\Theta}|Q_n(\theta;\hat{\psi}_{nk}) - Q_n(\theta;\psi_0)| = o_p(1)$.

Step 2. Uniform Law of Large Numbers: $\sup_{\theta\in\Theta}|Q(\theta;\psi_0) - Q_n(\theta;\psi_0)| = o_p(1)$.
Using similar arguments as above: $|Q_n(\theta;\psi_0) - \frac{1}{n}\sum_{t=1}^n\|\tilde{y}_t - y_t(\theta;\psi_0)\|^2_W| \leq \|W_n - W\|_{op}O_p(1) = o_p(1)$. Lemma B1 implies that, for each $\theta\in\Theta$, $\tilde{y}_t - y_t(\theta;\psi_0)$ is near-epoch dependent (NED) in L_q-norm with size $-b$ for $b > 2$. Theorems 18.8, 18.9 in Davidson (2021) imply that $\|\tilde{y}_t - y_t(\theta;\psi_0)\|^2_W$ is NED in $L_{q/2}$-norm with size $-b$. Thus, using Assumption 2 and Davidson (2021, Th18.6), it is an L_2-mixingale of size $-\min(b, r/(r-2)[1/2 - 2/q])$ and a weak law of large numbers applies: $\frac{1}{n}\sum_{t=1}^n\|\tilde{y}_t - y_t(\theta;\psi_0)\|^2_W \xrightarrow{p} \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^n\mathbb{E}[\|\tilde{y}_t - y_t(\theta;\psi_0)\|^2_W] = Q(\theta;\psi_0)$, which holds pointwise in θ. Take any two $\theta_1,\theta_2\in\Theta$, apply the mean-value theorem:

$$Q_n(\theta_1;\psi_0) - Q_n(\theta_2;\psi_0) = -(2/n)\sum_{t=1}^n(\tilde{y}_t - y_t(\tilde{\theta};\psi_0))'W\partial_\theta y_t(\tilde{\theta};\psi_0)[\theta_1 - \theta_2] + o_p(1),$$

where the $o_p(1)$ is due to $||W_n - W|| = o_p(1)$ and is uniform in θ as above. We have:

$$\mathbb{E}[\sup_{\theta\in\Theta} \|\tilde{y}_t - y_t(\tilde{\theta};\psi_0))' W \partial_\theta y_t(\tilde{\theta};\psi_0)\|_W] \leq (\mathbb{E}[\sup_{\theta\in\Theta} \|\tilde{y}_t - y_t(\theta;\psi_0)\|_W^2])^{1/2} (\mathbb{E}[\sup_{\theta\in\Theta} \|\partial_\theta y_t(\theta;\psi_0)\|_W^2])^{1/2},$$

which is bounded under Assumption 3. Then, uniformly in θ_1, θ_2: $|Q_n(\theta_1;\psi_0) - Q_n(\theta_2;\psi_0)| \leq O_p(1)\|\theta_1 - \theta_2\| + o_p(1)$. This implies stochastic equicontinuity and the uniform langle.

Step 3. Consistency. The objective $Q(\cdot;\psi_0)$ is continuous is θ and uniquely minimized at $\theta = \theta_0$. $Q_n(\cdot;\hat{\psi}_{nk})$ converges uniformly to $Q(\cdot;\psi_0)$ in probability. By standard arguments, this implies consistency: $\hat{\theta}_n \xrightarrow{p} \theta_0$. □

Proof of Theorem 2. By minimization, θ_0 and $\hat{\theta}_n$ satisfy:

$$\mathbb{E}\left(\partial_\theta y_t(\theta_0;\psi_0)' W[\tilde{y}_t - y_t(\theta_0;\psi_0)]\right) = 0, \quad (1/n)\sum\nolimits_{t=1}^n \partial_\theta y_t(\hat{\theta}_n;\hat{\psi}_{nk})' W_n[\tilde{y}_t - y_t(\hat{\theta}_n;\hat{\psi}_{nk})] = 0.$$

The sample score can be decomposed into:

$$\begin{aligned} 0 &= (1/n)\sum\nolimits_{t=1}^n \partial_\theta y_t(\hat{\theta}_n;\hat{\psi}_{nk})' W_n[\tilde{y}_t - y_t(\hat{\theta}_n;\hat{\psi}_{nk})] \\ &= (1/n)\sum\nolimits_{t=1}^n \partial_\theta y_t(\hat{\theta}_n;\hat{\psi}_{nk})' W_n[\tilde{y}_t - y_t(\theta_0;\hat{\psi}_{nk})] && \text{(A)} \\ &+ (1/n)\sum\nolimits_{t=1}^n \partial_\theta y_t(\hat{\theta}_n;\hat{\psi}_{nk})' W_n[y_t(\theta_0;\hat{\psi}_{nk}) - y_t(\hat{\theta}_n;\hat{\psi}_{nk})]. && \text{(B)} \end{aligned}$$

Define $\theta_n(\omega) = \omega\theta_0 + (1-\omega)\hat{\theta}_n$ for any $\omega \in [0,1]$. An integration and change of variable implies: $(B) = -\frac{1}{n}\sum_{t=1}^n \int_0^1 \partial_\theta y_t(\hat{\theta}_n;\hat{\psi}_{nk})' W_n \partial_\theta y_t(\theta_n(\omega);\hat{\psi}_{nk}) d\omega[\hat{\theta}_n - \theta_0]$. Likewise: $(A) = \frac{1}{n}\sum_{t=1}^n \partial_\theta y_t(\theta_0;\hat{\psi}_{nk})' W_n[\tilde{y}_t - y_t(\theta_0;\hat{\psi}_{nk})] + \frac{1}{n}\sum_{t=1}^n \int_0^1 ([\tilde{y}_t - y_t(\theta_0;\hat{\psi}_{nk})]' W_n \otimes I)\partial_\theta G_t(\theta_n(\omega);\hat{\psi}_{nk}) d\omega[\hat{\theta}_n - \theta_0]$. The expansions of (A) and (B) imply:

$$\begin{aligned} &\hat{\theta}_n - \theta_0 = M_n^{-1}\left((1/n)\sum\nolimits_{t=1}^n \partial_\theta y_t(\theta_0;\hat{\psi}_{nk})' W_n[\tilde{y}_t - y_t(\theta_0;\hat{\psi}_{nk})]\right), \\ &M_n = \frac{1}{n}\sum_{t=1}^n \int_0^1 \{\partial_\theta y_t(\hat{\theta}_n;\hat{\psi}_{nk})' W_n \partial_\theta y_t(\theta_n(\omega);\hat{\psi}_{nk}) + ([y_t(\theta_0;\hat{\psi}_{nk}) - \tilde{y}_t]' W_n \otimes I)\partial_\theta G_t(\theta_n(\omega);\hat{\psi}_{nk}) d\omega\}. \end{aligned}$$

Step 1. Consistency of M_n.

It was shown in the proof of Theorem 1 that $\sup_{t=1,\dots,n}\|y_t(\theta_0;\hat{\psi}_{nk}) - y_t(\theta_0;\psi_0)\| = o_p(1)$ and similar derivations yield $\sup_{t=1,\dots,n}\|\partial_\theta y_t(\theta_0;\hat{\psi}_{nk}) - \partial_\theta y_t(\theta_0;\psi_0)\| = o_p(1)$. The absolute summability of the third derivative (Assumption 3), $\sup_{t=k+1,\dots,n}\|\hat{e}_t - e_t\| = o_p(1)$, and $\|\tilde{\Sigma}_{nk} - \tilde{\Sigma}\| = O_p(n^{-1/2})$ imply $\sup_{\theta\in\Theta}\sup_{t=k+1,\dots,n}\|\partial_\theta G_t(\theta;\hat{\psi}_{nk}) - \partial_\theta G_t(\theta;\psi_0)\| = o_p(1)$. Also, $\sup_{\omega\in[0,1]}\|\theta_n(\omega) - \theta_0\| = o_p(1)$, by consistency of $\hat{\theta}_n$. The first $t = 1,\dots,k$ can be handled as in the proof of Theorem 1. Thus, $(1/n)\sum_{t=1}^{n}\int_0^1 \Big([\tilde{y}_t - y_t(\theta_0;\hat{\psi}_{nk})]'W_n \otimes I\Big)\,\partial_\theta G_t(\theta_n(\omega);\hat{\psi}_{nk})d\omega = (1/n)\sum_{t=1}^{n}\left([\tilde{y}_t - y_t(\theta_0;\psi_0)]'W \otimes I\right)\partial_\theta G_t(\theta_0;\psi_0) + o_p(1)$. Assumption 3 implies $\partial_\theta G_t(\theta_0;\psi_0)$ is NED in L_q-norm with size $-b \leq -2$. By Lemma B1, $\tilde{y}_t - y_t(\theta_0;\psi_0)$ and $\partial_\theta y_t(\theta_0;\psi_0)$ are also NED. Hence, $\left([\tilde{y}_t - y_t(\theta_0;\psi_0)]'W \otimes I\right)\partial_\theta G_t(\theta_0;\psi_0)$ is NED in $L_{q/2}$-norm with size $-b$. Therefore, a weak law of large numbers applies to the second term in the definition of M_n. Likewise, $(1/n)\sum_{t=1}^{n}\int_0^1 \partial_\theta y_t(\hat{\theta}_n;\hat{\psi}_{nk})'W_n\partial_\theta y_t(\theta_n(\omega);\hat{\psi}_{nk})d\omega = (1/n)\sum_{t=1}^{n}\partial_\theta y_t(\theta_0;\psi_0)'W\partial_\theta y_t(\theta_0;\psi_0) + o_p(1)$, an average of a NED process in $L_{q/2}$-norm with size $-b$. A weak law of large numbers applies to this term as well. Altogether: $M_n \xrightarrow{p} M$.

Step 2. Asymptotic Normality.

Let $\psi_k = (\tilde{\mu}', \text{vech}(\tilde{\Sigma})', \text{vec}(\Psi_1)', \dots, \text{vec}(\Psi_k)')'$; recall that the estimates are given by $\hat{\psi}_{nk} = (\tilde{\mu}_n', \text{vech}(\tilde{\Sigma}_{nk})', \text{vec}(\hat{\Psi}_1)', \dots, \text{vec}(\hat{\Psi}_k)')'$ with $\tilde{\mu}_n$ equal to the sample average. Let $u_t = y_t(\theta_0;\psi_0) - \tilde{y}_t$, and $u_{t,k} = y_t(\theta_0;\psi_k) - \tilde{y}_t$. Note that Lemma 3 implies that $\|\hat{\psi}_{nk} - \psi_k\|_\infty = O_p(\sqrt{\log(n)/n})$. An integration identity and a change of variable with respect to ψ imply:

$$
\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\partial_\theta y_t(\theta_0;\hat{\psi}_{nk})'W_n[\tilde{y}_t - y_t(\theta_0;\hat{\psi}_{nk})] \\
&= -\frac{1}{n}\sum_{t=1}^{n}\partial_\theta y_t(\theta_0;\psi_k)'W_n u_{t,k} && \text{(C)} \\
&\quad -\frac{1}{n}\sum_{t=1}^{n}\int_0^1 \left(u_{t,k}'W_n \otimes I\right)\partial_\psi G_t(\theta_0;\psi_{nk}(\omega))d\omega[\hat{\psi}_{nk} - \psi_k] && \text{(D)} \\
&\quad -\frac{1}{n}\sum_{t=1}^{n}\int_0^1 \partial_\theta y_t(\theta_0;\hat{\psi}_{nk})'W_n\partial_\psi y_t(\theta_0;\psi_{nk}(\omega))d\omega[\hat{\psi}_{nk} - \psi_k], && \text{(E)}
\end{aligned}
$$

where $\psi_{nk}(\omega) = \omega\hat{\psi}_{nk} + (1-\omega)\psi_k$ are intermediate values of $\hat{\psi}_{nk}$ and ψ_k.

Because $W_n = W + o_p(1)$ and W is invertible: $W_n = W(I_d + o_p(1))$. This implies: $(C) = -\left[\frac{1}{n}\sum_{t=1}^{n}\partial_\theta y_t(\theta_0;\psi_k)'W u_{t,k}\right](I_d + o_p(1))$. We now show:

$$
(C) = -(1/n)\sum_{t=1}^{n}\partial_\theta y_t(\theta_0;\psi_k)'W u_{t,k} + o_p(n^{-1/2}).
$$

For this, note that $\partial_\theta y_t(\theta_0;\psi_k)'$ and $u_{t,k}$ are NED in L_q-norm with size $-b$. Their product is NED in $L_{q/2}$-norm with size $-b$, with absolutely summable autocovariances, following from the same arguments as in the proof of Theorem 1. This means $\left[\frac{1}{n}\sum_{t=1}^{n}\partial_\theta y_t(\theta_0;\psi_k)'W u_{t,k}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\partial_\theta y_t(\theta_0;\psi_k)'W u_{t,k}\right] + O_p(n^{-1/2})$, using Chebyshev's inequality. Next, because $u_{t,k} - u_t = y_t(\theta_0;\psi_k) - y_t(\theta_0;\psi_0)$, we have: $(\mathbb{E}[\|u_{t,k} - u_t\|^2])^{1/2} \leq \sum_{j=0}^{\infty}\|\Lambda_j(\theta_0)\|_{op}\|P(\theta_0;\tilde{\Sigma})\|_\infty(\mathbb{E}[\|e_{t,k} - e_t\|^2])^{1/2}$ for $t = k+1,\dots,n$, where $(\mathbb{E}[\|e_{t,k} - e_t\|^2])^{1/2} \leq \sum_{j=k+1}^{\infty}\|\Psi_j\|_{op}(\mathbb{E}[\|\tilde{y}_t - \tilde{\mu}\|^2])^{1/2} = o(n^{-1/2})$, from the condition in Lemma 3. Similar derivations also imply $(\mathbb{E}[\|\partial_\theta y_t(\theta_0;\psi_k) - \partial_\theta y_t(\theta_0;\psi_0)\|^2])^{1/2} = o(n^{-1/2})$, for $t = k+1,\dots,n$. Apply the Cauchy-Schwarz inequality to find $\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\partial_\theta y_t(\theta_0;\psi_k)'Wu_{t,k}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\partial_\theta y_t(\theta_0;\psi_0)'Wu_t\right] + O(k/n) + o(n^{-1/2}) = o(n^{-1/2})$ since the moments are bounded uniformly for $t = 1,\dots,k$. Putting everything together: $(C) = -\frac{1}{n}\sum_{t=1}^{n}\partial_\theta y_t(\theta_0;\psi_k)'Wu_{t,k} + o_p(n^{-1/2})$. For notation, let $S_{1,t} = \partial_\theta y_t(\theta_0;\psi_k)'Wu_{t,k}$. Also, define the sample mean: $S_{1,n} = 1/n\sum_{t=1}^{n} S_{1,t}$.

Derivations for (D) and (E) have a similar outline, the following will consider (E). The derivation for (A1)-(C1) in the proof of Theorem 1 imply that $\sup_{t=k+1,\dots,n}\|\partial_\theta y_t(\theta_0;\hat\psi_{nk}) - \partial_\theta y_t(\theta_0;\psi_k)\| = O_p(\sqrt{k\log(n)}n^{1/q-1/2})$ which implies:[12]

$$(E) = -\frac{1}{n}\sum_{t=1}^{n}\partial_\theta y_t(\theta_0;\psi_k)'W_n\partial_\psi y_t(\theta_0;\psi_k)(\hat\psi_{nk}-\psi_k) + O_p(\max[\sqrt{k}\log(n)n^{1/q-1}, \frac{\sqrt{\log(n)}}{n^{3/2}}k^{1+1/q}]),$$

the last term is $o_p(n^{-1/2})$ because $q > 8$ and $k = o(n^{1/3})$.

The next step is to look at $\partial_\psi y_t(\theta_0;\psi_k)$ more closely in order to derive consistency for (E). For this, we will evaluate $\partial_{\tilde\mu} y_t(\theta_0;\psi_k)$, $\partial_{\phi_j} y_t(\theta_0;\psi_k)$, and $\partial_{\text{vech}(\tilde\Sigma)} y_t(\theta_0;\psi_k)$ separately.

Recall that $e_{t,k} = \tilde y_t - \tilde\mu - \sum_{j=1}^{k}\Psi_j(\tilde y_{t-j-1} - \tilde\mu)$ with $e_{t,k} = 0$ for $t \leq 0$ and $\tilde y_{t-j-1} - \tilde\mu = 0$ for $t-j-1 \leq 0$. The coupled sample is constructed as $y_t(\theta;\psi_k) = \mu(\theta) + \sum_{j=0}^{\infty}\Lambda_j(\theta)P(\theta;\tilde\Sigma)e_{t-j,k}$. Thus, the derivative with respect to $\tilde\mu$ is given by: $\partial_{\tilde\mu_n} e_{t,k} = I_d - \sum_{j=1}^{k}\Psi_j = I_d - \sum_{j=1}^{\infty}\Psi_j - \sum_{j=k+1}^{\infty}\Psi_j = I_d - \sum_{j=1}^{\infty}\Psi_j + o_p(n^{-1/2})$, uniformly in t, using the condition $\sqrt{n}\sum_{j=k+1}^{\infty}\Psi_j = o(1)$. The absolute summability of the Λ_j then implies that: $\partial_{\tilde\mu} y_t(\theta_0;\psi_k) = I + \sum_{j=0}^{\infty}\Lambda_j(\theta)P(\theta;\tilde\Sigma)(I_d - \sum_{\ell=1}^{\infty}\Psi_\ell) + o(n^{-1/2})$, uniformly in t, here using $\sqrt{n}\sum_{j=k+1}^{\infty}\sup_{\theta\in\Theta}\|\Lambda_j(\theta)\|_{op} = o(1)$. Let $\phi_j = \text{vec}(\Psi_j)$ for $j = 1,\dots$ and re-write: $e_{t,k} = \tilde y_t - \tilde\mu - \sum_{j=1}^{k}\left((\tilde y_{t-j-1} - \tilde\mu)'\otimes I\right)\phi_j$. The partial derivative is $\partial_{\phi_j} e_{t,k} = -(\tilde y_{t-j-1} - \tilde\mu)'\otimes I$ for $t-j-1 \in \{1,\dots,n\}$ and $\partial_{\phi_j} e_{t,k} = 0$ for $t-j-1 \leq 0$ and $j \in \{1,\dots,k\}$. Then, using $\sqrt{n}\sum_{\ell=k+1}^{\infty}\sup_{\theta\in\Theta}\|\Lambda_\ell(\theta)\|_{op} = o(1)$ and $\mathbb{E}(\|\tilde y_t\|^q) < \infty$, we get $\sup_{t=k+1,\dots,n}\|\partial_{\phi_j} y_t(\theta;\psi_k) + \sum_{\ell=0}^{\infty}(\tilde y_{t-\ell-j-1} - \tilde\mu)'\otimes[\Lambda_\ell(\theta)P(\theta;\tilde\Sigma)]\| = o_p(n^{1/q-1/2})$, here using the actual $\tilde y_{t-\ell-j-1}$ for all t, j, ℓ. Using similar derivations as for Theorem 1, $\sup_{t=1,\dots,k}\|\partial_{\phi_j} y_t(\theta;\psi_k) + \sum_{\ell=0}^{\infty}(\tilde y_{t-\ell-j-1} - \tilde\mu)'\otimes[\Lambda_\ell(\theta)P(\theta;\tilde\Sigma)]\| = o_p(n^{1/q-1/2}) + O_p(k^{1/q})$, since $e_{t,k}$ sets $\tilde y_{t-\ell-j-1} - \tilde\mu = 0$ for $t-\ell-j-1 \leq 0$ and e_t does not. Note that $P(\theta;\cdot)$ is twice continuously differentiable under the constant rank assumption for $\Sigma(\theta)$ and $\tilde\Sigma$ invertible. Similar to the above, $\sup_{t=k+1,\dots,n}\|\partial_{\text{vech}(\tilde\Sigma)} y_t(\theta;\psi_k) - \sum_{j=0}^{\infty}(e'_{t-j}\otimes\Lambda_j(\theta))\partial_{\text{vech}(\tilde\Sigma)}\text{vec}[P(\theta;\tilde\Sigma)]\| = o_p(n^{1/q-1/2})$ and

$$\sup_{t=1,\dots,k}\|\partial_{\text{vech}(\tilde\Sigma)} y_t(\theta;\psi_k) - \sum_{j=0}^{\infty}(e'_{t-j}\otimes\Lambda_j(\theta))\partial_{\text{vech}(\tilde\Sigma)}\text{vec}[P(\theta;\tilde\Sigma)]\| = o_p(n^{1/q-1/2}) + O_p(k^{1/q}).$$

[12] y_t is computed using $e_{t,k}$ so that the additional $k^{1/q}$-term due to $e_t - e_{t,k}$ for $t \leq k$ is negligible, as in the proof of Theorem 1.

Let $y_t(\theta;\psi_{k0})$ and $\partial_\psi y_t(\theta;\psi_{k0})$ be as in Lemma B1. Recall $\hat{\psi}_{nk}-\psi_k = O_p(\sqrt{\log(n)/n})$, then:

$$\frac{1}{n}\sum_{t=1}^{n}\partial_\theta y_t(\theta_0;\psi_k)'W\partial_\psi y_t(\theta_0;\psi_k)(\hat{\psi}_{nk}-\psi_k) = \frac{1}{n}\sum_{t=1}^{n}\partial_\psi y_t(\theta_0;\psi_k)'W\partial_\theta y_t(\theta_0;\psi_{k0})(\hat{\psi}_{nk}-\psi_k)+o_p(n^{-1/2}),$$

using $o_p(n^{1/q-1/2}) = o_p(1/\sqrt{\log(n)})$, $k^{1+1/q}\sqrt{\log(n)}/n = o(n^{-1/2})$ and similar derivations as in the proof of Theorem 1. Given the stated assumption, similar derivations to (B1)-(D1) in Theorem 1 for $\partial_\theta y_t(\theta;\psi_k)$ further imply:

$$\frac{1}{n}\sum_{t=1}^{n}\partial_\theta y_t(\theta_0;\psi_k)'W\partial_\psi y_t(\theta_0;\psi_k)(\hat{\psi}_{nk}-\psi_k) = \frac{1}{n}\sum_{t=1}^{n}\partial_\theta y_t(\theta_0;\psi_0)'W\partial_\psi y_t(\theta_0;\psi_{k0})(\hat{\psi}_{nk}-\psi_k)+o_p(n^{-1/2}).$$

The next step is to show: $\frac{1}{n}\sum_{t=1}^{n}\partial_\theta y_t(\theta_0;\psi_0)'W\partial_\psi y_t(\theta_0;\psi_{k0}) = \mathbb{E}[\partial_\theta y_t(\theta_0;\psi_0)'W\partial_\psi y_t(\theta_0;\psi_{k0})]+O_p(kn^{-1/2})$. The proof for this part is long as it involves non-standard arguments.

As shown in Lemma B1, $\partial_\theta y_t(\theta;\psi_0)$ and $\partial_\mu y_t(\theta;\psi_{k0})$, $\partial_{\psi_j} y_t(\theta;\psi_{k0})$, and $\partial_{\text{vec}(\tilde{\Sigma})} y_t(\theta;\psi_{k0})$ are NED in L_q-norm with size $-b$, for each $j = 1,\ldots$. Consequently, $\partial_\theta y_t(\theta;\psi_0)'W\partial_\mu y_t(\theta;\psi_{k0})$ and $\partial_\theta y_t(\theta;\psi_0)'W\partial_{\text{vec}(\tilde{\Sigma})} y_t(\theta;\psi_{k0})$ are NED in L_p-norm with size $-b$ (Davidson, 2021, Th18.9), with $b \geq 2$ for any $p < q/2$. They are $L_{q/2}$-mixingales of size $-\min(b, a(1/p-2/q)) = 1/2$. Pick $p = 2$ and $q \geq 8$ implies $\min(b, a(1/p-2/q)) \geq \min(2, 2(1/2-2/8)) = 1/2$ so that the autocovariances are absolutely summable (Davidson, 2021, Th17.16). Therefore, a weak law of large numbers applies to the sample average of $\partial_\theta y_t(\theta;\psi_0)'W\partial_\mu y_t(\theta;\psi_{k0})$ and $\partial_\theta y_t(\theta;\psi_0)'W\partial_{\text{vec}(\tilde{\Sigma})} y_t(\theta;\psi_{k0})$, using Chebyshev's inequality.

We also need to derive a weak law of large numbers for $\partial_\theta y_t(\theta;\psi_0)'W\partial_{\phi_j} y_t(\theta;\psi_{k0})$, with $j = 1,\ldots,k$. This is nonstandard because k diverges to infinity. Recall from Lemma B1 that for all $j = 1,\ldots,k$, $\partial_{\phi_j} y_t(\theta_0;\psi_{k0}) = -\sum_{\ell=0}^{\infty}[\tilde{y}_{t-\ell-j-1}-\tilde{\mu}]'\otimes\Lambda_j(\theta_0)P(\theta_0;\tilde{\Sigma})$, where $[\tilde{y}_{t-\ell-j-1}-\tilde{\mu}] \in \mathcal{F}_{t-j-1}$. Apply the law of iterated expectations:

$$\begin{aligned}&\mathbb{E}[\partial_\theta y_t(\theta_0;\psi_k)'W\partial_{\phi_j} y_t(\theta_0;\psi_{k0})] = \mathbb{E}\left[\mathbb{E}\{\partial_\theta y_t(\theta_0;\psi_k)'W\partial_{\phi_j} y_t(\theta_0;\psi_{k0})|\mathcal{F}_{t-j-1}\}\right]\\ &= -\mathbb{E}\left[\{\sum\nolimits_{\ell=j+1}^{\infty}(e'_{t-\ell}\otimes I)\partial_\theta\text{vec}[\Lambda_\ell(\theta_0)P(\theta_0;\tilde{\Sigma})]\}'W\{\sum\nolimits_{\ell=0}^{\infty}[\tilde{y}_{t-\ell-j-1}-\tilde{\mu}]'\otimes\Lambda_j(\theta_0)P(\theta_0;\tilde{\Sigma})\}\right].\end{aligned}$$

Takes norms on both sides, and apply the Cauchy-Schwarz inequality to find:

$$\|\mathbb{E}[\partial_\theta y_t(\theta_0;\psi_k)'W\partial_{\phi_j} y_t(\theta_0;\psi_{k0})]\| \leq \|e_t\|_2\|\tilde{y}_t\|_2\|P(\theta_0;\tilde{\Sigma})\|_\infty\sum\nolimits_{\ell=0}^{\infty}\|\tilde{\Lambda}_\ell(\theta_0)P(\theta_0;\tilde{\Sigma})\|_\infty C(j+1)^{-b-\varepsilon},$$

using Assumption 3, where $\|e_t\|_2 = (\mathbb{E}[\|e_t\|^2])^{1/2}$ and $b+\varepsilon > 2$. Take $1 \leq j \leq k$ and $s \geq 1$, let $w_t = \text{vec}\{\partial_\theta y_t(\theta;\psi_0)' W \partial_{\phi_j} y_t(\theta;\psi_{k0})\}$. From the bound above, we have:

$$\begin{aligned}\|\mathbb{E}[(w_{t+s} - \mathbb{E}[w_{t+s}])w_t']\| &= \|\mathbb{E}[\mathbb{E}(w_{t+s} - \mathbb{E}[w_{t+s}]|\mathcal{F}_t)w_t']\| \\ &\leq \|e_t\|_4 \|\tilde{y}_t\|_4 \|P(\theta_0;\tilde{\Sigma})\|_\infty \sum\nolimits_{\ell=0}^{\infty} \|\tilde{\Lambda}_\ell(\theta_0)P(\theta_0;\tilde{\Sigma})\|_\infty C(j+1)^{-b-\varepsilon}\|w_t\|_2,\end{aligned}$$

using the same approach, where $\|w_t\|_2$ can be bounded independently of j. Use this bound for $s = 0,\dots,8j$. Note that $(8j+1)(j+1)^{-(b+\varepsilon)} = o(1)$ as $j \to \infty$, so that the contribution of this bound is uniformly finite in $j \geq 1$.

Take $s > 8j$, we now bound $\|\mathbb{E}(w_{t+s} - \mathbb{E}[w_{t+s}]|\mathcal{F}_t)\|_p$ for some $p \geq 2$. Using a similar approach as Davidson (2021, p374), divide it into two terms: $\|\mathbb{E}(w_{t+s} - \mathbb{E}[w_{t+s}]|\mathcal{F}_t)\|_p \leq \|\mathbb{E}(w_{t+s} - \mathbb{E}[w_{t+s}|\mathcal{F}_{t+s-m}^{t+s+m}]|\mathcal{F}_t)\|_p + \|\mathbb{E}(\mathbb{E}[w_{t+s}] - \mathbb{E}[w_{t+s}|\mathcal{F}_{t+s-m}^{t+s+m}]|\mathcal{F}_t)\|_p$ for any $1 \leq m \leq s$. Since $\mathbb{E}[w_{t+s}|\mathcal{F}_{t+s-m}^{t+s+m}]$ is α-mixing with size $-a$, $\|\mathbb{E}(\mathbb{E}[w_{t+s}] - \mathbb{E}[w_{t+s}|\mathcal{F}_{t+s-m}^{t+s+m}]|\mathcal{F}_t)\|_p \leq (1+s-m)^{-a(1/p-1/r)}\|w_t\|_r$, for $r > p$. This is absolutely summable for $m = [s/2]$, over $s = 1,\dots$ for any $1/p - 1/r \geq 1/2$ since $a > 2$ (Assumption 2). Next, $\|\partial_\theta y_{t+s}(\theta;\psi_0) - \mathbb{E}[\partial_\theta y_{t+s}(\theta;\psi_0)|\mathcal{F}_{t+s-m}^{t+s+m}]\|_{2p} \leq \|\sum_{\ell=m+1}^{\infty}(I \otimes e_{t-\ell})\partial_\theta \text{vec}[P(\theta;\tilde{\Sigma})\Lambda_\ell(\theta)]\|_{2p} \leq (1+m)^{-(b+\varepsilon)}C[\|P(\theta;\tilde{\Sigma})\|_\infty + \|\partial_\theta \text{vec}[P(\theta;\tilde{\Sigma})]\|_\infty]\|e_t\|_{2p}$. This is absolutely summable over $s \geq 1$ with $m = [s/2]$. Using similar arguments, $\|\tilde{y}_{t+s-j-\ell} - \mathbb{E}[\tilde{y}_{t+s-j-\ell}|\mathcal{F}_{t+s-m}^{t+s+m}]\|_{2p} \leq C(1+[m-j-\ell]^+)^{-(b+\varepsilon)}\|e_t\|_{2p}$, where $[m-j-\ell]^+ = \max(0, m-j-\ell) \geq \max(0, 3/4m-\ell)$ for $s > 8j$ and $m = [s/2]$. Using the formula for $\partial_{\phi_j} y_t(\theta;\psi_{k0})$ this yields: $\|\partial_{\phi_j} y_t(\theta;\psi_{k0}) - \mathbb{E}[\partial_{\phi_j} y_t(\theta;\psi_{k0})|\mathcal{F}_{t+s-m}^{t+s+m}]\|_{2p} \leq \sum_{\ell=0}^{\infty} C^2(1+\ell)^{-(b+\varepsilon)}(1+[3/4m-\ell]^+)^{-(b+\varepsilon)}\|e_t\|_{2p}$. For $\ell = 0,\dots,[m/2] = [s/4]$, the partial sum is bounded by: $\sum_{\ell=0}^{\infty} C^2(1+\ell)^{-(b+\varepsilon)}(1+[1/4m]^+)^{-(b+\varepsilon)}\|e_t\|_{2p}$ which is absolutely summable in s. For $\ell > [m/2]$, the remainder is bounded by: $(1+[m/2])^{-(b+\varepsilon)/2}\sum_{\ell=0}^{\infty} C^2(1+\ell)^{-(b+\varepsilon)/2}\|e_t\|_{2p}$, also absolutely summable because $(b+\varepsilon)/2 > 1$. Then, we can bound $\|\mathbb{E}[w_{t+s} - \mathbb{E}(w_{t+s})|\mathcal{F}_t]\|_p \leq C_p(1+s)^{-(b+\varepsilon)/2}$, where C_p depends on p via $\|e_t\|_{2p}$. This is absolutely summable.

Then, use Hölder's inequality with $1 = 1/p + 1/r$ such that $\|e_t\|_{2p}$ and $\|w_t\|_r$ are finite, and $s > 8j$: $\|\mathbb{E}[(w_{t+s} - \mathbb{E}[w_{t+s}])w_t']\| \leq C_p(1+s)^{-(b+\varepsilon)/2}\|w_t\|_r$, which is absolutely summable. The sum over $s \leq 8j$ is of order $j(1+j)^{-(b+\varepsilon)}$, which is bounded. Thus the autocovariances are absolutely summable and the resulting sum is bounded uniformly in j. Chebyshev's inequality can be applied uniformly in j and yields the rate: $\sup_{j=1,\dots,k}\left\|(1/n)\sum_{t=1}^n \partial_\theta y_t(\theta;\psi_0)'W\partial_{\psi_j}y_t(\theta;\psi_{k0}) - \mathbb{E}[\partial_\theta y_t(\theta;\psi_0)'W\partial_{\psi_j}y_t(\theta;\psi_{k0})]\right\| \leq O_p(kn^{-1/2})$. As $\|\hat{\psi}_{nk} - \psi_k\| \leq O_p(\sqrt{\log(n)/n})$ and $O_p(kn^{-1/2})O_p(\sqrt{\log(n)/n}) = o_p(n^{-1/2})$, this implies:

$$\frac{1}{n}\sum_{t=1}^{n}\partial_\theta y_t(\theta_0;\psi_k)'W\partial_\psi y_t(\theta_0;\psi_k)(\hat{\psi}_{nk}-\psi_k) = \mathbb{E}\left[\partial_\theta y_t(\theta_0;\psi_0)'W\partial_\psi y_t(\theta_0;\psi_{k0})\right](\hat{\psi}_{nk}-\psi_k) + o_p(n^{-1/2}).$$

This completes the proof for (E). Similar derivations can be applied to (D). Combining them:

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\Big[\left(u'_{t,k}W_n\otimes I\right)\partial_\psi G_t(\theta_0;\bar{\psi}_{nk})+\partial_\theta y_t(\theta_0;\hat{\psi}_{nk})'W_n\partial_\psi y_t(\theta_0;\tilde{\psi}_{nk})\Big]\,[\hat{\psi}_{nk}-\psi_k]\\
&\quad=\mathbb{E}\left[\left(u'_tW\otimes I\right)\partial_\psi G_t(\theta_0;\psi_{k0})+\partial_\theta y_t(\theta_0;\psi_k)'W\partial_\psi y_t(\theta_0;\psi_{k0})\right](\hat{\psi}_{nk}-\psi_k)+o_p(n^{-1/2}),
\end{aligned}$$

where the expected matrix has full rank for $k\geq\underline{k}$, sufficiently large, by assumption.

The next step is to derive a CLT for the leading term on the right hand side of the preceding expression, where $\hat{\psi}_{nk}-\psi_k$ has increasing dimension. Note:

$$\begin{aligned}
&\mathbb{E}[\partial_\theta y_t(\theta_0;\psi_k)'W\partial_\psi y_t(\theta_0;\psi_{k0})](\hat{\psi}_{nk}-\psi_k)\\
=&\mathbb{E}[\partial_\theta y_t(\theta_0;\psi_k)'W\partial_{\tilde{\mu}}y_t(\theta_0;\psi_{k0})](\tilde{y}_n-\tilde{\mu})+\mathbb{E}[\partial_\theta y_t(\theta_0;\psi_k)'W\partial_{\text{vech}\tilde{\Sigma}}y_t(\theta_0;\psi_{k0})]\text{vech}(\tilde{\Sigma}_n-\tilde{\Sigma})\\
+&\mathbb{E}[\partial_\theta y_t(\theta_0;\psi_k)'W\partial_{\phi_1}y_t(\theta_0;\psi_{k0})](\hat{\phi}_1-\phi_1)+\cdots+\mathbb{E}[\partial_\theta y_t(\theta_0;\psi_k)'W\partial_{\phi_k}y_t(\theta_0;\psi_{k0})](\hat{\phi}_k-\phi_k).
\end{aligned}$$

The partial derivatives $\mathbb{E}[\partial_\theta y_t(\theta_0;\psi_k)'W\partial_{\phi_j}y_t(\theta_0;\psi_{k0})]$ are absolutely summable, as seen from the upper bound derived above. Similar arguments imply that the same holds for $\mathbb{E}[(u'_tW\otimes I)\partial_{\phi_j}G_t(\theta_0;\psi_{k0})]$ since $\partial_{\phi_j}G_t(\theta_0;\psi_{k0})\in\mathcal{F}_{t-j}$ as well, $\mathbb{E}(u_t-\mathbb{E}[u_t]|\mathcal{F}_{t-j})$ are absolutely summable, and $\mathbb{E}[\partial_{\phi_j}G_t(\theta_0;\psi_{k0})]=0$. As a result: $\|D_{\theta,\psi}(k)\|\leq c_2<\infty$ for some constant $c_2>0$. Together with the singular value condition, this implies that: $0<c_1\leq\sigma_{\min}[D_{\theta,\psi}(k)]\leq\sigma_{\max}[D_{\theta,\psi}(k)]\leq c_2<\infty$, for all $k\geq\underline{k}\geq 1$; that is, for k large enough $D_{\theta,\psi}(k)$ is a matrix with d_θ rows, each with norm bounded away from zero and infinity.

The following relies on Lewis and Reinsel (1985, Th2), which holds when the innovations are non-iid, under the stated assumptions.[13] Let $\tilde{Y}_{t-1,k}=(\tilde{y}_{t-1}-\tilde{\mu},\ldots,\tilde{y}_{t-k}-\tilde{\mu})'$, with mean zero, and $\Gamma_k=\mathbb{E}[\tilde{Y}_{t,k}\tilde{Y}'_{t,k}]$ is the autocovariance matrix. For any sequence of vectors $l(k)$, such that $0<c_1\leq\|l(k)\|\leq c_2<\infty$:

$$\sqrt{n}l(k)'[(\hat{\phi}_1,\ldots,\hat{\phi}_k)-(\phi_1,\ldots,\phi_k)]'=l(k)'\text{vec}\left[(1/\sqrt{n})\sum\nolimits_{t=1}^{n}e_t\tilde{Y}'_{t-1,k}\Gamma_k^{-1}\right]+o_p(1),$$

where the sum is taken from $t=1$ rather than $t=1+k$, the difference being $\sqrt{n}$-negligible. Note that Γ_k^{-1} is finite and uniformly bounded in $k\geq 1$ (Lewis and Reinsel, 1985).

The next step is to derive the dependence properties of the right-hand-side of the last display. Because e_t is the Wold innovation, it is a martingale difference sequence. A central limit Theorem for $\hat{\phi}_{nk}$ is usually derived from this property. However, here $\hat{\psi}_{nk}$ also involves $\tilde{y}_t$ and $e_te'_t$, which are not martingale differences. Using the notation of Lewis and

[13] See e.g. Hannan and Deistler (2012, Ch7), Kuersteiner (2005, Th2.6), Gonçalves and Kilian (2007, LemA.6) and Lütkepohl (2005, Ch15).

Reinsel (1985), let $\Gamma(j) = \mathbb{E}[(\tilde{y}_t - \tilde{\mu})(\tilde{y}_{t-j} - \tilde{\mu})']$. The first set of rows of Γ_k is given by $\Gamma(0), \Gamma(1), \ldots, \Gamma(k-1)$, the second row consists of $\Gamma(1)', \Gamma(0), \Gamma(1), \Gamma(2), \ldots, \Gamma(k-2)$, etc. Using the white noise property of the innotations, we have:

$$\begin{aligned}\|\Gamma(j)\| = \|\mathbb{E}\left([\sum\nolimits_{\ell=j}^{\infty} \tilde{\Lambda}_j e_{t-\ell}][\sum\nolimits_{\ell=0}^{\infty} \tilde{\Lambda}_j e_{t-\ell-j}]'\right)\| &\leq \sum\nolimits_{\ell=j}^{\infty} \|\tilde{\Lambda}_j\|_{op} \|e_t\|^2 \|\tilde{y}_t - \tilde{\mu}\|_2 \\ &\leq C\|e_t\|^2 \|\tilde{y}_t - \tilde{\mu}\|_2 (j-1)^{-(b+\varepsilon)}.\end{aligned}$$

This implies that $\|\Gamma(j)\|_\infty \leq C_1 j^{-(b+\varepsilon)}$, for some constant C_1 and the elements of Γ_k decay polynomially away from the diagonal, i.e. $(\Gamma_k)_{\ell,s} \leq C_2(1 + |\ell - s|)^{-(b+\varepsilon)}$ for some constant C_2 and for all $1 \leq \ell, s \leq k$, for all $k \geq 1$. Proposition 3 and the "window lemma" in Jaffard (1990) then imply:[14] $(\Gamma_k^{-1})_{\ell,s} \leq C_3(1 + |\ell - s|)^{-(b+\varepsilon)}$, for some other constant C_3, with the same polynomial rate of decay. Denote by $(\Gamma_k^{-1})_j$, $1 \leq j \leq k$, the j-th block of columns of the matrix Γ_k^{-1}; that is for $j = 1$, $(\Gamma_k^{-1})_1$ contains columns 1 to $\dim(\tilde{y}_t)$. Let $l_j(k)$ denote the coefficients of $l(k)$ which are multiplied by $\hat{\phi}_j - \phi_j$ above. Then, for each $j = 1, \ldots, k$:

$$\sqrt{n} l_j(k)'(\hat{\phi}_j - \phi_j) = l_j(k)' \text{vec}\left[(1/\sqrt{n}) \sum\nolimits_{t=1}^{n} e_t \tilde{Y}'_{t-1,k} (\Gamma_k^{-1})_j\right] + o_p(1),$$

where $\|l_j(k)\| \leq (1 + j)^{-(b+\varepsilon)}$, up to some constant, as shown above. The next step is to investigate the dependence properties of each $e_t \tilde{Y}'_{t-1,k}(\Gamma_k^{-1})_j$, $1 \leq j \leq k$. Recall that $\tilde{Y}_{t-1,k} = (\tilde{y}_{t-1} - \tilde{\mu}, \ldots, \tilde{y}_{t-k} - \tilde{\mu})$, $\mathcal{F}_{t-m}^{t+m}$ is the filtration generated from $(e_{t-m}, \ldots, e_{t+m})$. Because $e_t \in \mathcal{F}_{t-m}^{t+m}$, we can write using Hölder's inequality:

$$\|e_t \tilde{y}_{t-j} - \mathbb{E}(e_t \tilde{y}_{t-j} | \mathcal{F}_{t-m}^{t+m})\|_{p/2} \leq \|e_t\|_p \|\tilde{y}_{t-j} - \mathbb{E}(\tilde{y}_{t-j} | \mathcal{F}_{t-m}^{t+m})\|_p \leq d\|e_t\|_p \nu((m-j)^+),$$

where d and ν are the NED coefficients satisfying: $\|\tilde{y}_{t-j} - \mathbb{E}(\tilde{y}_t | \mathcal{F}_{t-m}^{t+m})\|_p \leq d\nu(m)$, for $m \geq 0$ and $(m-j)^+ = \max(m-j, 0)$. The NED coefficients are derived in the proof of Lemma B1. Their decay factor satisfies $\nu(m) \leq (1+m)^{-(b+\varepsilon)}$. Together, we get:

$$\begin{aligned}&\left\| l(k)' \text{vec}\left[e_t \tilde{Y}'_{t-1,k}(\Gamma_k^{-1}) - \mathbb{E}[e_t \tilde{Y}'_{t-1,k}(\Gamma_k^{-1}) | \mathcal{F}_{t-m}^{t+m}]\right] \right\|_{p/2} \\ \leq &\sum\nolimits_{j=1}^{k} \left\| l_j(k)' \text{vec}\left[e_t \tilde{Y}'_{t-1,k}(\Gamma_k^{-1})_j - \mathbb{E}[e_t \tilde{Y}'_{t-1,k}(\Gamma_k^{-1})_j | \mathcal{F}_{t-m}^{t+m}]\right] \right\|_{p/2} \\ \leq &C_5 \sum\nolimits_{j=1}^{k} \sum\nolimits_{\ell=1}^{k} (1+j)^{-(b+\varepsilon)} (1 + |\ell - j|)^{-(b+\varepsilon)} (1 + (m-\ell)^+)^{-(b+\varepsilon)},\end{aligned}$$

[14] His result applies to the infinite-dimensional operator Γ_∞ directly. Extend Γ_k to a banded infinite-dimensional operator $\Gamma_{k,\infty}$ equal to Γ_k on the diagonal everywhere and 0 elsewhere. $\Gamma_{k,\infty}$ is a "window" approximation of Γ_∞, with a window of size k. Elements of Γ_∞^{-1} and $\Gamma_{k,\infty}^{-1} - \Gamma_\infty^{-1}$ have polynomial decay. The operators $\Gamma_{k,\infty}^{-1}$ are equal Γ_k^{-1} on the restricted finite-dimensional subspace where Γ_k is defined.

for some constant C_5. We now split the double sum into three terms and study them separately. The first term is:

$$\begin{aligned}&\sum\nolimits_{j=1}^{k}\sum\nolimits_{\ell=1}^{[m/2]}(1+j)^{-(b+\varepsilon)}(1+|\ell-j|)^{-(b+\varepsilon)}(1+(m-\ell)^{+})^{-(b+\varepsilon)}\\ \leq&(1+[m/2])^{-(b+\varepsilon)}\sum\nolimits_{j=1}^{\infty}\sum\nolimits_{\ell=-\infty}^{\infty}(1+j)^{-(b+\varepsilon)}(1+|\ell|)^{-(b+\varepsilon)},\end{aligned}$$

where the right-hand side is summable since $b+\varepsilon>2$. The second term is:

$$\begin{aligned}&\sum\nolimits_{j=1}^{[m/4]}\sum\nolimits_{\ell=[m/2]+1}^{k}(1+j)^{-(b+\varepsilon)}(1+|\ell-j|)^{-(b+\varepsilon)}(1+(m-\ell)^{+})^{-(b+\varepsilon)}\\ \leq&(1+[m/4])^{-(b+\varepsilon)/2}\sum\nolimits_{j=1}^{\infty}\sum\nolimits_{\ell=-\infty}^{\infty}(1+j)^{-(b+\varepsilon)}(1+|\ell|)^{-(b+\varepsilon)/2},\end{aligned}$$

where the right-hand side is summable since $(b+\varepsilon)/2>1$. The third term is:

$$\begin{aligned}&\sum\nolimits_{j=1+[m/4]}^{k}\sum\nolimits_{\ell=[m/2]+1}^{k}(1+j)^{-(b+\varepsilon)}(1+|\ell-j|)^{-(b+\varepsilon)}(1+(m-\ell)^{+})^{-(b+\varepsilon)}\\ &\leq\sum\nolimits_{j=1+[m/4]}^{\infty}\sum\nolimits_{\ell=-\infty}^{\infty}(1+j)^{-(b+\varepsilon)}(1+|\ell|)^{-(b+\varepsilon)}\\ &\leq\int_{[m/4]}^{\infty}(1+x)^{-(b+\varepsilon)}dx\sum\nolimits_{\ell=-\infty}^{\infty}(1+|\ell|)^{-(b+\varepsilon)}\leq\frac{1}{b+\varepsilon-1}(1+[m/4])^{1-(b+\varepsilon)}\sum\nolimits_{\ell=-\infty}^{\infty}(1+|\ell|)^{-(b+\varepsilon)}.\end{aligned}$$

From these, we get that the sum of the three terms is bounded above by:

$$\left\|l(k)'\text{vec}\left[e_t\tilde{Y}_{t-1,k}'(\Gamma_k^{-1})-\mathbb{E}[e_t\tilde{Y}_{t-1,k}'(\Gamma_k^{-1})|\mathcal{F}_{t-m}^{t+m}]\right]\right\|_{p/2}\leq C_6(1+m)^{-(b+\varepsilon)/2},$$

since $b\geq 2$ implies $1-b\leq -b/2$, it is NED in $L_{p/2}$-norm with size $-b/2\leq -1$. Similarly,

$$-(1/n)\sum\nolimits_{t=1}^{n}D_{\theta,\psi}(k)((\tilde{y}_t-\tilde{\mu})',\text{vec}[e_t\tilde{Y}_{t-1,k}'\Gamma_k^{-1}]',\text{vech}[e_te_t'-\tilde{\Sigma}]')'=(D)+(E)+o_p(n^{-1/2})$$

is NED in $L_{q/2}$-norm with size $-\min(b-1,a)$ where a is the mixing size of e_t in Assumption 2. Altogether, we get that:

$$\sqrt{n}(\hat{\theta}_n-\theta_0)=-(1/\sqrt{n})\sum\nolimits_{t=1}^{n}M^{-1}[\partial_\theta y_t(\theta_0;\psi_k)'Wu_{t,k}+D_{\theta,\psi}(k)Z_{t,k}]+o_p(1),$$

where the leading term is NED in $L_{q/2}$-norm, $q/2>2$, with size $-\min(b/2,a)$, $\min(b/2,a)>1/2$ on e_t which is strongly-mixing with size $-a$, $a>r/(r-2)$ for $r>4$. The NED derivations imply that $V_{n,k}=O(1)$ is bounded. With the normalization $V_{n,k}^{-1/2}$, the conditions for Corollary 4.2 in Wooldridge and White (1988) hold and: $\sqrt{n}V_{n,k}^{-1/2}(\hat{\theta}_n-\theta_0)\xrightarrow{d}\mathcal{N}(0,I)$. □

Supplement to "Fitting Dynamically Misspecified Models: An Optimal Transportation Approach"

Jean-Jacques Forneron[*] Zhongjun Qu[†]

July 18, 2025

This Supplemental Material consists of Appendices D, E, F, G, H, I to the main text.

[*] Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215 USA. Email: jjmf@bu.edu, Website: http://jjforneron.com

[†] Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215 USA. Email: qu@bu.edu.

Appendix D Additional details for Algorithm 1

The following describes an implementation of Algorithm 1 which combines the Bootstrap filter (Chopin and Papaspiliopoulos, 2020, Ch10.3.1) with optimal transport.

Algorithm 3 An Implementation of the Optimal Transport Filter

```
1:  procedure OTF
    Inputs: 1) Sample: $\tilde{y}_1, \dots, \tilde{y}_n$, predictive distribution $\tilde{p}(\tilde{y}_t|\tilde{y}_{t-1}, \dots)$.
            2) Model: $p(y_t, z_t|z_{t-1}; \theta)$. Initial beliefs $z_0 \sim p_{0|0}(z_0)$. Number of draws $B$.
    Outputs: 1) Mapped data $y_1, \dots, y_n$.
             2) Filtered states $z_{t|t} \sim p(z_t|y_t, \dots, y_1)$.
    Initialize: For $b = 1, \dots, B$, do 1) draw $z_0^b \sim p_{0|0}(z_0; \theta)$, 2) weight $w_0^b = 1/B$.
2:      for $t \in \{1, \dots, n\}$ do
3:          if $\text{ESS}(w_{t-1}^{1:B}) < \text{ESS}_{\min}$ then
4:              resample $(z_{t-1}^b)_{b=1,\dots,B}$ with replacement and weight $w_{t-1}^b$, set $w_{t-1}^b = 1/B$.
5:          end if
6:          Predict: Draw $(y_t^b, z_t^b) \sim p(y_t, z_t|z_{t-1}^b)$ and $\tilde{y}_t^b \sim \tilde{p}(\tilde{y}_t|\tilde{y}_{t-1})$.
7:          Transport plan: Find a joint distribution $\pi_{t|t-1} = (p_{i,j})_{1 \leq i,j \leq B}$ which solves

                $$\min_{p_{i,j}} \left\{ \sum_{i,j} p_{i,j} \|y_t^i - \tilde{y}_t^j\|^2 - \varepsilon \sum_{i,j} p_{i,j} \log(p_{i,j}) \right\}$$

    where $0 \leq p_{i,j} \leq 1$, $\sum_i p_{i,j} = 1/B$, $\sum_j p_{i,j} = w_{t-1}^i$.
8:          Barycentric projection: find $j^\star$ such that $\|\tilde{y}_t - \tilde{y}_t^{j^\star}\|$ is minimal. Compute $p_i =$
    $p_{i,j^\star}/[\sum_{i=1}^B p_{i,j^\star}]$. Compute $y_t = \sum_{i=1}^B p_i y_t^i$.
9:          Update: Set $w_t^b = w_{t-1}^b p(y_t|z_t^b)$. Normalize $w_t^b = w_t^b/[\sum_i w_t^i]$.
10:     end for
11: end procedure
```

Algorithm 3 combine particle filtering steps with an additional draw $\tilde{y}_t^j$ and an optimal transport projection (steps 7,8) to construct the model-consistent sample $(y_t)_{t=1,\dots,n}$. Step 7 can be approximated using the Sinkhorn algorithm (Peyré and Cuturi, 2019, Ch4).

Appendix E Proofs for Lemmas and Preliminary Results

Proof of Lemma 1. Condition i) implies Assumption 2 (i). For Assumption 2 (ii), $\mu(\cdot)$ is continuous on Θ compact, hence bounded. Then, $\Lambda_j(\cdot) = A(\cdot)C(\cdot)^j K(\cdot)$ is the product of $A(\cdot)C(\cdot)$, continuous, with $K(\cdot)$. The mapping $\theta \to \Sigma(\theta)$ is continuously differentiable, Σ has constant rank, so the pseudo-inverse $\Sigma(\cdot)^\dagger$ is continuously differentiable with bounded derivative (Magnus and Neudecker, 2019, Prop8.2). This implies $V(\cdot)$ is continuously differentiable by the Implicit Function Theorem, and $K(\cdot)$ is continuously differentiable as a product of continuously differentiable matrices. Hence, $A(\cdot)$, $C(\cdot)$, $K(\cdot)$ are continuously differentiable on Θ. We then have $\|\Lambda_j(\cdot)\|_{op} \leq \|A(\cdot)\|_{op}\|C(\cdot)\|^j_{op}\|K(\cdot)\|_{op}$. Condition (iii) implies $\sup_{\theta\in\Theta}\|C(\theta)\|_{op} \leq \overline{c}$ for some $\overline{c} \in [0,1)$. Then, we get: $\sum_{j=0}^{\infty}\sup_{\theta\in\Theta}\|\Lambda_j(\theta)\|_{op} \leq \sup_{\theta\in\Theta}\|A(\theta)\|_{op}\sup_{\theta\in\Theta}\|K(\theta)\|_{op}\sum_{j=0}^{\infty}\overline{c}^j < \infty$. For the last Assumption 2 (iii), direct differentiation with respect to θ_ℓ yields $\partial_{\theta_\ell}\Lambda_j(\theta) = \partial_{\theta_\ell}A(\theta)C(\theta)^jK(\theta) + A(\theta)C(\theta)^j\partial_{\theta_\ell}K(\theta) + A(\theta)\partial_{\theta_\ell}C(\theta)C(\theta)^{j-1}K(\theta) + \cdots + A(\theta)C(\theta)^{j-1}\partial_{\theta_\ell}C(\theta)K(\theta)$. Then, as above: $\|\partial_{\theta_\ell}\Lambda_j(\theta)\|_{op} \leq \overline{c}^{j-1}\sup_{\theta\in\Theta}(\|\partial_\theta\text{vec}[K(\theta)]\|_\infty + (j-1)\|K(\theta)\|_\infty\|\partial_\theta\text{vec}[C(\theta)]\|_\infty)$ is summable since $\overline{c} < 1$. Equivalence between norms implies $\sup_{\theta\in\Theta}\|\partial_\theta\text{vec}[\Lambda_j(\theta)]\|_\infty \leq \mathrm{C}\sum_{\ell=1}^{d_\theta}\sup_{\theta\in\Theta}\|\partial_{\theta_\ell}\Lambda_j(\theta)\|_{op}$, where C only depends on the size of the matrix, which is summable. Similar derivations yield the results for derivatives of order $s = 2, 3$. □

Proof of Lemma 2. Take any $\theta_0 \in \Theta$. Because $A(\theta_0) \geq 0$, there exists U_0 and U_1 with sizes $m \times d$ and $r \times d$, respectively, where $m + r = d$, such that:

$$A(\theta_0)U_0 = 0_{d,m}, \quad U_0'U_0 = I_m \tag{E.1}$$

$$U_1'A(\theta_0)U_1 > 0, \quad U_1'A(\theta_0)^\dagger A(\theta_0)U_1 > 0, \quad U_1'U_1 = I_r, \quad U_0'U_1 = 0_{r,m}. \tag{E.2}$$

They can be found using the eigendecomposition:

$$A(\theta_0) = U\begin{pmatrix}\Lambda(\theta_0) & 0_{r,m}\\ 0_{m,r} & 0_{m,m}\end{pmatrix}U',$$

where $UU' = U'U = I_d$, $\Lambda_0 > 0$ is diagonal, and $U = (U_0, U_1)$. These are not unique but their spans, and the associated projection matrices, are unique. Note that $U_1'A(\theta_0)U_1 = \Lambda(\theta_0)$ with $0 < \underline{\lambda} \preceq \Lambda(\theta_0) \preceq \overline{\lambda} < \infty$. Define: $U_0(\theta) = [I_d - A(\theta)[A(\theta)'A(\theta)]^\dagger A(\theta)]U_0(\theta_0)$, $U_1(\theta) = [A(\theta)[A(\theta)'A(\theta)]^\dagger A(\theta)]U_1(\theta_0)$. They are continuously differentiable in θ because $\text{rank}[A(\theta)'A(\theta)] = r$ is constant (using Magnus and Neudecker, 2019, Prop8.2). By construction: $U_1(\theta)'U_0(\theta) = 0$ for $\forall\theta \in \Theta$. Because $A(\theta)$ is Hermitian, we have $A(\theta)[A(\theta)'A(\theta)]^\dagger A(\theta) = A(\theta)A(\theta)^\dagger$. Then, by the identities for the pseudo-inverse and $A(\theta_0)U_0 = 0_{d,m}$, we have $U_0(\theta_0)'U_0(\theta_0) = I_m$. By construction, $A(\theta)[I_d - A(\theta)A(\theta)^\dagger] = 0_{d,d}$, which implies $A(\theta)U_0(\theta) = 0_{m,n}$ for all θ. By the definition of a singular value: $\lambda_{\min}[U(\theta)'U(\theta)] = \sigma_{\min}[U(\theta)]^2$. Weyl's inequality for singular values (Horn and Johnson, 1991, Th3.3.16) implies: $\sigma_{\min}[U_0(\theta)] \geq \sigma_{\min}[U_0(\theta_0)] - \|U_0(\theta_0) - U_0(\theta)\|_{op} = 1 - \|U_0(\theta_0) - U_0(\theta)\|_{op}$, since $U_0(\theta_0)'U_0(\theta_0) = I_m$. Next, $\theta \to A(\theta)A(\theta)^\dagger$ is s-times continuously differentiable on Θ, so it is globally Lipschitz continuous with constant $0 \leq L_A < \infty$. This implies:

$$\|U_0(\theta_0) - U_0(\theta)\|_{op} = \|[A(\theta_0)A^{\dagger}(\theta_0) - A(\theta_0)A^{\dagger}(\theta_0)]U_0\|_{op} \leq L_A\|\theta - \theta_0\|.$$

Pick $0 < \delta \leq [2L_A]^{-1}$, then for all $\|\theta - \theta_0\| \leq \delta$, we have: $\sigma_{\min}[U_0(\theta)] \geq 1/2 > 0 \Rightarrow \lambda_{\min}[U_0(\theta)'U_0(\theta)] \geq 1/4 > 0$. By composition and invertibility, $\theta \to \tilde{U}_0(\theta) = U_0(\theta)[U_0(\theta)'U_0(\theta)]^{-1/2}$ is s-times continously differentiable, $\tilde{U}_0(\theta)'\tilde{U}_0(\theta) = I_m$, and $U_1(\theta)'\tilde{U}_0(\theta) = 0_{r,m}$. For the same δ, $\tilde{U}_1(\theta)$ constructed the same way is s-times continously differentiable with $\tilde{U}_1(\theta)'\tilde{U}_1(\theta) = I_r$ and $\tilde{U}_1(\theta)'\tilde{U}_0(\theta) = 0_{r,m}$. It remains to show $\lambda_{\min}[\tilde{U}_1(\theta)'A(\theta)\tilde{U}_1]$ is bounded below. Apply Weyl's inequality:

$$\lambda_{\min}[\tilde{U}_1(\theta)'A(\theta)\tilde{U}_1] \geq \lambda_{\min}[\tilde{U}_1(\theta_0)'A(\theta_0)\tilde{U}_1(\theta_0)] - \|\tilde{U}_1(\theta)'A(\theta)\tilde{U}_1 - \tilde{U}_1(\theta_0)'A(\theta_0)\tilde{U}_1(\theta_0)\|_{op},$$

where the last term is s-time continuously differentiable on $\mathcal{B}_\delta(\theta_0)$, hence Lipschitz continuous with finite constant $L_{U,A}$. Eventually, we get: $\lambda_{\min}[\tilde{U}_1(\theta)'A(\theta)\tilde{U}_1] \geq \lambda_{\min}[\tilde{U}_1(\theta_0)'A(\theta_0)\tilde{U}_1(\theta_0)] - L_{U,A}\|\theta - \theta_0\|$. By choosing $0 < \delta \leq \min[(2L_A)^{-1}, \underline{\lambda}[2L_{U,A}]^{-1}]$, we find: $\lambda_{\min}[\tilde{U}_1(\theta)'A(\theta)\tilde{U}_1] \geq \underline{\lambda}/2 > 0$. Now we have $M(\theta) = (U_1(\theta), U_0(\theta))$ continuously differentiable on $\mathcal{B}_\delta(\theta_0)$, invertible and $M(\theta)M(\theta)' = I_d$. By composition, $M(\theta)'A(\theta)M(\theta) = \text{blockdiag}(B(\theta), 0_{m,m})$ is continuously differentiable and $0 < \underline{\lambda}/2 \preceq B(\theta) \preceq \overline{\lambda} < \infty$. Then, for $\mathcal{B}_\delta(\theta_0)$, $A(\theta)^{1/2} = M(\theta)\text{blockdiag}(B(\theta)^{1/2}, 0_{m,m})M(\theta)'$ is s-times continuously differentiable by composition and $B(\theta)$ strictly positive definite.

Since Θ is compact and finite-dimensional, we can take a finite δ-cover $\{\theta_1, \dots, \theta_N\}$ of Θ. The mapping $\theta \to A(\theta)^{1/2}$ is continuously differentiable on each $\mathcal{B}_\delta(\theta_j) \cap \Theta$ and $\Theta = \cup_{j=1}^N \mathcal{B}_\delta(\theta_j) \cap \Theta$ so, by the gluing lemma (Lee, 2010, Lem3.23), $\theta \to A(\theta)^{1/2}$ is s-times continuously differentiable on Θ. □

Proof of Lemma 3. The existence of the infinite order VAR representation follows from the properties of the VMA polynomial stated in Assumption 3(i); $\det(\sum_{j=0}^{\infty} \Psi_j z^j) \neq 0$ for any $|z| \leq 1$ is a result of the same property of the VMA polynomial.

Lemma 3(i) is proved in Hannan and Deistler (2012, Th7.4.5); estimating the VAR coefficients on de-meaned data does not affect their results (Kuersteiner, 2005, Lem4.1). Hannan and Deistler (2012) require that $\lim_{s \to \infty} \mathbb{E}(e_t e_t' - \tilde{\Sigma}|\mathcal{F}_{t-s}) = 0$, this is implied by the mixing condition and Davidson (2021, Th15.2). Lemma 3(ii) is shown in Lütkepohl (2005, Prop15.1-15.3). For Lemma 3(iii), note that Assumption 2 (i) and the mixing condition implies $\tilde{y}_t$ is Near-Epoch Dependent with size $-b$ in L_q-norm for $q \geq 8$ (Lemma B1). Its autocovariances are absolutely summable, using Theorem 17.16 with Corollary 18.7 in (Davidson, 2021). This and Chebyshev's inequality imply Lemma 3(iii) for $\tilde{y}_n$. Finally, $e_t e_t'$ is α-mixing with size $-a$ and finite fourth moment. Thus its autocovariances are absolutely summable. This and Chebyshev's inequality yields Lemma 3(iii) for $\tilde{\Sigma}_n$. □

Proof of Lemma B1. The first part of the Lemma follows from Davidson (2021, Example 18.3) with a multivariate one-sided VMA(∞) process. For the second part, use the double series representation: $\partial_{\phi_j} y_t(\theta; \psi_{k0}) = -\sum_{\ell=0}^{\infty}\sum_{s=0}^{\infty}[e_{t-\ell-j-1-s}'\tilde{\Lambda}_s'] \otimes [\Lambda_\ell(\theta)P(\theta;\tilde{\Sigma})]$. Recall that an absolutely summable double series can be reparameterized as an absolutely summable single series. A single series is a one-sided VMA(∞) representation, which in our setting can be used to derive NED properties. It is sufficient to upper-bound the coefficients of the latter and bound their rate of decay to get the result. Using $\|e_t\|_q = [\mathbb{E}(\|e_t\|^q)]^{1/q}$ and $\|A \otimes B\|_{op} = \|A\|_{op}\|B\|_{op}$, we have, for any θ, by the triangle inequality followed by the Cauchy-Schwarz inequality:

$$\|\partial_{\phi_j} y_t(\theta;\psi_{k0})\|_q \leq \left[\sum_{s=0}^{\infty}\|\tilde{\Lambda}_s\|_{op}\right]\sum_{\ell=0}^{\infty}\sup_{\theta\in\Theta}\|\Lambda_\ell(\theta)\|_{op}\sup_{\theta\in\Theta}\|P(\theta;\tilde{\Sigma})\|_{op}\|e_{t-\ell-j-1-s}\|_q \leq C < \infty.$$

The majoration absolute summability and the NED property in L_q-norm. Also, the summation starting with $m+1$ satisfies: $\left[\sum_{s=0}^{\infty}\|\tilde{\Lambda}_s\|_{op}\right]\sum_{\ell=m+1}^{\infty}\sup_{\theta\in\Theta}\|\Lambda_\ell(\theta)\|_{op}\sup_{\theta\in\Theta}\|P(\theta;\tilde{\Sigma})\|_{op} \leq Cm^{-(b+\varepsilon)}$, implying that the size is $-b$. The constant rank assumption for $\Sigma(\theta)$ and invertibility condition for $\tilde{\Sigma}$ imply continuous differentiability of $P(\cdot;\cdot)$ using Lemma 2 for the matrix square root. The VMA(∞) representation then implies the NED property. □

Appendix F Proofs for the Specification Test

Lemma F2 (Strong Approximation). *Let $Z_{t,k}$ and $S_{n,k}$ be as Theorem 2. Suppose the Assumptions in Theorem 3 holds. Then, there exists $\mathcal{Z}_{n,k} \sim \mathcal{N}(0, S_{n,k}/n)$ such that: $\sqrt{n}\|\overline{Z}_{n,k} - \mathcal{Z}_{n,k}\| = o_p([\log(n)]^{-2})$.*

Proof of Lemma F2. There are four main steps: Step 1. Use the NED properties to approximate $Z_{t,k}$ by a strong-mixing sequence, in blocks of observations; Step 2. Use coupling results for strong-mixing sequences to further approximate the blocks with iid blocks; Step 3. Apply Yurinskii's method to approximate the iid blocks with Gaussian random variables; Step 4. Adjust the covariance of the Gaussian from Step 3. to get the coupling with the same covariance structure.

Step 1. Approximate $Z_{t,k}$ by a strong-mixing process $Z^m_{t,k}$. Recall: $Z_{t,k} = ((\tilde{y}_t - \tilde{\mu})', \mathrm{vec}[e_t\tilde{Y}'_{t-1,k}\Gamma_k^{-1}]', \mathrm{vech}[e_te_t']')'$, with $\mathbb{E}[Z_{t,k}] = 0$, a column vector of dimension $d(k) = O(k)$. Note that when analyzing the properties of $Z_{t,k}$ below, we often consider its elements $(\tilde{y}_t - \tilde{\mu})$, $\mathrm{vec}[e_t\tilde{Y}_{t-1,k}$, and $\mathrm{vech}[e_te_t']')'$ separately.

For any m satisfying $m \geq 2k$ and $m/n = o(1)$, define $Z^m_{t,k} = \mathbb{E}[Z_{t,k}|\mathcal{F}^t_{t-m}]$, a strong-mixing sequence with coefficients $\alpha_m(s) = \alpha([s-m]^+)$, with $\alpha(\cdot)$ the mixing coefficients of e_t. Note that $\mathbb{E}[e_te_t'|\mathcal{F}^t_{t-m}] = e_te_t'$. Since Γ_k^{-1} is bounded, we only need to bound $\|e_t\tilde{y}_{t-j} - e_t\mathbb{E}[\tilde{y}_{t-j}|\mathcal{F}_{t-m}]\|_{q/2}$ for $j \in \{1, \ldots, k\}$. For this, we have $\|\tilde{y}_{t-j} - \mathbb{E}[\tilde{y}_{t-j}|\mathcal{F}_{t-m}]\|_q \leq \sum_{\ell \geq m-k} \|\tilde{\Lambda}_\ell\|_{op}\|e_t\|_q \leq C[m-k]^{-(b+\varepsilon)}\|e_t\|_q$, uniformly in $j \leq k$; consequently, $\|e_t\tilde{y}_{t-j} - e_t\mathbb{E}[\tilde{y}_{t-j}|\mathcal{F}_{t-m}]\|_{q/2} \leq C[m-k]^{-(b+\varepsilon)}\|e_t\|_q^2$ using the Cauchy-Schwarz inequality.

Note that $\mathbb{E}[\overline{Z}_{n,k} - \mathbb{E}(\overline{Z}^m_{n,k})] = 0$, and $\mathbb{E}[\mathrm{vec}(e_t\tilde{Y}'_{t-1,k})|\mathcal{F}_{t-1}] = 0$ so that the elements of $Z_{t,k}$ and $Z^m_{t,k}$ corresponding to $\mathrm{vec}(e_t\tilde{Y}'_{t-1,k}\Gamma_k^{-1})$ are serially uncorrelated. By the NED and mixing properties, the autocovariances of $\tilde{y}_t$ and $\mathrm{vech}[e_te_t']$ are absolutely summable. For each row ℓ of $\overline{Z}_{t,k}$ we have: $\mathrm{var}[\overline{Z}_{n,k,\ell} - \overline{Z}^m_{n,k,\ell}] = 0$ for indices corresponding to $\mathrm{vech}[e_te_t']$, and $\mathrm{var}[\overline{Z}_{n,k,\ell} - \overline{Z}^m_{n,k,\ell}] = \frac{1}{n}\mathrm{var}[Z_{t,k,\ell} - Z^m_{t,k,\ell}] \leq n^{-1}C[m-k]^{-2(b+\varepsilon)}\|e_t\|_4^2$ for indices corresponding to $\mathrm{vec}(e_t\tilde{Y}'_{t-1,k}\Gamma_k^{-1})$. For indices ℓ corresponding to $\tilde{y}_t$, the autocovariance of order $s \geq 1$ satisfies: $|\mathrm{cov}(Z_{t,k,\ell} - Z^m_{t,k,\ell}, Z_{t-s,k,\ell} - Z^m_{t-s,k,\ell})| \leq \{|\mathrm{cov}(Z_{t,k,\ell} - Z^m_{t,k,\ell}, Z_{t-s,k,\ell} - Z^m_{t-s,k,\ell})|\}^{1/2}[\|Z_{t,k,\ell} - Z^m_{t,k,\ell}\|_2\|Z_{t,k,\ell} - Z^m_{t,k,\ell}\|_2]^{1/2} = \{|\mathrm{cov}(Z_{t,k,\ell} - Z^m_{t,k,\ell}, Z_{t-s,k,\ell} - Z^m_{t-s,k,\ell})|\}^{1/2}\|Z_{t,k,\ell} - Z^m_{t,k,\ell}\|_2$.[1] Next, $|\mathrm{cov}(Z_{t,k,\ell} - Z^m_{t,k,\ell}, Z_{t-s,k,\ell} - Z^m_{t-s,k,\ell})| = |\mathbb{E}[\mathbb{E}\{Z_{t,k,\ell} - Z^m_{t,k,\ell}|\mathcal{F}_{t-s}\}(Z_{t-s,k,\ell} - Z^m_{t-s,k,\ell})]| \leq \|\mathbb{E}\{Z_{t,k,\ell} - Z^m_{t,k,\ell}|\mathcal{F}_{t-s}\}\|_2\|Z_{t,k,\ell} - Z^m_{t,k,\ell}\|_2 \leq 2C(1+s)^{-(b+\varepsilon)}\|e_t\|_2\|Z_{t,k,\ell} - Z^m_{t,k,\ell}\|_2$, using derivations from Lemma B1. Taking the summation, we get:

$$\sum_{s=1}^{\infty} |\mathrm{cov}(Z_{t,k,\ell} - Z^m_{t,k,\ell}, Z_{t-s,k,\ell} - Z^m_{t-s,k,\ell})| \leq [\sqrt{2C}\sum_{s \geq 1}(1+s)^{-(b+\varepsilon)/2}\|e_t\|_2^{1/2}]\|Z_{t,k,\ell} - Z^m_{t,k,\ell}\|_2^{3/2},$$

which is a $O([m-k]^{-(3/2)(b+\varepsilon)})$ since $(b+\varepsilon)/2 \geq 1 + \varepsilon/2$. From this, we get that $\mathrm{var}[\overline{Z}_{n,k,\ell} - \overline{Z}^m_{n,k,\ell}] = O([m-k]^{-(3/2)(b+\varepsilon)}n^{-1})$.

In sum, we have $\max_{\ell=1,\ldots,d(k)} \mathbb{E}[\|\overline{Z}_{n,k,\ell} - \overline{Z}^m_{n,k,\ell}\|_2] \leq O(n^{-1/2}m^{-(3/4)(b+\varepsilon)})$. Then Pisier (1983)'s and Markov's inequalities yield: $\sqrt{n}\|\overline{Z}_{n,k} - \overline{Z}^m_{n,k}\|_\infty = O_p(k^{1/2}m^{-(3/4)(b+\varepsilon)})$, which is a $o_p(1)$ for $m \geq 2k$ since $(3/4)b > 1/2$.

Step 2. Coupling for the strong-mixing process $Z^m_{t,k}$.

For this step, we will use the method of Bernstein sums. Using the notation of Davidson (2021, Ch15), the sample $(Z^m_{t,k})_{t=1,\ldots,n}$ will be divided into large blocks of size b_n, separated by small blocks of size l_n. We will set $2m \leq l_n = o(b_n)$ and $b_n = o(n)$; $r_n = [n/b_n]$ is the

[1] Here we use $|\mathrm{cov}(X,Y)| = |\mathrm{cov}(X,Y)|^{1/2}|\mathrm{cov}(X,Y)|^{1/2}$ and Cauchy–Schwarz $|\mathrm{cov}(X,Y)| \leq \|X\|_2\|Y\|_2$ if X, Y have mean zero.

number of large blocks. The last block has $n - r_n b_n - (r_n - 1)l_n = o(1)$ observations.

The first large block consists of $(Z^m_{1,k}, \dots, Z^m_{b_n,k})$, the second $(Z^m_{b_n+l_n+1,k}, \dots, Z^m_{2b_n+l_n+1,k})$. The dependence between the first and second block is given by $\alpha([l_n - m]) \leq \alpha(m) \to 0$, since the blocks are separated by $l_n \geq 2m$ time periods. For $i = 1, \dots, r_n$, let $X^m_{i,k} = \sum_{t=(i-1)b_n+(i-1)l_n+1}^{ib_n+(i-1)l_n} Z^m_{t,k}$ denote the partial sum over the i-th large block.

For each row of $X^m_{i,k}$, say the ℓ-th row, $X^m_{i,k,\ell}$, we can apply Theorem 1 in Peligrad (2002), to find $\tilde{X}^m_{i,k,\ell} \overset{d}{=} X^m_{i,k,\ell}$, iid over $i = 1, \dots, r_n$, such that $\mathbb{E}|X^m_{i,k,\ell} - \tilde{X}^m_{i,k,\ell}| \leq \sqrt{\alpha(l_n - m)} \|X^m_{i,k,\ell}\|_2$. Being a sum of b_n terms with absolutely summable autocovariances (cf. step 1), $\|X^m_{i,k,\ell}\|_2 = O(\sqrt{b_n})$. Then, we have $\mathbb{E}|\sum_{i=1}^{r_n}[X^m_{i,k,\ell} - \tilde{X}^m_{i,k,\ell}]| \leq O(r_n\sqrt{b_n \alpha(l_n - m)})$ by the triangle inequality on the r_n blocks.

The difference $\|n\overline{Z}^m_{n,k,\ell} - \sum_{i=1}^{r_n} X^m_{i,k,\ell}\|_2$ is of order $O(\sqrt{n - r_n b_n})$, as there are $n - r_n b_n$ observations in the smaller and last blocks of observations with absolutely summable autocovariances. These inequality holds uniformly in $\ell = 1, \dots, d(k)$, so this implies:

$$\begin{aligned}
&\mathbb{E}\left[\max_{\ell=1,\dots,d(k)} |n\overline{Z}^m_{n,k,\ell} - \sum_{i=1}^{r_n} \tilde{X}^m_{i,k,\ell}|\right] \\
&\leq \mathbb{E}\left[\max_{\ell=1,\dots,d(k)} |\sum_{i=1}^{r_n} X^m_{i,k,\ell} - \sum_{i=1}^{r_n} \tilde{X}^m_{i,k,\ell}|\right] + \mathbb{E}\left[\max_{\ell=1,\dots,d(k)} |\sum_{i=1}^{r_n} X^m_{i,k,\ell} - n\overline{Z}^m_{n,k,\ell}|^2\right]^{1/2} \\
&\leq O(kr_n\sqrt{b_n\alpha(l_n - m)}) + O(k^{1/2}\sqrt{n - r_n b_n}),
\end{aligned}$$

Let $\tilde{X}^m_{n,k} = 1/n \sum_{i=1}^{r_n} \tilde{X}^m_{i,k}$, then:

$$\sqrt{n}\|\overline{Z}^m_{n,k} - \tilde{X}^m_{n,k}\|_\infty \leq O_p(\max[n^{-1/2}kr_n\sqrt{b_n\alpha(l_n - m)}, (k/n)^{1/2}\sqrt{n - r_n b_n}]),$$

where $(k/n)^{1/2}\sqrt{n - r_n b_n} = O(\sqrt{kr_n l_n / n})$ by construction.

Step 3. Gaussian Approximation. Here, the idea is to apply Yurinskii's coupling to the iid sequence $\tilde{X}^m_{i,k}$, see Pollard (2002, Ch10) for an introduction. For this, we need a bound on $\mathbb{E}|\tilde{X}^m_{i,k,\ell}|^3$, this will rely on moment bounds for strong-mixing random variables in Rio (1999, Ch2). From step 2. we have $\tilde{X}^m_{i,k,\ell} \overset{d}{=} X^m_{i,k,\ell} = \sum_{t=(i-1)b_n+(i-1)l_n}^{ib_n+(i-1)l_n} Z^m_{t,k,\ell}$, where $Z^m_{t,k,\ell}$ are strong-mixing with coefficients $\alpha([\cdot - m]^+) = \alpha_m(\cdot)$ to simplify notation below. Using the notation from Rio (1999), let $\alpha_m^{-1}(u) = \inf\{j \in \mathbb{N}, \alpha_m(j) \leq u\} = \sum_{j=0}^{\infty} \mathbb{1}_{u<\alpha_m(j)}$ for any

$u \in [0,1]$. Rio (1999, Th2.2) implies:[2]

$$\mathbb{E}[|X^m_{i,k,\ell}|^4] \leq 12 a_4 b_n^2 \left[\int_0^1 \min[\alpha_m^{-1}(u), b_n] Q_m^2(u) du \right]^2 + 12 b_4 b_n \int_0^1 \min[\alpha_m^{-1}(u), b_n]^3 Q_m^4(u) du,$$

where Q_m is the quantile function of $|Z^m_{t,k,\ell}|$ and a_4, b_4 are universal constants. Take $p \in \{2,4\}$, we now want to bound:

$$\begin{aligned}\int_0^1 \min[\alpha_m^{-1}(u), b_n]^{p-1} Q_m^p(u) du &\leq [\int_0^1 [\alpha_m^{-1}(u)]^{\vartheta(p-1)} du]^{1/\vartheta} [\int_0^1 Q_m(u)^{p\vartheta/(\vartheta-1)} du]^{(\vartheta-1)/\vartheta} \\ &= [\int_0^1 [\alpha_m^{-1}(u)]^{\vartheta(p-1)} du]^{1/\vartheta} \|Z^m_{t,k,\ell}\|^p_{p\vartheta/(\vartheta-1)},\end{aligned}$$

using $\int_0^1 Q_m(u)^p du = \mathbb{E}[|Z^m_{t,k,\ell}|^p]$. Apply inequality (C.5) in Rio (1999, p156) to find:

$$\begin{aligned}\int_0^1 [\alpha_m^{-1}(u)]^{\vartheta(p-1)} du &\leq 2 \sum_{j=0}^{\infty} (j+1)^{\vartheta(p-1)-1} \alpha_m(j) \\ &= 2 \sum_{j=0}^{m-1} (j+1)^{\vartheta(p-1)-1} \alpha(0) + 2 \sum_{j=0}^{\infty} (m+j+1)^{\vartheta(p-1)-1} \alpha(j) \\ &\leq 2 \frac{\alpha(0)}{\vartheta(p-1)-1} m^{\vartheta(p-1)} + 2C(1+m)^{\vartheta(p-1)-1} \sum_{j=0}^{\infty} (1+j)^{\vartheta(p-1)-1-a-\varepsilon},\end{aligned}$$

if $\alpha(j) \leq C(1+j)^{-(a+\varepsilon)}$. Take $\vartheta = 2$, since $a \geq 6$ the series on the right-hand-side is summable. Then we get $\int_0^1 [\alpha_m^{-1}(u)]^{\vartheta(p-1)} du \leq C_{a,p} m^{2(p-1)}$, for a constant which depends on $a + \varepsilon$, C and p. Apply this to the fourth moment bound:

$$\mathbb{E}[|X^m_{i,k,\ell}|^4] \leq 12 a_4 C_{a,2} \|Z^m_{t,k,\ell}\|_4^4 (m b_n)^2 + 12 b_4 C_{a,4}^{1/2} \|Z^m_{t,k,\ell}\|_8^4 b_n m^3 \leq O(\max[(m b_n)^2, b_n m^3]),$$

and note that $\|Z^m_{t,k,\ell}\|_8$ is finite if $\|e_t\|_{16}$ is finite. Let $L^m_{n,k} = \sum_{i=1}^{r_n} \mathbb{E}[\|n^{-1/2} X^m_{i,k}\|^4] = O(n^{-2} k r_n \max[(m b_n)^2, b_n m^3])$ and $V^m_k = \text{var}[n^{-1/2} \tilde{X}^m_{i,k}]$. By Theorem 6 in Zaitsev (2013), there exists $\mathcal{Z}^m_{i,k} \overset{iid}{\sim} \mathcal{N}(0, V^m_k)$ such that: $\sqrt{n} \|\tilde{X}^m_{n,k} - \mathcal{Z}^m_{n,k}\|_4 \leq C_0 16 [L^m_{n,k}]^{1/4}$, where $\mathcal{Z}^m_{n,k} = 1/n \sum_{i=1}^{r_n} \mathcal{Z}^m_{i,k}$ so that $\sqrt{n} \mathcal{Z}^m_{n,k} \sim \mathcal{N}(0, r_n/n V^m_k)$.

Step 4. Overall Approximation Error. Now, we need to combine the steps to examine

[2] The bound is only available for even moments $2p$ with $p \geq 1$.

how well $\sqrt{n}\overline{Z}^m_{n,k}$ is approximated by the Gaussian vector $\sqrt{n}\mathcal{Z}^m_{n,k}$:

$$\begin{aligned}\sqrt{n}\|\overline{Z}_{n,k} - \mathcal{Z}^m_{n,k}\| &\leq \sqrt{n}\|\overline{Z}_{n,k} - \overline{Z}^m_{n,k}\| + \sqrt{n}\|\overline{Z}^m_{n,k} - \tilde{X}^m_{n,k}\| + \sqrt{n}\|\tilde{X}^m_{n,k} - \mathcal{Z}^m_{n,k}\| \\ &\leq O_p\Big(\max\Big[k^{1/2}m^{-3/4(b+\varepsilon)}, (n/b_n)^{1/2}k[1+l_n-m]^{-(a+\varepsilon)/2}, \\ &\qquad (kl_n/b_n)^{1/2}, k^{1/2}n^{-1/4}\max[m^{1/2}b_n^{1/4}, m^{3/4}]\Big]\Big),\end{aligned}$$

using $\alpha(j) \leq (1+j)^{-(a+\varepsilon)}$ and $r_n \leq n/b_n$. Set $k = n^{1-\delta_1}$, $m = 2n^{1-\delta_2}$, $b_n = n^{1-\delta_3}$ with $1 > \delta_1 \geq \delta_2 > \delta_3 > 0$ and $l_n = 2m$. The main idea here is to check, given the restrictions on k, b, and a, whether there are feasible choices of (δ_2, δ_3) such that the above approximation error is negligible. For this, we have

$$\begin{aligned}\sqrt{n}\|\overline{Z}_{n,k} - S^m_{n,k}\| &\leq \sqrt{n}\|\overline{Z}_{n,k} - \overline{Z}^m_{n,k}\| + \sqrt{n}\|\overline{Z}^m_{n,k} - \tilde{X}^m_{n,k}\| + \sqrt{n}\|\tilde{X}^m_{n,k} - S^m_{n,k}\| \\ &\leq O_p\Big(\max\Big[n^{1/2-3/4(b+\varepsilon)-\delta_1/2+3/4(b+\varepsilon)\delta_2}, n^{1-(a+\varepsilon)/2+\delta_3/2-\delta_1+\delta_2(a+\varepsilon)/2}, \\ &\qquad n^{(1-(\delta_1+\delta_2)+\delta_3)/2}, n^{1-(\delta_1+\delta_2)/2}\max[n^{-\delta_3/4}, n^{-\delta_2/4}]\Big]\Big),\end{aligned}$$

which is a $o_p([\log(n)]^{-2})$ if:

$$\begin{aligned} &\delta_1 + \frac{3(b+\varepsilon)}{2} > 1 + \frac{3(b+\varepsilon)}{2}\delta_2, && \delta_1 + \frac{a+\varepsilon}{2} > 1 + \frac{a+\varepsilon}{2}\delta_2 + \frac{1}{2}\delta_3 \\ &\delta_1 + \delta_2 > 1 + \delta_3, && \delta_1 + \delta_2 + \frac{1}{2}\delta_3 > 2.\end{aligned}$$

The last row of inequalities implies $1 > \delta_1 \geq \delta_2 \geq \delta_3 > 2/3$ and $\delta_1 > 1 + \delta_3 - \delta_2$. The top left inequality is not binding for $b + \varepsilon > 2$, and the top right inequality further yields: $\delta_1 > [5 - (a+\varepsilon)]/3$ which is not binding for $a + \varepsilon > 6$. This implies that there is a feasible solution for which $\delta_1 > 3/4 > 2/3$, i.e. $k = o(n^{1/4})$, such that $\sqrt{n}\|\overline{Z}_{n,k} - S^m_{n,k}\|_\infty = o_p(n^{-\delta})$ for some $\delta \in (0,1)$ which depends on (a, b, ε). It can be found by minimizing the rates above with respect to $(\delta_1, \delta_2, \delta_3)$ over the feasible set.

The covariance matrix of $\sqrt{n}\overline{Z}_{n,k}$ is given by $S_{n,k} = \text{var}[n^{-1/2}\sum_{t=1}^n Z_{t,k}]$, whereas the variance of $\sqrt{n}\mathcal{Z}^m_{n,k}$ is, by construction, equal to $\text{var}[n^{-1/2}\sum_{i=1}^{r_n}\tilde{X}^m_{i,k}] = n^{-1}r_n\text{var}[X^m_{i,k}] := S^m_{n,k}$ by independence and equality in distribution. Let $X_{i,k} = \sum_{t=(i-1)b_n+(i-1)l_n}^{i_b+(i-1)l_n} Z_{t,k}$, we have:

$$\|S_{n,k} - S^m_{n,k}\| \leq \|S_{n,k} - n^{-1}r_n\text{var}[X_{i,k}]\| + n^{-1}r_n\|\text{var}[X_{i,k}] - \text{var}[X^m_{i,k}]\|.$$

Starting with the first term on the right-hand-side, standard calculations imply for $\Gamma_{k,j} =$

$\mathbb{E}(Z_{t,k}Z'_{t-j,k})$:

$$\begin{aligned} S_{n,k} - n^{-1}r_n\text{var}[X_{i,k}] &= \Gamma_{k,0} + \sum_{j=1}^{n-1}\frac{n-j}{n}\left[\Gamma_{k,j}+\Gamma'_{k,j}\right] - \frac{r_n b_n}{n}\left(\Gamma_{k,0}+\sum_{j=1}^{b_n-1}\frac{b_n-j}{b_n}\left[\Gamma_{k,j}+\Gamma'_{k,j}\right]\right) \\ &= \sum_{j=1}^{b_n-1}(b_n^{-1}-n^{-1})j\left[\Gamma_{k,j}+\Gamma'_{k,j}\right] \qquad \text{(S1)} \\ &+ \sum_{j=b_n+1}^{n-1}\frac{n-j}{n}\left[\Gamma_{k,j}+\Gamma'_{k,j}\right] \qquad \text{(S2)} \\ &+ [b_n r_n/n-1]\left(\Gamma_{k,0}+\sum_{j=1}^{b_n-1}\frac{b_n-j}{b_n}\left[\Gamma_{k,j}+\Gamma'_{k,j}\right]\right). \qquad \text{(S3)} \end{aligned}$$

Begin with (S2). Recall from step 1 that elements of $Z_{t,k}$ corresponding to $\text{vec}(e_t\tilde{Y}'_{t-1,k}\Gamma_k^{-1})$ are serially uncorrelated. The remaining terms correspond to $\tilde{y}_t-\tilde{\mu}$ and $\text{vech}[e_te'_t-\tilde{\Sigma}]$. Using previous calculations, for any $j\geq 1$: $\|\Gamma_{k,j}\| = \|\mathbb{E}(Z_{t,k}|\mathcal{F}_{t-j})Z'_{t-j,k}\| \leq \|\mathbb{E}(Z_{t,k}|\mathcal{F}_{t-j})\|_2\|Z_{t,k}\|_2$. There are two bounds to compute here: $\|\mathbb{E}(\tilde{y}_t-\tilde{\mu}|\mathcal{F}_{t-j})\| \leq C(1+j)^{-(b+\varepsilon)}\|e_t\|_2$ and $\|\mathbb{E}(\text{vech}[e_te'_t-\tilde{\Sigma}]|\mathcal{F}_{t-j})\|_2 \leq 6d^2(1+j)^{-(a+\varepsilon)3/8}\|e_t\|_{16}$ using Davidson (2021, Th15.2) with $p=2$ and $r=8$; $d=\dim(e_t)$. Using these two bounds, we find:

$$\begin{aligned} \|(S2)\| &\leq 4[C+6d^2]\|e_t\|_{16}\|Z_{t,k}\|_2\sum_{j=b_n+1}^{\infty}(1+j)^{-\min((3/8)[a+\varepsilon],(b+\varepsilon))} \\ &\leq \frac{4[C+6d^2]\|e_t\|_{16}\|Z_{t,k}\|_2 b_n^{1-\min((3/8)[a+\varepsilon],(b+\varepsilon))}}{\min((3/8)[a+\varepsilon],(b+\varepsilon))}, \end{aligned}$$

which is a $o(b_n^{-5/4}k^{1/2})$ since $a,b\geq 6$ and $\varepsilon>0$ by assumption. Since $b_n>k$, this implies that $\|(S2)\|=o(k^{-3/4})$. For (S1), using $b_n<n$ and the same bounds:

$$\|(S1)\| \leq 4b_n^{-1}4[C+6d^2]\|e_t\|_{16}\|Z_{t,k}\|_2\sum_{j=1}^{\infty}(1+j)^{1-\min(3/8[a+\varepsilon],(b+\varepsilon))} = O(b_n^{-1}k^{1/2}).$$

It is possible to pick $b_n^{-1}=o(k^{-1}[\log(n)]^{-6})$, from the constraints above. For (S3), recall that $n=b_nr_n+(r_n-1)l_n+o(1)$ so that $b_nr_n/n-1=(r_n-1)/nl_n+o(n^{-1})=O(l_n/b_n)+o(n^{-1})$. Looking at the feasible values for $(\delta_1,\delta_2,\delta_3)$ above, we have $1-\delta_1<\delta_2-\delta_3$ which implies

$l_n/b_n = o(k^{-1}[\log(n)]^{-6})$ is feasible, which implies:

$$\|(S3)\| \leq o(k^{-1}[\log(n)]^{-6}) \sum_{j=0}^{\infty} \|\Gamma_{k,j}\| = o(k^{-1/2}[\log(n)]^{-6}).$$

Altogether, we find:

$$\|S_{n,k} - n^{-1} r_n \mathrm{var}(X_{i,k})\| \leq o_p(k^{-1/2}[\log(n)]^{-6}).$$

Now, we consider $n^{-1} r_n \|\mathrm{var}[X_{i,k}] - \mathrm{var}[X^m_{i,k}]\| \leq 2n^{-1} r_n \|\mathbb{E}[(X_{i,k} - X^m_{i,k})(X_{i,k} - X^m_{i,k})']\| \leq n^{-1} r_n b_n^2 \|Z_{t,k} - Z^m_{t,k}\|_2 \|Z_{t,k} + Z^m_{t,k}\|_2 \leq O(b_n k^{1/2} m^{-3/4(b+\varepsilon)})$. Using the strategy from above, set $b_n = n^{1-\delta_3}$, $k = n^{1-\delta_1}$, and $m = n^{1-\delta_2}$. Using the inequalities $\delta_1 > 3/4$, $\delta_3 > 2/3$, $b + \varepsilon > 4$, we get an upper bound $\delta_2 < 61/72$, which is within the feasible set so that we can further set b_n, m, such that $n^{-1} r_n \|\mathrm{var}[X_{i,k}] - \mathrm{var}[X^m_{i,k}]\| = o_p([\log(n)]^{-6})$. This implies that overall we have: $\|S_{n,k} - S^m_{n,k}\| = o_p([\log(n)]^{-6})$.

We can write $\sqrt{n}\mathcal{Z}^m_{n,k} = (S^m_{n,k})^{1/2}\mathcal{Z}_k$ where $\mathcal{Z}_k \sim \mathcal{N}(0, I)$ and let $\sqrt{n}\mathcal{Z}_{n,k} = (S_{n,k})^{1/2}\mathcal{Z}_k \sim \mathcal{N}(0, S_{n,k})$ have the desired covariance structure. Now, apply the inequality $\|S_1^{1/2} - S_2^{1/2}\| \leq \|S_1 - S_2\|^{1/2}$ and Hölder's inequality to find $\|[(S^m_{n,k})^{1/2} - (S_{n,k})^{1/2}]\mathcal{Z}_{n,k}\| \leq \|S^m_{n,k} - S_{n,k}\|^{1/2}\|\mathcal{Z}_{n,k}\|_\infty$, which is less than $[\log(n)]^{-3}\|\mathcal{Z}_{n,k}\|_\infty$ with probability approaching one since $\|S^m_{n,k} - S_{n,k}\| = o_p([\log(n)]^{-6})$. Combine this with Chernoff's bound:

$$\mathbb{P}\Big(\|[(S^m_{n,k})^{1/2} - (S_{n,k})^{1/2}]\mathcal{Z}_{n,k}\| > [\log(n)]^{-2}\Big) \leq 2d(k)\exp\left(-[\log(n)]^2\right) + o(1) = o(1),$$

since $d(k) = O(k)$ and $\log(k) = o([\log(n)]^2)$. This implies that $\|[(S^m_{n,k})^{1/2} - (S_{n,k})^{1/2}]\mathcal{Z}_{n,k}\| = o_p([\log(n)]^{-2})$. Combining all the results together, we find: $\sqrt{n}\|\overline{Z}_{n,k} - \tilde{Z}_{n,k}\| = o_p([\log(n)]^{-2})$, where $\sqrt{n}\tilde{Z}_{n,k} \sim \mathcal{N}(0, S_{n,k})$ as desired, which implies the strong approximation result. □

Proof of Theorem 3. For correctly specified models, we have $\tilde{y}_t = y_t(\theta_0; \psi_0)$ and thus

$$nQ_n(\hat{\theta}_n; \hat{\psi}_{nk}) = \sum_{t=1}^{n} (y_t(\hat{\theta}_n; \hat{\psi}_{nk}) - y_t(\theta_0; \psi_0))' W_n (y_t(\hat{\theta}_n; \hat{\psi}_{nk}) - y_t(\theta_0; \psi_0)). \tag{F.3}$$

To simplify notation, define $\partial_\theta \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\hat{\theta}_n, \theta_0; \psi) = \int_0^1 \partial_\theta y_t(\omega\hat{\theta}_n + (1-\omega)\theta_0; \psi) d\omega$, for any ψ, and $\partial_\psi \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\theta; \hat{\psi}_{nk}, \psi_k) = \int_0^1 \partial_\theta y_t(\theta; \omega\hat{\psi}_{nk} + (1-\omega)\psi_k) d\omega$, for any θ. The proof consists of five steps as follows.

Step 1. Expansion for each $y_t(\hat{\theta}_n; \hat{\psi}_{nk}) - y_t(\theta_0; \psi_0)$. For each $t = 1, \dots, n$, we have:

$$y_t(\hat{\theta}_n; \hat{\psi}_{nk}) - y_t(\theta_0; \psi_k) = \partial_\psi \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\theta_0; \hat{\psi}_{nk}, \psi_k)(\hat{\psi}_{nk} - \psi_k) + \partial_\theta \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\hat{\theta}_n, \theta_0; \hat{\psi}_{nk})(\hat{\theta}_n - \theta_0).$$

Before expanding on (F.3), we express $\hat{\theta}_n - \theta_0$ in terms of $\hat{\psi}_{nk} - \psi_k$. We begin with the following representation, derived in the proof of Theorem 2:

$$\hat{\theta}_n - \theta_0 = M_n^{-1}[(C) + (D) + (E)],$$

where (D) and (E) involve $\hat{\psi}_{nk} - \psi_k$, while (C) only involves $u_{t,k} = y_t(\theta_0; \psi_k) - \tilde{y}_t$. Let $D_n = -\frac{1}{n} \sum_{t=1}^n (u'_{t,k} W_n \otimes I) \partial_\psi \bar{\boldsymbol{G}}_{\boldsymbol{t}}(\theta_0; \hat{\psi}_{nk}, \psi_k)$ and $E_n = -\frac{1}{n} \sum_{t=1}^n \partial_\theta y_t(\theta_0; \hat{\psi}_{nk})' W_n \partial_\psi \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\theta_0; \hat{\psi}_{nk}, \psi_k)$. Then, $\hat{\theta}_n - \theta_0 = M_n^{-1}(C) + M_n^{-1}[D_n + E_n](\hat{\psi}_{nk} - \psi_k)$. From these expressions, we derive:

$$\begin{aligned} y_t(\hat{\theta}_n; \hat{\psi}_{nk}) - y_t(\theta_0; \psi_0) = {} & \{\partial_\psi y_t(\theta_0; \tilde{\psi}_{k,t}) + \partial_\theta y_t(\tilde{\theta}_{n,t}; \hat{\psi}_{nk}) M_n^{-1}[D_n + E_n]\}(\hat{\psi}_{nk} - \psi_k) \\ & + \partial_\theta y_t(\tilde{\theta}_{n,t}; \hat{\psi}_{nk}) M_n^{-1}(C) + y_t(\theta_0; \psi_k) - y_t(\theta_0; \psi_0), \end{aligned}$$

where $y_t(\theta_0; \psi_k) - y_t(\theta_0; \psi_0) = u_{t,k}$ because the model is correctly specified.

We now establish an upper bound for (C) to be used subsequently. For $t = k+1, \dots, n$: $u_{t,k} = y_t(\theta_0; \psi_k) - y_t(\theta_0; \psi_0) = \sum_{\ell=0}^\infty \Lambda_\ell(\theta_0) P(\theta_0; \tilde{\Sigma})[e_{t,k} - e_t]$, where $e_{t,k} = \tilde{y}_t - \tilde{\mu} - \sum_{j=1}^k \Psi_j[\tilde{y}_{t-j} - \tilde{\mu}] = e_t + \sum_{j=k+1}^\infty \Psi_j[\tilde{y}_{t-j} - \tilde{\mu}]$. Assumption 5 implies $\sqrt{n}[\log(n)]^2(\mathbb{E}[\|e_t - e_{t,k}\|^2])^{1/2} = o(1)$, and in combination with Assumption 3 (ii), it yields $\sqrt{n}[\log(n)]^2(\mathbb{E}[\|u_t - u_{t,k}\|^2])^{1/2} = o(1)$. Thus, for (C), we have:

$$\begin{aligned} \|(C)\| &\leq \|W_n - W\|_{op}(1/n) \sum_{t=1}^n \|\partial_\theta y_t(\theta_0; \psi_k)\|_\infty \|u_{t,k}\| + \overline{\lambda}_W (1/n) \sum_{t=1}^n \|\partial_\theta y_t(\theta_0; \psi_k)\|_\infty \|u_{t,k}\| \\ &\leq [1 + o_p(1)][o_p(n^{-1/2}[\log(n)]^{-2}) + O_p(kn^{-1})] = o_p(n^{-1/2}[\log(n)]^{-2}), \end{aligned}$$

using the Cauchy-Schwarz inequality and the above results. The first $t = 1, \dots, k$ terms in each summation have finite moments, contributing $O_p(k/n)$ to the average, as in the proofs of Theorems 1 and 2, where $k = o(n^{1/4})$. They yield the $O_p(kn^{-1})$ on the right hand side.

Step 2. Expanding $nQ_n(\hat{\theta}_n; \hat{\psi}_{nk})$ to show: $nQ_n(\hat{\theta}_n; \hat{\psi}_{nk}) = n(\hat{\psi}_{nk} - \psi_k)' M_k (\hat{\psi}_{nk} - \psi_k) + o_p(1)$, with M_k defined below in equation (M_k).

Square and sum the terms from Step 1:

$$nQ_n(\hat{\theta}_n;\hat{\psi}_{nk}) = n(\hat{\psi}_{nk}-\psi_k)'M_{n,k}(\hat{\psi}_{nk}-\psi_k) \tag{Q1}$$

$$+\, 2(\hat{\psi}_{nk}-\psi_k)'\sum_{t=1}^{n} B_{n,t}'W_n\{\partial_\theta \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\hat{\theta}_n,\theta_0;\hat{\psi}_{nk})M_n^{-1}(C)+u_{t,k}\} \tag{Q2}$$

$$+\sum_{t=1}^{n}\{\partial_\theta \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\hat{\theta}_n,\theta_0;\hat{\psi}_{nk})M_n^{-1}(C)+u_{t,k}\}'W_n\{\partial_\theta \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\hat{\theta}_n,\theta_0;\hat{\psi}_{nk})M_n^{-1}(C)+u_{t,k}\} \tag{Q3}$$

where $B_{n,t} = \partial_\psi \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\theta_0;\hat{\psi}_{nk},\psi_k)+\partial_\theta \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\hat{\theta}_n,\theta_0;\hat{\psi}_{nk})M_n^{-1}[D_n+E_n]$, and $M_{n,k} = \frac{1}{n}\sum_{t=1}^{n} B_{n,t}'W_nB_{n,t}$.

Next, we will show that (Q2) and (Q3) are negligible. Recall that $\|\hat{\psi}_{nk}-\psi_k\|_\infty \leq O_p(\sqrt{\log(n)/n})$ (cf. Lemma 3), and $\|(C)\| \leq o_p(n^{-1/2}[\log(n)]^{-2})$. For (Q2):

$$(Q2) = 2\sqrt{n}(\hat{\psi}_{nk}-\psi_k)'\Big\{\frac{1}{n}\sum_{t=1}^{n} B_{n,t}'W_n\partial_\theta y_t(\tilde{\theta}_{n,t};\hat{\psi}_{nk})\Big\}M_n^{-1}\sqrt{n}(C)$$

$$+\sqrt{n}(\hat{\psi}_{nk}-\psi_k)'\Big\{\frac{1}{n}\sum_{t=1}^{n} B_{n,t}'W_n\sqrt{n}u_{t,k}\Big\}.$$

The proof of Theorem 2 implies $||\frac{1}{n}\sum_{t=1}^{n} B_{n,t}'W_n\partial_\theta \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\hat{\theta}_n,\theta_0;\hat{\psi}_{nk})|| = ||\mathbb{E}[\{\partial_\psi y_t(\theta_0;\psi_k) + \partial_\theta y_t(\theta_0;\psi_k)M^{-1}\{\mathbb{E}[(u_{t,k}'W\otimes I)\partial_\psi G_t(\theta_0;\psi_k)+\partial_\theta y_t(\theta_0;\psi_k)\}'W\partial_\theta y_t(\theta_0;\psi_k)]\}]|| + o_p(1)$. The expectation on the right hand side is absolutely summable, hence bounded. Thus, for the first part of (Q2), we get: $\|\sqrt{n}(\hat{\psi}_{nk}-\psi_k)'\Big\{\frac{1}{n}\sum_{t=1}^{n} B_{n,t}'W_n\partial_\theta \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\hat{\theta}_n,\theta_0;\hat{\psi}_{nk})\Big\}M_n^{-1}\sqrt{n}(C)\| \leq O_p(\sqrt{\log(n)})o_p([\log(n)]^{-2}) = o_p([\log(n)]^{-3/2})$. For the second part of (Q2), following the proof of Theorems 1 and 2, we have $||\frac{1}{n}\sum_{t=1}^{n} B_{n,t}'W_n\sqrt{n}u_{t,k}|| = ||\frac{1}{n}\sum_{t=1}^{n}[\partial_\psi y_t(\theta_0;\psi_k) + \partial_\theta y_t(\theta_0;\psi_k)M^{-1}\{\mathbb{E}[(u_{t,k}'W\otimes I)\partial_\psi G_t(\theta_0;\psi_k)+\partial_\theta y_t(\theta_0;\psi_k)'W\partial_\psi y_t(\theta_0;\psi_k)]\}]'W\sqrt{n}u_{t,k}||+o_p(1)$. From the expression for $\partial_{\phi_j}y_t(\theta_0;\psi_k)$ in Lemma B1, the summability conditions in Assumption 3, and the moment condition in Assumption 2, we have: $\mathbb{E}[\|\partial_{\phi_j}y_t(\theta_0;\psi_k)\|^q]$ is bounded uniformly in $j \geq 1$ for $q = 2r > 8$, and $\mathbb{E}[\|\partial_\psi y_t(\theta_0;\psi_k)\|_\infty^2] \leq O(k^{2/q})$. We have $\|\partial_\psi y_t(\theta_0;\psi_k)'W\sqrt{n}u_{t,k}\| \leq \overline{\lambda}_W\|\partial_\psi y_t(\theta_0;\psi_k)\|_\infty\|\sqrt{n}u_{t,k}\|$ so that with the Cauchy-Schwarz inequality: $\mathbb{E}[\|\partial_\psi y_t(\theta_0;\psi_k)'W\sqrt{n}u_{t,k}\|] \leq O(k^{1/q})o_p(n^{-1/(4r)}[\log(n)]^{-2})$, where $k = o(n^{1/4})$ by Assumption 5. Since $\partial_\theta y_t(\theta_0;\psi_k)$ is of fixed dimension, the same holds. Eventually, we get: $\|\sqrt{n}(\hat{\psi}_{nk}-\psi_k)'\Big\{\frac{1}{n}\sum_{t=1}^{n} B_{n,t}'W_n\sqrt{n}u_{t,k}\Big\}\| \leq O_p(\sqrt{\log(n)})o_p([\log(n)]^{-2}) = o_p([\log(n)]^{-3/2})$. Overall, we find $(Q2) = o_p([\log(n)]^{-3/2})$. Following the same steps as for (Q2), we can show that $\|(Q3)\| \leq o_p([\log(n)]^{-3/2})$. Overall, we get: $nQ_n(\hat{\theta}_n;\hat{\psi}_{nk}) = n(\hat{\psi}_{nk}-\psi_k)'M_{n,k}(\hat{\psi}_{nk}-\psi_k)+o_p([\log(n)]^{-3/2})$. The last part of this step is to show $M_{n,k} = M_k + o_p([\log(n)]^{-1})$. This is similar to the derivations for M_n in the proof of Theorem 2, but now the matrix has $O(k^2)$

elements instead of a fixed dimension. Write $M_{n,k}$ as follows:

$$\begin{aligned}
M_{n,k} = {} & \frac{1}{n}\sum_{t=1}^{n} \partial_\psi \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\theta_0; \hat{\psi}_{nk}, \psi_k)' W_n \partial_\psi \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\theta_0; \hat{\psi}_{nk}, \psi_k) && \text{(M1)} \\
& + [D_n + E_n]' M_n^{-1} \left\{ \frac{1}{n}\sum_{t=1}^{n} \partial_\theta \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\hat{\theta}_n, \theta_0; \hat{\psi}_{nk})' W_n \partial_\theta \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\hat{\theta}_n, \theta_0; \hat{\psi}_{nk}) \right\} M_n^{-1} [D_n + E_n] && \text{(M2)} \\
& + [D_n + E_n]' M_n^{-1} \left\{ \frac{1}{n}\sum_{t=1}^{n} \partial_\theta \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\hat{\theta}_n, \theta_0; \hat{\psi}_{nk})' W_n \partial_\psi \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\theta_0; \hat{\psi}_{nk}, \psi_k) \right\} && \text{(M3)} \\
& + \left\{ \frac{1}{n}\sum_{t=1}^{n} \partial_\psi \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\theta_0; \hat{\psi}_{nk}, \psi_k)' W_n \partial_\theta \bar{\boldsymbol{y}}_{\boldsymbol{t}}(\hat{\theta}_n, \theta_0; \hat{\psi}_{nk}) \right\} M_n^{-1} [D_n + E_n]. && \text{(M4)}
\end{aligned}$$

Note that $M_{n,k}$ depends on estimates $\hat{\theta}_n, \hat{\psi}_{nk}$, and $\bar{\psi}_k = \omega\hat{\psi}_{nk} + (1-\omega)\psi_k$, while M_k depends on the true values θ_0 and ψ_k. We now show that the former parameter values can be replaced by the later with a bounded error term. Specifically, for any intermediate value $\bar{\psi}_k$ of $\hat{\psi}_{nk}$ and ψ_k, we have, for $t \geq k+1$: $\partial_{\tilde{\mu}} \text{vec}[\partial_{\tilde{\mu}} y_t(\theta; \bar{\psi}_k)] = 0$, $\partial_{\phi_j} \text{vec}[\partial_{\tilde{\mu}} y_t(\theta; \bar{\psi}_k)] = -I_d \otimes [\sum_{\ell=0}^{\infty} \Lambda_\ell(\theta_0) P(\theta_0; \bar{\Sigma})]$, which is bounded; $\partial_{\text{vech}[\tilde{\Sigma}]} \text{vec}[\partial_{\tilde{\mu}} y_t(\theta; \bar{\psi}_k)] = [I_d - \sum_{\ell=1}^{k} \bar{\Psi}_\ell]' \otimes \left[\sum_{\ell=0}^{t-1} \Lambda_\ell(\theta_0)\right] \partial_{\text{vech}} \text{vec}[P(\theta_0; \bar{\Sigma})]$, which is also bounded with probability approaching 1 since $\|\bar{\psi}_k - \psi_k\|_\infty = O_p(\sqrt{\log(n)/n})$. For any $1 \leq j, \ell \leq k$, $\partial_{\phi_j} \text{vec}[\partial_{\phi_\ell} y_t(\theta; \bar{\psi}_k)] = 0$ and $\partial_{\text{vech}[\tilde{\Sigma}]_i} \text{vec}[\partial_{\phi_j} y_t(\theta; \bar{\psi}_k)] = -\sum_{\ell=0}^{t-1} [\tilde{y}_{t-\ell-j-1} - \tilde{\mu}]' \otimes [\Lambda_\ell(\theta_0) \partial_{\text{vech}[\tilde{\Sigma}]_i} P(\theta_0; \bar{\Sigma})]$. Using $(AC) \otimes (BD) = (A \otimes B)(C \otimes D)$ for conformable matrices, we get $\partial_{\text{vech}[\tilde{\Sigma}]_i} \text{vec}[\partial_{\phi_j} y_t(\theta; \bar{\psi}_k)] = \{-\sum_{\ell=0}^{t-1} [\tilde{y}_{t-\ell-j-1} - \tilde{\mu}]' [\Lambda_\ell(\theta_0)\} \{I_d \otimes \partial_{\text{vech}[\tilde{\Sigma}]_i} P(\theta_0; \bar{\Sigma})]\}$, and thus $\|\partial_{\text{vech}[\tilde{\Sigma}]_i} \text{vec}[\partial_{\phi_j} y_t(\theta; \bar{\psi}_k)]\| \leq \| \sum_{\ell=0}^{t-1} [\tilde{y}_{t-\ell-j-1} - \tilde{\mu}]' [\Lambda_\ell(\theta_0)\| \sup_{\tilde{\Sigma}} \partial_{\text{vech}[\tilde{\Sigma}]_i} \|P(\theta; \tilde{\Sigma})\|_\infty$, which has bounded q-th moment uniformly in t. The same steps reveal that $\|\partial_{\text{vech}[\tilde{\Sigma}]_i} \text{vec}[\partial_{\text{vech}[\tilde{\Sigma}]} y_t(\theta; \bar{\psi}_k)]\|$ also has bounded q-th moment, uniformly in t. Let $\Psi_{nk} = \{\bar{\psi}_k, \|\bar{\psi}_k - \psi_k\|_\infty \leq n^{-1/2} \log(n)\}$. The above results and Pisier (1983)'s inequality imply:[3] $(\mathbb{E}[\sup_{\bar{\psi}_k \in \Psi_{nk}} \|\partial_\psi^2 y_t(\theta_0; \bar{\psi}_k)\|_\infty^q])^{1/q} \leq O(k^{1/q})$, as the number of non-zero derivatives is linear in k. We now apply this upper bound to analyze the following decomposition for (M1), which holds with probability approaching 1:

$$\begin{aligned}
& \|(M1) - \frac{1}{n}\sum_{t=1}^{n} \partial_\psi y_t(\theta_0; \psi_k)' W_n \partial_\psi y_t(\theta_0; \psi_k)\| \\
\leq & 2\lambda_{\max}[W_n] \|\hat{\psi}_{nk} - \psi_k\|_\infty \frac{1}{n}\sum_{t=1}^{n} \left\{ \Big[\sup_{\bar{\psi}_k \in \Psi_{nk}} \|\partial_\psi^2 y_t(\theta_0; \bar{\psi}_k)\|_\infty \Big] \Big[\sup_{\bar{\psi}_k \in \Psi_{nk}} \|\partial_\psi y_t(\theta_0; \bar{\psi}_k)\|_\infty \Big] \right\}.
\end{aligned}$$

[3] Pisier (1983): $(\mathbb{E}[\sup_{j=1,\ldots,k} |X_j|^q])^{1/q} \leq k^{1/q} (\sup_{j=1,\ldots,k} \mathbb{E}[|X_j|^q])^{1/q}$ for any random variables $X_1, \ldots, X_k$ with finite q-th moment.

Among the right hand side terms, $\|\hat{\psi}_{nk}-\psi_k\|_\infty$ is $O_p(n^{-1/2}k^{-1/2})$, $(\mathbb{E}[\|\partial^2_\psi y_t(\theta_0;\bar{\psi}_k)\|^q_\infty])^{1/q}$ is $O(k^{1/q})$ as we have shown, and $(\mathbb{E}[\|\partial_\psi y_t(\theta_0;\bar{\psi}_k)\|^q_\infty])^{1/q} = O(k^{1/q})$ using similar derivations as for $\partial^2_\psi y_t(\theta_0;\bar{\psi}_k)$. Applying these results and the Cauchy-Schwarz inequality, the right-hand side of the inequality above is $O_p(n^{-1/2}k^{1/2}k^{1/2}\log(n)) = o_p(n^{-1/4}\log(n))$, since $k = o(n^{1/4})$ by Assumption. Similar results hold for (M2), (M3), and (M4) using the fact that $\|\hat{\theta}_n-\theta_0\| \leq n^{-1/2}\log(n)$ with probability approaching 1 (cf. Theorem 2). These results imply that replacing $\hat{\theta}_n, \hat{\psi}_{nk}$, and $\bar{\psi}_k = \omega\hat{\psi}_{nk}+(1-\omega)\psi_k$ by their true values θ_0,, ψ_k, and ψ_k impacts $M_{n,k}$ by no more than $o_p(n^{-1/4}\log(n))$.

An additional difference between $M_{n,k}$ and M_k is that $M_{n,k}$ depends on W_n, while M_k depends on W. Using similar arguments as above, the effect of this substitution is at most $o_p(n^{-1/4}\log(n))$ because $\|W_n - W\| = O_p(n^{-1/2})$. Therefore, we have: $M_{n,k} = \bar{M}_{n,k} + o_p(n^{-1/4}\log(n))$, with

$$
\begin{aligned}
\bar{M}_{n,k} = {} & (1/n)\sum\nolimits_{t=1}^{n} \partial_\psi y_t(\theta_0;\psi_k)' W \partial_\psi y_t(\theta_0;\psi_k) && \text{(M1')}\\
& + [D_n+E_n]' M_n^{-1}\{(1/n)\sum\nolimits_{t=1}^{n} \partial_\theta y_t(\theta_0;\psi_k)' W \partial_\theta y_t(\theta_0;\psi_k)\} M_n^{-1}[D_n+E_n] && \text{(M2')}\\
& + [D_n+E_n]' M_n^{-1}\{(1/n)\sum\nolimits_{t=1}^{n} \partial_\theta y_t(\theta_0;\psi_k)' W \partial_\psi y_t(\theta_0;\tilde{\psi}_{k,t})\} && \text{(M3')}\\
& + \{(1/n)\sum\nolimits_{t=1}^{n} \partial_\psi y_t(\theta_0;\psi_k)' W \partial_\theta y_t(\theta_0;\psi_k)\} M_n^{-1}[D_n+E_n]. && \text{(M4')}\\
& + o_p(n^{-1/4}\log(n)).
\end{aligned}
$$

This implies: $nQ_n(\hat{\theta}_n;\hat{\psi}_{nk}) = n(\hat{\psi}_{nk}-\psi_k)'\bar{M}_{n,k}(\hat{\psi}_{nk}-\psi_k) + o_p([\log(n)]^{-3/2})$.

Now, we derive a law of large numbers for (M1')-(M4'). We will only consider (M1') since the others are similar. For this, we first derive a bound for $\partial_{\phi_j} y_t(\theta_0;\psi_k) - \partial_{\phi_j} y_t(\theta_0;\psi_{k0})$, so that we eventually can use $\mathbb{E}[\partial_\psi y_t(\theta_0;\psi_{k0})' W \partial_\psi y_t(\theta_0;\psi_{k0})]$ to approximate (M1'). Note that $\partial_{\phi_j} y_t(\theta_0;\psi_k)$ sets $\tilde{y}_t - \tilde{\mu} = 0$ for $t \leq 0$, while the latter uses the actual, unobserved, $\tilde{y}_t - \tilde{\mu}$ for $t \leq 0$. As a result, the latter is stationary. For $t \geq 2k+1$ and any $j \in \{1,\dots,k\}$, we have

$$(\mathbb{E}[\|\partial_{\phi_j} y_t(\theta_0;\psi_k) - \partial_{\phi_j} y_t(\theta_0;\psi_{k0})\|^2])^{1/2} \leq \sum\nolimits_{\ell=k+1}^{\infty} \|\Lambda_j(\theta_0)\|_\infty \|P(\theta_0;\tilde{\Sigma})\|_\infty (\mathbb{E}[\|\tilde{y}_{t-j-\ell}\|^2])^{1/2} = o(n^{-1/2}),$$

uniformly in j. Also, $(\mathbb{E}[\|\partial_\psi y_t(\theta_0;\psi_k) - \partial_\psi y_t(\theta_0;\psi_{k0})\|^2_\infty])^{1/2} = o(n^{-1/2}k^{1/2})$. By the Cauchy-Schwarz inequality:

$$\frac{1}{n}\sum_{t=1}^{n} \partial_\psi y_t(\theta_0;\psi_k)' W \partial_\psi y_t(\theta_0;\psi_k) = \frac{1}{n}\sum_{t=1}^{n} \partial_\psi y_t(\theta_0;\psi_{k0})' W \partial_\psi y_t(\theta_0;\psi_{k0}) + o_p(n^{-1/2}k) = o_p(n^{-1/4}).$$

For any $m \geq 2k$, let $\mathcal{F}_{t-m}^{t}$ be the σ-field generated by $(e_t, \ldots, e_{t-m})$. We have:

$$\partial_{\phi_j} y_t(\theta_0; \psi_{k0}) - \mathbb{E}[\partial_{\phi_j} y_t(\theta_0; \psi_{k0}) | \mathcal{F}_{t-m}^{t}] = \sum_{\ell=0}^{\infty} \Lambda_\ell(\theta_0) P(\theta_0; \tilde{\Sigma}) \sum_{s=0}^{\infty} \tilde{\Lambda}_s [e_{t-s-j-\ell} - \mathbb{E}(e_{t-s-j-\ell} | \mathcal{F}_{t-m}^{t})],$$

for any $j \in \{1, \ldots, k\}$, where $[e_{t-s-j-\ell} - \mathbb{E}(e_{t-s-j-\ell} | \mathcal{F}_{t-m}^{t})] = 0$ for all $s + j + \ell \leq m$. Since $m \geq 2k$ and $j \leq k$, this representation holds for all $j + \ell \leq m - k$ with $m - k \geq k \to \infty$. Using this representation, an upper bound can be derived as:

$$\begin{aligned}
&(\mathbb{E}[\|\partial_{\phi_j} y_t(\theta_0; \psi_{k0}) - \mathbb{E}[\partial_{\phi_j} y_t(\theta_0; \psi_{k0}) | \mathcal{F}_{t-m}^{t}]\|^q])^{1/q} \\
&\leq \|P(\theta_0; \tilde{\Sigma})\|_\infty \sum_{\ell=0}^{\infty} \{\|\Lambda_\ell(\theta_0)\|_\infty \sum_{s=(m-k)-\ell}^{\infty} \|\tilde{\Lambda}_s\|_\infty\} \|e_t\|_q \\
&\leq \|P(\theta_0; \tilde{\Sigma})\|_\infty \Big\{ \underbrace{\sum_{\ell=0}^{[(m-k)/2]} \{\|\Lambda_\ell(\theta_0)\|_\infty \sum_{s=k-\ell}^{\infty} \|\tilde{\Lambda}_s\|_\infty\}}_{\leq \sum_{\ell=0}^{\infty} \|\Lambda_\ell(\theta_0)\|_\infty C[(m-k)/2]^{-(b+\varepsilon)}} + \underbrace{\sum_{\ell=[(m-k)/2]+1}^{\infty} \|\Lambda_\ell(\theta_0)\|_\infty \{ \sum_{s=(m-k)-\ell}^{\infty} \|\tilde{\Lambda}_s\|_\infty\}}_{\leq \sum_{s=0}^{\infty} \|\tilde{\Lambda}_s\|_\infty C[(m-k)/2]^{-(b+\varepsilon)}} \Big\} \|e_t\|_q,
\end{aligned}$$

which is a $O(m^{-(b+\varepsilon)}) \leq O(k^{-(b+\varepsilon)})$. Similar derivations apply to $\partial_{\tilde{\mu}} y_t(\theta_0; \psi_{k0})$ and $\partial_{\text{vech}[\tilde{\Sigma}]} y_t(\theta_0; \psi_{k0})$. Altogether, we get: $(\mathbb{E}[\|\partial_\psi y_t(\theta_0; \psi_{k0}) - \mathbb{E}[\partial_\psi y_t(\theta_0; \psi_{k0}) | \mathcal{F}_{t-m}^{t}]\|_\infty^q])^{1/q} \leq O(k^{-(b+\varepsilon)+1/q})$. Using Cauchy-Schwarz inequality, we find:

$$(M1') = (1/n) \sum\nolimits_{t=1}^{n} \mathbb{E}[\partial_\psi y_t(\theta_0; \psi_{k0}) | \mathcal{F}_{t-m}^{t}]' W \mathbb{E}[\partial_\psi y_t(\theta_0; \psi_{k0}) | \mathcal{F}_{t-m}^{t}] + O_p(k^{1-(b+\varepsilon)}),$$

and $O_p(k^{1-(b+\varepsilon)}) = o_p([\log(n)]^{-2})$ since $b \geq 2$ and $[\log(n)]^2/k = o(1)$. Because e_t are strong-mixing with coefficients $\alpha(\cdot)$, and $B_{t,m} = \mathbb{E}[\partial_\psi y_t(\theta_0; \psi_{k0}) | \mathcal{F}_{t-m}^{t}]' W \mathbb{E}[\partial_\psi y_t(\theta_0; \psi_{k0}) | \mathcal{F}_{t-m}^{t}]$ is a function of $(e_t, \ldots, e_{t-m})$, it is mixing with coefficients $\alpha([\cdot - m]^+)$, where $[j - m]^+ = \max(j - m, 0)$. For any scalar element ℓ, we can bound its autocovariances as: $|\text{cov}(B_{t,m,\ell}, B_{t-s,m,\ell})| \leq 6\alpha([s-m]^+)^{1-1/p-1/r} \|B_{t,m,\ell}\|_p \|B_{t,m,\ell}\|_r$, using Davidson (2021, Cor15.3). Since $B_{t,m,\ell}$ has $q/2$ finite moments, we can set $p = r = 4$ in the above inequality. Note that $\|B_{t,m,\ell}\|_4$ is bounded uniformly in ℓ, m, k. Apply Chebyshev's inequality:

$$\begin{aligned}
\text{var}\left(\frac{1}{n} \sum_{t=1}^{n} B_{t,m,\ell}\right) &\leq \frac{\text{var}(B_{t,m,\ell})}{n} + \frac{2}{n} \sum_{s=1}^{n-1} |\text{cov}(B_{t,m,\ell}, B_{t-s,m,\ell})| \\
&\leq \frac{\text{var}(B_{t,m,\ell})}{n} + \frac{m}{n} 12\alpha(0)^{1/2} + \frac{12}{n} \sum_{s=m+1}^{n-1} \alpha(s-m)^{1/2} \|B_{t,m,\ell}\|_4^2 \leq O(k/n),
\end{aligned}$$

where the last inequality holds because $\text{var}(B_{t,m,\ell})$ is finite and, because e_t is strong-mixing with size $-a$ where $a > 2$, the $\alpha(s)^{1/2}$ are summable over $s \geq 1$. Since $B_{t,m}$ has a $O(k^2)$ elements, we get: $\|\frac{1}{n}\sum_{t=1}^{n} B_{t,m} - \mathbb{E}(B_{t,m})\|_\infty \leq O_p(k^{3/2}n^{-1/2}) = o_p([\log(n)]^{-2})$. This implies

$$(M1') = \mathbb{E}\{\mathbb{E}[\partial_\psi y_t(\theta_0;\psi_{k0})|\mathcal{F}^t_{t-m}]'W\mathbb{E}[\partial_\psi y_t(\theta_0;\psi_{k0})|\mathcal{F}^t_{t-m}]\} + o_p([\log(n)]^{-2}).$$

The right hand side satisfies:

$$\begin{aligned}
&\|\mathbb{E}\{\mathbb{E}[\partial_\psi y_t(\theta_0;\psi_{k0})|\mathcal{F}^t_{t-m}]'W\mathbb{E}[\partial_\psi y_t(\theta_0;\psi_{k0})|\mathcal{F}^t_{t-m}]\} - \mathbb{E}[\partial_\psi y_t(\theta_0;\psi_{k0})'W\partial_\psi y_t(\theta_0;\psi_{k0})]\| \\
&\leq \|\mathbb{E}\{\|(\partial_\psi y_t(\theta_0;\psi_{k0}) - \mathbb{E}[\partial_\psi y_t(\theta_0;\psi_{k0})|\mathcal{F}^t_{t-m}])'W(\partial_\psi y_t(\theta_0;\psi_{k0}) + \mathbb{E}[\partial_\psi y_t(\theta_0;\psi_{k0})])\|\} \\
&\leq \overline{\lambda}_W O(k^{1-(b+\varepsilon)})O(k^{1/2}) = o(k^{-1/2}) = o([\log(n)]^{-2}).
\end{aligned}$$

Putting everything together, we find the desired result:

$$(M1') = \mathbb{E}[\partial_\psi y_t(\theta_0;\psi_{k0})'W\partial_\psi y_t(\theta_0;\psi_{k0})] + o_p([\log(n)]^{-2}).$$

Similar derivations apply to (M2'), (M3'), (M4') so that we have:

$$nQ_n(\hat\theta_n;\hat\psi_{nk}) = n(\hat\psi_{nk} - \psi_k)'M_k(\hat\psi_{nk} - \psi_k) + o_p([\log(n)]^{-1}),$$

where M_k equals

$$\mathbb{E}\Big[\big(\partial_\psi y_t(\theta_0;\psi_{k0}) + \partial_\theta y_t(\theta_0;\psi_k)M^{-1}J_k\big)'W\big(\partial_\psi y_t(\theta_0;\psi_{k0}) + \partial_\theta y_t(\theta_0;\psi_k)M^{-1}J_k\big)\Big], \qquad (M_k)$$

with $J_k = D_k + E_k$, $D_k = -\mathbb{E}[(u_t'W \otimes I)\partial_\psi G_t(\theta_0;\psi_{k0})] = 0$, since $u_t = 0$, and $E_k = -\mathbb{E}[\partial_\theta y_t(\theta_0;\psi_{k0})'W\partial_\psi y_t(\theta_0;\psi_{k0})]$.

Step 3. Further expanding $nQ_n(\hat\theta_n;\hat\psi_{nk})$ to obtain: $nQ_n(\hat\theta_n;\hat\psi_{nk}) = n\bar{Z}_{n,k}'M_k\bar{Z}_{n,k} + o_p(1)$, with $\bar{Z}_{n,k} = 1/n\sum_{t=1}^{n} Z_{k,t}$ for $Z_{k,t}$ defined in Theorem 2.

In this step, the goal is to replace $(\hat\psi_{nk} - \psi_k)$ with $\bar{Z}_{n,k}$ in the final approximation of $nQ_n(\hat\theta_n;\hat\psi_{nk})$ in step 2. The main difference with the proof of Theorem 2, which relied on existing results, is that we need more refined result regarding the order of the difference $(\hat\psi_{nk} - \psi_k) - \bar{Z}_{n,k}$. For $\tilde\mu_n$ and vech$(\tilde\Sigma_{nk})$ this is immediate, so the main focus is on the autoregressive coefficients:

$$(\hat\phi_1,\ldots,\hat\phi_k) - (\phi_1,\ldots,\phi_k) = (1/n)\sum_{t=1}^{n} \mathrm{vec}\left[e_{t,k}\tilde{Y}_{t-1,nk}'\hat\Gamma_{nk}^{-1}\right],$$

where $\tilde{Y}_{t-1,nk} = ((\tilde{y}_{t-1} - \tilde\mu_n)',\ldots,(\tilde{y}_{t-k} - \tilde\mu_n)')'$ and $\hat\Gamma_{nk} = (1/n)\sum_{t=1}^{n}\tilde{Y}_{t-1,nk}\tilde{Y}_{t-1,nk}'$.

Let $\tilde{Y}_{t-1,k}$ and $\hat\Gamma_k$ be the same as $\tilde{Y}_{t-1,nk}$ and $\hat\Gamma_{nk}$, but with $\tilde\mu$ replacing $\tilde\mu_n$. We want to show that the $\tilde{Y}_{t-1,nk}$, $\hat{\Gamma}_{nk}$, and $e_{t,k}$ in the above displayed expression can be replaced by $\tilde{Y}_{t-1,k}$, $\hat{\Gamma}_k$, and e_t with negligible error. To this end, note that $\|\tilde{Y}_{t-1,nk} - \tilde{Y}_{t-1,k}\|_\infty = \|\tilde{\mu}_n - \tilde{\mu}\|_\infty = O_p(n^{-1/2})$ since $\tilde{y}_t$ is NED in L_2-norm with appropriate size. Likewise, for $1_k = (1, \dots, 1)$ of size k: $\hat{\Gamma}_{nk} = \hat{\Gamma}_k + [1/n \sum_{t=1}^n \tilde{Y}_{t-1,k}][(\tilde{\mu}_n - \tilde{\mu}) \otimes 1_k']'$ so that $\|\hat{\Gamma}_{nk} - \hat{\Gamma}_k\|_\infty = O_p(n^{-1})$. Also, using similar projection and mixing arguments as in step 2, we can show that for $m - k \geq k$ we have $\|\hat{\Gamma}_k - \Gamma_k\|_\infty \leq O_p(k(m/n)^{1/2}) + O_p(km^{-(b+\varepsilon)})$ since it has $O(k^2)$ elements, all with finite $q/4 \geq 4$ moments. Then, using an inequality between operator and sup-norm:[4] $\|\hat{\Gamma}_k - \Gamma_k\|_{op} \leq O_p(k^2(m/n)^{1/2}) + O_p(k^2 m^{-(b+\varepsilon)}) = o_p([\log(n)]^{-2})$ by setting $m = 2k$ and using the Assumptions. Apply Weyl's inequality (Horn and Johnson, 1991, Th3.3.16): $\lambda_{\min}(\hat{\Gamma}_{nk}) \geq \lambda_{\min}(\Gamma_k) - \|\hat{\Gamma}_{nk} - \Gamma_k\|_{op} = \lambda_{\min}(\Gamma_k) - o_p([\log(n)]^{-2})$. This means that $\|\hat{\Gamma}_{nk}^{-1}\|_{op}$ is bounded with probability approaching one.

Using this results, we can examine the effects of the substitutions. When substituting $\tilde{Y}_{t-1,nk}$ for $\tilde{Y}_{t-1,k}$, we have: $\frac{1}{n}\sum_{t=1}^n e_{t,k}\tilde{Y}_{t-1,nk}'\hat{\Gamma}_{nk}^{-1} = \frac{1}{n}\sum_{t=1}^n e_{t,k}\tilde{Y}_{t-1,k}'\hat{\Gamma}_{nk}^{-1} + O_p(n^{-1})$ since $\frac{1}{n}\sum_{t=1}^n e_{t,k} = \frac{1}{n}\sum_{t=1}^n e_t + o_p(n^{-1/2}) = O_p(n^{-1/2})$, using previous derivations and the strong-mixing properties of e_t. Next, for $t \geq k+1$: $\|e_{t,k} - e_t\|_p \leq \sum_{t=k+1}^\infty \|\Psi_j\|_{op}\|\tilde{y}_t - \tilde{\mu}\|_p = o_p([nk]^{-1/2}\log(n)^{-2})$ since $\sqrt{nk}\sum_{t=k+1}^\infty \|\Psi_j\|_{op} = o_p([\log(n)]^{-2})$. Also, $(\mathbb{E}[\|\tilde{Y}_{t-1,k}\|_\infty^2])^{1/2} \leq O(k^{1/2}) = o(n^{1/16}\sqrt{\log(n)})$. Thus, when substituting $e_{t,k}$ for e_t, we have : $\frac{1}{n}\sum_{t=1}^n e_{t,k}\tilde{Y}_{t-1,k}'\hat{\Gamma}_{nk}^{-1} = \frac{1}{n}\sum_{t=1}^n e_t\tilde{Y}_{t-1,k}'\hat{\Gamma}_{nk}^{-1} + o_p(n^{-7/8}) + o_p(k^{1/2}[nk]^{-1/2}[\log(n)]^{-2}) = o_p(n^{-1/2}[\log(n)]^{-2})$. The $o_p(n^{-7/8})$ term is due to the summation from $t = 1$ to k. Finally, we substitute $\hat{\Gamma}_{nk}^{-1}$ for Γ_k^{-1}. Using projection arguments as above: $\|\frac{1}{n}\sum_{t=1}^n e_t\tilde{Y}_{t-1,k}'\|_\infty \leq O_p(k^{1/2}(1/n)^{1/2}) + O_p(k^{1/2}m^{-(b+\varepsilon)})$. It is possible here to replace the (m/n) term with $1/n$ because of a martingale property $\mathbb{E}(e_t\tilde{Y}_{t-1,k}|\mathcal{F}_{t-1}) = 0$, so that the autocovariances, even after projection, are zero. Recall that $\hat{\Gamma}_{nk}^{-1} - \Gamma_k^{-1} = \hat{\Gamma}_{nk}^{-1}[\Gamma_k - \hat{\Gamma}_{nk}]\Gamma_k^{-1} = O_p(k^2(m/n)^{1/2}) + O_p(k^2 m^{-(b+\varepsilon)})$ since $\|\hat{\Gamma}_{nk}^{-1}\|_{op} \leq \|\Gamma_k^{-1}\|_{op} + o_p(1) = O_p(1)$. Thus: $\frac{1}{n}\sum_{t=1}^n e_t\tilde{Y}_{t-1,k}'\hat{\Gamma}_{nk}^{-1} - \frac{1}{n}\sum_{t=1}^n e_t\tilde{Y}_{t-1,k}'\hat{\Gamma}_k^{-1} = O_p(k^{5/2}m^{1/2}/n) + O_p(k^{5/2}m^{-(b+\varepsilon)})$. Pick $m = \max(2k, n^{1/4}[\log(n)]^{-1}) = o(n^{1/4})$ to find $k^{5/2}m^{1/2}/n = o(n^{-9/16}[\log(n)]^{-3})$ and $k^{5/2}m^{-(b+\varepsilon)} = o(n^{-19/16}[\log(n)]^{-6})$ for $b \geq 6$. Altogether, we get: $\sqrt{n}\|(\hat{\psi}_n - \psi_k) - \overline{Z}_{n,k}\|_\infty = o_p([\log(n)]^{-2})$.

Next, show that $(M_k)_{k\geq 1}$ is a sequence of bounded operators, i.e. $\|M_k\|_{op} \leq c$ for all $k \geq 1$. We will focus on (M1'), as the derivations for the others are similar. The elements of $\mathbb{E}[\partial_\psi y_t(\theta_0;\psi_{k0})' W \partial_\psi y_t(\theta_0;\psi_{k0})]$ take the form $\mathbb{E}[\partial_{\phi_\ell} y_t(\theta_0;\psi_{k0})' W \partial_{\phi_j} y_t(\theta_0;\psi_{k0})]$ or involve derivatives with respect to $\tilde{\mu}$ or vech$[\tilde{\Sigma}]$]. For $\ell < j$, apply the law of iterated expectations against the filtration $\mathcal{F}_{t-j}$. We get: $\mathbb{E}[\partial_{\phi_\ell} y_t(\theta_0;\psi_{k0})' W \partial_{\phi_j} y_t(\theta_0;\psi_{k0})] = \mathbb{E}[\mathbb{E}[\partial_{\phi_\ell} y_t(\theta_0;\psi_{k0})|\mathcal{F}_{t-j}]' W \partial_{\phi_j} y_t(\theta_0;\psi_{k0})]$. Repeating the derivations from the proof of The-

[4] For a matrix A of size $d \times d$, $\|A\|_{op} \leq d\|A\|_\infty$; here $d = O(k)$

orem 2 to bound $\|\mathbb{E}[\partial_{\phi_\ell} y_t(\theta_0;\psi_{k0})|\mathcal{F}_{t-j}]\|_2$ and applying the Cauchy-Schwarz inequality, we get: $\|\mathbb{E}[\partial_{\phi_\ell} y_t(\theta_0;\psi_{k0})' W \partial_{\phi_j} y_t(\theta_0;\psi_{k0})]\| \leq C_M(1+|\ell-j|)^{-(b+\varepsilon)}$, for some constant C_M which does not depend on k. Since $b \geq 2$, these upper bounds are absolutely summable over ℓ, the sum is bounded over j - and vice-versa. By Schur's Lemma (Jaffard, 1990, Lem1), this implies that $\|\mathbb{E}[\partial_{\phi_1,\ldots,\phi_k} y_t(\theta_0;\psi_{k0})' W \partial_{\phi_1,\ldots,\phi_k} y_t(\theta_0;\psi_{k0})]\|_{op} \leq C_M \sup_{j\geq 1}\sum_{\ell=1}^{\infty}(1+|\ell-j|)^{-(b+\varepsilon)} \leq C_M \sum_{\ell=-\infty}^{+\infty}(1+|\ell|)^{-(b+\varepsilon)}$. Similar bounds apply to the other derivatives, and as a result, $\|M_k\|_{op}$ is uniformly bounded. Now we have:

$$\begin{aligned}
|n(\hat{\psi}_n-\psi_k)'M_k(\hat{\psi}_n-\psi_k) - n\overline{Z}'_{n,k}M_k\overline{Z}_{n,k}| &= n|([\hat{\psi}_n-\psi_k]-\overline{Z}_{n,k})'M_k([\hat{\psi}_n-\psi_k]+\overline{Z}_{n,k})|,\\
&\leq n\|[\hat{\psi}_n-\psi_k]-\overline{Z}_{n,k}\|_\infty \|M_k\|_{op}[\|\hat{\psi}_n-\psi_k\|+\|\overline{Z}_{n,k}\|]\\
&\leq o_p([\log(n)]^{-2})O_p(\sqrt{k\log(n)}) = o_p(k^{1/2}[\log(n)]^{-3/2}),
\end{aligned}$$

where the last inequality holds because $\sqrt{n}\|\overline{Z}_{n,k}\|_\infty = O_p(k^{1/2})$, each element having finite and bounded second moment, and $\sqrt{n}\|\hat{\psi}_n-\psi_k\| \leq O(k^{1/2})\|\hat{\psi}_n-\psi_k\|_\infty$ using an inequality between the ℓ_2 and ℓ_∞ norms. This allows to conclude this step with:

$$nQ_n(\hat{\theta}_n;\hat{\psi}_{nk}) = n\overline{Z}'_{n,k}M_k\overline{Z}_{n,k} + o_p(k^{1/2}[\log(n)]^{-3/2}).$$

Step 4. Obtaining the strong Approximation: $n\overline{Z}'_{n,k}M_k\overline{Z}_{n,k} = n\mathcal{Z}'_{n,k}M_k\mathcal{Z}_{n,k} + o_p(1)$, for some $\sqrt{n}\mathcal{Z}_{n,k} \sim \mathcal{N}(0,S_{n,k})$.

In Lemma F2, we constructed a normally distributed random vector $\mathcal{Z}_{n,k}$ such that: $\sqrt{n}\|\overline{Z}_{n,k}-\mathcal{Z}_{n,k}\| = o_p([\log(n)]^{-2})$. Given that the elements of $\sqrt{n}\overline{Z}_{n,k}$ and $\sqrt{n}\mathcal{Z}_{n,k}$ have finite second moments, we also have: $\sqrt{n}\|\overline{Z}_{n,k}+\mathcal{Z}_{n,k}\|_\infty \leq O_p(k^{1/2})$, using Pisier's inequality. Now, using Hölder's inequality, we find:

$$\begin{aligned}
|n\overline{Z}'_{n,k}M_k\overline{Z}_{n,k} - n\mathcal{Z}'_{n,k}M_k\mathcal{Z}_{n,k}| &= |\sqrt{n}[\overline{Z}_{n,k}-\mathcal{Z}_{n,k}]'M_k\sqrt{n}[\overline{Z}_{n,k}+\mathcal{Z}_{n,k}]|\\
&\leq \sqrt{n}\|\overline{Z}_{n,k}-\mathcal{Z}_{n,k}\| \times \|M_k\sqrt{n}[\overline{Z}_{n,k}+\mathcal{Z}_{n,k}]\|_\infty\\
&= o_p(k^{1/2}[\log(n)]^{-2}),
\end{aligned}$$

since $\|M_k\|_\infty$ is uniformly bounded in k. Now we can conclude that: $nQ_n(\hat{\theta}_n;\hat{\psi}_{nk}) = \sqrt{n}\mathcal{Z}'_{n,k}M_k\sqrt{n}\mathcal{Z}_{n,k} + o_p(k^{1/2}[\log(n)]^{-3/2})$.

Step 5. Validating the rejection rate.

It is possible to re-write the leading terms in the preceding equation as follows (Buckley and Eagleson, 1988, p152): $n\mathcal{Z}'_{n,k}M_k\mathcal{Z}_{n,k} = \sum_{j=1}^{d(k)}\lambda_j(S_{n,k}^{1/2}M_kS_{n,k}^{1/2})W_j$, where $(W_1,\ldots,W_{d(k)})$ are iid χ^2_1 distributed and $\lambda_j(S_{n,k}^{1/2}M_kS_{n,k}^{1/2})$ are the eigenvalues of $S_{n,k}^{1/2}M_kS_{n,k}^{1/2}$, with $d(k) =$

$\dim(\psi_k)$. Then, we have $\mathbb{E}[n\mathcal{Z}_{n,k}' M_k \mathcal{Z}_{n,k}] = \sum_{j=1}^{d(k)} \lambda_j(S_{n,k}^{1/2} M_k S_{n,k}^{1/2}) = \text{trace}(S_{n,k}M_k)$ and $\text{var}[n\mathcal{Z}_{n,k}' M_k \mathcal{Z}_{n,k}] = 3\sum_{j=1}^{d(k)} \lambda_j^2(S_{n,k}^{1/2} M_k S_{n,k}^{1/2}) = 3\text{trace}([S_{n,k}M_k]^2)$. Both $S_{n,k}$ and M_k have eigenvalues bounded above by Schur's Lemma. This implies that the two traces are at most $O(k)$. If, in addition, $\text{trace}(S_{n,k}M_k) \geq O(k)$ and $\text{trace}([S_{n,k}M_k]^2) \geq O(k)$ then the Paley–Zygmund inequality implies:

$$\begin{aligned}\mathbb{P}\left(n\mathcal{Z}_{n,k}' M_k \mathcal{Z}_{n,k} > k^{1/2}\right) &\geq \left(1 - \frac{k^{1/2}}{\text{trace}(S_{n,k}M_k)}\right)^2 \frac{[\text{trace}(S_{n,k}M_k)]^2}{3\text{trace}([S_{n,k}M_k]^2) + [\text{trace}(S_{n,k}M_k)]^2} \\ &\geq \left(1 - O(k^{-1/2})\right)^2 \frac{1}{1 + O(k^{-1})} = 1 - o(1),\end{aligned}$$

as $n \to \infty$ so that the $o_p(k^{1/2}[\log(n)]^{-3/2})$ term is negligible. We can write:

$$\mathbb{P}\Big(nQ_n(\hat{\theta}_n; \hat{\psi}_{nk}) > c_{n,k}(1-\alpha)\Big) = \mathbb{P}\left(n\mathcal{Z}_{n,k}' M_k \mathcal{Z}_{n,k} > c_{n,k}(1-\alpha) + o_p(c_{n,k}(1-\alpha))\right) = \alpha + o(1),$$

which concludes the proof. □

Lemma F3. *Suppose the conditions for Theorem 3 are satisfied but the model is misspecified, i.e. for θ_0 defined in Theorem 1 $Q(\theta_0; \psi_0) > 0$, then:* $\lim_{n\to\infty} \mathbb{P}\Big(nQ_n(\hat{\theta}_n; \hat{\psi}_{nk}) > c_{n,k}(\alpha)\Big) = 1$, *where $c_{n,k}$ is defined in Theorem 3.*

Proof of Lemma F3: Under a fixed alternative, $Q_n(\hat{\theta}_n; \hat{\psi}_{nk}) \xrightarrow{p} Q(\theta_0; \psi_0) > 0$ by uniform convergence, derived in the proof of Theorem 1. Derivations in the proof of Theorem 3 imply $n\mathcal{Z}_{n,k}' M_k \mathcal{Z}_{n,k} \leq O_p(k) = o_p(n)$ so that $\mathbb{P}\Big(nQ_n(\hat{\theta}_n; \hat{\psi}_{nk}) > c_{n,k}(1-\alpha)\Big) = \mathbb{P}\Big(Q_n(\hat{\theta}_n; \hat{\psi}_{nk}) > o_p(1)\Big) \to 1$. This is the desired result. □

Appendix G Additional Details for Section 2

The auxiliary model used for the OTF is given by:

$$\tilde{y}_t = \delta_0 + \delta_1 t + \delta_2 \cos(2\pi t/n) + \dots + \delta_5 \cos(2\pi 4t/n) + \eta_t, \quad \eta_t = \rho_1 \eta_{t-1} + \dots \rho_4 \eta_{t-4} + e_t.$$

The model is estimated in two steps, first the data is de-trended using OLS. Then an AR(4) is fitted to the residuals $\hat{\eta}_t$ using OLS. The residuals $\hat{e}_t$ are then used in Algorithm 2. The model is then estimated by minimizing the loss Q_n as described in the main text.

Appendix H Pen & Pencil Example: MA Processes

This Appendix illustrates OT estimation for a tractable MA process. The model and data generating process are given by:

$$y_t = \mu + \xi_t + \lambda_1 \xi_{t-1}, \quad \tilde{y}_t = \tilde{\mu} + e_t + \sum\nolimits_{j=1}^{\infty} \tilde{\lambda}_j e_{t-j},$$

where $\xi_t \sim (0, \sigma^2)$ and $e_t \sim (0, \tilde{\sigma}^2)$ are serially uncorrelated. The transportation matrix $P = \sigma/\tilde{\sigma}$ is the ratio of the two standard deviations, and the coupling is given by $y_t = \mu + Pe_t + P\lambda_1 e_{t-1}$. Let $\theta = (\mu, \lambda_1, \sigma)$. Since the shocks are serially uncorrelated, we have:

$$Q(\theta; \psi_0) = \mathbb{E}(|y_t(\theta) - \tilde{y}_t|^2) = |\mu - \tilde{\mu}|^2 + \tilde{\sigma}^2 |1 - \sigma/\tilde{\sigma}|^2 + \tilde{\sigma}^2 |\tilde{\lambda}_1 - \sigma/\tilde{\sigma}\lambda_1|^2 + \sum\nolimits_{j=2}^{\infty} \tilde{\sigma}^2 \tilde{\lambda}_j^2,$$

which is minimized at $\theta_0 = (\tilde{\mu}, \tilde{\lambda}_1, \tilde{\sigma})$. Therefore, the transportation maintains the mean of the series, its first-order covariance, and the standard deviation of the shocks.

More generally, if the model for y_t is MA(q) with $q \geq 1$: $y_t = \mu + \xi_t + \sum_{j=1}^{q} \lambda_j \xi_{t-j}$, then $\theta_0 = (\mu, \lambda_1, \dots, \lambda_q, \sigma) = (\tilde{\mu}, \tilde{\lambda}_1, \dots, \tilde{\lambda}_q, \tilde{\sigma}_q)$ matches the mean and the first q MA coefficients. In particular, if the true DGP is an AR(1) with AR coefficient $\tilde{\rho}$, then its MA coefficients satisfy $\tilde{\lambda}_j = \tilde{\rho}^j$ for each j; y_t captures the impulse response of $\tilde{y}_t$ at horizons $h = 1$ to $h = q$.

Appendix I Additional Details for Section 7

I.1 Small New-Keynesian Model

Data. Estimation and testing in Section 7.1 use linearly detrended US log GDP, annualized inflation, and interest rates for the period 1960Q1-2007Q4.

I.2 Smets-Wouters Model

Data. We use US data from the same sample period of 1960:I-2007:IV as above.

I.3 Affine Term Structure Model

Data. Estimation relies on monthly data for 6 zero-coupon bond yields, with maturities of 1 month and 1-5 years. The bond yields (1-5 years) are obtained from the Fama CRSP zero-coupon files, and the 1-month yields come from the Fama CRSP Treasury Bill files. The sample period is from December 1952 to December 2000, as in Ang and Piazzesi (2003).

Table I1: LS Model Parameter Interpretation and Prior Weights.

		Determinacy Regime				Indeterminacy Regime			
			Prior Distribution				Prior Distribution		
θ	Parameter Interpretation	Bounds	DENSITY	MEAN	SD	Bounds	DENSITY	MEAN	SD
τ^{-1}	risk aversion	[1, 10]	Gamma	2.00	0.50	[1, 10]	Gamma	2.00	0.50
r^*	steady-state real interest rate	[1, 4]	Gamma	2.00	1.00	[1, 4]	Gamma	2.00	1.00
κ	Phillips curve slope	[0.1, 1]	Gamma	0.50	0.20	[0.1, 1]	Gamma	0.50	0.20
ψ_1	inflation target	[1.1, 5]	Gamma	1.10	0.50	[0.1,0.9]	Gamma	1.10	0.50
ψ_2	output target	[0, 0.99]	Gamma	0.25	0.13	[0.01,5]	Gamma	0.25	0.13
ρ_r	interest rate smoothing	[0.01, 0.9]	Beta	0.50	0.20	[0.01,0.9]	Beta	0.50	0.20
ρ_g	exog spending AR	[0.01, 0.99]	Beta	0.70	0.10	[0.01,0.99]	Beta	0.70	0.10
ρ_z	technology shock AR	[0.01, 0.99]	Beta	0.70	0.10	[0.01,0.99]	Beta	0.70	0.10
σ_r	monetary policy shock SD	[0.01, 3]	IGamma	0.31	0.16	[0.01, 3]	IGamma	0.31	0.16
σ_g	exog spending SD	[0.01, 3]	IGamma	0.38	0.20	[0.01, 3]	IGamma	0.38	0.20
σ_z	technology shock SD	[0.01, 3]	IGamma	1.00	0.52	[0.01, 3]	IGamma	1.00	0.52
ρ_{gz}	exog spending-technology cor	[-0.9, -0.9]	Normal	0.00	0.40	[-0.9, -0.9]	Normal	0.00	0.40
$M_{r\epsilon}$	sunspot-monetary coef	–	–	–	–	[-3,3]	Normal	0.00	1.00
$M_{g\epsilon}$	sunspot-exog spending coef	–	–	–	–	[-3,3]	Normal	0.00	1.00
$M_{z\epsilon}$	sunspot-technology coef	–	–	–	–	[-3,3]	Normal	0.00	1.00
σ_ϵ	sunspot shock SD	–	–	–	–	[0.01,3]	IGamma	0.25	0.13
π^*	steady-state inflation	[2, 10]	Gamma	4.00	2.00	[2,10]	Gamma	4.00	2.00

Legend: The prior follows Prior 1 specification of Lubik and Schorfheide (2004).

Table I2: LS Model: Parameter Estimates, Specification Test ($k = 2$ lags)

Parameter Estimates		Determinacy (Full Sample)			Indeterminacy (Pre-Volcker)			Determinacy (Post-Volcker)		
θ	Parameter Interpretation	EST	SD$_c$	SD$_r$	EST	SD$_c$	SD$_r$	EST	SD$_c$	SD$_r$
τ^{-1}	risk aversion	2.45	3.86	1.93	1.25	2.06	0.63	1.40	1.08	2.13
r^*	steady state real interest rate	1.86	0.23	0.30	0.98	0.58	0.48	2.44	0.59	4.91
κ	Phillips curve slope	0.49	0.73	0.41	0.58	2.07	0.30	0.46	0.49	46.68
ψ_1	inflation target	1.21	0.20	0.42	0.67	0.09	0.13	1.73	0.64	70.25
ψ_2	output target	0.15	0.71	0.76	0.15	0.67	0.16	0.18	1.45	192.48
ρ_r	interest rate smoothing	0.66	0.07	0.22	0.43	0.18	0.15	0.73	0.10	1.77
ρ_g	exog spending AR	0.88	0.05	0.03	0.72	0.37	0.11	0.91	0.06	0.79
ρ_z	technology shock AR	0.82	0.05	0.03	0.79	0.11	0.11	0.83	0.05	0.33
σ_r	monetary policy shock SD	0.28	0.05	0.05	0.23	0.07	0.06	0.26	0.07	16.21
σ_g	exog spending SD	0.16	0.04	0.05	0.29	0.87	0.16	0.19	0.09	0.56
σ_z	technology shock SD	1.33	0.27	0.18	1.07	0.52	0.29	0.91	0.21	9.41
ρ_{gz}	exog spending-technology cor	0.90	0.26	0.16	0.10	1.83	0.75	0.33	0.62	15.67
$M_{r\epsilon}$	sunspot-monetary coef	–	–	–	0.42	1.40	0.76	–	–	–
$M_{g\epsilon}$	sunspot-exog spending coef	–	–	–	-1.77	4.20	0.96	–	–	–
$M_{z\epsilon}$	sunspot-technology coef	–	–	–	0.67	0.81	0.19	–	–	–
σ_ϵ	sunspot shock SD	–	–	–	0.09	3.23	0.85	–	–	–
π^*	steady state inflation	4.04	0.62	0.61	5.11	1.69	1.50	3.79	0.74	1.13
Specification Test		STAT	10%	5%	STAT	10%	5%	STAT	10%	5%
All		121.6	66.5	89.7	58.2	63.6	97.0	69.4	89.0	127.0
Output		65.5	45.6	63.4	43.7	30.5	47.8	37.3	67.3	98.6
Inflation		33.1	12.2	15.6	7.0	21.4	32.7	12.8	12.7	16.6
Interest Rate		23.1	12.2	16.7	7.6	14.1	22.0	19.3	15.8	22.7

Legend: EST: parameter estimates $\hat{\theta}_n$. SD$_c$: standard errors assuming correct model specification. SD$_r$: misspecification-robust standard errors. STAT: $nQ_n(\hat{\theta}_n; \hat{\psi}_{nk})$. 10%, 5%: critical values for specification test at corresponding significance levels. All: specification test on all variables. Output, Inflation, Interest Rate: specification test on individual variables. $n = 192, 78, 114$ for the full, pre and post-Volcker samples.

Table I3: SW Model: parameters, bounds, and prior distribution

θ	Parameter Interpretation	Bounds	Prior Distribution		
			DENSITY	MEAN	SD
ρ_{ga}	Corr.: tech. and exog. spending shocks	[0.01, 2]	Normal	0.50	0.25
μ_w	Wage mark-up shock MA	[0.01, 0.99]	Beta	0.50	0.20
μ_p	Price mark-up shock MA	[0.01, 0.99]	Beta	0.50	0.20
α	Share of capital in production	[0.01, 1]	Normal	0.30	0.05
ψ	Elast. of capital utilization adjustment cost	[0.01, 1]	Beta	0.50	0.15
φ	Investment adjustment cost	[3,15]	Normal	4.00	1.50
σ_c	Elast. of inertemporal substitution	[1, 3]	Normal	1.50	0.38
λ	Habit persistence	[0.001, 0.99]	Beta	0.70	0.10
ϕ_p	Fixed costs in production	[1, 3]	Normal	1.25	0.13
ι_w	Wage indexation	[0.01, 0.99]	Beta	0.50	0.15
ξ_w	Wage stickiness	[0.5, 0.95]	Beta	0.50	0.10
ι_p	Price indexation	[0.01, 0.99]	Beta	0.50	0.15
ξ_p	Price stickiness	[0.1, 0.95]	Beta	0.50	0.10
σ_l	Labor supply elasticity	[1, 10]	Normal	2.00	0.75
r_π	Taylor rule: inflation weight	[1, 3]	Normal	1.50	0.25
$r_{\Delta y}$	Taylor rule: output gap change weight	[0.001, 0.5]	Normal	0.13	0.05
r_y	Taylor rule: output gap weight	[0.001, 0.5]	Normal	0.13	0.05
ρ	Taylor rule: interest rate smoothing	[0.5, 0.975]	Beta	0.75	0.10
ρ_a	Productivity shock AR	[0.01, 0.99]	Beta	0.50	0.20
ρ_b	Risk premium shock AR	[0.01, 0.99]	Beta	0.50	0.20
ρ_g	Exogenous spending shock AR	[0.01, 0.99]	Beta	0.50	0.20
ρ_i	Investment shock AR	[0.01, 0.99]	Beta	0.50	0.20
ρ_r	Monetary policy shock AR	[0.01, 0.99]	Beta	0.50	0.20
ρ_p	Price mark-up shock AR	[0.01, 0.99]	Beta	0.50	0.20
ρ_w	Wage mark-up shock AR	[0.001, 0.99]	Beta	0.50	0.20
σ_a	Productivity shock std. dev.	[0.01, 3]	IGamma	0.10	2.00
σ_b	Risk premium shock std. dev.	[0.025, 5]	IGamma	0.10	2.00
σ_g	Exogenous spending shock std. dev.	[0.01, 3]	IGamma	0.10	2.00
σ_i	Investment shock std. dev.	[0.01, 3]	IGamma	0.10	2.00
σ_r	Monetary policy shock std. dev.	[0.01, 3]	IGamma	0.10	2.00
σ_p	Price mark-up shock std. dev.	[0.01, 3]	IGamma	0.10	2.00
σ_w	Wage mark-up shock std. dev.	[0.01, 3]	IGamma	0.10	2.00
$\overline{\gamma}$	Trend growth: real GDP, Infl., Wages	[0.1, 0.8]	Normal	0.40	0.10
r	Discount rate	[0.01, 2]	Gamma	0.25	0.10
$\overline{\pi}$	Steady state inflation rate	[0.1, 2]	Gamma	0.62	0.10
$\overline{l}$	Steady state hours worked	[-10,10]	Normal	0.00	2.00

Legend: Prior distributions are taken from Smets and Wouters (2007) Dynare code.

Table I4: SW: Estimates, Rejection Rates (Monte Carlo)

θ	TRUE	MEAN	STD	REJ$_c$	REJ$_r$	LEN$_c$	LEN$_r$
ρ_{ga}	0.58	0.49	0.17	0.07	0.01	0.94	2.31
μ_w	0.90	0.75	0.11	0.07	0.02	0.37	1.05
μ_p	0.82	0.68	0.11	0.01	0.00	0.78	2.80
α	0.23	0.24	0.02	0.02	0.00	0.11	0.31
ψ	0.50	0.36	0.14	0.12	0.01	0.85	2.84
φ	6.15	4.80	1.09	0.11	0.01	8.03	19.52
σ_c	1.51	1.34	0.13	0.17	0.01	0.75	2.88
λ	0.71	0.69	0.07	0.03	0.00	0.32	1.37
ϕ_p	1.67	1.44	0.07	0.17	0.03	0.64	1.24
ι_w	0.53	0.60	0.12	0.00	0.00	1.31	2.87
ξ_w	0.78	0.68	0.09	0.10	0.01	0.44	1.50
ι_p	0.26	0.35	0.12	0.05	0.00	0.79	2.90
ξ_p	0.68	0.66	0.07	0.03	0.01	0.30	0.87
σ_l	2.27	1.52	0.46	0.04	0.00	4.67	17.77
r_π	2.04	1.71	0.18	0.11	0.01	1.53	2.80
$r_{\Delta y}$	0.21	0.17	0.03	0.03	0.00	0.24	0.69
r_y	0.10	0.08	0.03	0.10	0.01	0.18	0.43
ρ	0.83	0.76	0.06	0.04	0.00	0.28	0.65
ρ_a	0.97	0.93	0.09	0.04	0.01	0.07	0.36
ρ_b	0.28	0.37	0.14	0.12	0.01	0.43	1.10
ρ_g	0.97	0.83	0.14	0.24	0.04	0.30	0.84
ρ_i	0.70	0.68	0.08	0.07	0.03	0.34	0.55
ρ_r	0.17	0.38	0.17	0.18	0.04	0.83	1.33
ρ_p	0.96	0.94	0.04	0.01	0.03	0.12	0.26
ρ_w	0.96	0.89	0.10	0.00	0.01	0.18	0.33
σ_a	0.46	0.40	0.12	0.21	0.06	0.35	0.71
σ_b	0.23	0.17	0.04	0.53	0.07	0.11	0.29
σ_g	0.50	0.37	0.06	0.71	0.20	0.16	0.46
σ_i	0.41	0.37	0.07	0.20	0.04	0.23	0.34
σ_r	0.22	0.15	0.04	0.48	0.13	0.14	0.27
σ_p	0.12	0.09	0.02	0.14	0.02	0.11	0.29
σ_w	0.29	0.25	0.03	0.28	0.02	0.13	0.30
$\overline{\gamma}$	0.45	0.44	0.03	0.34	0.05	0.06	0.13
r	0.12	0.22	0.07	0.06	0.01	0.43	1.30
$\overline{\pi}$	0.69	0.67	0.18	0.17	0.13	0.62	0.69
$\bar{l}$	1.35	1.39	1.02	0.17	0.13	3.06	3.64

Legend: $n = 192$, $k = 4$, 200 Monte Carlo replications. Rejection rates for specification test (5% level): 0.02 for all variables, and $0.31, 0.10, 0.15, 0.02, 0.02, 0.06$, and 0.04, respectively

Table I5: SW: Estimates, Standard Errors under Stochastic Singularity

	6 shocks			5 shocks			4 shocks		
	EST	SD_1	SD_r	EST	SD_1	SD_r	EST	SD_1	SD_r
ρ_{ga}	0.45	0.24	2.70	0.47	0.18	0.25	0.50	0.12	0.17
μ_w	0.62	0.38	1.33	–	–	–	–	–	–
μ_p	0.98	0.06	0.91	0.81	0.15	0.26	–	–	–
α	0.25	0.03	0.11	0.29	0.03	0.09	0.28	0.02	0.23
ψ	0.63	0.31	0.80	0.45	0.35	0.37	0.34	0.35	0.26
φ	3.00	1.39	7.25	3.96	4.98	7.07	3.85	3.68	7.71
σ_c	1.37	0.28	2.16	1.00	0.02	0.05	1.00	0.04	0.04
λ	0.36	0.10	1.27	0.99	0.01	0.01	0.99	0.02	0.01
ϕ_p	1.55	0.23	0.52	1.39	0.20	0.73	1.32	0.13	1.48
ι_w	0.74	0.22	1.28	0.44	0.20	1.69	0.61	0.22	0.96
ξ_w	0.83	0.23	0.23	0.50	0.13	0.40	0.50	0.10	0.29
ι_p	0.20	0.15	1.98	0.41	0.20	0.79	0.27	0.11	0.55
ξ_p	0.61	0.09	0.65	0.54	0.12	0.41	0.35	0.09	0.92
σ_l	1.00	3.27	5.48	1.00	1.61	3.42	1.00	2.34	2.92
r_π	1.72	1.03	0.53	1.87	0.87	0.68	1.87	0.73	1.23
$r_{\Delta y}$	0.12	0.08	0.62	0.15	0.15	0.41	0.15	0.35	0.78
r_y	0.12	0.13	0.06	0.11	0.16	0.51	0.12	0.51	1.01
ρ	0.89	0.09	0.27	0.74	0.12	0.19	0.73	0.15	0.24
ρ_a	0.87	0.06	0.88	0.99	0.01	0.01	0.99	0.01	0.04
ρ_b	–	–	–	–	–	–	–	–	–
ρ_g	0.82	0.09	0.41	0.95	0.04	0.04	0.94	0.04	0.07
ρ_i	0.53	0.11	0.28	0.70	0.14	0.18	0.68	0.13	0.15
ρ_r	0.61	0.16	1.56	0.95	0.04	0.05	0.95	0.04	0.03
ρ_p	0.80	0.13	2.12	0.97	0.04	0.05	–	–	–
ρ_w	0.94	0.08	0.07	–	–	–	–	–	–
σ_a	0.36	0.06	0.40	0.42	0.07	0.19	0.47	0.07	0.32
σ_b	–	–	–	–	–	–	–	–	–
σ_g	0.35	0.03	0.08	0.41	0.05	0.08	0.40	0.05	0.10
σ_i	0.38	0.07	0.24	0.38	0.07	0.10	0.41	0.06	0.11
σ_r	0.07	0.03	0.47	0.06	0.05	0.05	0.06	0.04	0.07
σ_p	0.23	0.03	0.22	0.13	0.02	0.13	–	–	–
σ_w	0.04	0.04	0.12	–	–	–	–	–	–
$\overline{\gamma}$	0.45	0.01	0.05	0.46	0.02	0.04	0.46	0.02	0.05
r	0.12	0.13	0.91	0.21	0.08	0.12	0.22	0.08	0.13
$\overline{\pi}$	0.78	0.26	0.24	0.85	0.25	0.26	0.84	0.23	0.25
$\bar{l}$	0.21	0.89	0.90	0.62	0.91	1.03	0.58	0.85	1.39

Legend: The following shocks, and their associated parameters, are suppressed in the following order: risk premium (ρ_b, σ_b), wage markup $(\mu_w, \rho_w, \sigma_w)$, and price markup $(\mu_p, \rho_p, \sigma_p)$.

Table I6: SW model: Properties of Filtered Shock Processes

(a) Cross Correlation Between Shock Processes

				TRUE			
	TFP	Risk	Spending	Investment	Monetary	Price	Wage
TFP	1.00						
Risk	0.00	1.00					
Spending	0.37	0.00	1.00				
Investment	0.00	0.00	0.00	1.00			
Monetary	0.00	0.00	0.00	0.00	1.00		
Price	0.00	0.00	0.00	0.00	0.00	1.00	
Wage	0.00	0.00	0.00	0.00	0.00	0.00	1.00
				KALMAN FILTER			
	TFP	Risk	Spending	Investment	Monetary	Price	Wage
TFP	1.00						
Risk	-0.48	1.00					
Spending	0.58	-0.35	1.00				
Investment	0.16	-0.11	0.08	1.00			
Monetary	0.03	0.28	-0.00	0.40	1.00		
Price	-0.04	-0.08	-0.17	0.27	0.20	1.00	
Wage	0.19	-0.09	0.17	0.14	0.00	-0.02	1.00
				OT FILTER			
	TFP	Risk	Spending	Investment	Monetary	Price	Wage
TFP	1.00						
Risk	0.07	1.00					
Spending	0.41	0.05	1.00				
Investment	-0.12	0.03	-0.05	1.00			
Monetary	-0.00	0.03	0.03	-0.01	1.00		
Price	-0.02	0.04	-0.06	-0.00	-0.05	1.00	
Wage	0.12	-0.01	0.08	-0.01	-0.00	-0.01	1.00

(b) First-Order Autocorrelation of Shock Processes

	TFP	Risk	Spending	Investment	Monetary	Price	Wage
TRUE	0.94	0.68	0.86	0.47	0.47	0.42	0.04
KALMAN FILTER	0.97	0.62	0.97	0.83	0.09	-0.04	0.00
OT FILTER	0.91	0.66	0.84	0.45	0.49	0.29	0.05

Legend: Panel (a) presents pairwise sample and model-implied (TRUE) correlations among the 7 shock processes. The first 5 processes are AR(1) and the remaining 2 are ARMA(1,1). Panel (b) sample and model-implied (TRUE) first-order autocorrelations of the seven filtered shocks.