The SR Approach: A New Estimation Procedure for Non-Linear and Non-Gaussian Dynamic Term Structure Models*

Martin M. Andreasen† Aarhus University and CREATES

Bent Jesper Christensen Aarhus University and CREATES

August 2, 2014

Abstract

This paper suggests a new approach for estimating linear and non-linear dynamic term structure models with latent factors. We impose no distributional assumptions on the factors which therefore may be non-Gaussian. The novelty of our approach is to use many observables (yields or bond prices) in the cross-section dimension. This implies that the latent factors can be determined quite accurately by a sequence of cross-section regressions. We also show how output from these regressions can be used to obtain model parameters by a two- or three-step moment-based estimation procedure.

Keywords: Bond data, GMM, Non-linear filtering, Non-linear least squares, SMM.

JEL: C10, C30

*We thank Juan Rubio-Ramirez, Jesus Fernandez-Villaverde, Mikhail Chernov, George Kapetanios, Dennis Kristensen, and Anders Bredahl Kock for useful comments and discussions. We also thank Ron Gallant (the editor) and two anonymous referees for useful comments. Both authors gratefully acknowledge financial support from the Danish Center for Scientific Computing (DCSC). The authors also appreciate financial support to CREATES - Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation.

†Corresponding author: Department of Economics and Business, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus V, Denmark. Email: mandreasen@econ.au.dk. Telephone number: +45 8716 5982.

1 Introduction

Most term structure models rely on latent (i.e. unobserved) factors when pricing bonds (see Dai & Singleton (2000), Duffee (2002), among others). These factors often simplify the model, but they also make the estimation more difficult and time consuming, in particular for bond data. It is therefore common practice to extract 5 to 10 zero-coupon yields from bond prices, and use only these yields to estimate dynamic term structure models (DTSMs) (see Duan & Simonato (1999), de Jong (2000), among others).

The present paper deviates from the common practice of using relatively few observables in the cross-section dimension when estimating DTSMs. That is, we suggest to use a large number of observables (yields or bond prices), and we argue that this may in fact simplify estimation. This is because latent factors can be estimated quite accurately by a sequence of cross-section regressions when there are many observables in each time period. We refer to this procedure as the "regression filter" because it provides estimates of the latent factors, like any other filter. For non-linear models with normally distributed measurement errors, we show that the regression filter converges to the optimal estimates when the number of observables in the cross-section dimension tends to infinity. For linear and Gaussian models, this means that the regression filter converges to the Kalman smoother.

We next show how output from the regression filter may be used to determine parameters in DTSMs by a two- or three-step estimation procedure. For models with no-arbitrage restrictions, our approach implies that the risk-neutral parameters are estimated in the first step from a pooled non-linear cross-section regression, whereas parameters for the market price of risk are obtained in the second step by either Generalized Method of Moments (GMM) or Simulated Method of Moments (SMM). In models with linear factor dynamics under the physical measure, this second step reduces to a simple OLS estimation of a VAR system with generated regressors. Our third estimation step is optional and only introduced to increase the efficiency of the estimates in the first step. Given that we rely on sequential regressions to estimate DTSMs, we name our estimation procedure the sequential regression (SR) approach.

The SR approach is shown to generate consistent and asymptotically normal estimators of the latent factors and model parameters. Estimation of the asymptotic covariance matrices relies only on first-order derivatives of the model, and this greatly simplifies the implementation. We also emphasize that the asymptotic results hold for non-linear models with potentially non-Gaussian factor dynamics and with weaker restrictions than those needed for likelihood inference. For instance, the SR approach

allows measurement errors in yields or bond prices to be autocorrelated, and no transition function for the latent factors is required to estimate these factors and a subset of the model parameters. In contrast, exact likelihood inference is in general unavailable for non-linear and/or non-Gaussian DTSMs. The generality of the SR approach is an important property because it greatly facilitates estimation of a wide class of DTSMs in a simple and robust manner.

Another benefit of the SR approach is its computational simplicity, which makes it feasible to estimate DTSMs directly on coupon bonds. This is desirable because much information may be lost when adopting the common practice of using only extracted zero-coupon yields. In contrast, using bond data and existing filtering methods makes the estimation very difficult because it induces a non-linear filtering problem without a known likelihood function, even for linear models. One option is to approximate the unknown likelihood function by sequential Monte Carlo methods, but this procedure is extremely time consuming and therefore rarely attempted. Another option is to use a non-linear extension of the Kalman filter and quasi-maximum likelihood, but the asymptotic distribution is in this case generally unknown. In comparison, the non-linear filtering problem induced by coupon bonds and the associated parameter estimation is easily addressed in the SR approach, which has known asymptotic properties. A second reason that makes it challenging to use coupon bonds directly for the estimation is simply that some bonds mature while new bonds are issued. Accounting for such an unbalanced data panel in existing filtering methods is complicated and therefore rarely pursued, but easy to accommodate in the regression filter.

Our SR approach is related to Joslin, Singleton & Zhu (2011), who show how Maximum Likelihood (ML) estimation of Gaussian affine term structure models (ATSMs) with observed factors may be implemented by a simple two-step procedure. Using a moment-based method, our SR approach extends this two-step procedure to potentially non-linear and non-Gaussian models where all observables are measured with errors. Contrary to the ML procedure in Joslin et al. (2011), the SR approach is not fully efficient, but Monte Carlo evidence suggests that the efficiency loss may be small. For a three-factor Gaussian ATSM, the SR approach basically delivers the same efficiency as ML when using just 25 observables in the cross-section dimension. In the case of a two-factor quadratic term structure model (QTSM) with unknown likelihood function, we also find that the same number of bonds hardly implies any efficiency loss for the SR approach compared to a Quasi Maximum Likelihood (QML) estimator based on the central difference Kalman filter of Norgaard, Poulsen & Ravn (2000). Our Monte Carlo results for the CIR model by Cox, Ingersoll & Ross (1985) with stochastic volatility are

also encouraging, as the SR approach achieves nearly the same efficiency for most of the parameters as the QML estimator of Duan & Simonato (1999) based on a modified Kalman filter.

The SR approach requires a large number of observables in each time period and is therefore naturally applied to bond prices. Another possibility is to include the entire yield curve in each time period. However, this must be done with some care because an estimated static yield curve with n parameters only has an information content corresponding to n interest rates, given knowledge of the functional form. From this perspective, it is preferable to use non-parametric estimates of the yield curve, instead of interest rates from parametric specifications, as in Nelson & Siegel (1987) and any of its extensions (see for instance Björk & Christensen (1999)). Existing non-parametric procedures to obtain the yield curve include the methods by Fama & Bliss (1987), Tanggard (1997), and Linton, Mammen, Nielsen & Tanggard (2001). Subject to this qualification, estimated zero-coupon yields may also be used in the SR approach.

The rest of the paper is organized as follows. Section 2 presents the class of DTSMs that can be estimated by the SR approach. The first two steps in the SR approach are presented in Section 3, and their asymptotic properties are discussed in Section 4. Section 5 extends the SR approach with an optional third step and derives its asymptotic properties. We relate our new estimation approach to the existing literature in Section 6, and the subsequent section examines the finite sample properties of the SR approach for a three-factor Gaussian ATSM, a two-factor QTSM, and the CIR model. Concluding comments are provided in Section 8. All proofs are deferred to the appendix.[1]

2 Dynamic term structure models

This section presents the class of DTSMs considered. Let $\mathbf{y}_t$ of dimension $n_{y,t} \times 1$ contain observed yields or bond prices at time t. The dimension of $\mathbf{y}_t$ is specified to be time-dependent to accommodate an unbalanced data panel. The observed factors are denoted $\mathbf{x}_{1,t}$, while $\mathbf{x}_{2,t}$ contains the latent factors. These vectors have dimensions $n_{x_1} \times 1$ and $n_{x_2} \times 1$, respectively. Jointly, we let $\mathbf{x}_t \equiv \begin{bmatrix} \mathbf{x}_{1,t}' & \mathbf{x}_{2,t}' \end{bmatrix}'$ with dimension $n_x \times 1$ where $n_x = n_{x_1} + n_{x_2}$. The class of DTSMs considered has the representation

$$\mathbf{y}_t = \mathbf{g}\left(\mathbf{x}_t; \boldsymbol{\theta}_1\right) + \mathbf{v}_t, \tag{1}$$

[1] In addition, a longer technical appendix is available on the home page of the corresponding author or on request.

which implies that yields or bond prices are potentially non-linear functions of the factors $\mathbf{x}_t$ and the model parameters $\boldsymbol{\theta}_1$. In relation to (1), the following notation is adopted:

$$\mathbf{g}\left(\mathbf{x}_t;\boldsymbol{\theta}_1\right) \equiv \left[\begin{array}{cccc} g_1\left(\mathbf{x}_t;\boldsymbol{\theta}_1\right) & g_2\left(\mathbf{x}_t;\boldsymbol{\theta}_1\right) & ... & g_{n_{y,t}}\left(\mathbf{x}_t;\boldsymbol{\theta}_1\right) \end{array}\right]',$$

and similarly for $\mathbf{y}_t$. The $\mathbf{g}$-function is model-specific and may be derived from no-arbitrage restrictions or determined from other principles as in Diebold, Rudebusch & Aruoba (2006). The vector $\mathbf{v}_t \equiv \{v_{t,j}\}_{j=1}^{n_{y,t}}$ accounts for measurement errors in the observables. For bond prices, such errors can be caused by i) non-synchronous trading, ii) rounding of market prices, and/or iii) bid-ask spreads. If estimated zero-coupon yields are used as observables, then $\mathbf{v}_t$ captures errors arising when constructing these yields. We refer to (1) as the measurement equations.

The law of motion for the factors under the physical probability measure is given by

$$\mathbf{x}_{t+1} = \mathbf{h}\left(\mathbf{x}_t, \mathbf{w}_{t+1}; \boldsymbol{\theta}_{11}, \boldsymbol{\theta}_2\right), \tag{2}$$

where $\mathbf{w}_t$ has dimension $n_w \times 1$ and are $\mathcal{IID}$ mean-zero disturbances. The factor dynamics implied by (2) may be subject to normalization restrictions to ensure model identification. Another possibility in models with no-arbitrage constraints is to impose the normalization restrictions on the factor dynamics under the risk-neutral probability measure (see for instance Singleton (2006)). The SR approach can accommodate both normalization schemes, but it will typically be easier to implement when the dynamics under the physical measure is unrestricted, as illustrated in Section 3.3.

The model is parameterized by $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ of dimensions $L_1 \times 1$ and $L_2 \times 1$, respectively. We further let $\boldsymbol{\theta}_1 \equiv \left[\begin{array}{cc} \boldsymbol{\theta}_{11}' & \boldsymbol{\theta}_{12}' \end{array}\right]'$ and $\boldsymbol{\theta}_2 \equiv \left[\begin{array}{cc} \boldsymbol{\theta}_{22}' & \boldsymbol{\theta}_{12}' \end{array}\right]'$ to highlight that some elements in $\boldsymbol{\theta}_1$ must be identified from the measurement equations in (1), i.e. $\boldsymbol{\theta}_{11}$, while elements in $\boldsymbol{\theta}_{12}$ also may be determined from the factor dynamics in (2) jointly with $\boldsymbol{\theta}_{22}$.[2] For DTSMs, $\boldsymbol{\theta}_1$ typically includes all risk-neutral parameters and $\boldsymbol{\theta}_2$ contains parameters specifying the market price of risk. For general equilibrium models, the market price of risk is often a function of $\boldsymbol{\theta}_1$, and $\boldsymbol{\theta}_{22}$ may therefore be empty.

We impose the following assumptions on the class of models considered:

$$E\left[\mathbf{v}_t \middle| \mathbf{z}_t\right] = \mathbf{0} \quad \text{for some } \boldsymbol{\theta}_1^o \in \boldsymbol{\Theta}_1 \text{ and } \mathbf{x}_{2,t}^o \in \mathcal{X}_{2,t} \text{ for } t = 1, 2, ..., T \tag{3}$$

[2] Our specification in (2) presents the typical situation where $\boldsymbol{\theta}_{12}$ is re-estimated from (2). As an alternative, we may only estimate $\boldsymbol{\theta}_{22}$ from (2) and condition on the estimates of $\boldsymbol{\theta}_{12}$ from the measurement equations. This alternative specification is obtained by replacing $\boldsymbol{\theta}_2$ by $\boldsymbol{\theta}_{22}$ and $\boldsymbol{\theta}_{11}$ by $\boldsymbol{\theta}_1$ in (2).

$$Cov\left(\mathbf{v}_t, \mathbf{w}_{t-k}\right) = \mathbf{0} \text{ for } k = \pm 0, 1, 2, ... \text{ and } t = 1, 2, ..., T \tag{4}$$

$$Var\left(\mathbf{v}_t | \, \mathbf{z}_t\right) = \mathbf{\Omega}_{\mathbf{v}}\left(\mathbf{z}_t; \boldsymbol{\gamma}\right) \text{ is finite and positive definite for } t = 1, 2, ..., T \tag{5}$$

$$\left\{v_{t,j}\right\}_{j=1}^{n_{y,t}} \text{ is } \phi\text{- or } \alpha\text{-mixing along } j \text{ as } n_{y,t} \longrightarrow \infty \text{ for } t = 1, 2, ..., T \tag{6}$$

$$E\left[\left|v_{t,j}^2\right|^{r+\delta} | \mathbf{z}_t\right] \leq M_v < \infty \text{ for } t = 1, 2, ..., T \text{ and all } j \tag{7}$$

$$\left(g_j\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o\right) - g_j\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}; \boldsymbol{\theta}_1\right)\right)^2 \leq D_{t,j} \text{ for all } j \text{ and } \left(\mathbf{x}_{2,t}, \boldsymbol{\theta}_1\right) \in \left(\mathcal{X}_{2,t}, \mathbf{\Theta}_1\right) \text{ for all } t, \tag{8}$$

where $E\left[\left|D_{t,j}\right|^{r+\delta} | \mathbf{z}_t\right] \leq M_D < \infty$ for $r \geq 1$ and any $\delta > 0$. Superscript "o" on $\boldsymbol{\theta}_1$ and $\mathbf{x}_{2,t}$ denote their true value in the population, while $\mathbf{\Theta}_1$ and $\mathcal{X}_{2,t}$ refer to the feasible compact domains, respectively. The vector $\mathbf{z}_t$ contains $\mathbf{x}_{1,t}$ and other exogenous variables used below to model potential time-variation in the covariance matrix for the measurement errors. Our first assumption in (3) implies that the model is correctly specified for the conditional mean. This is a weak requirement also imposed by all existing estimation methods. The second condition of no correlation between $\mathbf{v}_t$ and $\mathbf{w}_{t-k}$ at all leads and lags in (4) is common for DTSMs and only imposed in our setting to facilitate estimation of $\boldsymbol{\theta}_2$.

The requirements in (5) and (6) place minimal constraints on the covariance matrix for the measurement errors as we allow for time-variation in all variances and covariances. Any systematic variation in the covariance matrix for the measurement errors is assumed to be controlled by $\mathbf{z}_t$ and the parameter $\boldsymbol{\gamma}$, where $\mathbf{\Omega}_{\mathbf{v}}\left(\mathbf{z}_t; \boldsymbol{\gamma}\right)$ is continuous in $\boldsymbol{\gamma}$. To obtain consistency and asymptotic normality of the SR approach even if the functional form of $\mathbf{\Omega}_{\mathbf{v}}\left(\mathbf{z}_t; \boldsymbol{\gamma}\right)$ is misspecified, the condition in (6) imposes weak requirements on the cross-sectional persistence. To formally present this condition, we start by ordering the observables in every time period t along a given dimension. The most obvious possibility is to apply bond duration, but other ordering schemes may be applied. Persistence along this dimension is then specified to be either ϕ-mixing or α-mixing as in White & Domowitz (1984). That is, let $\mathcal{F}\left(t\right)_a^b = \sigma\left(v_{t,j}; a \leq j \leq b\right)$ denote the Borel σ-algebra of events generated by $v_{t,a}, v_{t,a+1}, ..., v_{t,b}$, and consider the functions

$$\phi_t\left(m\right) \equiv \sup_n \sup_{\left\{F \in \mathcal{F}(t)_{-\infty}^n, G \in \mathcal{F}(t)_{n+m}^{\infty} : P(F) > 0\right\}} \left|P\left(G|F\right) - P\left(G\right)\right|$$

$$\alpha_t\left(m\right) = \sup_n \sup_{\left\{F \in \mathcal{F}(t)_{-\infty}^n, G \in \mathcal{F}(t)_{n+m}^{\infty}\right\}} \left|P\left(F, G\right) - P\left(F\right) P\left(G\right)\right|$$

for $t = 1, 2, .., T$. Both $\phi_t\left(m\right)$ and $\alpha_t\left(m\right)$ measure the cross-sectional dependence between events

generated by the measurement errors when these events are separated by at least m observables in the given ordering scheme. If $\phi_t(m) \longrightarrow 0$ for $m \longrightarrow \infty$, the sequence $\{v_{t,j}\}_{j=1}^{n_{y,t}}$ is ϕ-mixing, and similarly for α-mixing. Intuitively, these mixing conditions constrain the cross-sectional dependence and ensure that two events sufficiently apart are asymptotically independent.[3] Restrictions on the persistence of $v_{t,j}$ in the time series dimension is not needed for the first step of the SR approach because it relies on cross-section inference. Accordingly, our assumptions on the measurement errors are clearly less restrictive than the standard specification of uncorrelated errors with a diagonal covariance matrix.

Finally, the two regularity conditions in (7) and (8) are fairly weak and only needed for consistency and asymptotic normality of the first step in the SR approach when $v_{t,j}$ displays heteroskedasticity and cross-sectional dependence.

3 The SR approach

This section presents the first two steps in the SR approach. Section 3.1 deals with the estimation of the latent factors, while Sections 3.2 and 3.3 present estimators of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively. The first two steps in the SR approach are finally summarized in Section 3.4.

3.1 Estimation of latent factors: the regression filter

The SR approach is based on the simple idea of suppressing the time dimension when estimating latent factors in a setting with many observables in the cross-section dimension. A formal justification for this approach is provided in Section 6.1. That is, we suggest to estimate latent factors by solving the following sequence of generalized non-linear least squares (GNLS) regressions

$$\hat{\mathbf{x}}_{2,t}(\boldsymbol{\theta}_1) = \arg\min_{\mathbf{x}_{2,t}\in\mathcal{X}_{2,t}} Q_t = \frac{1}{2n_{y,t}} \left(\mathbf{y}_t - \mathbf{g}(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}; \boldsymbol{\theta}_1)\right)' \boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{z}_t; \boldsymbol{\gamma})^{-1} \left(\mathbf{y}_t - \mathbf{g}(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}; \boldsymbol{\theta}_1)\right) \tag{9}$$

for $t = 1, 2, ..., T$. Here, $\boldsymbol{\theta}_1$ and $\boldsymbol{\gamma}$ are constants but we explain in the next section how to also estimate these parameters along with the latent factors. The estimated latent factors from the t'th regression are denoted $\hat{\mathbf{x}}_{2,t}(\boldsymbol{\theta}_1)$ because they are computed for a given $\boldsymbol{\theta}_1$. When solving for and analyzing the

[3] As shown in White & Domowitz (1984), ϕ- and α-mixing imply that the autocovariances tend to zero for arbitrarily long lags. The conditions derived below for consistency and asymptotic normality using ϕ- and α-mixing require that the autocovariances tend to zero as a power of m. Such a requirement is satisfied e.g. by stationary ARMA processes.

properties of $\hat{\mathbf{x}}_{2,t}(\boldsymbol{\theta}_1)$, it is often convenient to use the Cholesky decomposition

$$\boldsymbol{\Omega}_{\mathbf{v}}(\boldsymbol{\gamma}, \mathbf{z}_t) = \mathbf{S}_{\mathbf{v}}(\mathbf{z}_t; \boldsymbol{\gamma}) \mathbf{S}_{\mathbf{v}}(\mathbf{z}_t; \boldsymbol{\gamma})' \tag{10}$$

and rewrite (9) as

$$\hat{\mathbf{x}}_{2,t}(\boldsymbol{\theta}_1) = \arg \min_{\mathbf{x}_{2,t} \in \mathcal{X}_{2,t}} Q_t = \frac{1}{2n_{y,t}} \sum_{j=1}^{n_{y,t}} \left(\tilde{y}_{t,j}(\mathbf{z}_t; \boldsymbol{\gamma}) - \tilde{g}_j(\mathbf{x}_{2,t}, \mathbf{z}_t; \boldsymbol{\theta}_1, \boldsymbol{\gamma}) \right)^2, \tag{11}$$

where $\tilde{\mathbf{y}}_t(\mathbf{z}_t; \boldsymbol{\gamma}) \equiv \mathbf{S}_{\mathbf{v}}(\mathbf{z}_t; \boldsymbol{\gamma})^{-1} \mathbf{y}_t$ and $\tilde{\mathbf{g}}(\mathbf{x}_{2,t}, \mathbf{z}_t; \boldsymbol{\theta}_1, \boldsymbol{\gamma}) \equiv \mathbf{S}_{\mathbf{v}}(\mathbf{z}_t; \boldsymbol{\gamma})^{-1} \mathbf{g}(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}; \boldsymbol{\theta}_1)$. Hence, we obtain a standard non-linear least squares (NLS) regression which can be solved with fast optimizers such as the Levenberg-Marquardt method, the Gauss-Newton minimizer, or various modifications of these routines. Note that estimated factors from the previous time period $\hat{\mathbf{x}}_{2,t-1}(\boldsymbol{\theta}_1)$ can be used as good starting values for the optimizations in time period $t = 2, 3, ..., T$. When the **g**-function is linear in $\mathbf{x}_{2,t}$, the problem in (11) reduces to an ordinary least squares (OLS) regression with a closed-form solution.

We also note that $\{\hat{\mathbf{x}}_{2,t}(\boldsymbol{\theta}_1)\}_{t=1}^T$ do not use the transition equations for the latent factors. This has at least two interesting implications. Firstly, given a consistent estimator of $\boldsymbol{\theta}_1$, the estimated factors $\{\hat{\mathbf{x}}_{2,t}(\boldsymbol{\theta}_1)\}_{t=1}^T$ are consistent regardless of their law of motion. Secondly, the SR approach may be applied to models set in continuous and discrete time.

3.2 Estimation of $\boldsymbol{\theta}_1$

We suggest to estimate $\boldsymbol{\theta}_1$ by pooling all squared residuals from (9) and minimizing their sum with respect to $\boldsymbol{\theta}_1$, i.e.

$$\hat{\boldsymbol{\theta}}_1 = \arg \min_{\boldsymbol{\theta}_1 \in \Theta_1} Q = \frac{1}{2N} \sum_{t=1}^{T} \sum_{j=1}^{n_{y,t}} \left(\tilde{y}_{t,j}(\mathbf{z}_t; \hat{\boldsymbol{\gamma}}) - \tilde{g}_j(\hat{\mathbf{x}}_{2,t}(\boldsymbol{\theta}_1), \mathbf{z}_t; \boldsymbol{\theta}_1, \hat{\boldsymbol{\gamma}}) \right)^2, \tag{12}$$

where $N \equiv \sum_{t=1}^T n_{y,t}$ and $\hat{\boldsymbol{\gamma}}$ denotes a consistent estimator of $\boldsymbol{\gamma}$. The estimator in (12) is similar to standard GNLS with the exception that changes in $\boldsymbol{\theta}_1$ affect the **g**-function both directly *and* indirectly through the latent factors $\{\hat{\mathbf{x}}_{2,t}(\boldsymbol{\theta}_1)\}_{t=1}^T$. Hence, when solving the problem in (12) iteratively, the latent factors must be recomputed for each trial value of $\boldsymbol{\theta}_1$. We also note that $\hat{\boldsymbol{\theta}}_1$, like $\{\hat{\mathbf{x}}_{2,t}\}_{t=1}^T$, does not use transition equations for the factors and therefore is robust to a wide class of factor dynamics.

Consistent estimates of $\{\mathbf{\Omega}_{\mathbf{v}}(\mathbf{z}_t;\boldsymbol{\gamma})\}_{t=1}^{T}$ are needed to make the regression filter and hence (12) operational. To use existing results from the regression literature, let $\hat{\boldsymbol{\theta}}_1$ denote the consistent estimator of $\boldsymbol{\theta}_1$ from (12). A first stage regression with $\mathbf{\Omega}_{\mathbf{v}} = \mathbf{I}_{n_{y,t}}$ returns consistent estimates of $\hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1\right)$ and the residuals $\hat{v}_{t,j} \equiv y_{t,j} - g_j\left(\mathbf{x}_{1,t}, \hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1\right); \hat{\boldsymbol{\theta}}_1\right)$. We then suggest to model potential time-variation in the variances based on bond characteristics such as duration, liquidity, etc. in time period t and run the cross-section regressions

$$\log\left(\hat{v}_{j,t}^2\right) = \boldsymbol{\beta}_t'\mathbf{z}_{t,j} + \varepsilon_{t,j} \quad \text{for } t = 1, 2, ..., T, \tag{13}$$

where $\mathbf{z}_{t,j}$ contains n_z explanatory variables. The predicted variances are then $\exp\left\{\hat{\boldsymbol{\beta}}_t'\mathbf{z}_{t,j}\right\}$. For this specification we therefore have $\hat{\boldsymbol{\gamma}} \equiv \left[\begin{array}{cccc} \hat{\boldsymbol{\beta}}_1' & \hat{\boldsymbol{\beta}}_2' & ... & \hat{\boldsymbol{\beta}}_T' \end{array}\right]'$. If $\left\{\hat{\boldsymbol{\beta}}_t'\right\}_{t=1}^{T}$ display stability across time, it may be reasonable to let $\boldsymbol{\beta}$ be the same for all time periods and instead estimate (13) by pooled OLS. Any time-variation in the variances would then be captured by variation in $\mathbf{z}_t$. For this second specification, we also note that $\mathbf{z}_t$, in addition to any bond characteristics, may contain time series variables because the coefficients on such regressors can be identified from the time dimension. Obvious time series variables to include are the short interest rate and some of the observed factors in $\mathbf{x}_{1,t}$.

To model potential time-variation in the cross-correlations, one possibility is to assume that measurement errors for closely related bonds display stronger comovement than less similar bonds. This type of dependence can be modeled as an AR(1) process where the error-covariance in time period t depends on the difference in duration between bonds, i.e.

$$Cov\left(\hat{v}_{t,j}, \hat{v}_{t,i}\right) = \rho_t^{|d_{t,j} - d_{t,i}|}, \tag{14}$$

with $|\rho_t| < 1$ and $d_{t,j}$ denoting the duration of bond j in time period t. Hence, $\boldsymbol{\gamma}$ is here extended by $\{\rho_t\}_{t=1}^{T}$. It is obvious that more sophisticated specifications for the covariance structure could be considered, e.g. a general ARMA(p, q) specification, provided $\mathbf{\Omega}_{\mathbf{v}}(\mathbf{z}_t;\boldsymbol{\gamma})$ remains positive definite.

3.3 Estimation of $\boldsymbol{\theta}_2$

The parameters $\boldsymbol{\theta}_2$ appear in the law of motion for the factors, and we therefore suggest estimating them based on the observed factors $\{\mathbf{x}_{1,t}\}_{t=1}^{T}$ and the estimated latent factors $\{\hat{\mathbf{x}}_{2,t}\}_{t=1}^{T}$. Here, we need to account for the fact that $\{\hat{\mathbf{x}}_{2,t}\}_{t=1}^{T}$ are estimated. As shown in Section 4, the estimation uncertainty associated with $\hat{\mathbf{x}}_{2,t}$ is normally distributed when we use a sufficient number of observables, i.e. when

$n_{y,t} \longrightarrow \infty$. We thus get the standard additive measurement error case (see Fuller (1987))

$$\hat{\mathbf{x}}_{2,t} = \mathbf{x}_{2,t}^{o} + \mathbf{u}_{2,t} \quad \mathbf{u}_{2,t} \sim \mathcal{N}(\mathbf{0}, Var(\mathbf{u}_{2,t})) \text{ for } t = 1, 2, ..., T. \tag{15}$$

We also show in Section 4 how to estimate $Var(\mathbf{u}_{2,t})$ and some time series properties of $\mathbf{u}_{2,t}$. At this point, we simply note that $\boldsymbol{\theta}_2$ can be estimated by well-known moment matching methods because the distribution of $\mathbf{u}_{2,t}$ and some of its statistical properties are known. To realize this, let us start by considering a simple VAR system, before outlining the general case. That is, we consider the system

$$\mathbf{x}_{t+1} = \mathbf{h}_0 + \mathbf{h}_{\mathbf{x}}\mathbf{x}_t + \mathbf{w}_{t+1}, \tag{16}$$

where $\mathbf{w}_t \sim \mathcal{IID}(\mathbf{0}, Var(\mathbf{w}_{t+1}))$, $\boldsymbol{\theta}_{22} \equiv [\mathbf{h}_0, \mathbf{h}_{\mathbf{x}}]$, and $\boldsymbol{\theta}_{12} \equiv vech(Var(\mathbf{w}_{t+1}))$. This system cannot be used to estimate $\boldsymbol{\theta}_2 = \begin{bmatrix} \boldsymbol{\theta}_{22}' & \boldsymbol{\theta}_{12}' \end{bmatrix}'$ because $\mathbf{x}_t$ may contain unobserved elements. Instead, we rely on

$$\hat{\mathbf{x}}_{t+1} = \mathbf{h}_0 + \mathbf{h}_{\mathbf{x}}\hat{\mathbf{x}}_t + \hat{\mathbf{w}}_{t+1}, \tag{17}$$

where $\hat{\mathbf{x}}_t$ contains observed and estimated latent factors. Notice that $\hat{\mathbf{w}}_t$ denotes the innovation in (17) using the true value of $\boldsymbol{\theta}_2$ and $\hat{\mathbf{x}}_t$.

3.3.1 A VAR system with purely latent factors

We initially assume that all factors are unobserved, i.e. $\mathbf{x}_t \equiv \mathbf{x}_{2,t}$. Using standard moment conditions for VAR systems on (17), it easily follows that

$$\begin{bmatrix} E(\hat{\mathbf{w}}_{t+1}) \\ E(\hat{\mathbf{w}}_{t+1}\hat{\mathbf{x}}_t') \\ Var(\hat{\mathbf{w}}_{t+1}) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ Cov(\mathbf{u}_{t+1}, \mathbf{u}_t) - \mathbf{h}_{\mathbf{x}} Var(\mathbf{u}_t) \\ Var(\mathbf{w}_{t+1}) + Var(\mathbf{u}_t) + \mathbf{h}_{\mathbf{x}} Var(\mathbf{u}_t) \mathbf{h}_{\mathbf{x}}' - Cov(\mathbf{u}_{t+1}, \mathbf{u}_t) \mathbf{h}_{\mathbf{x}}' - \mathbf{h}_{\mathbf{x}} Cov(\mathbf{u}_t, \mathbf{u}_{t+1}) \end{bmatrix}.$$

Replacing population moments by their empirical counterparts, we consider the conditions

$$\mathbf{q}_T(\boldsymbol{\theta}_2) \equiv \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{q}_t(\boldsymbol{\theta}_2) = \mathbf{0}, \tag{18}$$

where

$$\mathbf{q}_t(\boldsymbol{\theta}_2) \equiv \begin{bmatrix} \hat{\mathbf{w}}_{t+1} \\ vec\left(\hat{\mathbf{w}}_{t+1}\hat{\mathbf{x}}_t' - Cov\left(\mathbf{u}_{t+1},\mathbf{u}_t\right) + \mathbf{h}_{\mathbf{x}} Var\left(\mathbf{u}_t\right)\right) \\ vech\left(\begin{array}{c} \hat{\mathbf{w}}_{t+1}\hat{\mathbf{w}}_{t+1}' - Var\left(\mathbf{w}_{t+1}\right) - Var\left(\mathbf{u}_t\right) - \mathbf{h}_{\mathbf{x}} Var\left(\mathbf{u}_t\right)\mathbf{h}_{\mathbf{x}}' \\ +Cov\left(\mathbf{u}_{t+1},\mathbf{u}_t\right)\mathbf{h}_{\mathbf{x}}' + \mathbf{h}_{\mathbf{x}} Cov\left(\mathbf{u}_t,\mathbf{u}_{t+1}\right) \end{array} \right) \end{bmatrix}. \tag{19}$$

Consistent estimates of $Var\left(\mathbf{u}_t\right)$, $Cov\left(\mathbf{u}_{t+1},\mathbf{u}_t\right)$, and $Cov\left(\mathbf{u}_t,\mathbf{u}_{t+1}\right)$ follow from the first step of the SR approach. Accordingly, $\boldsymbol{\theta}_2$ can be estimated consistently by GMM even though the factors display measurement errors. We also note that the moment conditions in (18), which just identify $\boldsymbol{\theta}_2$, have a simple closed-form solution when the factor dynamics are unrestricted. As shown in Appendix A.1, the estimates of $\mathbf{h}_{\mathbf{x}}$ and $\mathbf{h}_0$ can be obtained by running the modified regression

$$\begin{bmatrix} \hat{\mathbf{h}}_{\mathbf{x}} & \hat{\mathbf{h}}_0 \end{bmatrix} = \left(\sum_{t=1}^{T-1} \begin{bmatrix} \hat{\mathbf{x}}_{t+1}\hat{\mathbf{x}}_t' - \widehat{Cov}\left(\mathbf{u}_{t+1},\mathbf{u}_t\right) & \hat{\mathbf{x}}_{t+1} \end{bmatrix}\right)\left(\sum_{t=1}^{T-1} \begin{bmatrix} \hat{\mathbf{x}}_t\hat{\mathbf{x}}_t' - \widehat{Var}\left(\mathbf{u}_t\right) & \hat{\mathbf{x}}_t \\ \hat{\mathbf{x}}_t' & 1 \end{bmatrix}\right)^{-1}, \tag{20}$$

where all second moments require a correction as they are inflated by measurement errors. Given (20), we immediately have

$$\widehat{Var}\left(\mathbf{w}_{t+1}\right) = \frac{1}{T-1}\sum_{t=1}^{T-1}\left(\hat{\hat{\mathbf{w}}}_{t+1}\hat{\hat{\mathbf{w}}}_{t+1}' - \widehat{Var}\left(\mathbf{u}_t\right) - \hat{\mathbf{h}}_{\mathbf{x}}\widehat{Var}\left(\mathbf{u}_t\right)\hat{\mathbf{h}}_{\mathbf{x}}' + \widehat{Cov}\left(\mathbf{u}_{t+1},\mathbf{u}_t\right)\hat{\mathbf{h}}_{\mathbf{x}}' + \hat{\mathbf{h}}_{\mathbf{x}}\widehat{Cov}\left(\mathbf{u}_t,\mathbf{u}_{t+1}\right)\right) \tag{21}$$

where $\hat{\hat{\mathbf{w}}}_{t+1} = \hat{\mathbf{x}}_{t+1} - \hat{\mathbf{h}}_0 - \hat{\mathbf{h}}_{\mathbf{x}}\hat{\mathbf{x}}_t$.

3.3.2 A VAR system with latent and observed factors

We next consider the case where both observed and unobserved factors appear in the VAR system, i.e. $\mathbf{x}_t \equiv \begin{bmatrix} \mathbf{x}_{1,t}' & \mathbf{x}_{2,t}' \end{bmatrix}'$. If the observed factors are measured without errors, then $\boldsymbol{\theta}_2$ may still be estimated consistently based on (18). For the case with measurement errors in the observed factors, i.e. $\mathbf{x}_{1,t} = \mathbf{x}_{1,t}^o + \mathbf{u}_{1,t}$, the moment conditions in (18) are not applicable because the time series properties of $\mathbf{u}_{1,t}$ are generally unknown. To proceed we therefore impose the following assumptions:

$$Cov\left(\mathbf{u}_{1,\tau_1},\mathbf{u}_{2,\tau_2}\right) = \mathbf{0} \quad \text{for all values of } \tau_1 \text{ and } \tau_2 \tag{22}$$

$$Cov\left(\mathbf{u}_{1,\tau_1},\mathbf{w}_{\tau_2}\right) = \mathbf{0} \quad \text{for all values of } \tau_1 \text{ and } \tau_2 \tag{23}$$

$$Cov\left(\mathbf{u}_{1,t},\mathbf{u}_{1,\tau}\right) = \mathbf{0} \quad \text{for all values of } \tau \neq t \tag{24}$$

That is, $\mathbf{u}_{1,t}$ is uncorrelated with $\mathbf{u}_{2,t}$ and $\mathbf{w}_t$ at all leads and lags and does not display autocorrelation. We first consider the case where $Var\left(\hat{\mathbf{w}}_{t+1}\right)$ is known from the first step in the SR approach, i.e. $\hat{\boldsymbol{\theta}}_{12}$ is given. This is typically the case in models where all elements in $\mathbf{x}_t$ appear as factors in the term structure, i.e. when all risk is spanned by the yield curve. Given (22) to (24), we have

$$\begin{bmatrix} E\left(\hat{\mathbf{w}}_{t+1}\right) \\ E\left(\hat{\mathbf{w}}_{t+1}\hat{\mathbf{x}}_{t-1}'\right) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ Cov\left(\mathbf{u}_{t+1},\mathbf{u}_{t-1}\right) - \mathbf{h}_{\mathbf{x}}Cov\left(\mathbf{u}_t,\mathbf{u}_{t-1}\right) \end{bmatrix}, \tag{25}$$

where $\mathbf{u}_t \equiv \begin{bmatrix} \mathbf{u}_{1,t}' & \mathbf{u}_{2,t}' \end{bmatrix}'$. We therefore consider the moment conditions

$$\mathbf{q}_T\left(\boldsymbol{\theta}_{22},\hat{\boldsymbol{\theta}}_{12}\right) \equiv \frac{1}{T-2}\sum_{t=2}^{T-1}\mathbf{q}_t\left(\boldsymbol{\theta}_{22},\hat{\boldsymbol{\theta}}_{12}\right) = \mathbf{0}, \tag{26}$$

where $\mathbf{q}_t\left(\boldsymbol{\theta}_{22},\hat{\boldsymbol{\theta}}_{12}\right)$ is now defined as

$$\mathbf{q}_t\left(\boldsymbol{\theta}_{22},\hat{\boldsymbol{\theta}}_{12}\right) \equiv \begin{bmatrix} \hat{\mathbf{w}}_{t+1} \\ vec\left(\hat{\mathbf{w}}_{t+1}\hat{\mathbf{x}}_{t-1}' - Cov\left(\mathbf{u}_{t+1},\mathbf{u}_{t-1}\right) + \mathbf{h}_{\mathbf{x}}Cov\left(\mathbf{u}_t,\mathbf{u}_{t-1}\right)\right) \end{bmatrix}.$$

Only the lower right block of the auto-covariance matrix for $\mathbf{u}_t$ differs from zero given (22) to (24), and the non-zero block can be estimated consistently using output from the first step of the SR approach. Hence, $\mathbf{h}_0$ and $\mathbf{h}_{\mathbf{x}}$ may also be estimated consistently by GMM in this case. The system implied by (26) is just identified, and when the factor dynamics are unrestricted, Appendix A.2 shows that the closed-form solution is given by running the modified regression

$$\begin{aligned} \begin{bmatrix} \hat{\mathbf{h}}_{\mathbf{x}} & \hat{\mathbf{h}}_0 \end{bmatrix} &= \left(\sum_{t=2}^{T-1}\begin{bmatrix} \hat{\mathbf{x}}_{t+1}\hat{\mathbf{x}}_{t-1}' - \widehat{Cov}\left(\mathbf{u}_{t+1},\mathbf{u}_{t-1}\right) & \hat{\mathbf{x}}_{t+1} \end{bmatrix}\right) \\ &\quad \times \left(\sum_{t=2}^{T-1}\begin{bmatrix} \hat{\mathbf{x}}_t\hat{\mathbf{x}}_{t-1}' - \widehat{Cov}\left(\mathbf{u}_t,\mathbf{u}_{t-1}\right) & \hat{\mathbf{x}}_t \\ \hat{\mathbf{x}}_{t-1}' & 1 \end{bmatrix}\right)^{-1}. \end{aligned} \tag{27}$$

We finally study the case where $Var\left(\hat{\mathbf{w}}_{t+1}\right)$ is unknown and must be estimated along with $\mathbf{h}_0$ and $\mathbf{h}_{\mathbf{x}}$ in the second step of the SR approach. The last moment condition in (19) shows that this extension complicates the estimation problem because knowledge of the size of the measurement errors in the observed factors is necessary for a consistent estimator of $Var\left(\hat{\mathbf{w}}_{t+1}\right)$. In addition to (22) to (24), we therefore require $Var\left(\mathbf{u}_{1,t}\right)$ to be homoskedastic. To outline the suggested moment conditions, the

VAR system in (17) is written out more explicitly as

$$\begin{bmatrix} \hat{\mathbf{x}}_{1,t+1} \\ \hat{\mathbf{x}}_{2,t+1} \end{bmatrix} = \begin{bmatrix} \mathbf{h}_{0,1} \\ \mathbf{h}_{0,2} \end{bmatrix} + \begin{bmatrix} \mathbf{h}_{\mathbf{x}_1,\mathbf{x}_1} & \mathbf{h}_{\mathbf{x}_1,\mathbf{x}_2} \\ \mathbf{h}_{\mathbf{x}_2,\mathbf{x}_1} & \mathbf{h}_{\mathbf{x}_2,\mathbf{x}_2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}_{1,t} \\ \hat{\mathbf{x}}_{2,t} \end{bmatrix} + \begin{bmatrix} \hat{\mathbf{w}}_{1,t+1} \\ \hat{\mathbf{w}}_{2,t+1} \end{bmatrix}.$$

Given the assumptions imposed, we suggest to extend the moment conditions in (19) by $Cov\left(\hat{\mathbf{w}}_{1,t+1}, \hat{\mathbf{w}}_{1,t}\right)$ which has the population value $\mathbf{h}_{\mathbf{x}_1,\mathbf{x}_2} Cov\left(\mathbf{u}_{2,t}, \mathbf{u}_{2,t-1}\right) \mathbf{h}'_{\mathbf{x}_1,\mathbf{x}_2} - \mathbf{h}_{\mathbf{x}_1,\mathbf{x}_1} Var\left(\mathbf{u}_{1,t}\right)$. This additional moment restriction shows that $Var\left(\mathbf{u}_{1,t}\right)$ may be identified from $Cov\left(\hat{\mathbf{w}}_{1,t+1}, \hat{\mathbf{w}}_{1,t}\right)$ when $\mathbf{w}_{1,t}$ is $\mathcal{IID}$ because $Cov\left(\mathbf{u}_{2,t}, \mathbf{u}_{2,t-1}\right)$ is known from the first step of the SR approach. Accordingly, this extended set of moment conditions allows estimating $Var\left(\mathbf{u}_{1,t}\right)$, $\mathbf{h}_0$, $\mathbf{h}_{\mathbf{x}}$, and $vech\left(Var\left(\mathbf{w}_{t+1}\right)\right)$ using the same number of moments.

3.3.3 Estimation of $\boldsymbol{\theta}_2$ in the general case

GMM may also be applied to estimate $\boldsymbol{\theta}_2 = \begin{bmatrix} \boldsymbol{\theta}'_{22} & \boldsymbol{\theta}'_{12} \end{bmatrix}'$ when the factor dynamics are given by a non-linear system. To realize this, let $\mathbf{m}\left(\mathbf{x}_{1,t}, \hat{\mathbf{x}}_{2,t}; \hat{\boldsymbol{\theta}}_{11}, \boldsymbol{\theta}_2\right)$ be a vector of moment conditions with dimension $n_m \times 1$ where $n_m \geq L_2$ and $\hat{\boldsymbol{\theta}}_{11}$ is given from the first estimation step. These moments are assumed to uniquely identify $\boldsymbol{\theta}_2$. A vector $\mathbf{r}\left(\hat{\boldsymbol{\theta}}_{11}, \boldsymbol{\theta}_2\right)$ contains the corresponding population moments, and we have

$$E\left[\mathbf{q}_t\left(\hat{\boldsymbol{\theta}}_{11}, \boldsymbol{\theta}_2\right)\right] = \mathbf{0},$$

where $\mathbf{q}_t\left(\hat{\boldsymbol{\theta}}_{11}, \boldsymbol{\theta}_2\right) \equiv \mathbf{m}\left(\mathbf{x}_{1,t}, \hat{\mathbf{x}}_{2,t}; \hat{\boldsymbol{\theta}}_{11}, \boldsymbol{\theta}_2\right) - \mathbf{r}\left(\hat{\boldsymbol{\theta}}_{11}, \boldsymbol{\theta}_2\right)$. The GMM estimator of $\boldsymbol{\theta}_2$ is then given by

$$\hat{\boldsymbol{\theta}}_2 = \arg\min_{\boldsymbol{\theta}_2 \in \Theta_2} \left(\mathbf{q}_T\left(\hat{\boldsymbol{\theta}}_{11}, \boldsymbol{\theta}_2\right)\right)' \mathbf{W} \left(\mathbf{q}_T\left(\hat{\boldsymbol{\theta}}_{11}, \boldsymbol{\theta}_2\right)\right), \tag{28}$$

where $\mathbf{q}_T\left(\hat{\boldsymbol{\theta}}_{11}, \boldsymbol{\theta}_2\right) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{q}_t\left(\hat{\boldsymbol{\theta}}_{11}, \boldsymbol{\theta}_2\right)$ and $\mathbf{W}$ is some positive definite weighting matrix.

Closed-form solutions for the population moments may not be available for non-linear systems. This situation often occurs for transition functions set in continuous time if no discretization scheme is applied. Population moments may in this case be derived by simulation if we specify a probability distribution for the innovations to the latent factors and replace assumption (4) by the stronger requirement that $\mathbf{v}_t$ and $\mathbf{w}_{t-k}$ are independent for $k = 0, 1, 2, ..$ and $t = 1, 2, ..., T$. Then $\mathbf{r}\left(\hat{\boldsymbol{\theta}}_{11}, \boldsymbol{\theta}_2\right)$ is estimated by

$$\hat{\mathbf{r}}\left(\hat{\boldsymbol{\theta}}_{11}, \boldsymbol{\theta}_2\right) = \frac{1}{\tau T}\sum_{s=1}^{\tau T}\sum_{t=1}^{T} \mathbf{m}\left(\mathbf{x}^s_{1,t}, \hat{\mathbf{x}}^s_{2,t}; \hat{\boldsymbol{\theta}}_{11}, \boldsymbol{\theta}_2\right),$$

where τT denotes the number of simulations. In this case, $\boldsymbol{\theta}_2$ is estimated by SMM following Duffie & Singleton (1993). To compute (28), we abstract from measurement errors in $\mathbf{x}_{1,t}$ and suggest a conditional simulator of the form (see Carrasco & Florens (2002))

$$\begin{bmatrix} \mathbf{x}_{1,t+1}^s \\ \hat{\mathbf{x}}_{2,t+1}^s \end{bmatrix} = \mathbf{h}\left(\mathbf{x}_{1,t}, \hat{\mathbf{x}}_{2,t} - \mathbf{u}_{2,t}^s, \mathbf{w}_{t+1}^s; \boldsymbol{\theta}_{11}, \boldsymbol{\theta}_2\right) + \begin{bmatrix} \mathbf{0} \\ \mathbf{u}_{2,t+1}^s \end{bmatrix} \text{ for } s = 1, 2, ..., \tau, \text{ and } t = 1, 2, ..., T \tag{29}$$

where $\left\{\mathbf{w}_{t+1}^s\right\}_{s=1}^{\tau}$ and $\left\{\mathbf{u}_{2,t+1}^s\right\}_{s=1}^{\tau}$ are $\mathcal{IID}$ draws from their respective distributions.[4] This simulator has the advantage that it allows for time-varying distributions for the measurement errors based on (15). This is not the case with an unconditional simulator of the form

$$\begin{bmatrix} \mathbf{x}_{1,s+1} \\ \hat{\mathbf{x}}_{2,s+1} \end{bmatrix} = \mathbf{h}\left(\mathbf{x}_{1,s}, \hat{\mathbf{x}}_{2,s} - \mathbf{u}_{2,s}, \mathbf{w}_{s+1}; \boldsymbol{\theta}_{11}, \boldsymbol{\theta}_2\right) + \begin{bmatrix} \mathbf{0} \\ \mathbf{u}_{2,s+1} \end{bmatrix} \text{ for } s = 1, 2, ..., \tau T \tag{30}$$

because the distribution of $\mathbf{u}_{2,s}$ may be unknown for an arbitrary value of s. This unconditional simulator can therefore only be used if it is reasonable to assume that $\mathbf{u}_{2,t}$ has a common distribution for all t that can be used to generate draws for $\mathbf{u}_{2,s}$.

3.4 Summarizing the SR approach

Let us for the sake of clarity briefly summarize the two steps in the SR approach:

- Step 1: Obtain $\hat{\boldsymbol{\theta}}_1$ and $\left\{\hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1\right)\right\}_{t=1}^{T}$ from (12) using the regression filter.
- Step 2: Based on $\hat{\boldsymbol{\theta}}_1$ and $\left\{\mathbf{x}_{1,t}, \hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1\right)\right\}_{t=1}^{T}$, estimate $\boldsymbol{\theta}_2$ by GMM or SMM.

Note that the first step in the SR approach is similar to estimation of $\boldsymbol{\theta}_1$ by ML. That is, we run a filter to construct the objective function which is optimized with respect to a set of model parameters.

4 Asymptotic properties of the SR approach

This section derives the asymptotic properties of the estimated factors and model parameters in the two-step SR approach. For $\{\hat{\mathbf{x}}_{2,t}\}_{t=1}^{T}$ and $\hat{\boldsymbol{\theta}}_1$, this is done based on a fixed number of time periods when the number of observables in the cross-section tends to infinity in every time period. For $\boldsymbol{\theta}_2$,

[4] It is straightforward to generalize this conditional simulator (and the unconditional simulator in (30)) to allow for potential measurement errors in $\mathbf{x}_{1,t}$ if the probability distribution of $\mathbf{u}_{1,t}$ is available.

inference is conducted by letting the number of time periods tend to infinity. Thus, inference is done by sequentially letting the cross-sectional dimension tend to infinity and afterwards letting the time series dimension tend to infinity. A similar approach is used in Connor & Korajczyk (1986) when estimating a factor model based on the Arbitrage Pricing Theory.

All the derivations in this section are for uniquely identified DTSMs. For estimation of $\{\mathbf{x}_{2,t}\}_{t=1}^{T}$ and $\boldsymbol{\theta}_1$, this means

$$\sum_{t=1}^{T} E\left[\left(\tilde{g}_j\left(\mathbf{x}_{2,t}^{o}, \mathbf{z}_t; \boldsymbol{\theta}_1^{o}, \boldsymbol{\gamma}^*\right) - \tilde{g}_j\left(\mathbf{x}_{2,t}, \mathbf{z}_t; \boldsymbol{\theta}_1, \boldsymbol{\gamma}^*\right)\right)^2\right] > 0, \text{ for all } \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_1^{o} \text{ and } \mathbf{x}_2 \neq \mathbf{x}_2^{o}. \tag{31}$$

Here, $\mathbf{x}_2 \equiv \begin{bmatrix} \mathbf{x}_{2,1}' & \mathbf{x}_{2,2}' & ... & \mathbf{x}_{2,T}' \end{bmatrix}'$ and $\hat{\boldsymbol{\gamma}} \xrightarrow{a.s.} \boldsymbol{\gamma}^*$, where $\boldsymbol{\gamma}^*$ may differ from the true value $\boldsymbol{\gamma}^o$. We note that assumptions (3) and (5) allow for identification of $\boldsymbol{\theta}_1$ and $\{\mathbf{x}_{2,t}\}_{t=1}^{T}$ even if the covariance matrix for the measurement errors is misspecified. The requirement in (31) is therefore fairly weak but, for instance, rules out cases where $\mathbf{x}_{2,t}$ and $\boldsymbol{\theta}_1$ only enter as a product, i.e. if $g = \sum_{i=1}^{n_x} \theta_1(i)\, x_{2,t}(i)$. Linear factor models share the same problem, and additional assumptions are required to ensure identification (see for instance Stock & Watson (2002)). For estimation of $\boldsymbol{\theta}_2$, we assume moments for GMM or SMM that uniquely identify $\boldsymbol{\theta}_2$.

We begin by deriving the asymptotic properties of $\hat{\boldsymbol{\theta}}_1$ and $\{\hat{\mathbf{x}}_{2,t}\}_{t=1}^{T}$ in Section 4.1. The time series properties related to estimation uncertainty in the latent factors are derived in Section 4.2, while asymptotic properties of $\hat{\boldsymbol{\theta}}_2$ are presented in Section 4.3.

4.1 Results for $\{\hat{\mathbf{x}}_{2,t}\}_{t=1}^{T}$ and $\hat{\boldsymbol{\theta}}_1$

We begin with the following proposition, with the proof given in Appendix B:

Proposition 1 *The estimators of* $\{\mathbf{x}_{2,t}\}_{t=1}^{T}$ *in (9) and* $\boldsymbol{\theta}_1$ *in (12) are equivalent to joint estimation of* $\left(\boldsymbol{\theta}_1, \{\mathbf{x}_{2,t}\}_{t=1}^{T}\right)$ *from* $\min_{\left(\boldsymbol{\theta}_1, \{\mathbf{x}_{2,t}\}_{t=1}^{T}\right)} Q^{joint} = \frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{n_{y,t}}\left(\tilde{y}_{t,j}\left(\mathbf{z}_t; \hat{\boldsymbol{\gamma}}\right) - \tilde{g}_j\left(\mathbf{x}_{2,t}, \mathbf{z}_t; \boldsymbol{\theta}_1; \hat{\boldsymbol{\gamma}}\right)\right)^2$.

Proposition 1 states that the procedure described in the previous section to estimate $\{\mathbf{x}_{2,t}\}_{t=1}^{T}$ and $\boldsymbol{\theta}_1$ is equivalent to a joint estimation of $\{\mathbf{x}_{2,t}\}_{t=1}^{T}$ and $\boldsymbol{\theta}_1$. The SR approach can therefore be considered as a convenient numerical procedure for optimizing a high dimensional objective function with respect to the latent factors and the model parameters. Proposition 1 also implies that consistency and

asymptotic normality of $\{\hat{\mathbf{x}}_{2,t}\}_{t=1}^{T}$ and $\hat{\boldsymbol{\theta}}_1$ follow from standard results for NLS in White & Domowitz (1984). The conditions for consistency of $\{\hat{\mathbf{x}}_{2,t}\}_{t=1}^{T}$ and $\hat{\boldsymbol{\theta}}_1$ are stated in the next proposition:[5]

Proposition 2 *Given assumptions (3), (5) - (8), and (31), if either a)* $\phi_t(m) = O(m^{-\nu})$ *for* $\nu > r_1/(2r_1-1)$ *or b)* $\alpha_t(m) = O(m^{-\nu})$ *for* $\nu > r_1/(r_1-1)$*, where* $r_1 > 1$*, then* $(\hat{\boldsymbol{\theta}}_1, \{\hat{\mathbf{x}}_{2,t}\}_{t=1}^{T}) \xrightarrow{a.s.} \left(\boldsymbol{\theta}_1^o, \{\mathbf{x}_{2,t}^o\}_{t=1}^{T}\right)$ *as* $n_{y,t} \longrightarrow \infty$ *for* $t = 1, 2, ..., T$.

It is numerically challenging to apply the standard results for NLS when computing the asymptotic distributions of $\{\hat{\mathbf{x}}_{2,t}\}_{t=1}^{T}$ and $\hat{\boldsymbol{\theta}}_1$. For instance, in a DTSM with three latent factors and $T = 500$, the dimension of the covariance matrix would exceed $1,500$. When computing the asymptotic covariance matrix, we therefore find it numerically more convenient to exploit the structure of this matrix. This is done by first deriving the distribution of $\hat{\boldsymbol{\theta}}_1$. The covariance matrices of $\{\hat{\mathbf{x}}_{2,t}\}_{t=1}^{T}$ are then derived afterwards, with $\hat{\boldsymbol{\theta}}_1$ as a nuisance parameter.

4.1.1 The asymptotic distribution of $\hat{\boldsymbol{\theta}}_1$

We start by stacking the data column-wise and use $\mathcal{I}$ to index the $N \equiv \sum_{t=1}^{T} n_{y,t}$ elements in the pooled sample. Hence, when the subscript $i \in \mathcal{I}$ appears on a variable, we refer to the i'th observation in the pooled sample. As an illustration, the stacked data for all bond prices are denoted by $\mathbf{y} \equiv \left\{\{y_{t,1}\}_{t=1}^{T}, \{y_{t,2}\}_{t=1}^{T}, ..., \{y_{t,n_{y,t}}\}_{t=1}^{T}\right\}$ and y_i refers to the i'th bond price in this N-vector. Adopting this notation, the asymptotic distribution of $\boldsymbol{\theta}_1$ then follows from a mean-value expansion of the score function derived from (12). Given standard regularity conditions, we show in Appendix C that

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^o\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left(\mathbf{A}_o^{\boldsymbol{\theta}_1}\right)^{-1} \mathbf{B}_o^{\boldsymbol{\theta}_1} \left(\mathbf{A}_o^{\boldsymbol{\theta}_1}\right)^{-1}\right). \tag{32}$$

Here, $\mathbf{A}_o^{\boldsymbol{\theta}_1}$ is the expected value of the average Hessian matrix and $\mathbf{B}_o^{\boldsymbol{\theta}_1} \equiv Var\left(\frac{1}{\sqrt{N}} \sum_{t=1}^{N} \mathbf{s}_i^{\boldsymbol{\theta}_1}\right)$, where $\mathbf{s}_i^{\boldsymbol{\theta}_1}$ denotes the score of the i'th observation in the pooled sample. The distribution in (32) is derived for $n_{y,t} \to \infty$ in each time period, implying that $N \to \infty$. Note also that uncertainty from estimation of $\boldsymbol{\gamma}$ does not affect the distribution of $\hat{\boldsymbol{\theta}}_1$. This result is similar to a well-known property of weighted non-linear least squares with purely observed regressors (see for instance Wooldridge (2002)).

Before we present estimators of $\mathbf{A}_o^{\boldsymbol{\theta}_1}$ and $\mathbf{B}_o^{\boldsymbol{\theta}_1}$, consider the expression for the score function derived from (12)

$$\mathbf{s}_{t,j}^{\boldsymbol{\theta}_1} = -\sum_{t=1}^{T} \sum_{j=1}^{n_{y,t}} \tilde{v}_{t,j}(\mathbf{z}_t; \boldsymbol{\gamma}) \boldsymbol{\Psi}_{t,j}^{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1, \boldsymbol{\gamma}), \tag{33}$$

[5] We use the standard notation where a sequence b_m is at most of order m^{ν}, denoted $O(m^{\nu})$, if $b_m m^{-\nu}$ is bounded.

where

$$\boldsymbol{\Psi}_{t,j}^{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1,\boldsymbol{\gamma}) \equiv \frac{\partial \hat{\mathbf{x}}_{2,t}'(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1} \frac{\partial \tilde{g}_j(\hat{\mathbf{x}}_{2,t}(\boldsymbol{\theta}_1), \mathbf{z}_t; \boldsymbol{\theta}_1, \boldsymbol{\gamma})}{\partial \mathbf{x}_{2,t}(\boldsymbol{\theta}_1)} + \frac{\partial \tilde{g}_j(\hat{\mathbf{x}}_{2,t}(\boldsymbol{\theta}_1), \mathbf{z}_t; \boldsymbol{\theta}_1, \boldsymbol{\gamma})}{\partial \boldsymbol{\theta}_1} \tag{34}$$

and $\tilde{\mathbf{v}}_t(\mathbf{z}_t;\boldsymbol{\gamma}) \equiv \mathbf{S}_{\mathbf{v}}(\mathbf{z}_t;\boldsymbol{\gamma})^{-1}\mathbf{v}_t$, with $\tilde{v}_{t,j}(\mathbf{z}_t;\boldsymbol{\gamma})$ referring to the j'th observation in $\tilde{\mathbf{v}}_t(\mathbf{z}_t;\boldsymbol{\gamma})$.[6] The first term in (34) captures the indirect effect of changes in $\boldsymbol{\theta}_1$ that lead to changes in the latent factors $\mathbf{x}_{2,t}(\boldsymbol{\theta}_1)$ which in turn imply changes in the model observables. The second term in (34) is the direct effect of changes in $\boldsymbol{\theta}_1$ on model observables.

For the estimation of $\mathbf{B}_o^{\boldsymbol{\theta}_1}$, we consider an extension of the Newey-West estimator which is robust to heteroskedasticity, cross-section correlation, and autocorrelation in $v_{t,j}$. To present this estimator, let w_D denote the bandwidth for duration in the cross-section dimension and let w_T denote the corresponding bandwidth for the time series dimension. Following Pesaran & Tosetti (2011), who build on the work by Kelejian & Prucha (2007), we suggest the non-parametric estimator

$$\hat{\mathbf{B}}^{\boldsymbol{\theta}_1} = \frac{1}{N}\sum_{t=1}^{T}\sum_{j=1}^{n_{y,t}}\sum_{k_T=-w_T}^{w_T}\sum_{k_D=-w_D}^{w_D}\left(1-\frac{|k_T|}{1+w_T}\right)\left(1-\frac{|k_D|}{1+w_D}\right)\left(\hat{\boldsymbol{\Psi}}_{t,j}^{\boldsymbol{\theta}_1}\right)\left(\hat{\boldsymbol{\Psi}}_{t+k_T,j+k_D}^{\boldsymbol{\theta}_1}\right)'\widehat{\tilde{v}}_{t,j}\widehat{\tilde{v}}_{t+k_T,j+k_D},$$

where $\hat{\boldsymbol{\Psi}}_{t,j}^{\boldsymbol{\theta}_1} \equiv \boldsymbol{\Psi}_{t,j}^{\boldsymbol{\theta}_1}\left(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\gamma}}\right)$ and $\widehat{\tilde{v}}_{t,j} \equiv \tilde{y}_{t,j}(\mathbf{z}_t, \hat{\boldsymbol{\gamma}}) - \tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}, \mathbf{z}_t; \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\gamma}}\right)$. This estimator of $\mathbf{B}_o^{\boldsymbol{\theta}_1}$ is positive semi-definite, and consistency follows from the regularity conditions in Kelejian & Prucha (2007).[7]

A more efficient estimator of $\mathbf{B}_o^{\boldsymbol{\theta}_1}$ can be constructed if the covariance matrix for the measurement errors is correctly specified, i.e. if

$$Var(\mathbf{v}_t|\,\mathbf{z}_t) = \boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{z}_t;\boldsymbol{\gamma}^o) \quad \text{for some } \boldsymbol{\gamma}^o \in \Gamma \text{ for all } t = 1,2,...,T. \tag{35}$$

This means $\hat{\boldsymbol{\gamma}} \xrightarrow{a.s.} \boldsymbol{\gamma}^* = \boldsymbol{\gamma}^o$. Whenever we impose the assumption in (35), the decomposition of $\boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{z}_t;\boldsymbol{\gamma})$ in (10) is re-defined as $\boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{z}_t;\boldsymbol{\gamma}) \equiv \sigma_t^2\mathbf{V}(\mathbf{z}_t;\boldsymbol{\gamma})$ where $\mathbf{V}(\mathbf{z}_t;\boldsymbol{\gamma}) = \mathbf{S}_{\mathbf{v}}(\mathbf{z}_t;\boldsymbol{\gamma})\mathbf{S}_{\mathbf{v}}(\mathbf{z}_t;\boldsymbol{\gamma})'$. That is, we simply remove σ_t from the rotation matrix $\mathbf{S}_{\mathbf{v}}(\mathbf{z}_t;\boldsymbol{\gamma})$ in this case. Assumption (35) implies that the rotated model is conditionally homoskedastic and cross-sectionally uncorrelated, meaning

[6] Throughout, we adopt the short-hand notation $\frac{\partial Q(\hat{\theta})}{\partial \theta} = \frac{\partial Q(\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}}$ when evaluating partial derivatives, where Q is a function of θ.

[7] The optimal bandwidth for the class of estimators considered by Kelejian & Prucha (2007) is derived in Kim & Sun (2011).

that the variance of the score function may be estimated by

$$\hat{\mathbf{B}}_{\text{hom}}^{\boldsymbol{\theta}_1} = \sum_{t=1}^{T}\sum_{j=1}^{n_{y,t}} \left(\hat{\boldsymbol{\Psi}}_{t,j}^{\boldsymbol{\theta}_1}\right)\left(\hat{\boldsymbol{\Psi}}_{t,j}^{\boldsymbol{\theta}_1}\right)' \hat{\sigma}_t^2 + \frac{1}{N}\sum_{t=1}^{T}\sum_{j=1}^{n_{y,t}} \sum_{\substack{k_T=-w_T \\ k_T\neq 0}}^{w_T} \left(1-\frac{|k_T|}{1+w_T}\right)\left(\hat{\boldsymbol{\Psi}}_{t,j}^{\boldsymbol{\theta}_1}\right)\left(\hat{\boldsymbol{\Psi}}_{t+k_T,j}^{\boldsymbol{\theta}_1}\right)' \widehat{\tilde{v}}_{t,j}\widehat{\tilde{v}}_{t+k_T,j}, \tag{36}$$

when accounting for autocorrelation in $v_{t,j}$. In the absence of such autocorrelation, the measurement errors $\tilde{v}_{t,j}$ are uncorrelated in both the cross-section and time series dimension, and the variance of the score function further reduces to

$$\hat{\mathbf{B}}_{\text{hom},NoAuto}^{\boldsymbol{\theta}_1} = \frac{1}{N}\sum_{t=1}^{T}\sum_{j=1}^{n_{y,t}} \hat{\sigma}_t^2 \left(\hat{\boldsymbol{\Psi}}_{t,j}^{\boldsymbol{\theta}_1}\right)\left(\hat{\boldsymbol{\Psi}}_{t,j}^{\boldsymbol{\theta}_1}\right)'. \tag{37}$$

In relation to (36) and (37), we let

$$\hat{\sigma}_t^2 = \frac{1}{n_{y,t}-n_{x_2}}\sum_{j=1}^{n_{y,t}} \widehat{\tilde{v}}_{t,j}\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)^2 \quad \text{for } t=1,2,...,T. \tag{38}$$

Given standard regularity conditions, the estimator in (38) is shown to be consistent and asymptotically normal in Appendix D. Adopting the usual small sample correction for degrees of freedom, we obtain

$$\sqrt{n_{y,t}-n_{x_2}}\left(\hat{\sigma}_t^2-(\sigma_t^o)^2\right) \xrightarrow{d} \mathcal{N}\left(0, E\left[\tilde{v}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)^4\right]-(\sigma_t^o)^4\right) \tag{39}$$

for $n_{y,t}\to\infty$, where σ_t^o denotes the true value of σ_t. The variance of $\hat{\sigma}_t^2$ may be estimated by

$$\widehat{Var}\left(\hat{\sigma}_t^2\right) = \frac{1}{n_{y,t}-n_{x_2}}\sum_{j=1}^{n_{y,t}} \widehat{\tilde{v}}_{t,j}\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)^4 - \left(\hat{\sigma}_t^2\right)^2.$$

Accordingly, standard results for GNLS also hold in our case, even though $\boldsymbol{\theta}_1$ is estimated from $\left\{\{y_{t,j}\}_{j=1}^{n_{y,t}}\right\}_{t=1}^{T}$ whereas $\hat{\sigma}_t^2$ is estimated from the much smaller sample $\{y_{t,j}\}_{j=1}^{n_{y,t}}$. This difference is asymptotically unimportant because $n_{y,t}$ tending to infinity for any t clearly makes $N=\sum_{t=1}^{T} n_{y,t}$ tend to infinity.

Equation (33) implies that the Hessian matrix contains second-order derivatives of $\hat{\mathbf{x}}_{2,t}\left(\boldsymbol{\theta}_1\right)$ and $\tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}\left(\boldsymbol{\theta}_1\right),\mathbf{z}_t;\boldsymbol{\theta}_1,\boldsymbol{\gamma}\right)$ which rarely are available. Fortunately, these second-order derivatives are not present in the conditional expected Hessian matrix as $E\left[\mathbf{v}_t|\,\mathbf{z}_t\right]=\mathbf{0}$, and we can therefore estimate $\mathbf{A}^{\boldsymbol{\theta}_1}$ by

$$\hat{\mathbf{A}}^{\boldsymbol{\theta}_1} = \frac{1}{N}\sum_{t=1}^{T}\sum_{j=1}^{n_{y,t}} \left(\hat{\boldsymbol{\Psi}}_{t,j}^{\boldsymbol{\theta}_1}\right)\left(\hat{\boldsymbol{\Psi}}_{t,j}^{\boldsymbol{\theta}_1}\right)'. \tag{40}$$

As in GNLS without latent factors, the Hessian matrix can therefore be estimated solely based on first-order derivatives, and this greatly simplifies the implementation of the SR approach. It is finally worth noticing that if the conditions underlying (37) applies and $\sigma_t^2 = \sigma^2$ for all t, then $\mathbf{B}_o^{\boldsymbol{\theta}_1} = (\sigma^o)^2 \mathbf{A}_o^{\boldsymbol{\theta}_1}$. In this particular case, the asymptotic variance of $\sqrt{N}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^o\right)$ therefore simplifies to $(\sigma^o)^2 \left(\mathbf{A}_o^{\boldsymbol{\theta}_1}\right)^{-1}$ which may be estimated accordingly.

4.1.2 The asymptotic distribution of $\hat{\mathbf{x}}_{2,t}$

To present the asymptotic distribution of $\hat{\mathbf{x}}_{2,t}$, let $c_t \equiv n_{y,t}/N$ denote the fraction of observations from time period t in the pooled sample. For instance, $c_t = 1/T$ in a balanced panel. Furthermore, let $\mathbf{s}_{t,j}^{\mathbf{x}_2}$ denote the score function of $\mathbf{x}_{2,t}$ for the j'th observation, and let $\mathbf{B}_{t,o}^{\mathbf{x}_2} \equiv Var\left[\frac{1}{\sqrt{n_{y,t}}}\sum_{j=1}^{n_{y,t}} \mathbf{s}_{t,j}^{\mathbf{x}_2}\right]$ and $\mathbf{A}_{t,o}^{\mathbf{x}_2} \equiv E\left[\frac{1}{n_{y,t}}\sum_{j=1}^{n_{y,t}} \frac{\partial \mathbf{s}_{t,j}^{\mathbf{x}_2}}{\partial \mathbf{x}_{2,t}'}\right]$. Under standard regularity conditions and $n_{y,t} \longrightarrow \infty$ we show in Appendix E that

$$\sqrt{n_{y,t}}\left(\hat{\mathbf{x}}_{2,t} - \mathbf{x}_{2,t}^o\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left(\mathbf{A}_{t,o}^{\mathbf{x}_2}\right)^{-1} \mathbf{V}_t^{\mathbf{x}_2} \left(\mathbf{A}_{t,o}^{\mathbf{x}_2}\right)^{-1}\right),$$

where

$$\begin{aligned} \mathbf{V}_t^{\mathbf{x}_2} \equiv\ & \mathbf{B}_{t,o}^{\mathbf{x}_2} + c_t \mathbf{D}_{t,o}\left(\mathbf{A}_o^{\boldsymbol{\theta}_1}\right)^{-1} \mathbf{B}_o^{\boldsymbol{\theta}_1}\left(\mathbf{A}_o^{\boldsymbol{\theta}_1}\right)^{-1} \mathbf{D}_{t,o}' \\ & -\sqrt{c_t}\left(\left(\mathbf{C}_{t,o}^{\boldsymbol{\theta}_1,\mathbf{x}_2}\right)'\left(\mathbf{A}_o^{\boldsymbol{\theta}_1}\right)^{-1}\mathbf{D}_{t,o}' + \mathbf{D}_{t,o}\left(\mathbf{A}_o^{\boldsymbol{\theta}_1}\right)^{-1}\mathbf{C}_{t,o}^{\boldsymbol{\theta}_1,\mathbf{x}_2}\right). \end{aligned} \tag{41}$$

Here, $\mathbf{D}_{t,o} \equiv E\left[\frac{1}{n_{y,t}}\sum_{j=1}^{n_{y,t}} \frac{\partial \mathbf{s}_j^{\mathbf{x}_2}}{\partial \boldsymbol{\theta}_1'}\right]$ is the Jacobian with dimension $n_{x_2} \times L_1$ and $\mathbf{C}_{t,o}^{\boldsymbol{\theta}_1,\mathbf{x}_2} \equiv Cov\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{s}_i^{\boldsymbol{\theta}_1}, \frac{1}{\sqrt{n_{y,t}}}\sum_{j=1}^{n_{y,t}}\mathbf{s}_{t,j}^{\mathbf{x}_2}\right)$ with dimension $L_1 \times n_{x_2}$. The matrices $\mathbf{A}_o^{\boldsymbol{\theta}_1}$ and $\mathbf{B}_o^{\boldsymbol{\theta}_1}$ are those defined in Section 4.1.1. The first term in (41) represents the usual uncertainty associated with regression analysis, whereas the second and third terms in (41) are non-standard and capture the additional uncertainty from estimation of $\boldsymbol{\theta}_1$. Note that the uncertainty related to $\hat{\boldsymbol{\gamma}}$ does not appear in $\mathbf{V}_t^{\mathbf{x}_2}$.

Applying the results in Kelejian & Prucha (2007), each of the terms in (41) may be estimated by

$$\hat{\mathbf{B}}_t^{\mathbf{x}_2} = \frac{1}{n_{y,t}}\sum_{j=1}^{n_{y,t}}\sum_{k_D=-w_D}^{w_D}\left(1 - \frac{|k_D|}{1+w_D}\right)\left(\frac{\partial \widehat{\tilde{g}}_j}{\partial \mathbf{x}_{2,t}}\right)\left(\frac{\partial \widehat{\tilde{g}}_{j+k_D}}{\partial \mathbf{x}_{2,t}}\right)' \widehat{\tilde{v}}_{t,j}\widehat{\tilde{v}}_{t,j+k_D}$$

$$\hat{\mathbf{A}}_t^{\mathbf{x}_2} = \frac{1}{n_{y,t}}\sum_{j=1}^{n_{y,t}}\left(\frac{\partial \widehat{\tilde{g}}_j}{\partial \mathbf{x}_{2,t}}\right)\left(\frac{\partial \widehat{\tilde{g}}_j}{\partial \mathbf{x}_{2,t}}\right)'$$

$$\hat{\mathbf{D}}_t = \frac{1}{n_{y,t}} \sum_{j=1}^{n_{y,t}} \frac{\partial \widehat{\tilde{g}}_j}{\partial \mathbf{x}_{2,t}} \left(\frac{\partial \hat{\mathbf{x}}'_{2,t}}{\partial \boldsymbol{\theta}_1} \frac{\partial \widehat{\tilde{g}}_j}{\partial \mathbf{x}_{2,t}} + \frac{\partial \widehat{\tilde{g}}_j}{\partial \boldsymbol{\theta}_1} \right)'$$

$$\hat{\mathbf{C}}_t^{\boldsymbol{\theta}_1,\mathbf{x}_2} = \frac{1}{n_{y,t}} \sum_{j=1}^{n_{y,t}} \sum_{k_D=-w_D}^{w_D} \left(1 - \frac{|k_D|}{1+w_D}\right) \left(\frac{\partial \widehat{\tilde{g}}_j}{\partial \boldsymbol{\theta}_1} + \frac{\partial \hat{\mathbf{x}}'_{2,t}(\boldsymbol{\theta}_1^o)}{\partial \boldsymbol{\theta}_1} \frac{\partial \widehat{\tilde{g}}_j}{\partial \mathbf{x}_{2,t}(\boldsymbol{\theta}_1)} \right) \left(\frac{\partial \widehat{\tilde{g}}_{j+k_D}}{\partial \mathbf{x}_{2,t}} \right)' \widehat{\tilde{v}}_{t,j} \widehat{\tilde{v}}_{t,j+k_D}$$

where $\widehat{\tilde{g}}_j \equiv \tilde{g}_j \left(\hat{\mathbf{x}}_{2,t}, \mathbf{z}_t; \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\gamma}} \right)$. Note that the estimators of $\mathbf{B}_t^{\mathbf{x}_2}$ and $\mathbf{C}_t^{\boldsymbol{\theta}_1,\mathbf{x}_2}$ are robust to heteroskedasticity and correlation in the cross-section dimension. More efficient estimators of $\mathbf{B}_t^{\mathbf{x}_2}$ and $\mathbf{C}_t^{\boldsymbol{\theta}_1,\mathbf{x}_2}$ may be obtained if (35) holds and the model is conditionally homoskedastic and without cross-sectional correlation. As in Section 4.1.1, $\mathbf{S}_{\mathbf{v}}(\mathbf{z}_t;\boldsymbol{\gamma})$ is then defined without σ_t, and we get

$$\hat{\mathbf{B}}_{t,\text{hom}}^{\mathbf{x}_2} = \frac{\hat{\sigma}_t^2}{n_{y,t}} \sum_{j=1}^{n_{y,t}} \left(\frac{\partial \widehat{\tilde{g}}_j}{\partial \mathbf{x}_{2,t}} \right) \left(\frac{\partial \widehat{\tilde{g}}_j}{\partial \mathbf{x}_{2,t}} \right)',$$

$$\hat{\mathbf{C}}_{t,\text{hom}}^{\boldsymbol{\theta}_1,\mathbf{x}_2} = \frac{\hat{\sigma}_t^2}{n_{y,t}} \sum_{j=1}^{n_{y,t}} \left(\frac{\partial \hat{\mathbf{x}}'_{2,t}}{\partial \boldsymbol{\theta}_1} \frac{\partial \widehat{\tilde{g}}_j}{\partial \mathbf{x}_{2,t}} + \frac{\partial \widehat{\tilde{g}}_j}{\partial \boldsymbol{\theta}_1} \right) \left(\frac{\partial \widehat{\tilde{g}}_j}{\partial \mathbf{x}_{2,t}} \right)'.$$

4.2 The time series properties of $\mathbf{u}_{2,t}$

The time series properties of the uncertainty in the estimated factors can be derived from the approximated expression (see Appendix F)

$$\mathbf{u}_{2,t} \simeq \frac{1}{n_{y,t}} \left(\mathbf{A}_t^{\mathbf{x}_2}\right)^{-1} \sum_{j=1}^{n_{y,t}} \tilde{v}_{t,j} \frac{\partial \tilde{g}_{t,j}}{\partial \mathbf{x}_{2,t}}. \tag{42}$$

Based on this approximation, the conditional auto-covariance in $\mathbf{u}_{2,t}$ at time t is

$$Cov\left(\mathbf{u}_{2,t}, \mathbf{u}_{2,t-k} | \mathbf{z}_t\right) = \frac{1}{n_{y,t}} \frac{1}{n_{y,t-k}} \left(\mathbf{A}_t^{\mathbf{x}_2}\right)^{-1} \frac{\partial \tilde{\mathbf{g}}'_t}{\partial \mathbf{x}_{2,t}} E\left[\tilde{\mathbf{v}}_t \tilde{\mathbf{v}}'_{t-k} | \mathbf{z}_t\right] \frac{\partial \tilde{\mathbf{g}}_{t-k}}{\partial \left(\mathbf{x}_{2,t-k}\right)'} \left(\mathbf{A}_{t-k}^{\mathbf{x}_2}\right)^{-1}, \tag{43}$$

where $\frac{\partial \tilde{\mathbf{g}}'_t}{\partial \mathbf{x}_{2,t}} \equiv \left[\begin{array}{cccc} \frac{\partial \tilde{g}_{t,1}}{\partial \mathbf{x}_{2,t}} & \frac{\partial \tilde{g}_{t,2}}{\partial \mathbf{x}_{2,t}} & \ldots & \frac{\partial \tilde{g}_{t,n_{y,t}}}{\partial \mathbf{x}_{2,t}} \end{array} \right]$ with dimension $n_{x_2} \times n_{y,t}$ and $\frac{\partial \tilde{\mathbf{g}}_t}{\partial \mathbf{x}'_{2,t}}$ has dimension $n_{y,t} \times n_{x_2}$. We also observe that $E\left[\tilde{\mathbf{v}}_t \tilde{\mathbf{v}}'_{t-k} | \mathbf{z}_t\right]$ may be time-dependent, meaning that the conditional auto-covariance in $\mathbf{u}_{2,t}$ may be time-varying due to variation in $\left(\mathbf{A}_t^{\mathbf{x}_2}\right)^{-1} \frac{\partial \tilde{\mathbf{g}}'_t}{\partial \mathbf{x}_{2,t}^o}$ and $E\left[\tilde{\mathbf{v}}_t \tilde{\mathbf{v}}'_{t-k} | \mathbf{z}_t\right]$. If the measurement errors in the observables are uncorrelated across time, i.e. if $E\left[\tilde{\mathbf{v}}_t \tilde{\mathbf{v}}'_{t-k} | \mathbf{z}_t\right] = \mathbf{0}$, then $\mathbf{u}_{2,t}$ does not display auto-correlation.

We are unable to obtain a consistent estimator of $E\left[\tilde{\mathbf{v}}_t \tilde{\mathbf{v}}'_{t-k} | \mathbf{z}_t\right]$ without imposing some parametric structure on this auto-covariance matrix. Focus is therefore devoted to the case where the auto-covariance between $\mathbf{v}_t$ and $\mathbf{v}_{t-k}$ is time-invariant. Here, the conditional auto-covariance in (43) can

be estimated consistently by

$$\widehat{Cov}\left(\mathbf{u}_{2,t},\mathbf{u}_{2,t-k}|\mathbf{z}_t\right)=\frac{1}{n_{y,t}}\frac{1}{n_{y,t-k}}\left(\hat{\mathbf{A}}_t^{\mathbf{x}_2}\right)^{-1}\frac{\partial\widehat{\tilde{\mathbf{g}}}_t'}{\partial\mathbf{x}_{2,t}}E\left[\widehat{\tilde{\mathbf{v}}_t\tilde{\mathbf{v}}_{t-k}'}\right]\frac{\partial\widehat{\tilde{\mathbf{g}}}_{t-k}}{\partial\left(\mathbf{x}_{2,t-k}\right)'}\left(\hat{\mathbf{A}}_{t-k}^{\mathbf{x}_2}\right)^{-1},\tag{44}$$

where $E\left[\widehat{\tilde{\mathbf{v}}_t\tilde{\mathbf{v}}_{t-k}'}\right]=\frac{1}{T-k}\sum_{t=1+k}^{T}\widehat{\tilde{\mathbf{v}}}_t\widehat{\tilde{\mathbf{v}}}_{t-k}'$. Note that this conditional auto-covariance may still be time-varying if there is time-variation in $\left(\mathbf{A}_t^{\mathbf{x}_2}\right)^{-1}\frac{\partial\tilde{\mathbf{g}}_t'}{\partial\mathbf{x}_{2,t}}$. If the estimator of $E\left[\tilde{\mathbf{v}}_t\tilde{\mathbf{v}}_{t-k}'\right]$ is hard to implement due to several missing values, it may be easier to compute $E\left[\tilde{\mathbf{v}}_t\tilde{\mathbf{v}}_{t-k}'\right]$ based on the VAR system

$$\tilde{\mathbf{v}}_t=\mathbf{\Phi}\tilde{\mathbf{v}}_{t-1}+\boldsymbol{\epsilon}_{\tilde{\mathbf{v}},t}\tag{45}$$

with $\boldsymbol{\epsilon}_{\tilde{\mathbf{v}},t}\sim\mathcal{NID}\left(\mathbf{0},\mathbf{\Sigma}\right)$.[8] Standard results imply $E\left[\tilde{\mathbf{v}}_t\tilde{\mathbf{v}}_{t-k}'\right]=\mathbf{\Phi}^kVar\left(\tilde{\mathbf{v}}_t\right)$ where $vec\left(Var\left(\tilde{\mathbf{v}}_t\right)\right)=\left(\mathbf{I}-\mathbf{\Phi}\otimes\mathbf{\Phi}\right)^{-1}vec\left(\mathbf{\Sigma}\right)$. Based on (45) we therefore have

$$\widehat{Cov}\left(\mathbf{u}_{2,t},\mathbf{u}_{2,t-k}|\mathbf{z}_t\right)=\frac{1}{n_{y,t}}\frac{1}{n_{y,t-k}}\left(\hat{\mathbf{A}}_t^{\mathbf{x}_2}\right)^{-1}\frac{\partial\widehat{\tilde{\mathbf{g}}}_t'}{\partial\mathbf{x}_{2,t}}\hat{\mathbf{\Phi}}^k\widehat{Var}\left(\tilde{\mathbf{v}}_t\right)\frac{\partial\widehat{\tilde{\mathbf{g}}}_{t-k}}{\partial\left(\mathbf{x}_{2,t-k}\right)'}\left(\hat{\mathbf{A}}_{t-k}^{\mathbf{x}_2}\right)^{-1}.\tag{46}$$

Using either (44) or (46), the unconditional auto-covariance may then be estimated by

$$\widehat{Cov}\left(\mathbf{u}_{2,t},\mathbf{u}_{2,t-k}\right)=\frac{1}{T-k}\sum_{t=1+k}^{T}\widehat{Cov}\left(\mathbf{u}_{2,t},\mathbf{u}_{2,t-k}|\mathbf{z}_t\right).$$

Finally, the empirical support for the assumed absence of correlation between $\mathbf{v}_t$ and $\mathbf{w}_t$ at all leads and lags can be tested in a standard manner using $\hat{\mathbf{v}}_t$ and $\widehat{\mathbf{w}}_t$. Here, $\widehat{\mathbf{w}}_t$ is the innovation to the estimated factors using the estimated model parameters, i.e. $\hat{\mathbf{x}}_{t+1}=\mathbf{h}\left(\mathbf{x}_{1,t},\hat{\mathbf{x}}_{2,t},\widehat{\mathbf{w}}_{t+1};\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2\right)$.

4.3 Results for $\hat{\boldsymbol{\theta}}_2$

In deriving the asymptotic properties of $\hat{\boldsymbol{\theta}}_2$ we need to deal with two non-standard features. Firstly, estimation uncertainty is present in the latent factors and must be accounted for. Fortunately, the SR approach provides a consistent estimate of this uncertainty, and it is therefore straightforward to correct for it when setting up the moment condition, as shown in Section 3.3.

The second non-standard feature is the presence of $\hat{\boldsymbol{\theta}}_{11}$ in the moment conditions. Appendix G shows that the asymptotic distribution for $\boldsymbol{\theta}_2$ is unaffected by the fact that $\boldsymbol{\theta}_{11}$ is estimated, since inference on $\boldsymbol{\theta}_2$ is undertaken in the time series dimension (i.e. $T\longrightarrow\infty$) whereas inference for $\boldsymbol{\theta}_{11}$

[8] See Penzer & Shea (1997) for methods on estimating VAR systems with missing observations.

is carried out in the cross-section dimension (i.e. $n_{y_t} \to \infty$ for all t, implying $N \equiv \sum_{t=1}^{T} n_{y_t} \to \infty$). Hence, when T tends to infinity, N goes faster to infinity and $\boldsymbol{\theta}_{11}$ is estimated super-consistently, meaning that we can treat $\hat{\boldsymbol{\theta}}_{11}$ as known in the asymptotic distribution of $\hat{\boldsymbol{\theta}}_2$.[9] As a result, the conditions stated in Hansen (1982) and Duffie & Singleton (1993) for consistency and asymptotic normality of GMM and SMM for stationary and ergodic processes, respectively, also apply in our case. When relying on GMM based on the optimal weighting matrix, we therefore have

$$\sqrt{T}\left(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_2^o\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{V}_{\boldsymbol{\theta}_2}\right), \tag{47}$$

where $\mathbf{V}_{\boldsymbol{\theta}_2} \equiv \left(\mathbf{R}_o^{\boldsymbol{\theta}_2} \mathbf{S}_o^{-1} \left(\mathbf{R}_o^{\boldsymbol{\theta}_2}\right)'\right)^{-1}$, $\mathbf{R}_o^{\boldsymbol{\theta}_2} \equiv \frac{\partial \mathbf{q}_T(\boldsymbol{\theta}_{11}^o, \boldsymbol{\theta}_2^o)'}{\partial \boldsymbol{\theta}_2}$ and $\mathbf{S}_o \equiv \sum_{\nu=-\infty}^{\infty} E\left[\mathbf{q}_t\left(\boldsymbol{\theta}_{11}^o, \boldsymbol{\theta}_2^o\right) \mathbf{q}_{t-\nu}\left(\boldsymbol{\theta}_{11}^o, \boldsymbol{\theta}_2^o\right)'\right]$. Accordingly, the asymptotic distribution of $\hat{\boldsymbol{\theta}}_2$ in the SR approach may also be obtained solely based on first-order derivatives. If we instead use SMM, the covariance matrix in (47) must be multiplied by $1 + 1/\tau$ where τ controls the length of the simulated sample path (see (29) and (30)).

If the SR approach is used with relative few observations in the cross-section dimension, it may be desirable to extend (47) with a finite sample correction that accounts for estimation uncertainty in $\hat{\boldsymbol{\theta}}_{11}$. To derive this correction we draw on assumption (4) of no correlation between the measurement errors $\mathbf{v}_t$ and the factor innovations $\mathbf{w}_t$, implying that $\hat{\boldsymbol{\theta}}_2$ and $\hat{\boldsymbol{\theta}}_{11}$ are uncorrelated as stated below in Proposition 3. Including the finite sample correction, the co-variance matrix for $\boldsymbol{\theta}_2$ is then given by (see Appendix G)

$$Var\left(\hat{\boldsymbol{\theta}}_2\right) = \frac{\mathbf{V}_{\boldsymbol{\theta}_2}}{T} + \left(\mathbf{A}_o^{\boldsymbol{\theta}_2}\right)^{-1} \mathbf{G}_o^{\boldsymbol{\theta}_2, \boldsymbol{\theta}_{11}} \frac{\mathbf{V}_{\boldsymbol{\theta}_{11}}}{N} \left(\mathbf{G}_o^{\boldsymbol{\theta}_2, \boldsymbol{\theta}_{11}}\right)' \left(\mathbf{A}_o^{\boldsymbol{\theta}_2}\right)^{-1}, \tag{48}$$

where $\mathbf{A}_o^{\boldsymbol{\theta}_2} \equiv \mathbf{R}_o^{\boldsymbol{\theta}_2 \prime} \mathbf{W} \mathbf{R}_o^{\boldsymbol{\theta}_2}$, $\mathbf{R}_o^{\boldsymbol{\theta}_{11}} \equiv \frac{\partial \mathbf{q}_T(\boldsymbol{\theta}_{11}^o, \boldsymbol{\theta}_2^o)'}{\partial \boldsymbol{\theta}_{11}}$, and $\mathbf{G}_o^{\boldsymbol{\theta}_2, \boldsymbol{\theta}_{11}} \equiv \mathbf{R}_o^{\boldsymbol{\theta}_2 \prime} \mathbf{W} \mathbf{R}_o^{\boldsymbol{\theta}_{11}}$. Finally, $\mathbf{V}_{\boldsymbol{\theta}_{11}}/N$ denotes the covariance matrix for $\hat{\boldsymbol{\theta}}_{11}$ as derived in (32).[10]

[9] The same argument implies that we may treat $\hat{\boldsymbol{\theta}}_1$ as known for the purpose of computing the size of the measurement error in the latent factors when conducting time series inference in the second step of the SR approach. This is also seen from (41), as $c_t \longrightarrow 0$ for $T \longrightarrow \infty$ and $\mathbf{V}_t^{\mathbf{x}_2}$ therefore reduces to $\mathbf{B}_{t,o}^{\mathbf{x}_2}$.

[10] Consider the alternative situation mentioned in footnote 2, where only $\boldsymbol{\theta}_{22}$ is estimated in the second step conditioned on the estimates of $\boldsymbol{\theta}_{12}$ in the first step. The results derived in (47) and (48) also apply in this situation if we in these formulas replace $\boldsymbol{\theta}_2$ by $\boldsymbol{\theta}_{22}$ and $\boldsymbol{\theta}_{11}$ by $\boldsymbol{\theta}_1$.

5 Sequential identification: a third step in the SR approach

The SR approach relies on sequential identification which enables us to estimate DTSMs based on assumptions that are weaker than those commonly imposed in the literature. Another benefit from sequential identification relates to the use of numerical optimizers, which often show faster convergence and encounter less frequent problems with local optima when used separately for a subset of coefficients rather than simultaneously across all parameters. The main disadvantages of sequential identification are a potential loss of efficiency and any implementation issues faced by the researcher if some parameters can be identified in both the first and second step of the SR approach. To address the potential loss of efficiency, we next extend the SR approach with a third step which combines estimates from the previous two steps in an optimal way. In so doing, we also provide guidelines on how to deal with parameters that can be identified in both the first and second step of the SR approach.

We proceed as follows. Section 5.1 presents the third estimation step, while Section 5.2 derives its asymptotic properties.

5.1 The third estimation step

To formalize our discussion, recall that $\boldsymbol{\theta}_1 = \begin{bmatrix} \boldsymbol{\theta}_{11}' & \boldsymbol{\theta}_{12}' \end{bmatrix}'$ where $\boldsymbol{\theta}_{11}$ is only identified from the first step whereas $\boldsymbol{\theta}_{12}$ may be determined in both the first and second step of the SR approach. Here, $\boldsymbol{\theta}_{11}$ has dimension L_{11} and $\boldsymbol{\theta}_{12}$ has dimension L_{12}, implying $L_1 = L_{11} + L_{12}$. For Gaussian ATSMs, the covariance matrix for factor innovations belongs to $\boldsymbol{\theta}_{12}$ because it is required for derivation of the term structure and to determine the factor dynamics. A similar situation arises in the models of Dai & Singleton (2000), where parameters describing the stochastic volatility appear in the term structure and in the law of motion for the factors.[11]

The suggested procedure for dealing with this setting is to start by estimating $\boldsymbol{\theta}_{11}$ and $\boldsymbol{\theta}_{12}$ in the first step. These estimates are denoted by $\hat{\boldsymbol{\theta}}_1^{step1} \equiv \begin{bmatrix} \left(\hat{\boldsymbol{\theta}}_{11}^{step1}\right)' & \left(\hat{\boldsymbol{\theta}}_{12}^{step1}\right)' \end{bmatrix}'$. Based on $\hat{\boldsymbol{\theta}}_1^{step1}$ and $\left\{\mathbf{x}_{1,t}, \hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1^{step1}\right)\right\}_{t=1}^{T}$, we then estimate $\boldsymbol{\theta}_2 = \begin{bmatrix} \boldsymbol{\theta}_{22}' & \boldsymbol{\theta}_{12}' \end{bmatrix}'$ in the second step by GMM or SMM. These estimates are referred to as $\hat{\boldsymbol{\theta}}_2^{step2}$.[12] The third step in the SR approach starts by combining the two estimates of $\boldsymbol{\theta}_{12}$ using

$$\hat{\boldsymbol{\theta}}_{12}^{step3} = \boldsymbol{\Lambda}\hat{\boldsymbol{\theta}}_{12}^{step1} + (\mathbf{I} - \boldsymbol{\Lambda})\,\hat{\boldsymbol{\theta}}_{12}^{step2}. \tag{49}$$

[11] We are grateful to an anonymous referee for making this point.

[12] The order in which $\hat{\boldsymbol{\theta}}_{12}^{step1}$ and $\hat{\boldsymbol{\theta}}_{12}^{step2}$ are computed is optional when $\boldsymbol{\theta}_{12}$ only relates to observed factors.

To derive the optimal value of $\mathbf{\Lambda}$ we rely on the following proposition, with the proof provided in Appendix H:

Proposition 3 *Assumption (4) and consistency of* $\hat{\boldsymbol{\theta}}_1^{step1}$ *and* $\{\hat{\mathbf{x}}_{2,t}\}_{t=1}^T$ *imply* $Cov(\hat{\boldsymbol{\theta}}_1^{step1}, \hat{\boldsymbol{\theta}}_{12}^{step2}) = \mathbf{0}$ *and* $Cov(\hat{\boldsymbol{\theta}}_1^{step1}, \hat{\boldsymbol{\theta}}_{22}^{step2}) = \mathbf{0}$ *when* $n_{y,t} \longrightarrow \infty$ *for all* t.

This proposition clearly implies

$$Var\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right) = \mathbf{\Lambda} Var\left(\hat{\boldsymbol{\theta}}_{12}^{step1}\right)\mathbf{\Lambda}' + (\mathbf{I}-\mathbf{\Lambda})\, Var\left(\hat{\boldsymbol{\theta}}_{12}^{step2}\right)(\mathbf{I}-\mathbf{\Lambda})',$$

and Appendix I shows that $Var\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)$ is minimized when

$$\mathbf{\Lambda} = Var\left(\hat{\boldsymbol{\theta}}_{12}^{step2}\right)\left(Var\left(\hat{\boldsymbol{\theta}}_{12}^{step1}\right) + Var\left(\hat{\boldsymbol{\theta}}_{12}^{step2}\right)\right)^{-1}. \tag{50}$$

Hence, $\hat{\boldsymbol{\theta}}_{12}^{step3}$ forms the optimal combination between the cross-section information in $\hat{\boldsymbol{\theta}}_{12}^{step1}$ and the time series information in $\hat{\boldsymbol{\theta}}_{12}^{step2}$. This implies that $\hat{\boldsymbol{\theta}}_{12}^{step3}$ in general achieves higher efficiency than $\hat{\boldsymbol{\theta}}_{12}^{step1}$ and $\hat{\boldsymbol{\theta}}_{12}^{step2}$, unless $\mathbf{\Lambda} \approx \mathbf{0}$ or $\mathbf{\Lambda} \approx \mathbf{I}$, in which case $\hat{\boldsymbol{\theta}}_{12}^{step3}$ simply selects the most efficient among the two estimators of $\boldsymbol{\theta}_{12}$.

Based on the more accurate estimate of $\boldsymbol{\theta}_{12}$ in the third step it is natural to re-estimate $\boldsymbol{\theta}_{11}$. That is, the estimator of $\boldsymbol{\theta}_{11}$ in the third step is given by

$$\hat{\boldsymbol{\theta}}_{11}^{step3} = \arg\min_{\boldsymbol{\theta}_{11}\in\Theta_{11}} \frac{1}{2N}\sum_{t=1}^{T}\sum_{j=1}^{n_{y,t}}\left(\tilde{y}_{t,j}\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right) - \tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}\left(\boldsymbol{\theta}_{11},\hat{\boldsymbol{\theta}}_{12}^{step3}\right),\mathbf{z}_t;\boldsymbol{\theta}_{11},\hat{\boldsymbol{\theta}}_{12}^{step3},\hat{\boldsymbol{\gamma}}\right)\right)^2, \tag{51}$$

where we condition on $\hat{\boldsymbol{\theta}}_{12}^{step3}$ and only optimize across $\boldsymbol{\theta}_{11}$. As $\hat{\boldsymbol{\theta}}_{12}^{step3}$ generally achieves higher efficiency than $\hat{\boldsymbol{\theta}}_{12}^{step1}$, we expect $\hat{\boldsymbol{\theta}}_{11}^{step3}$ to be more efficient than $\hat{\boldsymbol{\theta}}_{11}^{step1}$.

To summarize, let us outline the SR approach when extended with the third estimation step:

- Step 1: Obtain $\hat{\boldsymbol{\theta}}_1^{step1}$ and $\left\{\hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1^{step1}\right)\right\}_{t=1}^T$ from (12).
- Step 2: Based on $\hat{\boldsymbol{\theta}}_1^{step1}$ and $\left\{\mathbf{x}_{1,t},\hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1^{step1}\right)\right\}_{t=1}^T$, compute $\hat{\boldsymbol{\theta}}_2^{step2} = \left[\begin{array}{cc}\left(\boldsymbol{\theta}_{22}^{step2}\right)' & \left(\boldsymbol{\theta}_{12}^{step2}\right)'\end{array}\right]'$ by GMM or SMM.
- Step 3: Use (49) and (50) to determine $\hat{\boldsymbol{\theta}}_{12}^{step3}$, and obtain $\hat{\boldsymbol{\theta}}_{11}^{step3}$ from (51).

After the third step, it is optional to re-run the second step using $\hat{\boldsymbol{\theta}}_1^{step3} \equiv \left[\begin{array}{cc} \left(\hat{\boldsymbol{\theta}}_{11}^{step3}\right)' & \left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)' \end{array} \right]'$, or even to iterate on the second and third steps until convergence is achieved for $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_1$.[13] Such additional steps may improve the finite sample performance of the SR approach but do not alter its asymptotic properties because each step provides consistent estimates.

5.2 Inference for the third estimation step

Deriving the asymptotic distribution of $\hat{\boldsymbol{\theta}}_1^{step3}$ is complicated by the fact that $\hat{\boldsymbol{\theta}}_1^{step3}$ combines a cross-section estimator ($\hat{\boldsymbol{\theta}}_1^{step1}$) and a time series estimator ($\hat{\boldsymbol{\theta}}_{12}^{step2}$). One possibility would therefore be to study the distribution of $\hat{\boldsymbol{\theta}}_1^{step3}$ when both the cross-section dimension $n_{y,t}$ and the time series dimension T simultaneously tend to infinity. For the estimation of $\boldsymbol{\theta}_1$, such double limit asymptotics would require stronger assumptions about the law of motion for the factors and the measurement errors than currently imposed in our cross-sectional framework. To avoid such additional assumptions, we adopt another approach where the probability distribution of $\hat{\boldsymbol{\theta}}_1^{step3}$ is obtained directly from the limiting distributions of $\hat{\boldsymbol{\theta}}_1^{step1}$ and $\hat{\boldsymbol{\theta}}_{12}^{step2}$ as derived in the first two steps of the SR approach. To illustrate our way of reasoning, consider $\hat{\boldsymbol{\theta}}_{12}^{step3}$ in (49). We know from the first step that $\hat{\boldsymbol{\theta}}_{12}^{step1}$ is approximately $\mathcal{N}\left(\boldsymbol{\theta}_{12}^o, \mathbf{V}_{\boldsymbol{\theta}_{12}}^{step1}/N\right)$. From the second step we have that $\hat{\boldsymbol{\theta}}_{12}^{step2}$ is approximately $\mathcal{N}\left(\boldsymbol{\theta}_{12}^o, \mathbf{V}_{\boldsymbol{\theta}_{12}}^{step2}/T\right)$. Conditioning on these distributions, we therefore conclude that $\hat{\boldsymbol{\theta}}_{12}^{step3}$ is also approximately normally distributed, i.e. $\hat{\boldsymbol{\theta}}_{12}^{step3} \sim \mathcal{N}\left(\boldsymbol{\theta}_{12}^o, Var\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)\right)$ where

$$Var\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right) \equiv \boldsymbol{\Lambda}\frac{\mathbf{V}_{\boldsymbol{\theta}_{12}}^{step1}}{N}\boldsymbol{\Lambda}' + (\mathbf{I}-\boldsymbol{\Lambda})\frac{\mathbf{V}_{\boldsymbol{\theta}_{12}}^{step2}}{T}(\mathbf{I}-\boldsymbol{\Lambda})'. \tag{52}$$

Given that $\mathbf{V}_{\boldsymbol{\theta}_{12}}^{step1}$ and $\mathbf{V}_{\boldsymbol{\theta}_{12}}^{step2}$ are unknown, the operational version of (52) applies the estimated covariances from the first and second step of the SR approach, i.e.

$$\widehat{Var}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right) = \hat{\boldsymbol{\Lambda}}\frac{\hat{\mathbf{V}}_{\boldsymbol{\theta}_{12}}^{step1}}{N}\hat{\boldsymbol{\Lambda}}' + \left(\mathbf{I}-\hat{\boldsymbol{\Lambda}}\right)\frac{\hat{\mathbf{V}}_{\boldsymbol{\theta}_{12}}^{step2}}{T}\left(\mathbf{I}-\hat{\boldsymbol{\Lambda}}\right)',$$

where $\hat{\boldsymbol{\Lambda}}$ is given by

$$\hat{\boldsymbol{\Lambda}} = \hat{\mathbf{V}}_{\boldsymbol{\theta}_{12}}^{step2}\left(\hat{\mathbf{V}}_{\boldsymbol{\theta}_{12}}^{step1}\frac{T}{N} + \hat{\mathbf{V}}_{\boldsymbol{\theta}_{12}}^{step2}\right)^{-1}.$$

Using similar arguments and a linear approximation of the relationship between $\hat{\boldsymbol{\theta}}_{11}^{step3}$ and $\hat{\boldsymbol{\theta}}_{12}^{step3}$,

[13] In this situation, the finite sample correction in (48) may be further refined because there may be a non-zero covariance between $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{q}_t\left(\boldsymbol{\theta}_{11}^o, \boldsymbol{\theta}_2^o\right)$ and $\hat{\boldsymbol{\theta}}_{11}^{step3}$, as $\hat{\boldsymbol{\theta}}_{11}^{step3}$ depends on time-series information through $\hat{\boldsymbol{\theta}}_{12}^{step3}$ when $\boldsymbol{\Lambda} \neq \mathbf{I}$.

we show in Appendix J that $\hat{\boldsymbol{\theta}}_{11}^{step3} \sim \mathcal{N}\left(\boldsymbol{\theta}_{11}^{o}, Var\left(\hat{\boldsymbol{\theta}}_{11}^{step3}\right)\right)$ where

$$Var\left(\hat{\boldsymbol{\theta}}_{11}^{step3}\right) \equiv \frac{\mathbf{V}_{\boldsymbol{\theta}_{11}}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right)}{N} + \mathbf{K}Var\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)\mathbf{K}' + \mathbf{K}\boldsymbol{\Lambda}\mathbf{U} + \mathbf{U}'\boldsymbol{\Lambda}'\mathbf{K}'. \tag{53}$$

Here, $\mathbf{U} \equiv Cov\left(\hat{\boldsymbol{\theta}}_{12}^{step1}, \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right)\right)$ has dimension $L_{12} \times L_{11}$ and $\mathbf{K} \equiv \partial \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right) / \partial \boldsymbol{\theta}_{12}'$ is of dimension $L_{11} \times L_{12}$. The first term in (53), i.e. $\mathbf{V}_{\boldsymbol{\theta}_{11}}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right)/N$, is the covariance matrix for $\hat{\boldsymbol{\theta}}_{11}^{step3}$ when we use the true (but infeasible) value of $\boldsymbol{\theta}_{12}$. The specific expression for $\mathbf{V}_{\boldsymbol{\theta}_{11}}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right)$ follows from (32) when used on the subset of $\boldsymbol{\theta}_1$ corresponding to $\boldsymbol{\theta}_{11}$. The remaining terms in (53) account for the additional uncertainty in $\hat{\boldsymbol{\theta}}_{11}^{step3}$ stemming from conditioning the estimation on $\hat{\boldsymbol{\theta}}_{12}^{step3}$ instead of the true value $\boldsymbol{\theta}_{12}^{o}$. Given standard regularity conditions, the operational version of (53) is

$$\widehat{Var}\left(\hat{\boldsymbol{\theta}}_{11}^{step3}\right) = \frac{\hat{\mathbf{V}}_{\boldsymbol{\theta}_{11}}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)}{N} + \hat{\mathbf{K}}\widehat{Var}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)\hat{\mathbf{K}}' + \hat{\mathbf{K}}\hat{\boldsymbol{\Lambda}}\hat{\mathbf{U}} + \hat{\mathbf{U}}'\hat{\boldsymbol{\Lambda}}'\hat{\mathbf{K}}'. \tag{54}$$

The matrix $\mathbf{K}$ may be estimated using $\hat{\mathbf{K}} = -\left(\hat{\mathbf{A}}_{step3}^{\boldsymbol{\theta}_{11}}\right)^{-1}\hat{\mathbf{J}}$, where $\hat{\mathbf{A}}_{step3}^{\boldsymbol{\theta}_{11}}$ follows from (40) when used on the subset of $\boldsymbol{\theta}_1$ corresponding to $\boldsymbol{\theta}_{11}$ and evaluated at $\hat{\boldsymbol{\theta}}_1^{step3}$. To provide the expression for $\hat{\mathbf{J}}$, let us decompose $\boldsymbol{\Psi}_{t,j}^{\boldsymbol{\theta}_1}\left(\boldsymbol{\theta}_1, \boldsymbol{\gamma}\right)$ in (34) as $\left[\begin{array}{cc} \boldsymbol{\Psi}_{t,j}^{\boldsymbol{\theta}_{11}}\left(\boldsymbol{\theta}_1, \boldsymbol{\gamma}\right)' & \boldsymbol{\Psi}_{t,j}^{\boldsymbol{\theta}_{12}}\left(\boldsymbol{\theta}_1, \boldsymbol{\gamma}\right)' \end{array}\right]'$. Hence, the score functions with respect to $\boldsymbol{\theta}_{11}$ and $\boldsymbol{\theta}_{12}$ at time t for the j'th observation are given by $\mathbf{s}_{t,j}^{\boldsymbol{\theta}_{11}} = \tilde{v}_{t,j}\left(\mathbf{z}_i; \boldsymbol{\gamma}\right)\boldsymbol{\Psi}_{t,j}^{\boldsymbol{\theta}_{11}}\left(\boldsymbol{\theta}_1, \boldsymbol{\gamma}\right)$ and $\mathbf{s}_{t,j}^{\boldsymbol{\theta}_{12}} = \tilde{v}_{t,j}\left(\mathbf{z}_i; \boldsymbol{\gamma}\right)\boldsymbol{\Psi}_{t,j}^{\boldsymbol{\theta}_{12}}\left(\boldsymbol{\theta}_1, \boldsymbol{\gamma}\right)$, respectively. Based on this notation, we have (see Appendix J)

$$\hat{\mathbf{J}} = \frac{1}{N}\sum_{t=1}^{T}\sum_{j=1}^{n_{y,t}}\boldsymbol{\Psi}_{t,j}^{\boldsymbol{\theta}_{11}}\left(\hat{\boldsymbol{\theta}}_1^{step3}, \hat{\boldsymbol{\gamma}}\right)\boldsymbol{\Psi}_{t,j}^{\boldsymbol{\theta}_{12}}\left(\hat{\boldsymbol{\theta}}_1^{step3}, \hat{\boldsymbol{\gamma}}\right)'. \tag{55}$$

It is also shown in Appendix J that the covariance term in (54) may be estimated by

$$\begin{aligned} \hat{\mathbf{U}} \;=\; & \frac{1}{N^2}\sum_{t=1}^{T}\sum_{j=1}^{n_{y,t}}\sum_{k_T=-w_T}^{w_T}\sum_{k_D=-w_D}^{w_D}\left(1-\frac{|k_T|}{1+w_T}\right)\left(1-\frac{|k_D|}{1+w_D}\right)\widehat{\tilde{v}}_{t,j}^{step1}\widehat{\tilde{v}}_{t+k_T,j+k_D}^{step3} \\ & \times\left[\hat{\boldsymbol{\Omega}}_{21}^{step1}\boldsymbol{\Psi}_{t,j}^{\boldsymbol{\theta}_{11}}\left(\hat{\boldsymbol{\theta}}_1^{step1}, \hat{\boldsymbol{\gamma}}\right) + \hat{\boldsymbol{\Omega}}_{22}^{step1}\boldsymbol{\Psi}_{t,j}^{\boldsymbol{\theta}_{12}}\left(\hat{\boldsymbol{\theta}}_1^{step1}, \hat{\boldsymbol{\gamma}}\right)\right]\boldsymbol{\Psi}_{t+k_T,j+k_D}^{\boldsymbol{\theta}_{11}}\left(\hat{\boldsymbol{\theta}}_1^{step3}, \hat{\boldsymbol{\gamma}}\right)'\left(\hat{\mathbf{A}}_{step3}^{\boldsymbol{\theta}_{11}}\right)^{-1}, \end{aligned} \tag{56}$$

which is robust to heteroskedasticity and correlation in both the cross-section and time series dimension. In (56), we let $\widehat{\tilde{v}}_{t,j}^{step1} \equiv \tilde{y}_{t,j}\left(\mathbf{z}_t, \hat{\boldsymbol{\gamma}}\right) - \tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1^{step1}\right), \mathbf{z}_t; \hat{\boldsymbol{\theta}}_1^{step1}, \hat{\boldsymbol{\gamma}}\right)$ and $\widehat{\tilde{v}}_{t,j}^{step3} \equiv \tilde{y}_{t,j}\left(\mathbf{z}_i, \hat{\boldsymbol{\gamma}}\right) - \tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1^{step3}\right), \mathbf{z}_t; \hat{\boldsymbol{\theta}}_1^{step3}, \hat{\boldsymbol{\gamma}}\right)$. The matrices $\hat{\boldsymbol{\Omega}}_{21}^{step1}$ and $\hat{\boldsymbol{\Omega}}_{22}^{step1}$ refer to the lower part of the inverse

of the Hessian matrix for $\boldsymbol{\theta}_1$ when evaluated at $\hat{\boldsymbol{\theta}}_1^{step1}$, i.e.

$$\begin{bmatrix} \hat{\boldsymbol{\Omega}}_{11}^{step1} & \hat{\boldsymbol{\Omega}}_{12}^{step1} \\ \hat{\boldsymbol{\Omega}}_{21}^{step1} & \hat{\boldsymbol{\Omega}}_{22}^{step1} \end{bmatrix} \equiv \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{s}_i^{\boldsymbol{\theta}_1} \left(\hat{\boldsymbol{\theta}}_1^{step1}, \hat{\boldsymbol{\gamma}} \right)}{\partial \boldsymbol{\theta}_1'} \right)^{-1}. \tag{57}$$

The estimator in (56) may be simplified in the absence of heteroskedasticity, cross-sectional correlation, and/or autocorrelation. When considering the homoskedastic version, we suggest replacing $\widehat{\widetilde{v}}_{t,j}^{step1} \widehat{\widetilde{v}}_{t,j}^{step3}$ by $\left(\hat{\sigma}_i^{step3} \right)^2$, where

$$\left(\hat{\sigma}_t^{step3} \right)^2 = \frac{1}{n_{y,t} - n_{x_2}} \sum_{j=1}^{n_{y,t}} \left(\widehat{\widetilde{v}}_{t,j}^{step3} \left(\mathbf{z}_t; \hat{\boldsymbol{\gamma}} \right) \right)^2 \quad \text{for } t = 1, 2, ..., T.$$

As was the case in the first two steps, we are also able to estimate all asymptotic covariance matrices based on first-order derivatives in the third step of the SR approach.

6 Relating the SR approach to the literature

This section compares the SR approach to the existing literature. Section 6.1 starts by relating the regression filter to other filtering methods, while Section 6.2 discusses how our estimator of $\boldsymbol{\theta}_1$ differs from the one in Pastorello, Patilea & Renault (2003). Section 6.3 compares our estimator of $\boldsymbol{\theta}_2$ to the work by Pan (2002), and we finally relate the SR approach to other regression-based estimation methods for DTSMs in Section 6.4.

6.1 The regression filter and other filtering methods

To compare the SR approach to other filtering methods, we start by imposing some additional assumptions that are standard in the filtering literature, although not necessary for the SR approach: i) $\mathbf{v}_t$ and $\mathbf{w}_t$ are independent, normally distributed, and have constant covariance matrices, ii) $\mathbf{x}_0$ is normally distributed, and iii) $\mathbf{x}_0$, $\mathbf{v}_t$, and $\mathbf{w}_t$ are mutually uncorrelated at all leads and lags. To reduce the notational burden, let $\mathbf{x}_{1,t}$ be empty, so that $\mathbf{x}_{2,t} = \mathbf{x}_t$. In this setup, the logarithm of the conditional probability density of $\mathbf{x}_{1:T} \equiv \{\mathbf{x}_t\}_{t=1}^T$ given $\mathbf{y}_{1:T} \equiv \{\mathbf{y}_t\}_{t=1}^T$, denoted $\log p\left(\mathbf{x}_{1:T} | \, \mathbf{y}_{1:T}\right)$, is

proportional to (see Durbin & Koopman (2001))

$$\begin{aligned}\tilde{Q}_p &= -\frac{1}{2}\sum_{t=1}^{T}\left[\mathbf{y}_t-\mathbf{g}\left(\mathbf{x}_t;\boldsymbol{\theta}_1\right)\right]'\boldsymbol{\Omega}_{\mathbf{v}}\left(\mathbf{z}_t;\boldsymbol{\gamma}\right)^{-1}\left[\mathbf{y}_t-\mathbf{g}\left(\mathbf{x}_t;\boldsymbol{\theta}_1\right)\right] \\ &\quad -\frac{1}{2}\sum_{t=1}^{T}\mathbf{w}_t'\left(Var\left(\mathbf{w}_t\right)\right)^{-1}\mathbf{w}_t-\frac{1}{2}\tilde{\mathbf{x}}_0'\mathbf{P}_0^{-1}\tilde{\mathbf{x}}_0.\end{aligned} \tag{58}$$

Here, $\tilde{\mathbf{x}}_0$ denotes the estimation error for the initial factors and $\mathbf{P}_0$ is the covariance matrix of $\tilde{\mathbf{x}}_0$. In the case of a fixed number of observables in each time period, we scale (58) by $-1/n_y$ and obtain

$$Q_p=\frac{1}{2n_y}\sum_{t=1}^{T}\sum_{j=1}^{n_y}\left(\tilde{y}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}\right)-\tilde{g}_j\left(\mathbf{x}_t,\mathbf{z}_t;\boldsymbol{\theta}_1,\boldsymbol{\gamma}\right)\right)^2+\frac{1}{2n_y}\left(\sum_{t=1}^{T}\mathbf{w}_t'\left(Var\left(\mathbf{w}_t\right)\right)^{-1}\mathbf{w}_t+\tilde{\mathbf{x}}_0'\mathbf{P}_0^{-1}\tilde{\mathbf{x}}_0\right), \tag{59}$$

using the notation for the rotated version of the model. We then impose the standard assumption in cross-section regressions that

$$\frac{1}{n_y}\sum_{j=1}^{n_y}\left(\tilde{y}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}\right)-\tilde{g}_j\left(\mathbf{x}_t,\mathbf{z}_t;\boldsymbol{\theta}_1,\boldsymbol{\gamma}\right)\right)^2\xrightarrow{p}E\left[\left(\tilde{y}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}\right)-\tilde{g}_j\left(\mathbf{x}_t,\mathbf{z}_t;\boldsymbol{\theta}_1,\boldsymbol{\gamma}\right)\right)^2\right]$$

for $n_y\longrightarrow\infty$. Accordingly, (59) converges to $\frac{1}{2}\sum_{t=1}^{T}E\left[\left(\tilde{y}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}\right)-\tilde{g}_j\left(\mathbf{x}_t,\mathbf{z}_t;\boldsymbol{\theta}_1,\boldsymbol{\gamma}\right)\right)^2\right]$ when n_y tends to infinity. This expression corresponds to the population terms minimized by the regression filter which therefore converges to the mode of $p\left(\mathbf{x}_{1:T}|\,\mathbf{y}_{1:T}\right)$. The values of $\mathbf{x}_{1:T}$ which maximize $p\left(\mathbf{x}_{1:T}|\,\mathbf{y}_{1:T}\right)$ are the most probable values for the latent factors, and these estimates are therefore optimal (Durbin & Koopman (2001)). Accordingly, the regression filter converges to the optimal estimates as the number of observables tends to infinity. For linear systems, this implies that the regression filter converges to the Kalman smoother which reports the mode of $p\left(\mathbf{x}_{1:T}|\,\mathbf{y}_{1:T}\right)$.[14]

The above argument also illustrates what the regression filter is missing in achieving efficiency — namely, smoothing of the estimated latent factors according to their transition equations. Ignoring this smoothing can be justified by i) many observables in the cross-section dimension or ii) small measurement errors, thus making the observables more informative about the latent factors.

The SR approach is also related to estimation of factors in standard or approximate factor models. For these models, Stock & Watson (2002) show how factors can be estimated consistently using the method of principal components when a large number of observables are available in each time period. This implies that the factor dynamics are estimated non-parametrically, as in the SR approach. A key

[14] Note that the derived results can be generalized to the case where $\mathbf{w}_t$ and $\mathbf{x}_0$ are not necessarily normally distributed.

difference between the SR approach and the method of principal components is that the latter only works for linear models with stationary time series. In contrast, the SR approach can easily handle non-linear models and potentially non-stationary time series.

6.2 The estimator for $\boldsymbol{\theta}_1$

Our suggested estimator of $\boldsymbol{\theta}_1$ in (12) is related to the work by Pastorello et al. (2003). One way to implement their backfitting estimator in our setting is to determine the latent factors by the regression filter in (9), and subsequently let the objective function for $\boldsymbol{\theta}_1$ be the sum of squared residuals from these regressions. In the framework of Pastorello et al. (2003), the estimated factors appear as nuisance parameters in the estimation of $\boldsymbol{\theta}_1$ based on

$$\hat{\boldsymbol{\theta}}_1^{p+1} = \arg\min_{\boldsymbol{\theta}_1 \in \Theta_1} \quad \frac{1}{2N} \sum_{t=1}^{T} \sum_{j=1}^{n_{y,t}} \left(\tilde{y}_{t,j} \left(\mathbf{z}_t; \boldsymbol{\gamma} \right) - \tilde{g}_j \left(\hat{\mathbf{x}}_{2,t} \left(\boldsymbol{\theta}_1^p \right), \mathbf{z}_t; \boldsymbol{\theta}_1, \boldsymbol{\gamma} \right) \right)^2 \quad \text{for } p = 1, 2, ..., P. \tag{60}$$

The important difference between this estimator and the one in (12) is that (60) only accounts for the direct effects of $\boldsymbol{\theta}_1$ on the $\mathbf{g}$-function when estimating $\hat{\boldsymbol{\theta}}_1^{p+1}$. That is, the indirect effect of $\boldsymbol{\theta}_1$ on the latent factors is not accounted for when computing $\hat{\boldsymbol{\theta}}_1^{p+1}$, making the optimization computationally simpler. The estimate $\hat{\boldsymbol{\theta}}_1^{p+1}$ gives rise to a new set of estimated factors $\left\{ \hat{\mathbf{x}}_{2,t} \left(\boldsymbol{\theta}_1^{p+1} \right) \right\}_{t=1}^{T}$ which can be substituted into (60) to find a new estimate of $\boldsymbol{\theta}_1$, denoted $\hat{\boldsymbol{\theta}}_1^{p+2}$. The idea behind the backfitting estimator is then to iterate this procedure until convergence, i.e. until $\hat{\boldsymbol{\theta}}_1^{p+2} \simeq \hat{\boldsymbol{\theta}}_1^{p+1}$.

Pastorello et al. (2003) state conditions to ensure this convergence, and they show consistency and asymptotic normality of the backfitting estimator. In establishing these results, they face the issue of nonadaptivity, meaning that the estimate of the latent factors (which depend on $\boldsymbol{\theta}_1$) prevents the econometrician from directly estimating $\boldsymbol{\theta}_1$ consistently. The issue is resolved by assuming that the mapping defined by (60) is contracting and has a unique fixed point at the true value of $\boldsymbol{\theta}_1$. As pointed out by Pastorello et al. (2003) and Sherman (2003), the requirements for consistency and asymptotic normality of iterated estimators are stronger than the conditions for optimization based estimators. In other words, our estimator in (12) requires weaker assumptions to ensure consistency and asymptotic normality than the backfitting estimator in (60). This is because our estimator in (12) does not face the nonadaptivity problem, as we account for the dependence of $\boldsymbol{\theta}_1$ on the latent factors during the optimization with respect to $\boldsymbol{\theta}_1$.

6.3 The estimator of $\boldsymbol{\theta}_2$

Our estimator of $\boldsymbol{\theta}_2$ is related to the Implied State (IS) GMM estimator by Pan (2002). In the IS-GMM estimator, n_x factors are backed out from the same number of observables which are assumed to be measured without errors. These factors are then used to evaluate a set of GMM moment conditions along the time series dimension, leading to estimates of the model parameters. Thus, our estimator of $\boldsymbol{\theta}_2$ supplements the IS-GMM estimator, as we consider the case where all observables are measured with errors.

6.4 Alternative regression-based estimation methods

We finally relate the SR approach to four papers which also propose estimating Gaussian ATSMs by regressions. The first method we consider is the one by Joslin et al. (2011), where K factors are recovered from K linear combinations of yields which are perfectly priced by the model. The authors propose a normalization of Gaussian ATSMs which has the convenient property that the parameters in the likelihood function can be decomposed into two subsets. The first subset describes the dynamics of the factors under the physical measure and their ML estimator is shown to coincide with the well-known OLS estimator for VAR systems. The second subset contains the risk-neutral coefficients along with the few remaining parameters which are readily found by numerical optimization. As shown by Joslin et al. (2011), the ability to separate the parameter space greatly simplifies estimation of Gaussian ATSMs. The SR approach allows for the same convenient decomposition of the parameters, and this makes the estimation problem straightforward to solve with standard optimizers. Unlike the SR approach, the separation of the parameter space and the simplified estimation procedure in Joslin et al. (2011) no longer apply when i) there are cross-restrictions between the factor dynamics under the physical and risk-neutral measure, ii) all yields are observed with errors, or iii) measurement errors in yields display autocorrelation.[15]

Another regression-based estimator for Gaussian ATSMs is the one by Hamilton & Wu (2012). They impose the same assumption as in Joslin et al. (2011) where K factors can be recovered from K linear combinations of yields which are perfectly priced. This implies that the reduced-form VAR system for the Gaussian ATSM can be estimated by OLS. Hamilton & Wu (2012) then propose a minimum chi-squared estimator (MCSE) where the structural parameters in the Gaussian ATSM are

[15] Although less computationally efficient, the normalization in Joslin et al. (2011) remains useful when all yields have measurement errors and the likelihood function is evaluated by the Kalman filter. We rely on this property in Section 7 when conducting a Monte Carlo study for a Gaussian ATSM with three factors.

easily inferred from the reduced-form OLS estimates. Contrary to the method by Joslin et al. (2011), the computational gain from the MCSE does not rely on a particular normalization of the Gaussian ATSM, and the MCSE is in this sense similar to the SR approach. We also note that the MSCE achieves the same asymptotic efficiency as ML, meaning that the methods in Hamilton & Wu (2012) and Joslin et al. (2011) are asymptotically equivalent and more efficient than the SR approach. This efficiency gain only holds in correctly specified models, while the SR approach remains robust to departures from several assumptions needed for likelihood inference, as mentioned in Section 2.

The methods by Joslin et al. (2011) and Hamilton & Wu (2012) have subsequently been extended with a small-sample bias correction by Bauer, Rudebusch & Wu (2012). They suggest an inverse bootstrap method which substantially reduces the negative bias in OLS estimates of the auto-correlation function for strongly persistent VAR systems. This bias correction is relevant for the second step of the SR approach and can be directly applied when all factors are observed, as assumed in Bauer et al. (2012). With latent factors in the model, the bootstrap procedure requires a minor modification to account for measurement errors in the estimated factors, but is also in this case straightforward to implement in the SR approach.

The third regression method we consider is the one by Adrian, Crump & Moench (2013). As in Joslin et al. (2011) and Hamilton & Wu (2012), the factors are assumed to be observed from linear combinations of yields. Adrian et al. (2013) then show how to estimate Gaussian ATSMs by running three linear regressions. Hence, Adrian et al. (2013) do not rely on any numerical optimization, but this property is only achieved because the restrictions implied by the Riccatti equations for zero-coupon bond prices are not imposed in the estimation. On the other hand, these restrictions are imposed in the SR approach and the two aforementioned regression methods. Similarly to our approach, the method in Adrian et al. (2013) allows for autocorrelation in the measurement errors, as seen in several affine DTSMs (Hamilton & Wu (2014)).

It is finally worth stressing that these alternative regression-based methods are applicable only for *linear* models. This is in contrast to the SR approach which applies to linear and *non-linear* DTSMs. Another notable benefit of the SR approach is that all yields or bond prices may enter in the expression for the latent factors. That is, we do not assume that K factors can be determined from K linear combinations of yields without pricing errors. The results by Hamilton & Wu (2014) strongly suggest that this assumption should be avoided because it is generally rejected by the data.

7 Finite sample properties of the SR approach

This section carries out three Monte Carlo experiments to analyze the finite sample properties of the SR approach. We first study the popular three-factor Gaussian ATSM in Section 7.1. Section 7.2 considers a two-factor QTSM to explore how the SR approach deals with nonlinearities in the measurement equations. To study how the SR approach performs when the latent factors display stochastic volatility, we finally consider the CIR model in Section 7.3. For each model, we first describe the design of the Monte Carlo experiment, before analyzing the performance of the SR approach for factor and parameter estimation, respectively. Since the results for the Gaussian ATSM largely generalize to the QTSM and the CIR model, we discuss the former in greatest detail.

7.1 ATSM: The study design

The likelihood function for Gaussian ATSMs can be evaluated by the Kalman filter, allowing us to compare the SR approach to ML.[16] As previously mentioned, likelihood inference for these models has been greatly simplified by Joslin et al. (2011), and we therefore adopt their notation and normalization. They consider a setting where the dynamics for the factors $\mathcal{P}_t$ under the physical measure $\mathbb{P}$ and under the risk-neutral measure $\mathbb{Q}$ are modeled as VAR(1) systems, i.e.

$$\Delta\mathcal{P}_t = \mathbf{K}_{0\mathcal{P}}^{\mathbb{P}} + \mathbf{K}_{1\mathcal{P}}^{\mathbb{P}}\mathcal{P}_{t-1} + \boldsymbol{\Sigma}_{\mathcal{P}}\boldsymbol{\epsilon}_t^{\mathbb{P}} \tag{61}$$

$$\Delta\mathcal{P}_t = \mathbf{K}_{0\mathcal{P}}^{\mathbb{Q}} + \mathbf{K}_{1\mathcal{P}}^{\mathbb{Q}}\mathcal{P}_{t-1} + \boldsymbol{\Sigma}_{\mathcal{P}}\boldsymbol{\epsilon}_t^{\mathbb{Q}} \tag{62}$$

where $\boldsymbol{\epsilon}_t^{\mathbb{P}}$ and $\boldsymbol{\epsilon}_t^{\mathbb{Q}}$ are $\mathcal{NID}(\mathbf{0}, \mathbf{I}_3)$. The process for the short rate r_t is given by

$$r_t = \rho_{0\mathcal{P}} + \boldsymbol{\rho}_{1\mathcal{P}}\mathcal{P}_t. \tag{63}$$

Adopting the normalization in Joslin et al. (2011), this model can be parameterized by $\boldsymbol{\lambda}^{\mathbb{Q}}$, $r_\infty^{\mathbb{Q}}$, $\mathbf{K}_{0\mathcal{P}}^{\mathbb{P}}$, $\mathbf{K}_{1\mathcal{P}}^{\mathbb{P}}$, and $\boldsymbol{\Sigma}_{\mathcal{P}}$.[17] Here, $\boldsymbol{\lambda}^{\mathbb{Q}}$ denotes the vector of eigenvalues of $\mathbf{K}_{1\mathcal{P}}^{\mathbb{Q}}$ and $r_\infty^{\mathbb{Q}}$ is the long-run mean of the short rate r_t under the $\mathbb{Q}$ measure. The continuously compounded yield on a zero-coupon bond

[16] Recall that a closed-form solution for the likelihood function is generally not available for other DTSMs with latent factors.

[17] The mapping between these parameters and the ones in (61) to (63) are given in Proposition 2 of Joslin et al. (2011).

maturing in k periods is given by

$$i_{t,k}\left(\mathcal{P}_t\right) = A_k + \mathbf{B}_k \mathcal{P}_t + v_{t,k}, \tag{64}$$

where A_k and $\mathbf{B}_k$ follow a set of Ricatti equations and $v_{t,k}$ are measurement errors. As in Joslin et al. (2011), the variance in the measurement errors is specified to be constant across yields and time, i.e. $Var\left(v_{t,k}\right) = R_v$ for all t and k. Joslin et al. (2011) estimate this model on monthly US data from January 1990 to December 2007 using the 10-year yield curve. For our Monte Carlo experiment, we apply the estimates from their preferred specification where the first three principal components are used to infer the factors.[18]

In our simulation experiment we apply log-transformed zero-coupon bond prices using $\log P_{t,k} = -i_{t,k}\left(\mathcal{P}_t\right) k$.[19] In relation to our general notation, for this Gaussian ATSM we have $y_{t,k} = \log P_{t,k}$, $g_k\left(\mathbf{x}_t\right) = -i_{t,k}\left(\mathbf{x}_t\right) k$, $\mathbf{x}_t = \mathcal{P}_t$, $\mathbf{h}_0 = \mathbf{K}_{0\mathcal{P}}^{\mathbb{P}}$, $\mathbf{h}_{\mathbf{x}} = \mathbf{I} + \mathbf{K}_{1\mathcal{P}}^{\mathbb{P}}$, $\mathbf{w}_t = \boldsymbol{\epsilon}_t^{\mathbb{P}}$, and $Var\left(\mathbf{w}_{t+1}\right) = \boldsymbol{\Sigma}_{\mathcal{P}} \boldsymbol{\Sigma}_{\mathcal{P}}'$. To make the study design as realistic as possible, we allow the maturities of these zero-coupon bonds to vary between time periods. That is, at one point in time we may have zero-coupon bonds with monthly maturities of (5, 20, 60, 80, 120), whereas in the next time period we may have zero-coupon bonds with monthly maturities of (6, 10, 50, 80, 100). The specific maturities available at a given point in time are derived by partitioning the 10-year yield curve into three equally sized segments according to maturity and then sampling randomly from each of these segments. This sampling procedure ensures that we always have bonds with short, medium, and long maturities, as in empirical data.

The length of the simulated time series is set to 480 periods, corresponding to 40 years of monthly data. As for the number of bonds in each time period, we examine the performance of likelihood inference by starting with a minimum of 5 bonds and then gradually increase this figure. For the SR approach, a minimum of 10 bonds is considered.

We explore two scenarios in our stimulation experiment. In the first scenario (Case 1), all bond prices are generated from interest rates that have measurement errors with a standard deviation of 10 basis points. In the second scenario (Case 2), all interest rates have measurement errors with a standard deviation of 20 basis points. For the version of the Gaussian ATSM considered, Joslin et al.

[18] This corresponds to the so-called RPC specification in Tables 2 to 4 of Joslin et al. (2011). In order to apply these estimates, we therefore rotate the factors in our Gaussian ATSM such that they correspond to the first three principal components in the data set by Joslin et al. (2011). For implemention of the model, we are grateful to Joslin et al. (2011) for making their data set and codes publicly available.

[19] This simulation procedure induces larger measurement errors in bonds with long maturities than in bonds with short maturities. To keep the simulation study as simple as possible, we take the structure of this heteroskedasticity to be known. This implies that the subsequent results based on bonds are equivalent to using interest rates $i_{t,k}$ directly.

(2011) estimate the measurement errors in interest rates to have a standard deviation of 5.5 basis points, and we therefore consider Case 1 as the most realistic scenario.[20]

7.1.1 ATSM: Factor estimation

This section compares the accuracy of the regression filter to the optimal estimates from the Kalman smoother. The root mean squared errors (RMSEs) for the two estimators are shown in Figure 1. For Case 1, the regression filter (red lines with circles) converges relatively quickly to the Kalman smoother (black lines with stars). Based on just 25 bonds, the regression filter is close to the Kalman smoother, and the two estimators are nearly identical with 50 bonds. It is even more interesting to compare the Kalman smoother based on just 5 to 10 bonds to the regression filter using many bonds, for instance 50 to 100. Such a comparison shows that the regression filter with more observables clearly outperforms the Kalman smoother, and that the gain in precision is roughly a 50% reduction in the RMSEs. Hence, factor estimation in Gaussian ATSMs based on a large number of observables and the inefficient regression filter is clearly preferred to using 5 to 10 observables and the optimal estimator.

Turning to Case 2 where all bond prices are measured less precisely than in Case 1, convergence of the regression filter to the Kalman smoother requires about 100 bonds and is thus slower than in Case 1. Hence, the speed of convergence for the regression filter to the optimal estimator is greater when bonds are measured more precisely, as argued in Section 6.1.

< Figure 1 about here >

The average number of seconds required to evaluate the regression filter and the Kalman smoother are displayed in the first row of Figure 2. We find a steep increase in the time requirement for the Kalman smoother using the algorithm by Durbin & Koopman (2001), whereas the number of bonds hardly affects the computing time of the regression filter. This implies that the regression filter is between 5 and 16 times faster to compute than the Kalman smoother.[21]

< Figure 2 about here >

To explore the effects of potential cross-correlation in the measurement errors of bond prices, we momentarily alter our setup and introduce correlation in $v_{t,k}$ based on the specification in (14) with $\rho_t = 0.9 \times R_v$ for all time periods. For small measurement errors (i.e., Case 1), Figure 3 shows that a version of the regression filter which ignores the cross-correlation (i.e., OLS) remains highly accurate in

[20] The size of the estimated measurement errors in the RPC model was kindly provided to us by Scott Joslin.

[21] Jungbacker & Koopman (Forthcoming) show that the computing time for the Kalman filter and the related smoother with many observables can be reduced by using a transformation of the state-space system. Thus, the steep increase in computing time for the Kalman smoother can probably be somewhat reduced if their method is applied in our case.

comparison to a Kalman smoother accounting for the cross-correlation. We next modify the regression filter to exploit the cross-correlation and perform generalized least squares (GLS) regressions, indicated by the blue lines with squares in Figure 3. As expected, the GLS regressions deliver more efficient estimates of the latent factors than OLS, where the biggest improvement appears for the curvature factor. Broadly similar results are found in Case 2 with larger measurement errors, and we also note that more than 100 bonds are needed in this case for convergence of the regression filter to the Kalman smoother.

< Figure 3 about here >

This simulation experiment leads us to conclude that the regression filter delivers comparable performance to the optimal estimator for Gaussian ATSM. In the most realistic case where bond prices are generated from interest rates with measurement errors having a standard deviation of 10 basis points (Case 1), just 25 to 50 observables are needed for convergence of the regression filter to the Kalman smoother. The regression filter is also found to be robust to cross-correlation in the measurement errors for bond prices, although the speed of convergence for the regression filter is somewhat slower in this setting.

7.1.2 ATSM: Parameter estimation

We next study the finite sample properties of the SR and ML approach for parameter estimation. To apply the SR approach on our Gaussian ATSM, we let $\boldsymbol{\theta}_{11} = \begin{bmatrix} \lambda_1^{\mathbb{Q}} & \lambda_2^{\mathbb{Q}} & \lambda_3^{\mathbb{Q}} & r_\infty^{\mathbb{Q}} \end{bmatrix}'$ and note that $\boldsymbol{\theta}_{12}$ contains all non-zero elements in $\boldsymbol{\Sigma}_{\mathcal{P}}$, i.e. $\boldsymbol{\theta}_{12} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{22} & \sigma_{23} & \sigma_{33} \end{bmatrix}'$. These ten parameters must be estimated in the first step of the SR approach, in addition to the variance of the measurement errors R_v. During the optimization of $\boldsymbol{\theta}_1 = \begin{bmatrix} \boldsymbol{\theta}_{11}' & \boldsymbol{\theta}_{12}' \end{bmatrix}'$, bond prices are computed using (64), where the loadings A_k and $\mathbf{B}_k$ are determined by the appropriate Ricatti equations. All remaining parameters appear in $\boldsymbol{\theta}_{22}$, i.e. $\boldsymbol{\theta}_{22} = \begin{bmatrix} \left(\mathbf{K}_{0\mathcal{P}}^{\mathbb{P}}\right)' & vec\left(\mathbf{K}_{1\mathcal{P}}^{\mathbb{P}}\right)' \end{bmatrix}'$. The elements in $\boldsymbol{\theta}_{22}$ are estimated by (20) as all factors are latent, and we use (21) when estimating $\boldsymbol{\theta}_{12}$ in the second step of the SR approach.[22] Unless otherwise stated, the results reported for the SR approach are from the three-step procedure, where $\{\hat{\mathbf{x}}_{2,t}\}_{t=1}^T$ and $\hat{\boldsymbol{\theta}}_{22}$ are re-estimated based on $\hat{\boldsymbol{\theta}}_1^{step3}$. The performance of all estimators are studied using 1,000 repetitions of samples of 480 observations, where we gradually increase the number of bond prices in each time period.

Preliminary findings for the SR and ML approaches show that the full model is unidentified when

[22] Note that $\hat{\boldsymbol{\theta}}_{22}$ may be computed without knowledge of $\hat{\boldsymbol{\theta}}_{12}$, and $\hat{\boldsymbol{\theta}}_{22}$ is therefore unaffected if $\boldsymbol{\theta}_{12}$ is not re-estimated in the second step.

$\lambda_2^{\mathbb{Q}}$ is very close to $\lambda_3^{\mathbb{Q}}$.[23] Throughout the Monte Carlo experiment we therefore impose the constraint that $\left|\lambda_2^{\mathbb{Q}} - \lambda_3^{\mathbb{Q}}\right| > 0.001$ to ensure global identification.[24]

To structure our presentation, we first study the estimates and related biases before discussing the efficiency of the two estimators. The performance of standard errors and Type I errors are analyzed afterwards. To conserve space, we only report results for the most realistic scenario, i.e. Case I, where bond prices are generated from interest rates with measurement errors having a standard deviation of 10 basis points.[25]

Estimates: Figure 4 reports the average estimates from the three-step SR approach and ML.[26] The first four charts display elements in $\boldsymbol{\theta}_{11}$ and show minor biases in the SR estimates with 10 to 20 bonds, but these biases disappear as we increase the number of bonds. The next six charts are related to $\boldsymbol{\theta}_{12}$ where the SR estimates from just 15 bonds broadly coincide with the ML estimates. The third step SR estimates of $\boldsymbol{\theta}_{12}$ rely on Proposition 3 which clearly applies in our case, as the average absolute correlation between $\hat{\boldsymbol{\theta}}_1^{step1}$ and $\left[\begin{array}{cc} \left(\hat{\boldsymbol{\theta}}_{22}^{step2}\right)' & \left(\hat{\boldsymbol{\theta}}_{12}^{step2}\right)' \end{array}\right]'$ is around 0.02 for the number of bonds considered in Figure 4. To obtain further insight into the close relationship between the SR and ML estimates of $\boldsymbol{\theta}_{12}$, we next note that $\hat{\boldsymbol{\Lambda}} \approx \mathbf{0}$ in the third step of the SR approach because $Var\left(\hat{\boldsymbol{\theta}}_{12}^{step1}\right)$ is much larger than $Var\left(\hat{\boldsymbol{\theta}}_{12}^{step2}\right)$. Hence, within our Gaussian ATSM, cross-section information is not helpful for estimating $\boldsymbol{\Sigma}_{\mathcal{P}}$, which is purely identified from the $\mathbb{P}$ dynamics. Joslin et al. (2011) draw the same conclusion, as their first stage estimate of $\boldsymbol{\Sigma}_{\mathcal{P}}$ from the VAR system hardly changes when including yields in their second estimation step.

The second part of Figure 4 relates to elements in $\boldsymbol{\theta}_{22}$. It is encouraging that the SR estimates from just 15 bonds broadly coincide with the ML estimates from the Kalman filter. Figure 4 also reports the infeasible ML estimates (denoted by tick pluses) when the factors are observed, or equivalently, when there is an infinite number of bonds available in each time period. Figure 4 shows that the SR estimates from just 15 bonds have converged to these optimal estimates. It is also worth noting that all SR and ML estimates of $\boldsymbol{\theta}_{22}$ display the well-known biases for highly persistent VAR systems, thus illustrating the potential benefit of using small-sample bias corrections as in Bauer et al. (2012).

[23] This empirical result is somewhat related to the finding in Hamilton & Wu (2012), showing that eigenvalues under the $\mathbb{Q}$ measure must be decreasing in models with observed factors to insure global identification. That is, $\lambda_1^{\mathbb{Q}} \geq \lambda_2^{\mathbb{Q}} \geq \lambda_3^{\mathbb{Q}}$ in a three-factor model. Although Hamilton & Wu (2012) do not rule out equality between the eigenvalues, it may nevertheless be empirically difficult to obtain joint identification of $\lambda_2^{\mathbb{Q}}$ and $\lambda_3^{\mathbb{Q}}$ in this case.

[24] For the reported results, this condition is rarely binding in the SR and ML approach when we use 20 or more bonds in the cross-section dimension.

[25] The Monte Carlo results for Case II are available on request.

[26] When reporting our results, we scale elements in $\boldsymbol{\theta}_{12}$ by $100 \times \sqrt{12}$ and elements in $\boldsymbol{\theta}_{22}$ by 12.

< Figure 4 about here >

The average time requirement for computing the SR and ML estimates are reported in the first row of Figure 5. It takes roughly 4 to 6 minutes to maximize the likelihood function, whereas it takes only 21 to 26 seconds to compute the SR estimates.[27] Accordingly, the SR approach is between 10 and 15 times faster to implement than ML. We also note that there exists an optimal number of bonds for both estimators if we solely want to minimize estimation time. The minimum is achieved for 10 bonds in the case of ML and 20 bonds when using the SR approach. Note also that including 20 bonds in the likelihood function is actually faster than using just 5 bonds, as is typical in the literature. To understand the effects driving these results, recall that the computational requirement for the Kalman and regression filter is monotonically increasing in the number of bonds. Hence, using more bonds can only reduce the estimation time because these bonds help regularizing the objective function, thus making it easier to optimize. We find that the second effect dominates with relatively few bonds, whereas the first effect dominates with many bonds.

< Figure 5 about here >

Efficiency: To evaluate the efficiency of the parameter estimates, Figure 6 plots the standard deviations of the Monte Carlo estimates. As expected, the SR estimates are generally less efficient compared to ML, but this efficiency loss gradually disappears as we increase the number of bonds.

< Figure 6 about here >

To understand the very encouraging performance of the SR approach, Figure 7 compares the accuracy of the first stage estimates $\hat{\boldsymbol{\theta}}_1^{step1}$ to those obtained in the third step $\hat{\boldsymbol{\theta}}_1^{step3}$ (indicated by blue lines with squares). Figure 7 confirms our expectation, as the third step clearly improves the efficiency of the SR approach. The improvement is particularly evident for $\lambda_2^{\mathbb{Q}}$, $\lambda_3^{\mathbb{Q}}$, $r_\infty^{\mathbb{Q}}$, and all elements of $\boldsymbol{\Sigma}_{\mathcal{P}}$. Hence, the third step in the SR approach works as intended and explains why the SR approach is close to being fully efficient in this particular application with just 25 bonds in each time period.

< Figure 7 about here >

Standard errors and Type I errors: Figure 8 displays the biases for the estimated standard errors. For comparability with ML, all SR estimates are computed without correlation in the cross-

[27] In optimizing the likelihood function, we use the infeasible ML estimates for the VAR system (i.e., assuming observed factors) as starting values for our non-linear optimizer. Even based on these good starting values, preliminary results showed that a gradient-based optimizer often was unable to find the global optimum. A similar finding is reported in Hamilton & Wu (2012). However, the Nelder-Mead optimizer displayed considerably better performance and was therefore chosen for the Monte Carlo experiment, although it is somewhat slower than gradient-based optimizers. For the SR estimates, a standard gradient-based optimizer displayed satisfactory performance and was therefore used.

sectional and time series dimension and with homoskedasticity where $\sigma_t = \sigma$ for all t. When calculating the standard errors for $\boldsymbol{\theta}_{22}$ and $\boldsymbol{\theta}_{12}$ in the second step of SR approach based on (47), we use one lag in the Newey-West estimator, as all remaining lags are zero in our case. The first part of Figure 8 reveals that the biases in the standard errors for $\boldsymbol{\theta}_1$ are small and essentially unbiased with 50 bonds. Some biases appear in the standard errors for $\boldsymbol{\theta}_{22}$, but the same is seen for the ML estimates.

< Figure 8 about here >

Figure 9 reports the rejection probabilities at a 5% significance level when testing separate null hypotheses that the estimated parameters equal their true values. For $\boldsymbol{\theta}_1$ in the first part of Figure 9, nearly all rejection probabilities using the SR approach are close to the desired 5% level, even for relatively few bonds in the cross-section dimension. The exceptions are $\lambda_2^{\mathbb{Q}}$, $\lambda_3^{\mathbb{Q}}$, and R_v, where the tests are slightly oversized with 10 to 20 bonds, but these rejection probabilities fall as we increase the number of observables, and are close to the desired 5% level with 50 bonds. The second part of Figure 9 shows that several rejection probabilities for $\boldsymbol{\theta}_{22}$ are slightly elevated in the SR approach, but the same is seen for the ML estimates based on the Kalman filter and the infeasible ML estimates.

< Figure 9 about here >

7.2 QTSM: The study design

In our discrete-time two-factor QTSM, the short rate r_t is quadratic in the latent factors $\mathbf{x}_t$, i.e.

$$r_t = \mathbf{x}_t' \boldsymbol{\Psi} \mathbf{x}_t. \tag{65}$$

We let $\boldsymbol{\Psi} = \mathbf{1}_{2\times 2}$ (i.e. a matrix of unit entries) such that the short rate is only driven by one factor whereas long-term interest rates are controlled by both factors. As explained by Kim & Singleton (2012), this specification allows the QTSM to enforce the zero lower bound and produce near constant short-term interest rates close to the bound while simultaneously generating volatile long-term interest rates. The dynamics for the factors under the $\mathbb{Q}$ measure is given by the VAR(1) system

$$\mathbf{x}_{t+1} = \boldsymbol{\Phi}\boldsymbol{\mu} + (\mathbf{I} - \boldsymbol{\Phi})\,\mathbf{x}_t + \boldsymbol{\Sigma}\boldsymbol{\varepsilon}_{t+1}^{\mathbb{Q}}, \tag{66}$$

with $\boldsymbol{\varepsilon}_{t+1}^{\mathbb{Q}} \sim \mathcal{NID}(\mathbf{0}, \mathbf{I}_2)$. The continuously compounded yield on a zero-coupon bond maturing in k periods is then

$$i_{t,k} = \tilde{A}_k + \tilde{\mathbf{B}}_k' \mathbf{x}_t + \mathbf{x}_t' \tilde{\mathbf{C}}_k \mathbf{x}_t + v_{t,k}, \tag{67}$$

where $\tilde{A}_k$, $\tilde{\mathbf{B}}_k$, and $\tilde{\mathbf{C}}_k$ follow a set of Ricatti equations derived in Realdon (2006) and $v_{t,k}$ are measurement errors. As in the Gaussian ATSM, we adopt an affine specification for the market price of risk, implying that the factor dynamics under the $\mathbb{P}$ measure may be expressed as

$$\mathbf{x}_{t+1} = \mathbf{h}_0 + \mathbf{h}_{\mathbf{x}}\mathbf{x}_t + \boldsymbol{\Sigma}\boldsymbol{\varepsilon}_{t+1}^{\mathbb{P}}, \tag{68}$$

with $\mathbf{h}_0$ and $\mathbf{h}_{\mathbf{x}}$ as free matrices and $\boldsymbol{\varepsilon}_{t+1}^{\mathbb{P}} \sim \mathcal{NID}(\mathbf{0}, \mathbf{I}_2)$. The model is normalized with unconstrained factor dynamics under the $\mathbb{P}$ measure by imposing the restrictions: i) $\boldsymbol{\mu} \geq \mathbf{0}$, ii) $\boldsymbol{\Phi}$ is diagonal with strictly increasing elements, and iii) $\boldsymbol{\Sigma}$ is triangular. As for the Gaussian ATSM, we let the variance in the measurement errors be constant across yields and time, i.e. $Var(v_{t,k}) = R_v$ for all t and k. Using the data set by Joslin et al. (2011), the remaining parameters are calibrated such that the model matches the average level and the standard deviation of the 10-year yield curve.

As for the Gaussian ATSM, we use log-transformed zero-coupon bond prices $\log P_{t,k}$ for the Monte Carlo experiment and simulate these bonds as described in Section 7.1. Given that exact likelihood inference for the QTSM is infeasible, we cannot compare the SR approach to the optimal smoother for the latent factors and the ML estimates. As an alternative benchmark, we consider the central difference Kalman filter (CDKF) by Norgaard et al. (2000) and the related smoother, which generalizes the Kalman filter to non-linear systems by using multivariate Sterling interpolations to approximate moments in the filtering equations. The CDKF is generally more accurate than the extended Kalman filter (EKF), implying that the quasi maximum likelihood (QML) estimator based on the CDKF tends to perform better than a QML estimator using the EKF (see Andreasen (2013)). Although the CDKF is relatively fast to execute, the QML estimation is nevertheless computationally involved, and this explains our choice of two factors in the QTSM.

7.2.1 QTSM: Factor estimation

The RMSEs for estimating the latent factors by the regression filter and the CDKF smoother are shown in Figure 10. For Case 1, the ability of the regression filter to estimate the level factor is close to the CDKF smoother even with just 10 to 15 bonds. The regression filter is less precise for the slope factor with 10 to 20 bonds when compared to the CDKF smoother, but this difference disappears with 50 bonds. In Case 2 with all bond prices measured less precisely than in Case 1, we find larger efficiency losses from the regression filter compared to the CDKF smoother with 10 to 20 bonds, but the two estimators display roughtly similar performance with 50 to 100 bonds.

< Figure 10 about here >

The average number of seconds needed to run the regression filter and the CDKF smoother are displayed in the second row of Figure 2. As for the Gaussian ATSM, we see a steep increase in the time requirement for the CDKF smoother, whereas the number of bonds hardly affects the non-linear regressions in the SR approach. Accordingly, the regression filter is between 2 and 14 times faster than the CDKF smoother.[28]

7.2.2 QTSM: Parameter estimation

To apply the SR approach for parameter estimation for the QTSM, we let $\boldsymbol{\theta}_{11} = \begin{bmatrix} \phi_1 & \phi_2 & \mu_1 & \mu_2 \end{bmatrix}'$ and $\boldsymbol{\theta}_{12} = \begin{bmatrix} \sigma_1 & \sigma_{12} & \sigma_2 \end{bmatrix}'$. These seven parameters are estimated in the first step of the SR approach along with the variance of the measurement errors R_v. All remaining parameters appear in $\boldsymbol{\theta}_{22}$, i.e. $\boldsymbol{\theta}_{22} = \begin{bmatrix} \mathbf{h}_0' & vec\,(\mathbf{h}_{\mathbf{x}})' \end{bmatrix}'$, which are estimated by (20), and $\boldsymbol{\theta}_{12}$ is obtained by (21) in the second step. The remaining details for the Monte Carlo experiment are identical to those described in Section 7.1.2, meaning that we for the QTSM also focus on Case 1 with bond prices generated from interest rates with measurement errors having a standard deviation of 10 basis points.

Estimates: Figure 11 shows that both the SR and the QML estimates of $\boldsymbol{\theta}_{11}$, $\boldsymbol{\theta}_{12}$, $\mathbf{h}_0$, and R_v are basically unbiased for any number of bonds. Only the estimates of $\mathbf{h}_{\mathbf{x}}$ display the usual biases for highly persistent VAR systems. In the third step of the SR approach, we find $\hat{\boldsymbol{\Lambda}} \approx \mathbf{0}$, implying that $\boldsymbol{\theta}_{12}$ also for this model mainly is determined from the time series dimension of the data.

< Figure 11 about here >

The average number of seconds for computing the SR and QML estimates are given in the second row of Figure 5. It takes about 3.5 to 6 minutes to maximize the quasi log-likelihood function, whereas the computing time for the SR estimates drops from 80 to 60 seconds as the number of bonds increases from 10 to 50. Hence, in this range, the effect of the increased information and therefore better behaved objective function more than outweighs the extra computational burden per iteration when adding more bonds. This implies that the SR approach is 3 to 6 times faster to execute for the QTSM when compared to the QML approach based on the CDKF.

[28] We conjecture that the computational gain for the regression filter would be even larger if compared to an approximation of the optimal smoother by importance sampling, because the computational requirements for importance sampling increase rapidly as the number of observables increases in each time period.

Efficiency: The SR estimates of $\boldsymbol{\theta}_{11}$ and R_v are nearly as accurate as those from the QML estimator according to Figure 12. With 10 to 20 bonds, we find some loss of efficiency for the SR estimates of $\boldsymbol{\theta}_{22}$ and $\boldsymbol{\theta}_{12}$, but it gradually disppears when increasing the number of bonds in the cross-section dimension. Unreported results reveal that this satisfying performance of the SR approach is once again due to the third estimation step.

< Figure 12 about here >

Standard errors and Type I errors: Figure 13 shows that the biases in the standard errors for $\boldsymbol{\theta}_{11}$, $\boldsymbol{\theta}_{12}$, $\mathbf{h}_0$, and R_v are small and essentially unbiased with 20 to 25 bonds in the SR approach. We find some biases in the standard errors for $\mathbf{h}_{\mathbf{x}}$, but the same is seen for both the QML and the infeasible ML estimates.[29]

< Figure 13 about here >

The rejection probabilities at a 5% significance level for testing the separate null hypotheses that the estimated parameters are equal to their true values is given in Figure 14. Nearly all probabilities using the SR approach are close to the desired 5% level, at least when including 20 or more bonds each time period. The only exception is for μ_1, but the same is seen for the QML estimates.

< Figure 14 about here >

7.3 CIR model: The study design

In the CIR model, the instantaneous short rate under the $\mathbb{Q}$ measure evolves as

$$dr_t = \kappa^{\mathbb{Q}} \left(\theta^{\mathbb{Q}} - r_t \right) dt + \beta \sqrt{r_t} dz_t^{\mathbb{Q}},$$

where $z_t^{\mathbb{Q}}$ is a standard Wiener process. The continuously compounded yield on a zero-coupon bond at maturity τ is then

$$i_t(\tau) = a(\tau) + b(\tau) r_t + v_t(\tau),$$

[29] For reliable performance of the standard errors in the SR approach, the tolerance for changes in the objective functions for the non-linear cross-section regressions must be sufficiently low compared to the stepsize for computing numerical derivatives with respect to $\boldsymbol{\theta}_1$. For the considered QTSM, we use a tolerance of 10^{-12} for the cross-section regressions and a stepsize of 10^{-7} for the derivatives with respect to $\boldsymbol{\theta}_1$.

where $a(\tau)$ and $b(\tau)$ are given in closed form and $v_t(\tau)$ are measurement errors. The market price of risk takes the form $\lambda\sqrt{r_t}/\beta$, implying that the dynamics for r_t under the $\mathbb{P}$ measure is

$$dr_t = \kappa(\theta - r_t)\,dt + \beta\sqrt{r_t}dz_t^{\mathbb{P}}, \tag{69}$$

with $\kappa \equiv \kappa^{\mathbb{Q}} - \lambda$ and $\theta \equiv \kappa^{\mathbb{Q}}\theta^{\mathbb{Q}}/\kappa$. To relate this process to the general results derived above, we consider the exact discrete representation of (69) which reads

$$r_{t+\tau} = \theta\left(1 - e^{-\kappa\tau}\right) + e^{-\kappa\tau}r_t + \beta\int_t^{t+\tau} e^{\kappa(u-t-\tau)}\sqrt{r_u}dz_u^{\mathbb{P}}.$$

Hence, in relation to (16), $h_0 \equiv \theta(1 - e^{-\kappa\tau})$, $h_x \equiv e^{-\kappa\tau}$, and $w_{t+\tau}(r_t;\beta,\kappa) \equiv \beta\int_t^{t+\tau} e^{\kappa(u-t-\tau)}\sqrt{r_u}dz_u^{\mathbb{P}}$ where $\tau = 1/12$ to ensure that one time period corresponds to one month. As for the QTSM, we calibrate parameters in the model such that it matches the average level and the standard deviation of the 10-year yield curve in the data set by Joslin et al. (2011).

As for the previous two models, we use log-transformed zero-coupon bond prices $\log P_t(\tau)$ for the Monte Carlo experiment. To avoid any discretization bias when simulating the factor dynamics, we use an exact simulation scheme for $(r_t)_{t\geq 0}$ (see Andersen, Jäckel & Kahl (2010)) and generate bond prices as described in Section 7.1. Exact likelihood inference for the CIR model is unfortunately not available given our assumption that all bond prices are measured with errors. Instead, the performance of the SR approach is compared to the quasi log-likelihood procedure by Duan & Simonato (1999), where the Kalman filter is modified to account for the stochastic volatility in r_t.

7.3.1 CIR model: Factor estimation

The RMSEs for estimating the short rate are shown in Figure 15. For Case 1, the ability of the regression filter to extract the short rate is nearly identical to the modified Kalman smoother even with just 5 to 10 bonds. In Case 2 with all bond prices measured less precisely than in Case 1, we find a small efficiency loss from the regression filter compared to the modified Kalman smoother with 10 to 20 bonds, but the two estimators are identical with 25 bonds. Hence, the performance of the SR approach does not deteriorate in the presence of stochastic volatility, as latent factors are obtained from static cross-section regressions.

< Figure 15 about here >

The average number of seconds used to run the regression filter and the modified Kalman smoother

are displayed in last row of Figure 2. We see a steep increase in the time requirement for the Kalman smoother, whereas the SR approach is broadly unaffected by the number of bonds. Thus, the regression filter is 10 to 40 times faster than the modified Kalman smoother.

7.3.2 CIR model: Parameter estimation

To estimate model parameters in the CIR model by the SR approach, we let $\boldsymbol{\theta}_{11} = \begin{bmatrix} \kappa^{\mathbb{Q}} & \theta^{\mathbb{Q}} \end{bmatrix}'$ and $\boldsymbol{\theta}_{12} = \begin{bmatrix} \beta \end{bmatrix}$, which are both estimated in the first step along with the variance of the measurement errors R_v. In the second step, we estimate κ and θ by (20) and use (21) to obtain β as $Var\left(w_{t+\tau}\right) = \beta^2\theta\left(1-e^{-2\kappa\tau}\right)/\left(2\kappa\right)$. Accordingly, all three parameters in the $\mathbb{P}$ dynamics of the short rate can be estimated in closed form by a regression that accounts for generated regressors. Only the market price of risk λ appears in $\boldsymbol{\theta}_{22}$, and we identify it from the difference between the estimated unconditional mean of the short rate under the $\mathbb{Q}$ and $\mathbb{P}$ measures, i.e. by $\hat{\lambda} = \hat{\kappa}^{\mathbb{Q}}\left(1-\hat{\theta}^{\mathbb{Q}}/\hat{\theta}\right)$.[30] The variance of this estimator is given by

$$Var\left(\hat{\lambda}\right) = \left(\hat{\kappa}^{\mathbb{Q}}\hat{\theta}^{\mathbb{Q}}\right)^2 Var\left(\frac{1}{\hat{\theta}}\right) \approx \left(\frac{\hat{\kappa}^{\mathbb{Q}}\hat{\theta}^{\mathbb{Q}}}{\hat{\theta}^2}\right)^2 Var\left(\hat{\theta}\right) \tag{70}$$

using the super-consistency of $\hat{\kappa}^{\mathbb{Q}}\hat{\theta}^{\mathbb{Q}}$ with respect to $\hat{\theta}$ and the delta-method for $Var\left(\frac{1}{\hat{\theta}}\right)$.[31] The remaining details for the Monte Carlo experiment are identical to those provided in Section 7.1.2, meaning that we for the CIR model also focus on Case 1 with bond prices generated from interest rates with measurement errors having a standard deviation of 10 basis points.

Estimates: The first row of Figure 16 shows that both estimators of $\hat{\kappa}^{\mathbb{Q}}$, $\hat{\theta}^{\mathbb{Q}}$, and R_v are unbiased for any number of bonds in the cross-section dimension. The SR estimates of β have a tiny positive bias in contrast to the QML estimates, but the SR approach has a smaller negative bias for λ than QML. Unreported results reveal that the mean value of Λ for combining the first and second step estimates of β in (49) is 0.25 with 5 bonds and increases concavely to 0.75 with 50 bonds. Hence, the presence of stochastic volatility means that $\hat{\beta}$ is not fully determined from the time-series dimension of the data, in contrast to $\boldsymbol{\Sigma}_{\mathcal{P}}$ in the Gaussian ATSM and $\boldsymbol{\Sigma}$ in the QTSM.

[30] It is also possible to identify the market price of risk from the difference between the degree of estimated mean reversion under the $\mathbb{Q}$ and $\mathbb{P}$ measures, i.e. as $\tilde{\lambda} = \kappa^{\mathbb{Q}} - \kappa$. However, $Var\left(\tilde{\lambda}\right)$ is much larger than $Var\left(\hat{\lambda}\right)$, and the well-known positive bias in κ for persistent processes furthermore induces a negative bias in $\tilde{\lambda}$.

[31] Extending (70) with the uncertainty attached to $\hat{\kappa}^{\mathbb{Q}}$ and $\hat{\theta}^{\mathbb{Q}}$ from the first step gives nearly identical results because $Var\left(\hat{\theta}\right)$ is much larger than $Var\left(\hat{\kappa}^{\mathbb{Q}}\right)$ and $Var\left(\hat{\theta}^{\mathbb{Q}}\right)$.

< Figure 16 about here >

The average computing time for the two estimators is given in the third row of Figure 5. It takes about 12 to 40 seconds to maximize the quasi log-likelihood function, whereas it takes 3 to 7 seconds to compute the SR estimates. Hence, the SR approach is 4.5 to 6 times faster than the modified Kalman filter in Duan & Simonato (1999).

Efficiency: The second row of Figure 16 suggests that the SR estimates of $\boldsymbol{\theta}_{11}$ nearly achieve the same precision as QML, but we see some loss of efficiency in the SR estimates of β and λ.

Standard errors and Type I errors: The SR estimates of $\boldsymbol{\theta}_1$ from the third step display positive biases in the standard errors, which induces too low rejection probabilities for the separate null hypotheses that the estimated parameters equal their true values (see the third and fourth row of Figure 16). On the other hand, standard errors from the first step of the SR approach are completely unbiased and their rejection probabilities are close to the desired 5% level. This suggests that well-known shortcomings of asymptotic inference for persistent processes given short samples ($T = 480$) in the second step of the SR approach carry over to the third step.

For the market price of risk λ in $\boldsymbol{\theta}_{22}$ we find negative biases in the standard errors and therefore too high rejection probabilities. These biases are also related to the second estimation step, as unreported results reveal a positive bias in $\hat{\theta}^{step2}$ and a negative bias in $Var\left(\hat{\theta}^{step2}\right)$. From (70), both biases help explain the low estimates of $Var\left(\hat{\lambda}\right)$.

8 Conclusion

This paper presents a new estimation approach for a wide class of non-linear DTSMs with latent factors. We impose no distributional assumptions on the factors which therefore may be non-Gaussian. The novelty of our approach is to use many observables (yields or bonds prices) in the cross-section dimension. This implies that the latent factors can be estimated quite accurately by a sequence of cross-section regressions. The proposed regression filter is shown to converge to the optimal estimator when the number of observables in the cross-section dimension tends to infinity. We also show how output from the regression filter may be used to estimate parameters in DTSMs by a two- or three-step moment-based procedure.

Three Monte Carlo experiments based on a three-factor Gaussian ATSM, a two-factor QTSM, and

the CIR model suggest that the finite sample properties of the SR estimates are well approximated by their asymptotic distribution. This is the case even with a relatively small number of observables in each time period. Moreover, the efficiency loss for the SR approach compared to ML or the QML estimators considered is found to be small, even with just 25 bonds in the cross-section dimension. This result is mainly due to the third step in the SR approach which increases the efficiency of the estimated risk-neutral parameters.

A Estimation of $\boldsymbol{\theta}_2$

A.1 Proof of (20)

We start by rewriting the system in (17) as

$$\underbrace{\begin{bmatrix} \hat{\mathbf{x}}_{t+1} \\ 1 \end{bmatrix}}_{\mathbf{a}_{t+1}} = \underbrace{\begin{bmatrix} \mathbf{h_x} & \mathbf{h}_0 \\ \mathbf{0} & 1 \end{bmatrix}}_{\mathbf{h_a}} \underbrace{\begin{bmatrix} \hat{\mathbf{x}}_t \\ 1 \end{bmatrix}}_{\mathbf{a}_t} + \underbrace{\begin{bmatrix} \hat{\mathbf{w}}_{t+1} \\ 0 \end{bmatrix}}_{\hat{\mathbf{w}}_{\mathbf{a},t+1}}.$$

The measurement errors to this transformed system are denoted $\mathbf{u}_{\mathbf{a},t} \equiv \begin{bmatrix} \mathbf{u}_t' & 0 \end{bmatrix}'$. Hence,

$$Var\left(\mathbf{u}_{\mathbf{a},t}\right) = \begin{bmatrix} Var\left(\mathbf{u}_t\right) & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix} \quad \text{and} \quad Cov\left(\mathbf{u}_{\mathbf{a},t+1}, \mathbf{u}_{\mathbf{a},t}\right) = \begin{bmatrix} Cov\left(\mathbf{u}_{t+1}, \mathbf{u}_t\right) & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix}.$$

The first two moment conditions (18) may now be expressed as

$$\frac{1}{T-1}\sum_{t=1}^{T} \hat{\mathbf{w}}_{\mathbf{a},t+1}\mathbf{a}_t' = \frac{1}{T-1}\sum_{t=1}^{T}\left(Cov\left(\mathbf{u}_{\mathbf{a},t+1}, \mathbf{u}_{\mathbf{a},t}\right) - \mathbf{h_a} Var\left(\mathbf{u}_{\mathbf{a},t}\right)\right). \tag{71}$$

We then note that

$$\frac{1}{T-1}\sum_{t=1}^{T} \hat{\mathbf{w}}_{\mathbf{a},t+1}\mathbf{a}_t' = \frac{1}{T-1}\sum_{t=1}^{T} \mathbf{a}_{t+1}\mathbf{a}_t' - \mathbf{h_a}\frac{1}{T-1}\sum_{t=1}^{T}\mathbf{a}_t\mathbf{a}_t'. \tag{72}$$

Combining (71) and (72), we get

$$\frac{1}{T-1}\sum_{t=1}^{T} \mathbf{a}_{t+1}\mathbf{a}_t' - \frac{1}{T-1}\sum_{t=1}^{T} Cov\left(\mathbf{u}_{\mathbf{a},t+1}, \mathbf{u}_{\mathbf{a},t}\right) = \mathbf{h_a}\frac{1}{T-1}\sum_{t=1}^{T}\mathbf{a}_t\mathbf{a}_t' - \mathbf{h_a}\frac{1}{T-1}\sum_{t=1}^{T} Var\left(\mathbf{u}_{\mathbf{a},t}\right),$$

from which we obtain the expression in (20).

A.2 Proof of (27)

We use the notation from Section A.1. Hence, the moment conditions in (25) may be expressed as

$$\frac{1}{T-2}\sum_{t=2}^{T-1} \hat{\mathbf{w}}_{\mathbf{a},t+1}\mathbf{a}_{t-1}' = \frac{1}{T-2}\sum_{t=2}^{T-1}\left(Cov\left(\mathbf{u}_{\mathbf{a},t+1}, \mathbf{u}_{\mathbf{a},t-1}\right) - \mathbf{h_a} Cov\left(\mathbf{u}_{\mathbf{a},t}, \mathbf{u}_{\mathbf{a},t-1}\right)\right). \tag{73}$$

We then note that

$$\frac{1}{T-2}\sum_{t=2}^{T-1} \hat{\mathbf{w}}_{\mathbf{a},t+1}\mathbf{a}_{t-1}' = \frac{1}{T-2}\sum_{t=2}^{T-1}\left(\mathbf{a}_{t+1} - \mathbf{h_a}\mathbf{a}_t\right)\mathbf{a}_{t-1}' = \frac{1}{T-2}\sum_{t=2}^{T-1}\mathbf{a}_{t+1}\mathbf{a}_{t-1}' - \mathbf{h_a}\frac{1}{T-2}\sum_{t=2}^{T-1}\mathbf{a}_t\mathbf{a}_{t-1}'. \tag{74}$$

Combining (73) and (74), we immediately obtain (27).

B Proof of Proposition 1

The first-order conditions for $\{\hat{\mathbf{x}}_{2,t}\}_{t=1}^{T}$ and $\hat{\boldsymbol{\theta}}_1$ implied by the SR approach are:

$$\frac{\partial Q_t}{\partial \mathbf{x}_{2,t}} = -\sum_{j=1}^{n_{y,t}}\left(\tilde{y}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}\right) - \tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}, \mathbf{z}_t; \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\gamma}}\right)\right)\frac{\partial \tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}, \mathbf{z}_t; \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\gamma}}\right)}{\partial \mathbf{x}_{2,t}} = \mathbf{0} \qquad \text{for } t = 1, 2, ..., T, \tag{75}$$

$$\begin{aligned}\frac{\partial Q\left(\boldsymbol{\theta}_1\right)}{\partial \boldsymbol{\theta}_1} &= -\sum_{t=1}^{T}\sum_{j=1}^{n_{y,t}}\left(\tilde{y}_{t,j}\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)-\tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1\right),\mathbf{z}_t;\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\gamma}}\right)\right)\times \\ &\quad \left(\frac{\partial \hat{\mathbf{x}}_{2,t}'\left(\hat{\boldsymbol{\theta}}_1\right)}{\partial \boldsymbol{\theta}_1}\frac{\partial \tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1\right),\mathbf{z}_t;\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\gamma}}\right)}{\partial \hat{\mathbf{x}}_{2,t}\left(\boldsymbol{\theta}_1\right)}+\frac{\partial \tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1\right),\mathbf{z}_t;\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\gamma}}\right)}{\partial \boldsymbol{\theta}_1}\right) \\ &= \mathbf{0}.\end{aligned} \tag{76}$$

Here, $\frac{\partial \hat{\mathbf{x}}_{2,t}'\left(\hat{\boldsymbol{\theta}}_1\right)}{\partial \boldsymbol{\theta}_1}$ has dimension $L_1 \times n_{x_2}$ and $\frac{\partial \tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1\right),\mathbf{z}_t;\hat{\boldsymbol{\theta}}_1,\boldsymbol{\gamma}\right)}{\partial \mathbf{x}_{2,t}\left(\boldsymbol{\theta}_1\right)}$ has dimension $n_{x_2}\times 1$. Furthermore,

$$\begin{aligned}&\frac{\partial Q\left(\boldsymbol{\theta}_1\right)}{\partial \boldsymbol{\theta}_1} = -\sum_{t=1}^{T}\frac{\partial \hat{\mathbf{x}}_{2,t}'\left(\hat{\boldsymbol{\theta}}_1\right)}{\partial \boldsymbol{\theta}_1}\sum_{j=1}^{n_{y,t}}\left(\tilde{y}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}\right)-\tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1\right),\mathbf{z}_t;\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\gamma}}\right)\right)\frac{\partial \tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1\right),\mathbf{z}_t;\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\gamma}}\right)}{\partial \hat{\mathbf{x}}_{2,t}\left(\boldsymbol{\theta}_1\right)} \\ &\quad -\sum_{t=1}^{T}\sum_{j=1}^{n_{y,t}}\left(\tilde{y}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}\right)-\tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1\right),\mathbf{z}_t;\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\gamma}}\right)\right)\frac{\partial \tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1\right),\mathbf{z}_t;\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\gamma}}\right)}{\partial \boldsymbol{\theta}_1} = \mathbf{0}\end{aligned}$$

$$\Updownarrow$$

$$\frac{\partial Q\left(\boldsymbol{\theta}_1\right)}{\partial \boldsymbol{\theta}_1} = -\sum_{t=1}^{T}\sum_{j=1}^{n_{y,t}}\left(\tilde{y}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}\right)-\tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1\right),\mathbf{z}_t;\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\gamma}}\right)\right)\frac{\partial \tilde{g}_j\left(\hat{\mathbf{x}}_{2,t}\left(\hat{\boldsymbol{\theta}}_1\right),\mathbf{z}_t;\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\gamma}}\right)}{\partial \boldsymbol{\theta}_1} = \mathbf{0}$$

when using (75). Hence, (75) and (76) are also the first-order conditions for the joint estimation problem.

C The asymptotic distribution of $\hat{\boldsymbol{\theta}}_1$

The first-order condition with respect to $\boldsymbol{\theta}_1$ is

$$\frac{\partial Q\left(\boldsymbol{\theta}_1\right)}{\partial \boldsymbol{\theta}_1} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{s}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i;\hat{\boldsymbol{\theta}}_1,\boldsymbol{\gamma}^*\right) = \mathbf{0} \tag{77}$$

where $\mathbf{r}_i \equiv \left(\mathbf{x}_{2,i},\mathbf{z}_i\right)$ using the index for the pooled sample, and $\mathbf{s}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i;\hat{\boldsymbol{\theta}}_1,\boldsymbol{\gamma}^*\right)$ denotes the score function with respect to $\boldsymbol{\theta}_1$. A mean-value expansion of this score function around $\boldsymbol{\theta}_1^o$ and $\boldsymbol{\gamma}^*$ implies

$$\sum_{i=1}^{N}\mathbf{s}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i;\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\gamma}}\right) = \sum_{i=1}^{N}\mathbf{s}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i;\boldsymbol{\theta}_1^o,\boldsymbol{\gamma}^*\right)+\sum_{i=1}^{N}\frac{\partial \mathbf{s}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i;\dot{\boldsymbol{\theta}}_1,\boldsymbol{\gamma}^*\right)}{\partial \boldsymbol{\theta}_1'}\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o\right). \tag{78}$$

Let $\boldsymbol{H}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i;\dot{\boldsymbol{\theta}}_1,\boldsymbol{\gamma}^*\right) \equiv \frac{\partial \mathbf{s}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i;\dot{\boldsymbol{\theta}}_1,\boldsymbol{\gamma}^*\right)}{\partial \boldsymbol{\theta}_1'}$ be the Hessian matrix evaluated at $\dot{\boldsymbol{\theta}}_1$. This notation indicates that each row of the Hessian matrix is evaluated at a potentially different convex combination of $\hat{\boldsymbol{\theta}}_1$ and $\boldsymbol{\theta}_1^o$. Using (77), condition (78) reduces to

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o\right) = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{H}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i;\dot{\boldsymbol{\theta}}_1,\boldsymbol{\gamma}^*\right)\right)^{-1}\left(\frac{-1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{s}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i;\boldsymbol{\theta}_1^o,\boldsymbol{\gamma}^*\right)\right).$$

If either $\phi_t\left(m\right) = O\left(m^{-\nu}\right)$ for $\nu > r_2/\left(r_2-1\right)$ and $t = 1,2,...,T$ or $\alpha_t\left(m\right) = O\left(m^{-\nu}\right)$ for $\nu > \max\left\{r_1/\left(r_1-1\right), r_2/\left(r_2-1\right)\right\}$ and $t = 1,2,...,T$, where $r_1 > 1$, it follows from White & Domowitz (1984) that $\sqrt{N}\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o\right)$ based on $\boldsymbol{\gamma}^*$ is asymptotically normal when $n_{y,t}\to\infty$ for all t and the covariance matrix is $\left(\mathbf{A}_o^{\boldsymbol{\theta}_1}\right)^{-1}\mathbf{B}_o^{\boldsymbol{\theta}_1}\left(\mathbf{A}_o^{\boldsymbol{\theta}_1}\right)^{-1}$. See White & Domowitz (1984) for the appropriate regularity conditions or the technical appendix accompanying the present paper. A straightforward application

of the asymptotic equivalence lemma reveals that the asymptotic distribution of $\sqrt{N}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^o\right)$ is unaffected if $\boldsymbol{\gamma}^*$ is replaced by $\hat{\boldsymbol{\gamma}}$.

D The asymptotic properties of $\hat{\sigma}_t^2$

This section shows consistency and asymptotic normality of $\hat{\sigma}_t^2$. In addition to the conditions ensuring consistency and normality of $\hat{\boldsymbol{\theta}}_1$ and $\{\hat{\mathbf{x}}_{2,t}\}_{t=1}^T$, we impose the following conditions:

1. $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}^o$ for some $\boldsymbol{\gamma}^o \in \Gamma$
2. $E\left[\mathbf{g}_{\mathbf{x}_2}\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o\right)' \mathbf{V}\left(\mathbf{z}_t; \boldsymbol{\gamma}^o\right)^{-1} \mathbf{g}_{\mathbf{x}_2}\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o\right)\right] < \infty$
3. $E\left[\mathbf{g}_{\mathbf{x}_2}\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o\right)' \mathbf{V}\left(\mathbf{z}_t; \boldsymbol{\gamma}^o\right)^{-1} \mathbf{g}_{\boldsymbol{\theta}_1}\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o\right)\right] < \infty$
4. $E\left[\mathbf{g}_{\boldsymbol{\theta}_1}\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o\right)' \mathbf{V}\left(\mathbf{z}_t; \boldsymbol{\gamma}^o\right)^{-1} \mathbf{g}_{\boldsymbol{\theta}_1}\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o\right)\right] < \infty$
5. $E\left[\mathbf{g}_{\mathbf{x}_2}\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o\right)' \mathbf{V}\left(\mathbf{z}_t; \boldsymbol{\gamma}^o\right)^{-1} \mathbf{g}_{\boldsymbol{\gamma}}\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o\right)\right] < \infty$
6. $E\left[\mathbf{g}_{\boldsymbol{\theta}_1}\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o\right)' \mathbf{V}\left(\mathbf{z}_t; \boldsymbol{\gamma}^o\right)^{-1} \mathbf{g}_{\boldsymbol{\gamma}}\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o\right)\right] < \infty$
7. $E\left[\mathbf{g}_{\boldsymbol{\gamma}}\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o\right)' \mathbf{V}\left(\mathbf{z}_t; \boldsymbol{\gamma}^o\right)^{-1} \mathbf{g}_{\boldsymbol{\gamma}}\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o\right)\right] < \infty$
8. $E\left[\left(\tilde{v}_{t,j}\left(\mathbf{z}_{t,j}; \boldsymbol{\gamma}^o\right)\right)^4\right] < \infty$

Here, $\mathbf{g}_{\mathbf{x}_2}\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o\right)$ has dimension $n_{y,t} \times n_{x_2}$, $\mathbf{g}_{\boldsymbol{\theta}_1}\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o\right)$ has dimension $n_{y,t} \times L_1$, and $\mathbf{g}_{\boldsymbol{\gamma}}\left(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o\right)$ has dimension $n_{y,t} \times n_{\gamma}$. Given that the covariance matrix for the measurement errors is correctly specified, we use the decomposition $\boldsymbol{\Omega}_{\mathbf{v}}\left(\mathbf{z}_t; \boldsymbol{\gamma}\right) = \sigma_t^2 \mathbf{V}\left(\mathbf{z}_t; \boldsymbol{\gamma}\right)$ with $\mathbf{V}\left(\mathbf{z}_t; \boldsymbol{\gamma}\right) \equiv \mathbf{S}_{\mathbf{v}}\left(\mathbf{z}_t; \boldsymbol{\gamma}\right) \mathbf{S}_{\mathbf{v}}\left(\mathbf{z}_t; \boldsymbol{\gamma}\right)'$. Hence, the rotation matrix $\mathbf{S}_{\mathbf{v}}\left(\mathbf{z}_t; \boldsymbol{\gamma}\right)$ does not contain σ_t^2.

D.1 Proof of consistency of $\hat{\sigma}_t^2$

A mean-value expansion of $\tilde{\mathbf{g}}\left(\mathbf{z}_t, \hat{\mathbf{x}}_{2,t}; \hat{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^o\right)$ in $\mathbf{x}_{2,t}$ and $\boldsymbol{\theta}_1$ implies

$$\begin{aligned}\tilde{\mathbf{g}}\left(\mathbf{z}_t, \hat{\mathbf{x}}_{2,t}; \hat{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^o\right) &= \tilde{\mathbf{g}}\left(\mathbf{z}_t, \mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o, \boldsymbol{\gamma}^o\right) + \tilde{\mathbf{g}}_{\mathbf{x}_2}\left(\mathbf{z}_t, \dot{\mathbf{x}}_{2,t}; \dot{\boldsymbol{\theta}}_1, \dot{\boldsymbol{\gamma}}\right)\left(\hat{\mathbf{x}}_{2,t} - \mathbf{x}_{2,t}^o\right) \\ &\quad + \tilde{\mathbf{g}}_{\boldsymbol{\theta}_1}\left(\mathbf{z}_t, \dot{\mathbf{x}}_{2,t}; \dot{\boldsymbol{\theta}}_1, \dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^o\right) + \tilde{\mathbf{g}}_{\boldsymbol{\gamma}}\left(\mathbf{z}_t, \dot{\mathbf{x}}_{2,t}; \dot{\boldsymbol{\theta}}_1, \dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^o\right),\end{aligned}$$

where the notation $\left(\dot{\mathbf{x}}_{2,t}, \dot{\boldsymbol{\theta}}_1, \dot{\boldsymbol{\gamma}}\right)$ indicates that each row of these matrices is evaluated at a potentially different convex combination of $\left(\hat{\mathbf{x}}_{2,t}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\gamma}}\right)$ and $\left(\mathbf{x}_{2,t}^o; \boldsymbol{\theta}_1^o, \boldsymbol{\gamma}^o\right)$. Simple albeit tedious algebra implies

$$
\begin{aligned}
\frac{1}{n_{y,t}}\widehat{\tilde{\mathbf{v}}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)'\widehat{\tilde{\mathbf{v}}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right) &= \frac{1}{n_{y,t}}\tilde{\mathbf{v}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)'\tilde{\mathbf{v}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right) \qquad (79)\\
&\quad -\frac{1}{n_{y,t}}2\widehat{\tilde{\mathbf{v}}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\mathbf{x}}_{2,t}-\mathbf{x}_{2,t}^o\right)\\
&\quad -\frac{1}{n_{y,t}}2\widehat{\tilde{\mathbf{v}}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\boldsymbol{\theta}_1}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o\right)\\
&\quad -\frac{1}{n_{y,t}}2\widehat{\tilde{\mathbf{v}}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\boldsymbol{\gamma}}\left(\mathbf{z}_t,\dot{\mathbf{x}}_{2,t};\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^o\right)\\
&\quad -\frac{1}{n_{y,t}}\left(\hat{\mathbf{x}}_{2,t}-\mathbf{x}_{2,t}^o\right)\tilde{\mathbf{g}}_{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\mathbf{x}}_{2,t}-\mathbf{x}_{2,t}^o\right)\\
&\quad -\frac{1}{n_{y,t}}2\left(\hat{\mathbf{x}}_{2,t}-\mathbf{x}_{2,t}^o\right)\tilde{\mathbf{g}}_{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\boldsymbol{\theta}_1}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o\right)\\
&\quad -\frac{1}{n_{y,t}}2\left(\hat{\mathbf{x}}_{2,t}-\mathbf{x}_{2,t}^o\right)\tilde{\mathbf{g}}_{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\boldsymbol{\gamma}}\left(\mathbf{z}_t,\dot{\mathbf{x}}_{2,t};\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^o\right)\\
&\quad -\frac{1}{n_{y,t}}\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o\right)'\tilde{\mathbf{g}}_{\boldsymbol{\theta}_1}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\boldsymbol{\theta}_1}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o\right)\\
&\quad -\frac{1}{n_{y,t}}2\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o\right)'\tilde{\mathbf{g}}_{\boldsymbol{\theta}_1}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\boldsymbol{\gamma}}\left(\mathbf{z}_t,\dot{\mathbf{x}}_{2,t};\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^o\right)\\
&\quad -\frac{1}{n_{y,t}}\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^o\right)'\tilde{\mathbf{g}}_{\boldsymbol{\gamma}}\left(\mathbf{z}_t,\dot{\mathbf{x}}_{2,t};\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\boldsymbol{\gamma}}\left(\mathbf{z}_t,\dot{\mathbf{x}}_{2,t};\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^o\right)\\
&\quad -\frac{1}{n_{y,t}}\left(\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)-\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)\right)'\left(\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)-\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)\right)\\
&\quad -\frac{1}{n_{y,t}}2\left(\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)-\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)\right)'\widehat{\tilde{\mathbf{v}}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)\\
&\quad -\frac{1}{n_{y,t}}2\left(\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)-\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)\right)'\tilde{\mathbf{g}}_{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\mathbf{x}}_{2,t}-\mathbf{x}_{2,t}^o\right)\\
&\quad -\frac{1}{n_{y,t}}2\left(\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)-\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)\right)'\tilde{\mathbf{g}}_{\boldsymbol{\theta}_1}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o\right)\\
&\quad -\frac{1}{n_{y,t}}2\left(\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)-\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)\right)'\tilde{\mathbf{g}}_{\boldsymbol{\gamma}}\left(\mathbf{z}_t,\dot{\mathbf{x}}_{2,t};\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^o\right).
\end{aligned}
$$

We then note that

$$
\begin{aligned}
\frac{1}{n_{y,t}}\tilde{\mathbf{v}}_t'\tilde{\mathbf{v}}_t &= \frac{1}{n_{y,t}}\sum_{j=1}^{n_{y,t}}\left(\tilde{v}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)\right)^2=\frac{1}{n_{y,t}}\left(\mathbf{S}_{\mathbf{v}}\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)^{-1}\mathbf{v}_t\right)'\left(\mathbf{S}_{\mathbf{v}}\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)^{-1}\mathbf{v}_t\right)\\
&= \frac{1}{n_{y,t}}\mathbf{v}_t'\mathbf{V}\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)^{-1}\mathbf{v}_t
\end{aligned}
$$

Moreover, $E\left[\frac{1}{n_{y,t}}\tilde{\mathbf{v}}_t'\tilde{\mathbf{v}}_t\right]=\frac{1}{n_{y,t}}E\left[\mathbf{v}_t'\mathbf{V}\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)^{-1}\mathbf{v}_t\right]=\frac{(\sigma_t^o)^2}{n_{y,t}}E\left[\mathbf{v}_t'\left((\sigma_t^o)^2\mathbf{V}\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)\right)^{-1}\mathbf{v}_t\right]=(\sigma_t^o)^2$. Thus, each term in $\frac{1}{n_{y,t}}\sum_{j=1}^{n_{y,t}}\left(\tilde{v}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)\right)^2$ has a mean value of $(\sigma_t^o)^2$. Hence, the law of large numbers implies $\frac{1}{n_{y,t}}\tilde{\mathbf{v}}_t'\tilde{\mathbf{v}}_t\xrightarrow{p}(\sigma_t^o)^2$ for $n_{y,t}\longrightarrow\infty$. Using the stated assumptions, it then follows that all remaining terms in (79) converge to zero. This proves consistency of $\hat{\sigma}_t^2=\frac{1}{n_{y,t}}\widehat{\tilde{\mathbf{v}}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)'\widehat{\tilde{\mathbf{v}}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)$, as claimed in Section 4.1.1.

D.2 Proof of asymptotic normality of $\hat{\sigma}_t^2$

To prove asymptotic normality, we scale the expression for $\frac{1}{n_{y,t}}\widehat{\tilde{\mathbf{v}}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)'\widehat{\tilde{\mathbf{v}}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)$ by $\sqrt{n_{y,t}}$. Thus,

$$
\begin{aligned}
&\sqrt{n_{y,t}}\left(\tfrac{1}{n_{y,t}}\widehat{\tilde{\mathbf{v}}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)'\widehat{\tilde{\mathbf{v}}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)-\left(\sigma_t^o\right)^2\right)=\sqrt{n_{y,t}}\left(\tfrac{1}{n_{y,t}}\tilde{\mathbf{v}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)'\tilde{\mathbf{v}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)-\left(\sigma_t^o\right)^2\right)\\
&\quad-\tfrac{1}{\sqrt{n_{y,t}}}2\widehat{\tilde{\mathbf{v}}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\mathbf{x}}_{2,t}-\mathbf{x}_{2,t}^o\right)\\
&\quad-\tfrac{1}{\sqrt{n_{y,t}}}2\widehat{\tilde{\mathbf{v}}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\boldsymbol{\theta}_1}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o\right)\\
&\quad-\tfrac{1}{\sqrt{n_{y,t}}}2\widehat{\tilde{\mathbf{v}}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\boldsymbol{\gamma}}\left(\mathbf{z}_t,\dot{\mathbf{x}}_{2,t};\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^o\right)\\
&\quad-\tfrac{1}{\sqrt{n_{y,t}}}\left(\hat{\mathbf{x}}_{2,t}-\mathbf{x}_{2,t}^o\right)\tilde{\mathbf{g}}_{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\mathbf{x}}_{2,t}-\mathbf{x}_{2,t}^o\right)\\
&\quad-\tfrac{1}{\sqrt{n_{y,t}}}2\left(\hat{\mathbf{x}}_{2,t}-\mathbf{x}_{2,t}^o\right)\tilde{\mathbf{g}}_{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\boldsymbol{\theta}_1}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o\right)\\
&\quad-\tfrac{1}{\sqrt{n_{y,t}}}2\left(\hat{\mathbf{x}}_{2,t}-\mathbf{x}_{2,t}^o\right)\tilde{\mathbf{g}}_{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\boldsymbol{\gamma}}\left(\mathbf{z}_t,\dot{\mathbf{x}}_{2,t};\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^o\right)\\
&\quad-\tfrac{1}{\sqrt{n_{y,t}}}\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o\right)'\tilde{\mathbf{g}}_{\boldsymbol{\theta}_1}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\boldsymbol{\theta}_1}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o\right)\\
&\quad-\tfrac{1}{\sqrt{n_{y,t}}}2\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o\right)'\tilde{\mathbf{g}}_{\boldsymbol{\theta}_1}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\boldsymbol{\gamma}}\left(\mathbf{z}_t,\dot{\mathbf{x}}_{2,t};\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^o\right)\\
&\quad-\tfrac{1}{\sqrt{n_{y,t}}}\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^o\right)'\tilde{\mathbf{g}}_{\boldsymbol{\gamma}}\left(\mathbf{z}_t,\dot{\mathbf{x}}_{2,t};\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)'\tilde{\mathbf{g}}_{\boldsymbol{\gamma}}\left(\mathbf{z}_t,\dot{\mathbf{x}}_{2,t};\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^o\right)\\
&\quad-\tfrac{1}{\sqrt{n_{y,t}}}\left(\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)-\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)\right)'\left(\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)-\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)\right)\\
&\quad-\tfrac{1}{\sqrt{n_{y,t}}}2\left(\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)-\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)\right)'\widehat{\tilde{\mathbf{v}}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)\\
&\quad-\tfrac{1}{\sqrt{n_{y,t}}}2\left(\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)-\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)\right)'\tilde{\mathbf{g}}_{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\mathbf{x}}_{2,t}-\mathbf{x}_{2,t}^o\right)\\
&\quad-\tfrac{1}{\sqrt{n_{y,t}}}2\left(\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)-\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)\right)'\tilde{\mathbf{g}}_{\boldsymbol{\theta}_1}\left(\dot{\mathbf{x}}_{2,t},\mathbf{z}_t;\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o\right)\\
&\quad-\tfrac{1}{\sqrt{n_{y,t}}}2\left(\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)-\tilde{\mathbf{y}}_t\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)\right)'\tilde{\mathbf{g}}_{\boldsymbol{\gamma}}\left(\mathbf{z}_t,\dot{\mathbf{x}}_{2,t};\dot{\boldsymbol{\theta}}_1,\dot{\boldsymbol{\gamma}}\right)\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^o\right).
\end{aligned}
$$

Consider
$\frac{1}{\sqrt{n_{y,t}}}\tilde{\mathbf{v}}_t'\tilde{\mathbf{v}}_t-\sqrt{n_{y,t}}\left(\sigma_t^o\right)^2=\frac{1}{\sqrt{n_{y,t}}}\sum_{i=1}^{n_{y,t}}\left(\tilde{v}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)\right)^2-\frac{n_{y,t}}{\sqrt{n_{y,t}}}\left(\sigma_t^o\right)^2=\frac{1}{\sqrt{n_{y,t}}}\sum_{i=1}^{n_{y,t}}\left(\left(\tilde{v}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)\right)^2-\left(\sigma_t^o\right)^2\right)$
where $\left\{\left(\tilde{v}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)\right)^2-\left(\sigma_t^o\right)^2\right\}_{j=1}^{n_{y,t}}$ is an iid sequence with mean value zero and a variance of

$$
E\left[\left(\left(\tilde{v}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)\right)^2-\left(\sigma_t^o\right)^2\right)^2\right]=E\left[\left(\tilde{v}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)\right)^4\right]-\sigma_{t,o}^4.
$$

By the central limit theorem we have

$$
\frac{1}{\sqrt{n_{y,t}}}\tilde{\mathbf{v}}_t'\tilde{\mathbf{v}}_t-\sqrt{n_{y,t}}\left(\sigma_t^o\right)^2\xrightarrow{d}\mathcal{N}\left(0,E\left[\left(\tilde{v}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)\right)^4\right]-\left(\sigma_t^o\right)^4\right)
$$

for $n_{y,t}\longrightarrow\infty$. Given the stated assumptions, all remaining terms converge to zero. In conclusion, we thus have

$$
\sqrt{n_{y,t}}\left(\hat{\sigma}_t^2-\left(\sigma_t^o\right)^2\right)\xrightarrow{d}\mathcal{N}\left(0,E\left[\left(\tilde{v}_{t,j}\left(\mathbf{z}_t;\boldsymbol{\gamma}^o\right)\right)^4\right]-\left(\sigma_t^o\right)^4\right)
$$

for $n_{y,t}\longrightarrow\infty$, as claimed in (39).

E The asymptotic distribution of the latent factors, $\hat{\mathbf{x}}_{2,t}$

The first-order condition with respect to $\mathbf{x}_{2,t}$ is $\sum_{j=1}^{n_{y,t}} \mathbf{s}_j^{\mathbf{x}_2}\left(\hat{\mathbf{x}}_{2,t}, \mathbf{z}_t; \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\gamma}}\right) = \mathbf{0}$, where

$$\mathbf{s}_j^{\mathbf{x}_2}\left(\mathbf{x}_{2,t}, \mathbf{z}_t; \boldsymbol{\theta}_1, \boldsymbol{\gamma}\right) \equiv -\left(\tilde{y}_{t,j}\left(\mathbf{z}_t; \boldsymbol{\gamma}\right) - \tilde{g}_j\left(\mathbf{x}_{2,t}, \mathbf{z}_t; \boldsymbol{\theta}_1, \boldsymbol{\gamma}\right)\right) \frac{\partial \tilde{g}_j\left(\mathbf{x}_{2,t}, \mathbf{z}_t; \boldsymbol{\theta}_1, \boldsymbol{\gamma}\right)}{\partial \mathbf{x}_{2,t}} \tag{80}$$

with dimension $n_{x_2} \times 1$. A mean-value expansion around $\mathbf{x}_{2,t}^o$, $\boldsymbol{\theta}_1^o$, and $\boldsymbol{\gamma}^*$ implies

$$\begin{aligned}\sum_{j=1}^{n_{y,t}} \mathbf{s}_j^{\mathbf{x}_2}\left(\hat{\mathbf{x}}_{2,t}, \mathbf{z}_t; \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\gamma}}\right) &= \sum_{j=1}^{n_{y,t}} \mathbf{s}_j^{\mathbf{x}_2}\left(\mathbf{x}_{2,t}^o, \mathbf{z}_t; \boldsymbol{\theta}_1^o, \boldsymbol{\gamma}^*\right) + \sum_{j=1}^{n_{y,t}} \frac{\partial \mathbf{s}_j^{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t}, \mathbf{z}_t; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right)}{\partial \mathbf{x}_{2,t}'}\left(\hat{\mathbf{x}}_{2,t} - \mathbf{x}_{2,t}^o\right) \\ &\quad + \sum_{j=1}^{n_{y,t}} \frac{\partial \mathbf{s}_j^{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t}, \mathbf{z}_t; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right)}{\partial \boldsymbol{\theta}_1'}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^o\right).\end{aligned}$$

Let $\mathbf{H}_j^{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t}, \mathbf{z}_t; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right) \equiv \frac{\partial \mathbf{s}_j^{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t}, \mathbf{z}_t; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right)}{\partial \mathbf{x}_t'}$ be the Hessian matrix for the latent factors in time period t for observation j. Further, let $\mathbf{D}_j\left(\dot{\mathbf{x}}_{2,t}, \mathbf{z}_t; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right) \equiv \frac{\partial \mathbf{s}_j^{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t}, \mathbf{z}_t; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right)}{\partial \boldsymbol{\theta}_1'}$ be the Jacobian of dimension $n_{x_2} \times L_1$. Using these definitions and the first-order condition, the mean-value expansion simplifies to

$$\begin{aligned}\sqrt{n_{y,t}}\left(\hat{\mathbf{x}}_{2,t} - \mathbf{x}_{2,t}^o\right) &= \left(\frac{1}{n_{y,t}} \sum_{j=1}^{n_{y,t}} \mathbf{H}_j^{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t}, \mathbf{z}_t; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right)\right)^{-1} \left(\frac{-1}{\sqrt{n_{y,t}}} \sum_{j=1}^{n_{y,t}} \mathbf{s}_j^{\mathbf{x}_2}\left(\mathbf{x}_{2,t}^o, \mathbf{z}_t; \boldsymbol{\theta}_1^o, \boldsymbol{\gamma}^*\right)\right) \\ &\quad - \left(\frac{1}{n_{y,t}} \sum_{j=1}^{n_{y,t}} \mathbf{H}_j^{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t}, \mathbf{z}_t; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right)\right)^{-1} \left(\frac{1}{n_{y,t}} \sum_{j=1}^{n_{y,t}} \mathbf{D}_j\left(\dot{\mathbf{x}}_{2,t}, \mathbf{z}_t; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right)\right) \sqrt{n_{y,t}}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^o\right).\end{aligned}$$

From Appendix C, we have

$$\sqrt{n_{y,t}}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^o\right) = \frac{\sqrt{n_{y,t}}}{\sqrt{N}} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{H}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right)\right)^{-1} \left(\frac{-1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{s}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i; \boldsymbol{\theta}_1^o, \boldsymbol{\gamma}^*\right)\right).$$

Hence,

$$\begin{aligned}\sqrt{n_{y,t}}\left(\hat{\mathbf{x}}_{2,t} - \mathbf{x}_{2,t}^o\right) &= \left(\frac{1}{n_{y,t}} \sum_{j=1}^{n_{y,t}} \mathbf{H}_j^{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t}, \mathbf{z}_t; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right)\right)^{-1} \left\{\left(\frac{-1}{\sqrt{n_{y,t}}} \sum_{j=1}^{n_{y,t}} \mathbf{s}_j^{\mathbf{x}_2}\left(\mathbf{x}_{2,t}^o, \mathbf{z}_t; \boldsymbol{\theta}_1^o, \boldsymbol{\gamma}^*\right)\right)\right. \\ &\quad \left. - \left(\frac{1}{n_{y,t}} \sum_{j=1}^{n_{y,t}} \mathbf{D}_j\left(\dot{\mathbf{x}}_{2,t}, \mathbf{z}_t; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right)\right) \frac{\sqrt{n_{y,t}}}{\sqrt{N}} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{H}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right)\right)^{-1} \left(\frac{-1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{s}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i; \boldsymbol{\theta}_1^o, \boldsymbol{\gamma}^*\right)\right)\right\}.\end{aligned}$$

We then note that $\frac{\sqrt{n_{y,t}}}{\sqrt{N}} = \sqrt{c_t}$. If either $\phi_t(m) = O\left(m^{-v}\right)$ for $\nu > r_2/(r_2 - 1)$ and $t = 1, 2, ..., T$ or $\alpha_t(m) = O\left(m^{-\nu}\right)$ for $\nu > \max\left\{r_1/(r_1 - 1), r_2/(r_2 - 1)\right\}$ and $t = 1, 2, ..., T$, where $r_1 > 1$, it follows from White & Domowitz (1984) that $\sqrt{n_{y,t}}\left(\hat{\mathbf{x}}_{2,t} - \mathbf{x}_{2,t}^o\right)$ based on $\boldsymbol{\gamma}^*$ is asymptotically normal when $n_{y,t} \to \infty$ and the covariance matrix is $\left(\mathbf{A}_{t,o}^{\mathbf{x}_2}\right)^{-1} \mathbf{V}_t^{\mathbf{x}_2} \left(\mathbf{A}_{t,o}^{\mathbf{x}_2}\right)^{-1}$. See White & Domowitz (1984) for the appropriate regularity conditions or the technical appendix accompanying the present paper. A straightforward application of the asymptotic equivalence lemma reveals that the asymptotic distribution of $\sqrt{n_{y,t}}\left(\hat{\mathbf{x}}_{2,t} - \mathbf{x}_{2,t}^o\right)$ is unaffected if $\boldsymbol{\gamma}^*$ is replaced by $\hat{\boldsymbol{\gamma}}$.

F The time series properties of $\mathbf{u}_{2,t}$

From the mean-value expansion in relation to $\hat{\mathbf{x}}_{2,t}$, we have

$\sqrt{n_{y,t}}\mathbf{u}_{2,t} \equiv \sqrt{n_{y,t}}\left(\hat{\mathbf{x}}_{2,t} - \mathbf{x}_{2,t}^o\right)$

$$= \left(\frac{1}{n_{y,t}}\sum_{j=1}^{n_{y,t}} \mathbf{H}^{\mathbf{x}_2}\left(\dot{\mathbf{x}}_{2,t}, \mathbf{z}_t; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right)\right)^{-1} \left\{ \frac{1}{\sqrt{n_{y,t}}}\sum_{j=1}^{n_{y,t}} \tilde{v}_{t,j}\left(\mathbf{z}_t; \boldsymbol{\gamma}^*\right) \frac{\partial \tilde{g}_j\left(\mathbf{x}^o_{2,t}, \mathbf{z}_t; \boldsymbol{\theta}^o_1, \boldsymbol{\gamma}^*\right)}{\partial \mathbf{x}^o_{2,t}} \right.$$
$$- \left(\frac{1}{n_{y,t}}\sum_{j=1}^{n_{y,t}} \mathbf{D}\left(\dot{\mathbf{x}}_{2,t}, \mathbf{z}_{t,j}; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right)\right) \sqrt{c_t} \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{H}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right)\right)^{-1}$$
$$\left. \times \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \tilde{v}_i\left(\mathbf{z}_i; \boldsymbol{\gamma}^*\right) \left(\frac{\partial \tilde{g}_i(\mathbf{x}_{2,i}(\boldsymbol{\theta}^o_1), \mathbf{z}_i; \boldsymbol{\theta}^o_1, \boldsymbol{\gamma}^*)}{\partial \boldsymbol{\theta}^o_1} + \frac{\partial \mathbf{x}'_{2,i}(\boldsymbol{\theta}^o_1)}{\partial \boldsymbol{\theta}^o_1} \frac{\partial \tilde{g}_i(\mathbf{x}_{2,i}(\boldsymbol{\theta}^o_1), \mathbf{z}_i; \boldsymbol{\theta}^o_1, \boldsymbol{\gamma}^*)}{\partial \mathbf{x}_{2,i}(\boldsymbol{\theta}^o_1)}\right)\right)\right\}.$$

The last component in this expression introduces autocorrelation into $\mathbf{u}_{2,t}$. However, this effect is very small for standard values of N and is therefore ignored in this approximated expression for $\mathbf{u}_{2,t}$. Note also that for estimating the unconditional correlation (for $T \longrightarrow \infty$), this effect tends to zero because $c_t \longrightarrow 0$ when $T \longrightarrow \infty$ as $\boldsymbol{\theta}_1$ is then estimated super-consistently. Hence,

$$\sqrt{n_{y,t}}\mathbf{u}_{2,t} \simeq \left(\mathbf{A}^{\mathbf{x}_2}_t\left(\dot{\mathbf{x}}_{2,t}, \mathbf{z}_t; \dot{\boldsymbol{\theta}}_1, \boldsymbol{\gamma}^*\right)\right)^{-1} \frac{1}{\sqrt{n_{y,t}}}\sum_{j=1}^{n_{y,t}} \tilde{v}_{t,j}\left(\mathbf{z}_t; \boldsymbol{\gamma}^*\right) \frac{\partial \tilde{g}_j\left(\mathbf{x}^o_{2,t}, \mathbf{z}_t; \boldsymbol{\theta}^o_1, \boldsymbol{\gamma}^*\right)}{\partial \mathbf{x}^o_{2,t}}$$

or simply $\sqrt{n_{y,t}}\mathbf{u}_{2,t} \simeq \left(\mathbf{A}^{\mathbf{x}_2}_t\right)^{-1} \frac{1}{\sqrt{n_{y,t}}} \frac{\partial \tilde{\mathbf{g}}'_t}{\partial \mathbf{x}^o_{2,t}} \tilde{\mathbf{v}}_t$, which is (42).

G The asymptotic distribution of the $\hat{\boldsymbol{\theta}}_2$

The first-order condition for the optimization problem in (28) can be written as

$$\mathbf{R}^{\boldsymbol{\theta}_2}\left(\hat{\boldsymbol{\theta}}_{11}, \hat{\boldsymbol{\theta}}_2\right)' \mathbf{W}\left(\mathbf{q}_T\left(\hat{\boldsymbol{\theta}}_{11}, \hat{\boldsymbol{\theta}}_2\right)\right) = \mathbf{0}, \tag{81}$$

where $\mathbf{R}^{\boldsymbol{\theta}_2}\left(\boldsymbol{\theta}_{11}, \boldsymbol{\theta}_2\right) \equiv \frac{\partial \mathbf{q}_T(\boldsymbol{\theta}_{11}, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}'_2}$ has dimension $n_m \times L_2$. A mean-value expansion of $\mathbf{q}_T\left(\hat{\boldsymbol{\theta}}_{11}, \hat{\boldsymbol{\theta}}_2\right)$ around $\boldsymbol{\theta}^o_{11}$ and $\boldsymbol{\theta}^o_2$ gives

$$\mathbf{q}_T\left(\hat{\boldsymbol{\theta}}_{11}, \hat{\boldsymbol{\theta}}_2\right) = \mathbf{q}_T\left(\boldsymbol{\theta}^o_{11}, \boldsymbol{\theta}^o_2\right) + \mathbf{R}^{\boldsymbol{\theta}_{11}}\left(\dot{\boldsymbol{\theta}}_{11}, \dot{\boldsymbol{\theta}}_2\right)\left(\hat{\boldsymbol{\theta}}_{11} - \boldsymbol{\theta}^o_{11}\right) + \mathbf{R}^{\boldsymbol{\theta}_2}\left(\dot{\boldsymbol{\theta}}_{11}, \dot{\boldsymbol{\theta}}_2\right)\left(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^o_2\right), \tag{82}$$

with $\mathbf{R}^{\boldsymbol{\theta}_{11}}\left(\boldsymbol{\theta}_{11}, \boldsymbol{\theta}_2\right) \equiv \frac{\partial \mathbf{q}_T(\boldsymbol{\theta}_{11}, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}'_{11}}$. Here, $\dot{\boldsymbol{\theta}}_{11}$ denotes that each row is evaluated at a potentially different convex combination between $\hat{\boldsymbol{\theta}}_{11}$ and $\boldsymbol{\theta}^o_{11}$. A similar notation applies for $\boldsymbol{\theta}_2$. Inserting (82) into (81) gives

$$\begin{aligned} & \mathbf{R}^{\boldsymbol{\theta}_2}\left(\hat{\boldsymbol{\theta}}_{11}, \hat{\boldsymbol{\theta}}_2\right)' \mathbf{W}\left(\frac{1}{\sqrt{T}}\textstyle\sum_{t=1}^{T} \mathbf{q}_t\left(\boldsymbol{\theta}^o_{11}, \boldsymbol{\theta}^o_2\right)\right) + \mathbf{R}^{\boldsymbol{\theta}_2}\left(\hat{\boldsymbol{\theta}}_{11}, \hat{\boldsymbol{\theta}}_2\right)' \mathbf{W}\mathbf{R}^{\boldsymbol{\theta}_{11}}\left(\dot{\boldsymbol{\theta}}_{11}, \dot{\boldsymbol{\theta}}_2\right)\left(\hat{\boldsymbol{\theta}}_{11} - \boldsymbol{\theta}^o_{11}\right)\sqrt{T} \\ & + \mathbf{R}^{\boldsymbol{\theta}_2}\left(\hat{\boldsymbol{\theta}}_{11}, \hat{\boldsymbol{\theta}}_2\right)' \mathbf{W}\mathbf{R}^{\boldsymbol{\theta}_2}\left(\dot{\boldsymbol{\theta}}_{11}, \dot{\boldsymbol{\theta}}_2\right)\left(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^o_2\right)\sqrt{T} \\ = \; & \mathbf{0}. \end{aligned}$$

When T tends to infinity, $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \mathbf{q}_t\left(\boldsymbol{\theta}^o_{11}, \boldsymbol{\theta}^o_2\right)$ is bounded in probability and $\mathbf{R}^{\boldsymbol{\theta}_2}\left(\hat{\boldsymbol{\theta}}_{11}, \hat{\boldsymbol{\theta}}_2\right)$ converges to $\mathbf{R}^{\boldsymbol{\theta}_2}_o \equiv \mathbf{R}^{\boldsymbol{\theta}_2}\left(\boldsymbol{\theta}^o_{11}, \boldsymbol{\theta}^o_2\right)$ given standard regularity conditions. Moreover, $\mathbf{R}^{\boldsymbol{\theta}_{11}}\left(\dot{\boldsymbol{\theta}}_{11}, \dot{\boldsymbol{\theta}}_2\right)$ converges to $\mathbf{R}^{\boldsymbol{\theta}_{11}}_o \equiv \mathbf{R}^{\boldsymbol{\theta}_{11}}\left(\boldsymbol{\theta}^o_{11}, \boldsymbol{\theta}^o_2\right)$ and $\mathbf{R}^{\boldsymbol{\theta}_2}\left(\dot{\boldsymbol{\theta}}_{11}, \dot{\boldsymbol{\theta}}_2\right)$ converges to $\mathbf{R}^{\boldsymbol{\theta}_2}_o$. Thus we have

$$\left(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^o_2\right)\sqrt{T} = \left(\mathbf{A}^{\boldsymbol{\theta}_2}_o\right)^{-1}\left\{-\mathbf{R}^{\boldsymbol{\theta}_2\prime}_o \mathbf{W}\left(\frac{1}{\sqrt{T}}\textstyle\sum_{t=1}^{T} \mathbf{q}_t\left(\boldsymbol{\theta}^o_{11}, \boldsymbol{\theta}^o_2\right)\right) - \mathbf{G}^{\boldsymbol{\theta}_2, \boldsymbol{\theta}_{11}}_o\left(\hat{\boldsymbol{\theta}}_{11} - \boldsymbol{\theta}^o_{11}\right)\sqrt{T}\right\}, \tag{83}$$

where $\mathbf{A}_o^{\boldsymbol{\theta}_2} \equiv \mathbf{R}_o^{\boldsymbol{\theta}_2\prime}\mathbf{W}\mathbf{R}_o^{\boldsymbol{\theta}_2}$ and $\mathbf{G}_o^{\boldsymbol{\theta}_2,\boldsymbol{\theta}_{11}} \equiv \mathbf{R}_o^{\boldsymbol{\theta}_2\prime}\mathbf{W}\mathbf{R}_o^{\boldsymbol{\theta}_{11}}$. The first step in the SR approach implies

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_{11} - \boldsymbol{\theta}_{11}^o\right) = \frac{\sqrt{T}}{\sqrt{N}}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{H}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i;\dot{\boldsymbol{\theta}}_1,\boldsymbol{\gamma}^*\right)\right)^{-1}\left(\frac{-1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i;\boldsymbol{\theta}_1^o,\boldsymbol{\gamma}^*\right)\right).$$

Here, we introduce f_t by $n_{y,t} = n_y f_t$ for $t = 1, 2, .., T$. By normalization $\sum_{t=1}^{T} f_t = T$, meaning that $N \equiv \sum_{t=1}^{T} n_{y,t} = \sum_{t=1}^{T} n_y f_t = n_y \sum_{t=1}^{T} f_t = n_y T$. Thus, we get

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_{11} - \boldsymbol{\theta}_{11}^o\right) = \sqrt{\frac{1}{n_y}}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{H}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i;\dot{\boldsymbol{\theta}}_1,\boldsymbol{\gamma}^*\right)\right)^{-1}\left(\frac{-1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i;\boldsymbol{\theta}_1^o,\boldsymbol{\gamma}^*\right)\right). \tag{84}$$

Inserting (84) in (83) gives

$$\begin{aligned}\left(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^o\right)\sqrt{T} \;=\; & \left(\mathbf{A}_o^{\boldsymbol{\theta}_2}\right)^{-1}\left\{-\mathbf{R}_o^{\boldsymbol{\theta}_2\prime}\mathbf{W}\left(\frac{1}{\sqrt{T}}\textstyle\sum_{t=1}^{T}\mathbf{q}_t\left(\boldsymbol{\theta}_{11}^o,\boldsymbol{\theta}_2^o\right)\right)\right\} \\ & -\left(\mathbf{A}_o^{\boldsymbol{\theta}_2}\right)^{-1}\mathbf{G}_o^{\boldsymbol{\theta}_2,\boldsymbol{\theta}_{11}}\left(\sqrt{\frac{1}{n_y}}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{H}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i;\dot{\boldsymbol{\theta}}_1,\boldsymbol{\gamma}^*\right)\right)^{-1}\left(\frac{-1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i;\boldsymbol{\theta}_1^o,\boldsymbol{\gamma}^*\right)\right)\right)\end{aligned} \tag{85}$$

In the first step of the SR approach, we let $n_{y,t} \longrightarrow \infty$ for $t = 1, 2, ..., T$, implying $n_y \longrightarrow \infty$. Hence, the second term in (85) drops out asymptotically and we get

$$\left(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^o\right)\sqrt{T} = \left(\mathbf{A}_o^{\boldsymbol{\theta}_2}\right)^{-1}\left\{-\mathbf{R}_o^{\boldsymbol{\theta}_2\prime}\mathbf{W}\left(\frac{1}{\sqrt{T}}\textstyle\sum_{t=1}^{T}\mathbf{q}_t\left(\boldsymbol{\theta}_{11}^o,\boldsymbol{\theta}_2^o\right)\right)\right\},$$

as for standard GMM estimation, implying $\hat{\boldsymbol{\theta}}_2 \sim \mathcal{N}\left(\boldsymbol{\theta}_2^o, \mathbf{V}_{\boldsymbol{\theta}_2}/T\right)$.

To derive the finite sample correction, we note that the variance of (83) for a given value of n_y is

$$T \times Var\left(\hat{\boldsymbol{\theta}}_2\right) = \mathbf{V}_{\boldsymbol{\theta}_2} + T \times \left(\mathbf{A}_o^{\boldsymbol{\theta}_2}\right)^{-1}\mathbf{G}_o^{\boldsymbol{\theta}_2,\boldsymbol{\theta}_{11}} Var\left(\hat{\boldsymbol{\theta}}_{11}\right)\left(\mathbf{G}_o^{\boldsymbol{\theta}_2,\boldsymbol{\theta}_{11}}\right)'\left(\mathbf{A}_o^{\boldsymbol{\theta}_2}\right)^{-1} \tag{86}$$

which simplifies to (48). Note that the covariances in (86) drop out due to Proposition 3.

H Proof of Proposition 3

We first recall that the asymptotic distribution of $\hat{\boldsymbol{\theta}}_1^{step1}$ and hence $\hat{\boldsymbol{\theta}}_{12}^{step1}$ follows from $\frac{\partial Q}{\partial \boldsymbol{\theta}_1} = -\sum_{t=1}^{T}\sum_{j=1}^{n_{y,t}}\tilde{v}_{t,j}\left(\mathbf{z}_t;\hat{\boldsymbol{\gamma}}\right)\left(\frac{\partial \tilde{g}_j(\hat{\mathbf{x}}_{2,t}(\boldsymbol{\theta}_1),\mathbf{z}_t;\boldsymbol{\theta}_1,\hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\theta}_1} + \frac{\partial \hat{\mathbf{x}}_{2,t}'(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1}\frac{\partial \tilde{g}_j(\hat{\mathbf{x}}_{2,t}(\boldsymbol{\theta}_1),\mathbf{z}_t;\boldsymbol{\theta}_1,\hat{\boldsymbol{\gamma}})}{\partial \hat{\mathbf{x}}_{2,t}(\boldsymbol{\theta}_1)}\right)$. Hence, the measurement errors $\tilde{v}_{t,j}$ when properly scaled determine the behavior of $\hat{\boldsymbol{\theta}}_1^{step1}$. The properties of $\hat{\boldsymbol{\theta}}_{12}^{step2}$ and $\hat{\boldsymbol{\theta}}_{22}^{step2}$ depend on $\mathbf{w}_t$ (with appropriate scaling), the estimation uncertainty in $\hat{\boldsymbol{\theta}}_{11}^{step1}$, and $\tilde{v}_{t,j}$ affecting the size of the measurement errors in any estimated latent factors. Consistency of the SR approach implies that these measurement errors tend to zero when $n_{y,t} \longrightarrow \infty$ for all t and so does the estimation uncertainty in $\hat{\boldsymbol{\theta}}_{11}^{step1}$. Hence , $\hat{\boldsymbol{\theta}}_1^{step1}$ is uncorrelated with $\hat{\boldsymbol{\theta}}_{22}^{step2}$ and $\hat{\boldsymbol{\theta}}_{12}^{step2}$ when $n_{y,t} \longrightarrow \infty$ for all t, given the assumption in (4).

I Optimal value of $\mathbf{\Lambda}$

Let $\mathbf{e}^i$ denote the i'th $1 \times L_{12}$ unit vector, where $e^i(i) = 1$ and $e^i(j) = 0$ for $j \neq i$. The variance of the i'th element in $\hat{\boldsymbol{\theta}}_{12}^{step3}$ is

$$Var\left(\hat{\boldsymbol{\theta}}_{12,i}^{step3}\right) = \mathbf{\Lambda}_i Var\left(\hat{\boldsymbol{\theta}}_{12}^{step1}\right)\mathbf{\Lambda}_i' + \left(\mathbf{e}^i - \mathbf{\Lambda}_i\right) Var\left(\hat{\boldsymbol{\theta}}_{12}^{step2}\right)\left(\mathbf{e}^i - \mathbf{\Lambda}_i\right)',$$

where $\mathbf{\Lambda}_i$ has dimension $1 \times L_{12}$ and refers to the i'th row of $\mathbf{\Lambda}$. The first-order condition for minimization of $Var\left(\hat{\boldsymbol{\theta}}_{12,i}^{step3}\right)$ is

$$\frac{\partial Var\left(\hat{\boldsymbol{\theta}}_{12,i}^{step3}\right)}{\partial \mathbf{\Lambda}_i} = 2Var\left(\hat{\boldsymbol{\theta}}_{12}^{step1}\right)\mathbf{\Lambda}_i' - 2Var\left(\hat{\boldsymbol{\theta}}_{12}^{step2}\right)\left(\mathbf{e}^i - \mathbf{\Lambda}_i\right)' = \mathbf{0}$$

or

$$\mathbf{\Lambda}_i' = \left(Var\left(\hat{\boldsymbol{\theta}}_{12}^{step1}\right) + Var\left(\hat{\boldsymbol{\theta}}_{12}^{step2}\right)\right)^{-1} Var\left(\hat{\boldsymbol{\theta}}_{12}^{step2}\right)\left(\mathbf{e}^i\right)'$$

Using the same condition for $i = 1, 2, ..., L_{12}$ we obtain

$$\mathbf{\Lambda}' = \left(Var\left(\hat{\boldsymbol{\theta}}_{12}^{step1}\right) + Var\left(\hat{\boldsymbol{\theta}}_{12}^{step2}\right)\right)^{-1} Var\left(\hat{\boldsymbol{\theta}}_{12}^{step2}\right)\mathbf{I},$$

which corresponds to (50) after transposing. Note finally that the first-order conditions are sufficient because $\partial^2 Var\left(\hat{\boldsymbol{\theta}}_{12,i}^{step3}\right)/\partial \mathbf{\Lambda}_i \partial \mathbf{\Lambda}_i'$ is positive definite, provided $Var\left(\hat{\boldsymbol{\theta}}_{12}^{step1}\right)$ and $Var\left(\hat{\boldsymbol{\theta}}_{12}^{step2}\right)$ are positive definite.

J Asymptotic distribution of $\hat{\boldsymbol{\theta}}_{11}^{step3}$

Using the same arguments as in the first step of the SR approach, cross-section inference when conditioning on the true value of $\boldsymbol{\theta}_{12}^o$ is given by

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^o\right) - \boldsymbol{\theta}_{11}^o\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{V}_{\boldsymbol{\theta}_{11}}^{step1}\left(\boldsymbol{\theta}_{12}^o\right)\right). \tag{87}$$

To derive the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)$ we apply a first-order approximation of $\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)$ around $\boldsymbol{\theta}_{12}^o$ and obtain

$$\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right) = \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^o\right) + \mathbf{K}\left(\hat{\boldsymbol{\theta}}_{12}^{step3} - \boldsymbol{\theta}_{12}^o\right), \tag{88}$$

where $\mathbf{K} \equiv \frac{\partial \hat{\boldsymbol{\theta}}_{11}^{step3}(\boldsymbol{\theta}_{12}^o)}{\partial \boldsymbol{\theta}_{12}'}$ is the Jacobian with dimension $L_{11} \times L_{12}$. Using (88) we have

$$\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right) - \boldsymbol{\theta}_{11}^o = \left(\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^o\right) - \boldsymbol{\theta}_{11}^o\right) + \mathbf{K}\left(\hat{\boldsymbol{\theta}}_{12}^{step3} - \boldsymbol{\theta}_{12}^o\right).$$

We know $\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^o\right)$ is asymptotically normal and so is $\hat{\boldsymbol{\theta}}_{12}^{step3}$, implying that $\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)$ is also asymptotically normal. Moreover, the mean value of $\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)$ is directly seen to be $\boldsymbol{\theta}_{11}^o$. To compute the asymptotic variance of $\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)$, we observe

$$Var\left(\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right) - \boldsymbol{\theta}_{11}^o\right) = Var\left(\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right) - \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^o\right) + \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^o\right) - \boldsymbol{\theta}_{11}^o\right)$$

$$= Var\left(\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right) - \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right)\right) + Var\left(\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right) - \boldsymbol{\theta}_{11}^{o}\right)$$
$$+Cov\left(\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right) - \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right), \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right) - \boldsymbol{\theta}_{11}^{o}\right)$$
$$+Cov\left(\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right) - \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right), \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right) - \boldsymbol{\theta}_{11}^{o}\right)'.$$

We next compute each of these terms. Firstly, using (88)
$Var\left(\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right) - \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right)\right) = Var\left(\mathbf{K}\left(\hat{\boldsymbol{\theta}}_{12}^{step3} - \boldsymbol{\theta}_{12}^{o}\right)\right) = \mathbf{K}Var\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)\mathbf{K}'.$

Secondly, from (87), we directly have
$Var\left(\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right) - \boldsymbol{\theta}_{11}^{o}\right) = \frac{\mathbf{V}_{\boldsymbol{\theta}_{11}}^{step3}(\boldsymbol{\theta}_{12}^{o})}{N}.$

Finally, for the covariance term
$Cov\left(\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right) - \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right), \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right) - \boldsymbol{\theta}_{11}^{o}\right)$

$$= Cov\left(\mathbf{K}\left(\hat{\boldsymbol{\theta}}_{12}^{step3} - \boldsymbol{\theta}_{12}^{o}\right), \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right) - \boldsymbol{\theta}_{11}^{o}\right)$$
because $\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right) - \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right) = \mathbf{K}\left(\hat{\boldsymbol{\theta}}_{12}^{step3} - \boldsymbol{\theta}_{12}^{o}\right)$ from (88)

$$= \mathbf{K}Cov\left(\boldsymbol{\Lambda}\hat{\boldsymbol{\theta}}_{12}^{step1} + (\mathbf{I}-\boldsymbol{\Lambda})\hat{\boldsymbol{\theta}}_{12}^{step2} - \boldsymbol{\theta}_{12}^{o}, \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right) - \boldsymbol{\theta}_{11}^{o}\right)$$
because $\hat{\boldsymbol{\theta}}_{12}^{step3} = \boldsymbol{\Lambda}\hat{\boldsymbol{\theta}}_{12}^{step1} + (\mathbf{I}-\boldsymbol{\Lambda})\hat{\boldsymbol{\theta}}_{12}^{step2}$

$$= \mathbf{K}\left(\boldsymbol{\Lambda}Cov\left(\hat{\boldsymbol{\theta}}_{12}^{step1}, \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right)\right) + (\mathbf{I}-\boldsymbol{\Lambda})Cov\left(\hat{\boldsymbol{\theta}}_{12}^{step2}, \hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right)\right)\right)$$

$$= \mathbf{K}\boldsymbol{\Lambda}\mathbf{U}$$
because $\hat{\boldsymbol{\theta}}_{12}^{step2}$ and the cross-section estimator $\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right)$ are uncorrelated. Thus, we have

$$Var\left(\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)\right) = \mathbf{K}Var\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)\mathbf{K}' + \frac{\mathbf{V}_{\boldsymbol{\theta}_{11}}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right)}{N} + \mathbf{K}\boldsymbol{\Lambda}\mathbf{U} + \mathbf{U}'\boldsymbol{\Lambda}'\mathbf{K}',$$

which is (53). We already have consistent estimators of $Var\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)$ and $\mathbf{V}_{\boldsymbol{\theta}_{11}}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right)$ using the results from the first two steps of the SR approach. We next derive consistent estimators of $\mathbf{K}$ and $\mathbf{U}$.

J.1 Estimation of K

We first note that $\hat{\boldsymbol{\theta}}_{11}^{step3}$ is determined from the first-order condition $\frac{1}{N}\sum_{i=1}^{N}\mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i; \hat{\boldsymbol{\theta}}_{11}^{step3}, \hat{\boldsymbol{\theta}}_{12}^{step3}, \hat{\boldsymbol{\gamma}}\right) = \mathbf{0}$, where $\mathbf{s}^{\boldsymbol{\theta}_{11}}$ is the score function for $\boldsymbol{\theta}_{11}$. A first-order approximation of this expression around $\boldsymbol{\theta}_{11}^{o}$, $\boldsymbol{\theta}_{12}^{o}$, and $\boldsymbol{\gamma}^{*}$ gives

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i; \boldsymbol{\theta}_{11}^{o}, \boldsymbol{\theta}_{12}^{o}, \boldsymbol{\gamma}^{*}\right) + \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i; \hat{\boldsymbol{\theta}}_{11}^{step3}, \hat{\boldsymbol{\theta}}_{12}^{step3}, \boldsymbol{\gamma}^{*}\right)}{\partial \boldsymbol{\theta}_{11}'}\left(\hat{\boldsymbol{\theta}}_{11}^{step3} - \boldsymbol{\theta}_{11}^{o}\right)$$
$$+\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i; \hat{\boldsymbol{\theta}}_{11}^{step3}, \hat{\boldsymbol{\theta}}_{12}^{step3}, \boldsymbol{\gamma}^{*}\right)}{\partial \boldsymbol{\theta}_{12}'}\left(\hat{\boldsymbol{\theta}}_{12}^{step3} - \boldsymbol{\theta}_{12}^{o}\right) + \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i; \hat{\boldsymbol{\theta}}_{11}^{step3}, \hat{\boldsymbol{\theta}}_{12}^{step3}, \boldsymbol{\gamma}^{*}\right)}{\partial \boldsymbol{\gamma}'}\left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^{*}\right) = \mathbf{0}.$$

For $N \longrightarrow \infty$, we have $\frac{1}{N}\sum_{i=1}^{N}\mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i; \boldsymbol{\theta}_{11}^{o}, \boldsymbol{\theta}_{12}^{o}, \hat{\boldsymbol{\gamma}}\right) \longrightarrow E\left[\mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i; \boldsymbol{\theta}_{11}^{o}, \boldsymbol{\theta}_{12}^{o}, \boldsymbol{\gamma}^{*}\right)\right] = \mathbf{0}$. It is also easy to verify that $\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i; \hat{\boldsymbol{\theta}}_{11}^{step3}, \hat{\boldsymbol{\theta}}_{12}^{step3}, \hat{\boldsymbol{\gamma}}\right)}{\partial \boldsymbol{\gamma}'} \xrightarrow{p} E\left[\frac{\partial \mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i; \boldsymbol{\theta}_{11}^{o}, \boldsymbol{\theta}_{12}^{o}, \boldsymbol{\gamma}^{*}\right)}{\partial \boldsymbol{\gamma}'}\right] = \mathbf{0}$. Let $\hat{\mathbf{A}}_{step3}^{\boldsymbol{\theta}_{11}} \equiv \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i; \hat{\boldsymbol{\theta}}_{11}^{step3}, \hat{\boldsymbol{\theta}}_{12}^{step3}, \hat{\boldsymbol{\gamma}}\right)}{\partial \boldsymbol{\theta}_{11}'}$ and $\hat{\mathbf{J}}_D \equiv \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i; \hat{\boldsymbol{\theta}}_{11}^{step3}, \hat{\boldsymbol{\theta}}_{12}^{step3}, \hat{\boldsymbol{\gamma}}\right)}{\partial \boldsymbol{\theta}_{12}'}$. We then have

$$\hat{\mathbf{A}}_{step3}^{\boldsymbol{\theta}_{11}}\left(\hat{\boldsymbol{\theta}}_{11}^{step3} - \boldsymbol{\theta}_{11}^{o}\right) + \hat{\mathbf{J}}_D\left(\hat{\boldsymbol{\theta}}_{12}^{step3} - \boldsymbol{\theta}_{12}^{o}\right) = \mathbf{0}$$

$\Updownarrow$

$$\left(\hat{\boldsymbol{\theta}}_{11}^{step3}-\boldsymbol{\theta}_{11}^{o}\right)=-\left[\hat{\mathbf{A}}_{step3}^{\boldsymbol{\theta}_{11}}\right]^{-1}\hat{\mathbf{J}}_{D}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}-\boldsymbol{\theta}_{12}^{o}\right).$$

Hence, $\mathbf{K}\equiv\frac{\partial\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)}{\partial\boldsymbol{\theta}_{12}'}=-\left[\hat{\mathbf{A}}_{step3}^{\boldsymbol{\theta}_{11}}\right]^{-1}\hat{\mathbf{J}}_{D}$. From the first step of the SR approach, we already have an estimator of $\hat{\mathbf{A}}_{step3}^{\boldsymbol{\theta}_{11}}$ using only first-order derivatives. To obtain an estimator of $\hat{\mathbf{J}}_{D}$ which only relies on first-order derivatives, consider the score function for $\boldsymbol{\theta}_{11}$, i.e. $\mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i,\boldsymbol{\theta}_1,\hat{\boldsymbol{\gamma}}\right)=-\tilde{v}_i\left(\mathbf{z}_i;\hat{\boldsymbol{\gamma}}\right)\boldsymbol{\Psi}_i^{\boldsymbol{\theta}_{11}}\left(\boldsymbol{\theta}_1,\boldsymbol{\gamma}\right)$. Standard arguments then imply

$$E\left[\frac{\partial\mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i,\boldsymbol{\theta}_1,\boldsymbol{\gamma}\right)}{\partial\boldsymbol{\theta}_{12}'}\right]=E\left[\boldsymbol{\Psi}_i^{\boldsymbol{\theta}_{11}}\left(\boldsymbol{\theta}_1,\boldsymbol{\gamma}\right)\boldsymbol{\Psi}_i^{\boldsymbol{\theta}_{12}}\left(\boldsymbol{\theta}_1,\boldsymbol{\gamma}\right)'\right],$$

and the stated estimator for $\hat{\mathbf{J}}$ in (55) follows directly.

J.2 Estimation of U

From the first step of the SR approach we have

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_1^{step1}-\boldsymbol{\theta}_1^{o}\right)=\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{H}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i;\dot{\boldsymbol{\theta}}_{11},\boldsymbol{\gamma}^{*}\right)\right)^{-1}\left(\frac{-1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{s}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i;\boldsymbol{\theta}_1^{o},\boldsymbol{\gamma}^{*}\right)\right).$$

Using (57) and $\mathbf{s}^{\boldsymbol{\theta}_1}\left(\mathbf{r}_i;\boldsymbol{\theta}_1^{o},\boldsymbol{\gamma}^{*}\right)\equiv\left[\begin{array}{cc}\mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i;\boldsymbol{\theta}_1^{o},\boldsymbol{\gamma}^{*}\right)' & \mathbf{s}^{\boldsymbol{\theta}_{12}}\left(\mathbf{r}_i;\boldsymbol{\theta}_1^{o},\boldsymbol{\gamma}^{*}\right)\end{array}\right]'$, we get in the first step

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{12}^{step1}-\boldsymbol{\theta}_{12}^{o}\right)=\left(\frac{-1}{\sqrt{N}}\sum_{i=1}^{N}\left[\hat{\boldsymbol{\Omega}}_{21}^{step1}\mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i;\hat{\boldsymbol{\theta}}_1^{step1},\hat{\boldsymbol{\gamma}}\right)+\hat{\boldsymbol{\Omega}}_{22}^{step1}\mathbf{s}^{\boldsymbol{\theta}_{12}}\left(\mathbf{r}_i;\hat{\boldsymbol{\theta}}_1^{step1},\hat{\boldsymbol{\gamma}}\right)\right]\right).$$

From the estimation of $\boldsymbol{\theta}_{11}$ in the third step of the SR approach we have

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\boldsymbol{\theta}_{12}^{o}\right)-\boldsymbol{\theta}_{11}^{o}\right)=\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{H}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i;\dot{\boldsymbol{\theta}}_{11},\boldsymbol{\theta}_{12}^{o},\boldsymbol{\gamma}^{*}\right)\right)^{-1}\left(\frac{-1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i;\boldsymbol{\theta}_{11}^{o},\boldsymbol{\theta}_{12}^{o},\boldsymbol{\gamma}^{*}\right)\right).$$

Recall that $\frac{1}{N}\sum_{i=1}^{N}\mathbf{H}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i;\hat{\boldsymbol{\theta}}_1^{step3},\hat{\boldsymbol{\gamma}}\right)=\hat{\mathbf{A}}_{step3}^{\boldsymbol{\theta}_{11}}$. Hence, we have

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{11}^{step3}\left(\hat{\boldsymbol{\theta}}_{12}^{step3}\right)-\boldsymbol{\theta}_{11}^{o}\right)=\frac{-1}{\sqrt{N}}\sum_{i=1}^{N}\left(\hat{\mathbf{A}}_{step3}^{\boldsymbol{\theta}_{11}}\right)^{-1}\mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i;\hat{\boldsymbol{\theta}}_1^{step3},\hat{\boldsymbol{\gamma}}\right).$$

Combining the two previous expressions, we obtain

$$\begin{aligned}\mathbf{U}=Cov(&\left(\tfrac{-1}{N}\sum_{i=1}^{N}\left[\hat{\boldsymbol{\Omega}}_{21}^{step1}\mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i;\hat{\boldsymbol{\theta}}_1^{step1},\hat{\boldsymbol{\gamma}}\right)+\hat{\boldsymbol{\Omega}}_{22}^{step1}\mathbf{s}^{\boldsymbol{\theta}_{12}}\left(\mathbf{r}_i;\hat{\boldsymbol{\theta}}_1^{step1},\hat{\boldsymbol{\gamma}}\right)\right]\right),\\&\tfrac{-1}{N}\sum_{i=1}^{N}\left(\hat{\mathbf{A}}_{step3}^{\boldsymbol{\theta}_{11}}\right)^{-1}\mathbf{s}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_i;\hat{\boldsymbol{\theta}}_1^{step3},\hat{\boldsymbol{\gamma}}\right))\end{aligned}$$

$$\begin{aligned}=&\tfrac{1}{N^2}\sum_{t_2=1}^{T}\sum_{j_2=1}^{n_{y,t}}\sum_{t_1=1}^{T}\sum_{j_1=1}^{n_{y,t}}\hat{\boldsymbol{\Omega}}_{21}^{step1}E\left(\mathbf{s}_{j_1}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_{t_1};\hat{\boldsymbol{\theta}}_1^{step1},\hat{\boldsymbol{\gamma}}\right)\mathbf{s}_{j_2}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_{t_2};\hat{\boldsymbol{\theta}}_1^{step3},\hat{\boldsymbol{\gamma}}\right)'\right)\left(\hat{\mathbf{A}}_{step3}^{\boldsymbol{\theta}_{11}}\right)^{-1}\\&+\tfrac{1}{N^2}\sum_{t_2=1}^{T}\sum_{j_2=1}^{n_{y,t}}\sum_{t_1=1}^{T}\sum_{j_1=1}^{n_{y,t}}\hat{\boldsymbol{\Omega}}_{22}^{step1}E\left(\mathbf{s}_{j_1}^{\boldsymbol{\theta}_{12}}\left(\mathbf{r}_{t_1};\hat{\boldsymbol{\theta}}_1^{step1},\hat{\boldsymbol{\gamma}}\right)\mathbf{s}_{j_2}^{\boldsymbol{\theta}_{11}}\left(\mathbf{r}_{t_2};\hat{\boldsymbol{\theta}}_1^{step3},\hat{\boldsymbol{\gamma}}\right)'\right)\left(\hat{\mathbf{A}}_{step3}^{\boldsymbol{\theta}_{11}}\right)^{-1}.\end{aligned}$$

The suggested estimator in (56) then follows immediately by standard arguments.

References

Adrian, T., Crump, R. K. & Moench, E. (2013), 'Pricing the term structure with linear regressions', *Journal of Financial Economics* **110**, 110–138.

Andersen, L. B. G., Jäckel, P. & Kahl, C. (2010), 'Simulation of square-root processes', *Encyclopedia of Quantitative Finance* pp. 1–10.

Andreasen, M. M. (2013), 'Non-linear DSGE models and the central difference kalman filter', *Journal of Applied Econometrics* **28**, 929–955.

Bauer, M. D., Rudebusch, G. D. & Wu, J. C. (2012), 'Correcting estimation bias in dynamic term structure models', *Journal of Business and Economic Statistics* **3**, 454–467.

Björk, T. & Christensen, B. J. (1999), 'Interest rate dynamics and consistent forward rate curves', *Mathematical Finance* **9**(4), 323–348.

Carrasco, M. & Florens, J.-P. (2002), 'Simulation-based method of moments and efficiency', *Journal of Business and Economic Statistics* **20**, 482–492.

Connor, G. & Korajczyk, R. A. (1986), 'Performance measurement with the arbitrage pricing theory', *Journal of Financial Economics* **15**, 373–394.

Cox, J. C., Ingersoll, J. E. & Ross, S. A. (1985), 'A theory of the term structure of interest rates', *Econometrica* **53**(2), 385–407.

Dai, Q. & Singleton, K. J. (2000), 'Specification analysis of affine term structure models', *Journal of Finance* **55**, 1946–1978.

de Jong, F. (2000), 'Time series and cross-section information in affine term-structure models', *Journal of Business and Economic Statistics* **18**, 300–314.

Diebold, F. X., Rudebusch, G. D. & Aruoba, S. B. (2006), 'The macroeconomy and the yield curve: A dynamic latent factor approach', *Journal of Econometrics* **131**, 309–338.

Duan, J.-C. & Simonato, J.-G. (1999), 'Estimating and testing exponential-affine term structure models by Kalman filter', *Review of Quantitative Finance and Accounting* **13**, 111–135.

Duffee, G. R. (2002), 'Term premia and interest rate forecasts in affine models', *The Journal of Finance* **57**(1), 405–443.

Duffie, D. & Singleton, K. J. (1993), 'Simulated moments estimation of markov models of asset prices', *Econometrica* **61**(4), 929–952.

Durbin, J. & Koopman, S. J. (2001), *Time Series Analysis by State Space Methods*, Oxford University Press, Oxford.

Fama, E. F. & Bliss, R. R. (1987), 'The information in long-maturity forward rates', *The American Economic Review* **77**(4), 680–692.

Fuller, W. A. (1987), *Measurement Error Models*, John Wiley and Sons, New York.

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

Hamilton, J. & Wu, J. C. (2014), 'Testable implications of affine-term-structure models', *Journal of Econometrics* **178**, 231–242.

Hansen, L. P. (1982), 'Large sample properties of generalized method of moments estimators', *Econometrica* **50**(4), 1029–1054.

Joslin, S., Singleton, K. J. & Zhu, H. (2011), 'A new perspective on gaussian dynamic term structure models', *The Review of Financial Studies* **24**, 926–970.

Jungbacker, B. & Koopman, S. J. (Forthcoming), 'Likelihood-based dynamic factor analysis for measurement and forecasting', *Econometrics Journal* **17**.

Kelejian, H. H. & Prucha, I. R. (2007), 'HAC estimation in a spatial framework', *Journal of Econometrics* **140**, 131–154.

Kim, D. H. & Singleton, K. J. (2012), 'Term structure models and the zero bound: An empirical investigation of japanese yields', *Journal of Econometrics* **170**, 32–49.

Kim, M. S. & Sun, Y. (2011), 'Spatial heteroskedasticity and autocorrelation consistent estimation of covariance matrix', *Journal of Econometrics* **160**, 349–371.

Linton, O., Mammen, E., Nielsen, J. P. & Tanggard, C. (2001), 'Yield curve estimation by kernel smoothing methods', *Journal of Econometrics* **105**, 185–223.

Nelson, C. R. & Siegel, A. F. (1987), 'Parsimonious modelling of yield curves', *The Journal of Business* **60**(4), 473–489.

Norgaard, M., Poulsen, N. K. & Ravn, O. (2000), 'Advances in derivative-free state estimation for nonlinear systems', *Automatica* **36:11**, 1627–1638.

Pan, J. (2002), 'The jump-risk premia implicit in options: Evidence from an integrated time-series study', *Journal of Financial Economics* **63**, 3–50.

Pastorello, S., Patilea, V. & Renault, E. (2003), 'Iterative and recursive estimation in structural nonadaptive models', *Journal of Business and Economic Statistics* **21**(4), 449–482.

Penzer, J. & Shea, B. (1997), 'The exact likelihood of an autoregressive-moving average model with incomplete data', *Biometrika* **84**, 919–928.

Pesaran, M. H. & Tosetti, E. (2011), 'Large panels with common factors and spatial correlation', *Journal of Econometrics* **161**, 182–202.

Realdon, M. (2006), 'Quadratic term structure models in discrete time', *Finance Research Letters* **3**, 277–289.

Sherman, R. P. (2003), 'Iterative and recursive estimation in structural nonadaptive models: Comment', *Journal of Business and Economic Statistics* **21**(4), 498–500.

Singleton, K. J. (2006), *Empirical Dynamic Asset Pricing*, Princeton University Press, Princeton.

Stock, J. H. & Watson, M. W. (2002), 'Forecasting using principal components from a large number of predictors', *Journal of the American Statistical Association* **97**(460), 1167–1179.

Tanggard, C. (1997), 'Nonparametric smoothing of yield curves', *Review of Quantitative Finance and Accounting* **9**, 251–267.

White, H. & Domowitz, I. (1984), 'Nonlinear regression with dependent observations', *Econometrica* **52**(1), 143–162.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, Cambridge.

Figure 1: ATSM: RMSEs for estimated latent factors

The RMSEs are scaled by 100 and calculated based on 500 repetitions of a sample of 480 observations. Black lines with stars refer to the Kalman smoother, and red lines with circles to the regression filter.

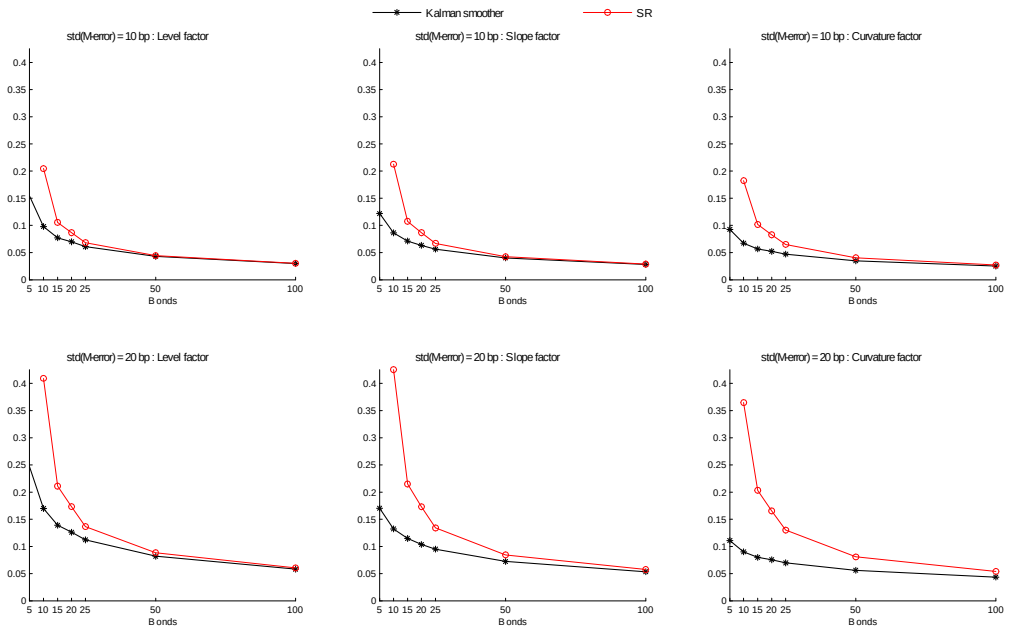

Figure 2: Factor estimation: Computing time

The first two charts in each row report the average number of seconds used to calculate the Kalman or CDKF smoother (the black line with stars) and the regression filter (the red line with circles). The charts in the final column display the ratio of the average number of seconds for the Kalman or CDKF smoother to the average number of seconds for the regression filter. The computations are carried out in MATLAB 2012a using an Intel(R) Core(TM) i5-2410 CPU with 2.30 GHz.

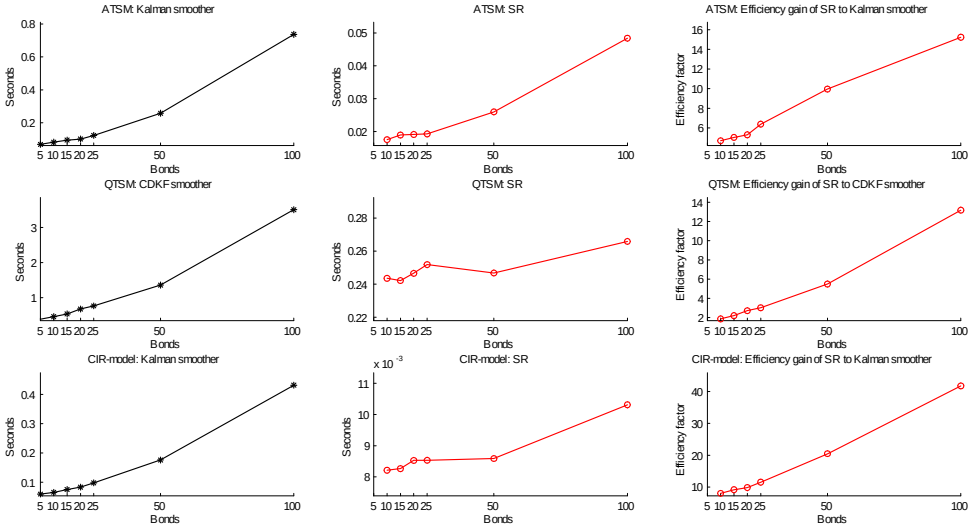

Figure 3: ATSM: RMSEs for estimated latent factors with cross-correlation
The RMSEs are scaled by 100 and calculated based on 500 repetitions of a sample of 480 observations. The cross-correlation in bonds is modelled as in Eq. (14) using $\rho_t = 0.90 \times R_v$ for all t. Black lines with stars refer to the Kalman smoother when accounting for the cross-correlation in bonds. Red lines with circles refer to the regression filter without accounting for the cross-correlation in bonds (OLS). Blue lines with squares refer to the regression filter when accounting for the cross-correlation in bonds (GLS).

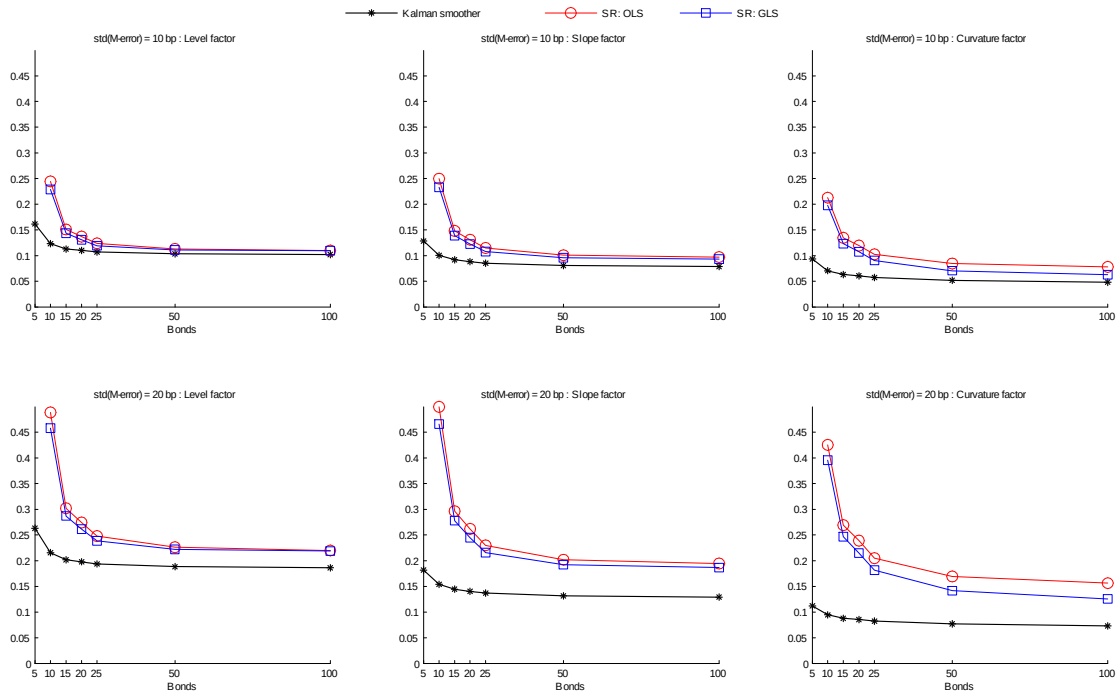

Figure 4: ATSM: mean value of estimated parameters

These results are calculated based on 1,000 repetitions of a sample of 480 observations and with measurement errors having a standard deviation of 10 basis points along the yield curve. Black lines with stars refer to the ML estimates, red lines with circles to the estimates from the three-step SR approach, and black pluses to the infeasible ML estimates where the factors are observed, or equivalently, when there is an infinite number of observables each time period.

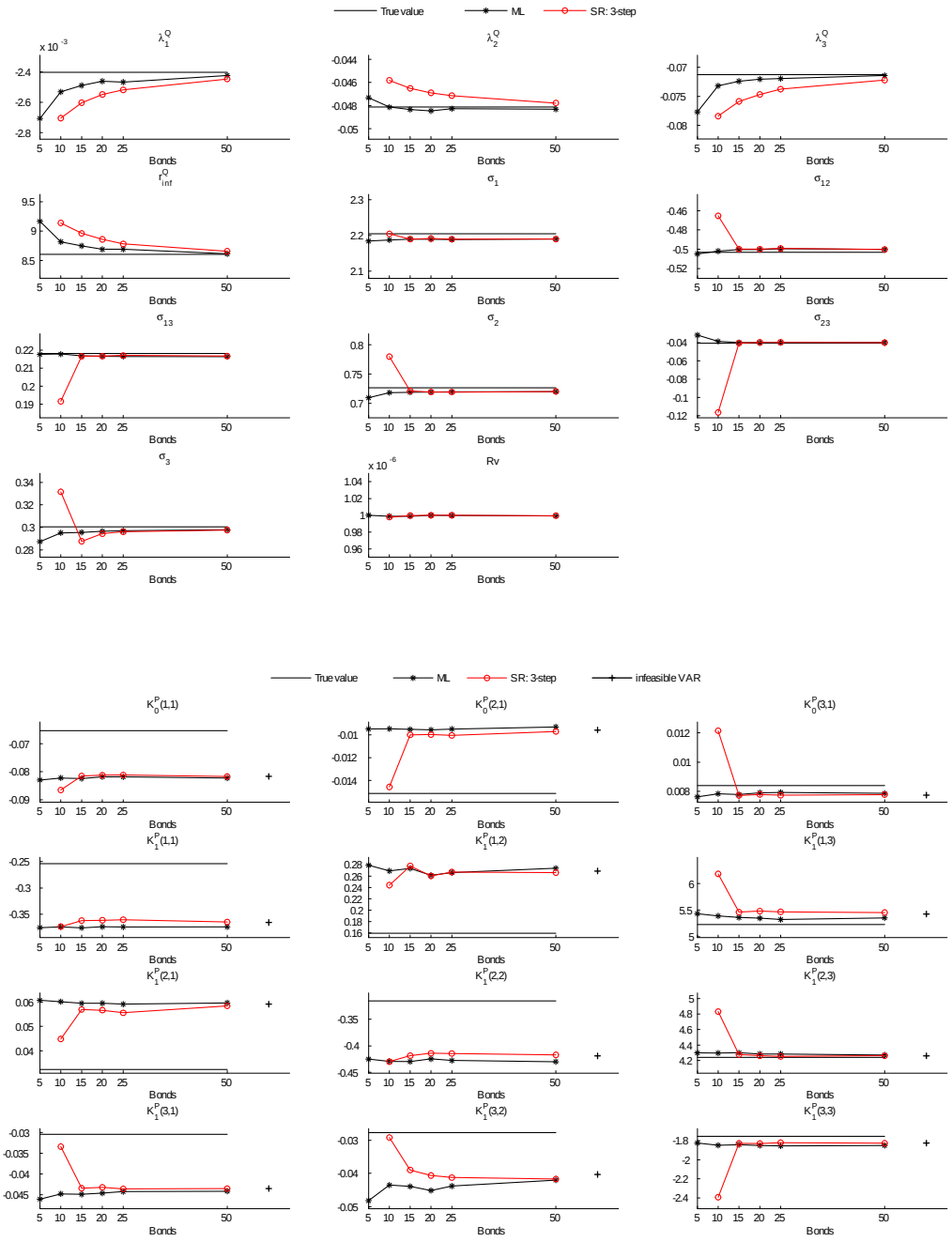

Figure 5: Parameter estimation: Computing time

These results are calculated based on 1,000 repetitions of a sample of 480 observations and with measurement errors having a standard deviation of 10 basis points along the yield curve. Black lines with stars refer to the ML or QML estimates computed using the Nelder-Mead optimizer, and red lines with circles to estimates from the three-step SR approach computing using a gradient-based optimizer. The computations are carried out in MATLAB 2012a using an Intel(R) Core(TM) i5-2410 CPU with 2.30 GHz.

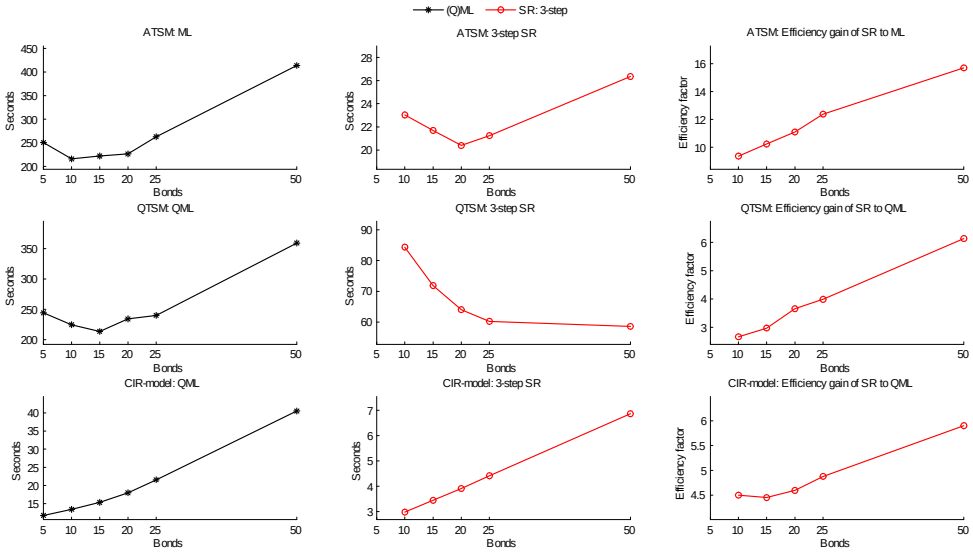

Figure 6: ATSM: Efficiency of parameter estimates

These results are calculated based on 1,000 repetitions of a sample of 480 observations and with measurement errors having a standard deviation of 10 basis points along the yield curve. Black lines with stars refer to the ML estimates, red lines with circles to the estimates from the three-step SR approach, and black pluses to the infeasible ML estimates where the factors are observed, or equivalently, when there is an infinite number of observables each time period.

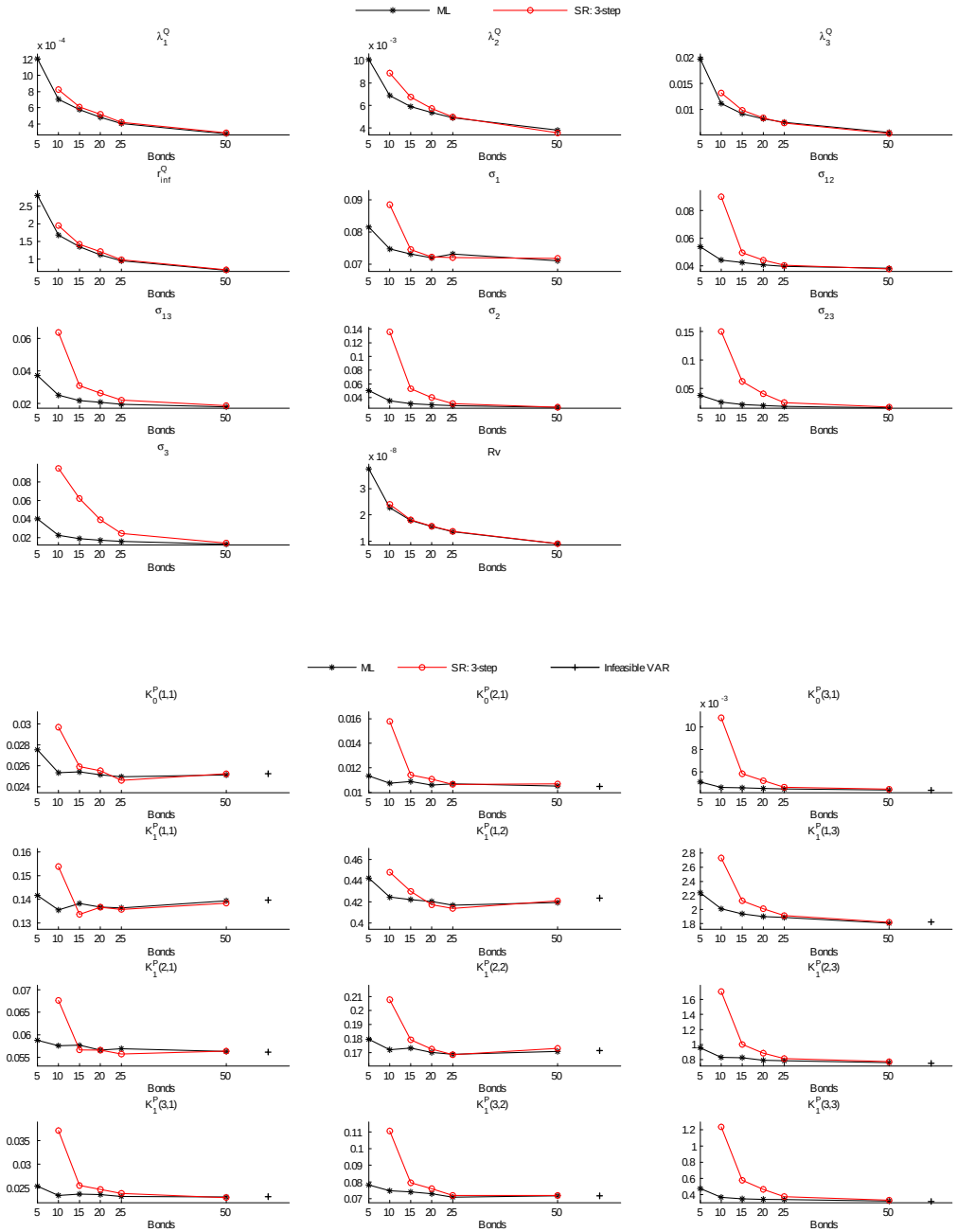

Figure 7: ATSM: Efficiency in the first and third step of the SR approach

These results are calculated based on 1,000 repetitions of a sample of 480 observations and with measurement errors having a standard deviation of 10 basis points along the yield curve. Black lines with stars refer to the ML estimates, red lines with circles to estimates from the third step in the SR approach, and blue lines with squares to estimates from the first step in the SR approach.

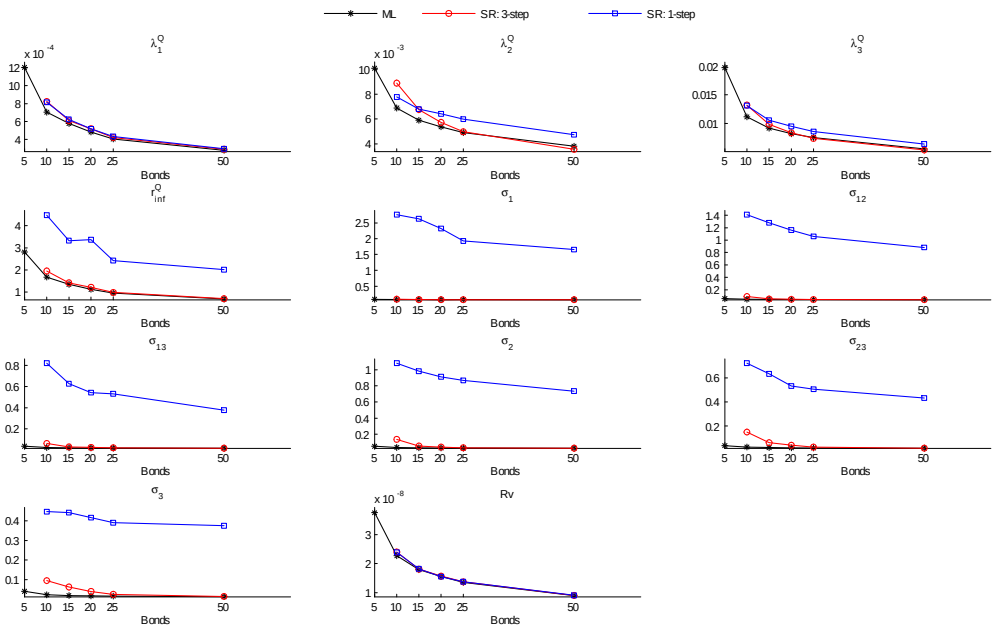

Figure 8: ATSM: Bias in estimated standard errors

These results are calculated based on 1,000 repetitions of a sample of 480 observations and with measurement errors having a standard deviation of 10 basis points along the yield curve. Black lines with stars refer to the ML estimates, red lines with circles to the estimates from the three-step SR approach for the homoskedastic case, and black pluses to the infeasible ML estimates where the factors are observed, or equivalently, when there is an infinite number of observables each time period.

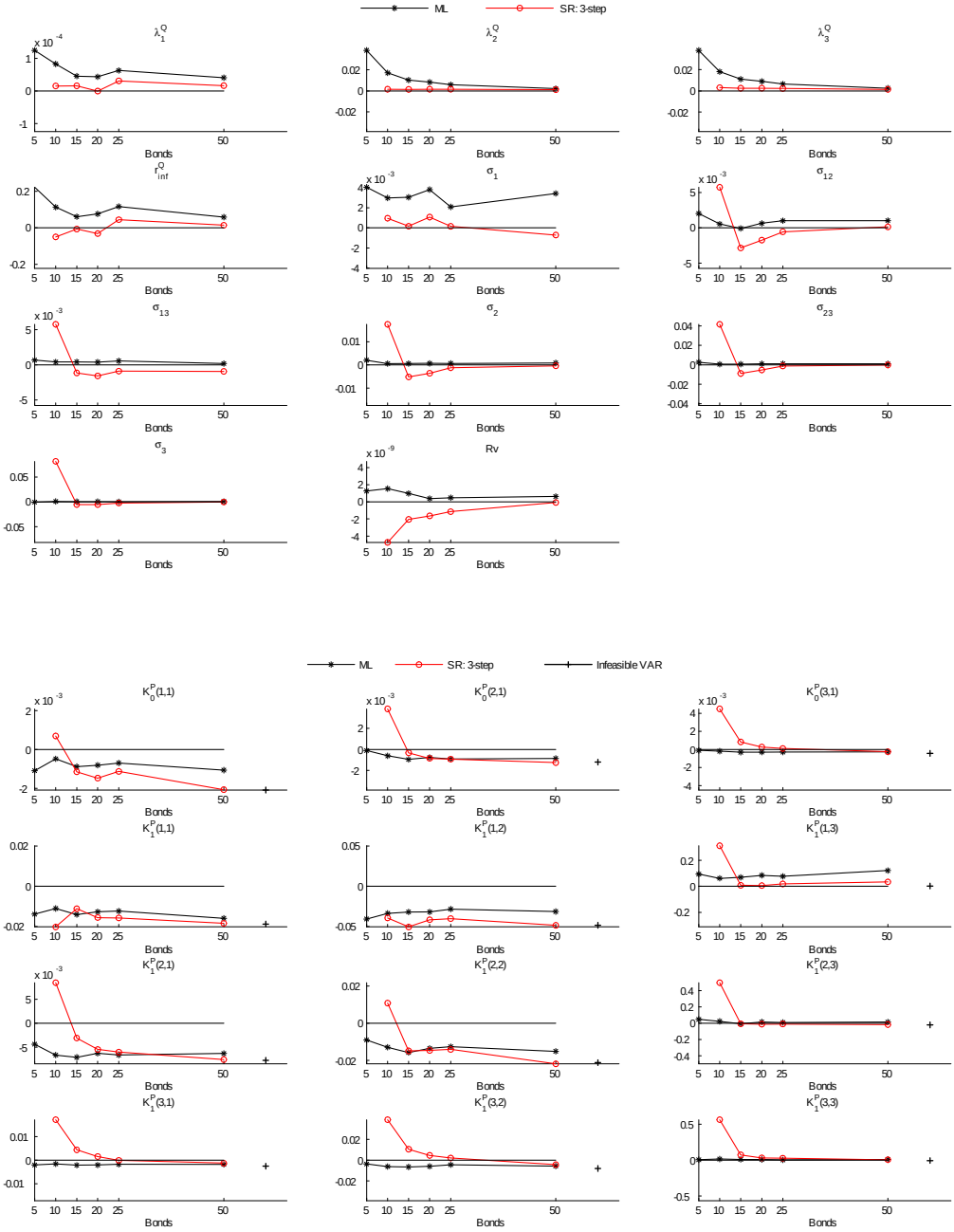

Figure 9: ATSM: Type I errors at the 5 percent level

These results are calculated based on 1,000 repetitions of a sample of 480 observations and with measurement errors having a standard deviation of 10 basis points along the yield curve. Black lines with stars refer to the ML estimates, red lines with circles to the estimates from the three-step SR approach for the homoskedastic case, and black pluses to the infeasible ML estimates where the factors are observed, or equivalently, when there is an infinite number of observables each time period.

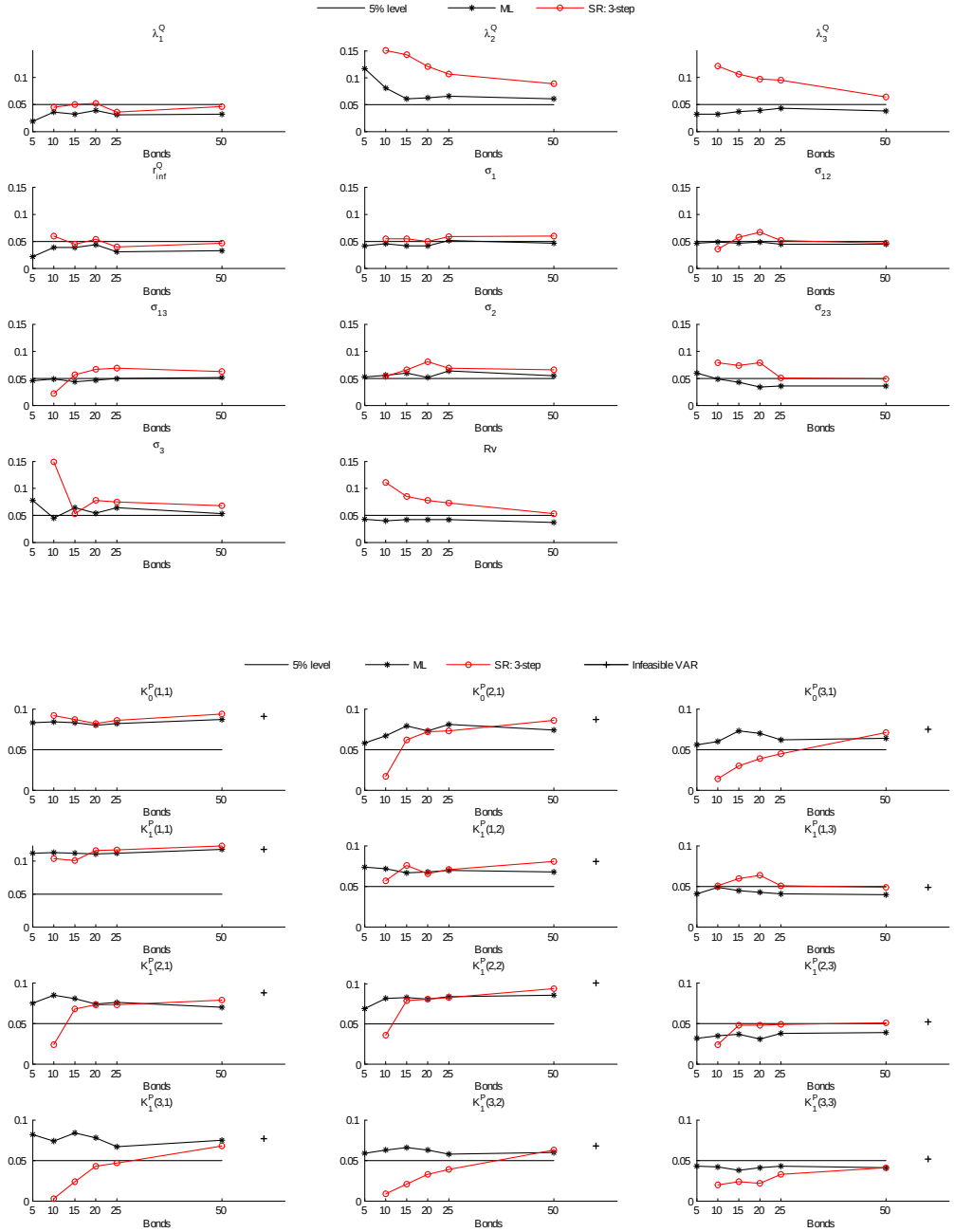

Figure 10: QTSM: RMSEs for estimated latent factors

The RMSEs are scaled by 100 and calculated based on 500 repetitions of a sample of 480 observations. Black lines with stars refer to the CDKF smoother, and red lines with circles to the regression filter.

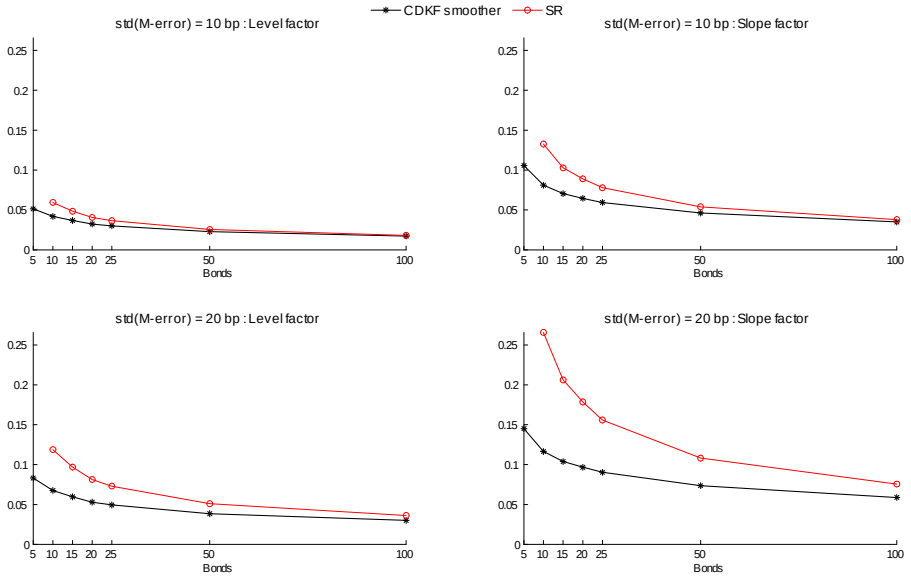

Figure 11: QTSM: mean value of parameter estimates

These results are calculated based on 1,000 repetitions of a sample of 480 observations and with measurement errors having a standard deviation of 10 basis points along the yield curve. Black lines with stars refer to the QML estimates, red lines with circles to the estimates from the three-step SR approach, and black pluses to the infeasible ML estimates where the factors are observed, or equivalently, when there is an infinite number of observables each time period.

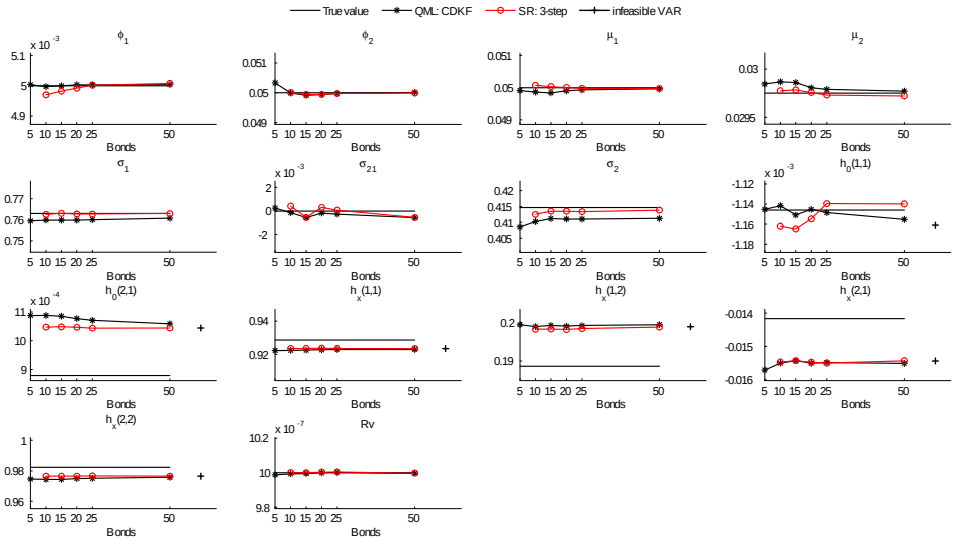

Figure 12: QTSM: Efficiency of parameter estimates

These results are calculated based on 1,000 repetitions of a sample of 480 observations and with measurement errors having a standard deviation of 10 basis points along the yield curve. Black lines with stars refer to the QML estimates, red lines with circles to the estimates from the three-step SR approach, and black pluses to the infeasible ML estimates where the factors are observed, or equivalently, when there is an infinite number of observables each time period.

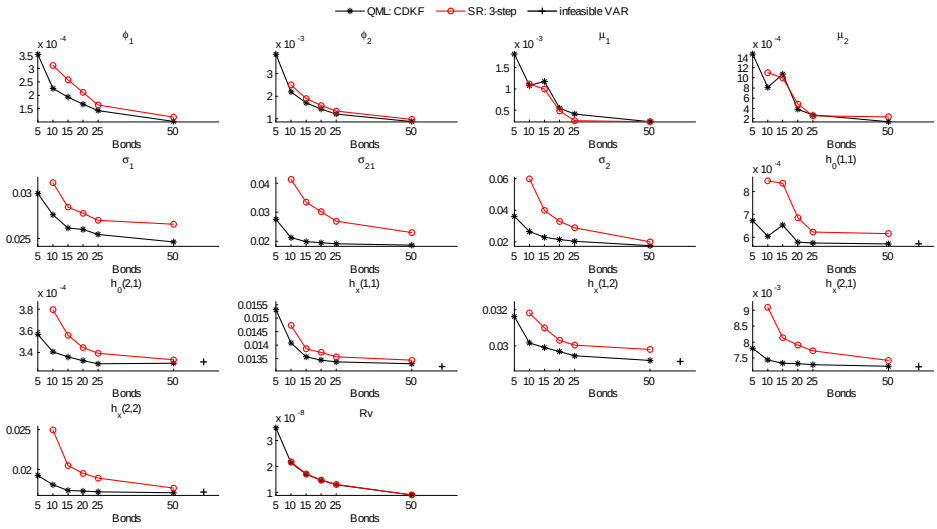

Figure 13: QTSM: Bias in estimated standard errors

These results are calculated based on 1,000 repetitions of a sample of 480 observations and with measurement errors having a standard deviation of 10 basis points along the yield curve. Black lines with stars refer to the QML estimates, red lines with circles to the estimates from the three-step SR approach for the homoskedastic case, and black pluses to the infeasible ML estimates where the factors are observed, or equivalently, when there is an infinite number of observables each time period.

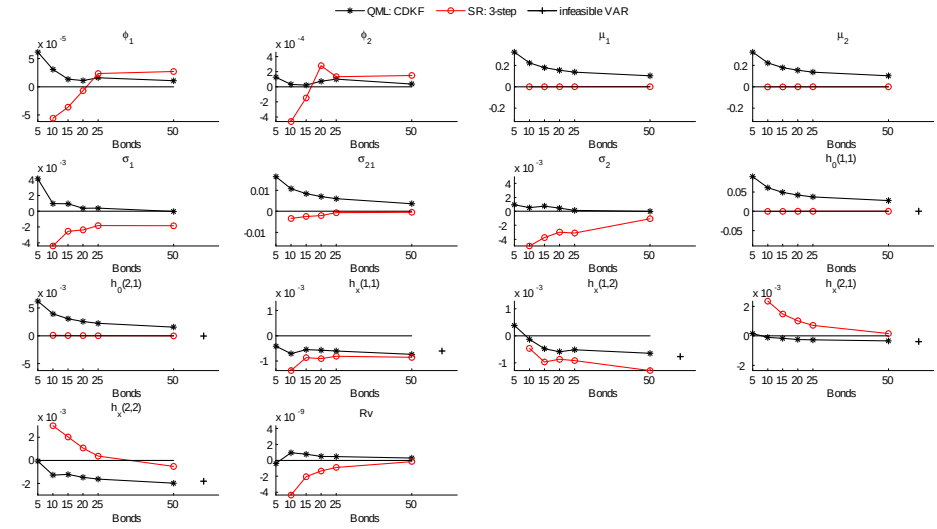

Figure 14: QTSM: Type I errors at the 5 percent level

These results are calculated based on 1,000 repetitions of a sample of 480 observations and with measurement errors having a standard deviation of 10 basis points along the yield curve. Black lines with stars refer to the QML estimates, red lines with circles to the estimates from the three-step SR approach for the homoskedastic case, and black pluses to the infeasible ML estimates where the factors are observed, or equivalently, when there is an infinite number of observables each time period.

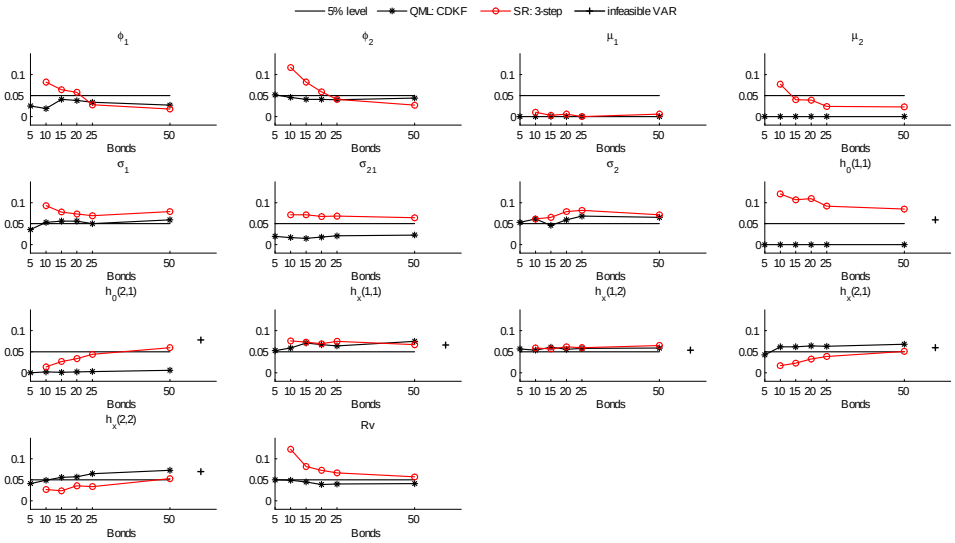

Figure 15: CIR model: RMSEs for estimated latent factors

The RMSEs are scaled by 100 and calculated based on 500 repetitions of a sample of 480 observations. Black lines with stars refer to the Kalman smoother, and red lines with circles to the regression filter.

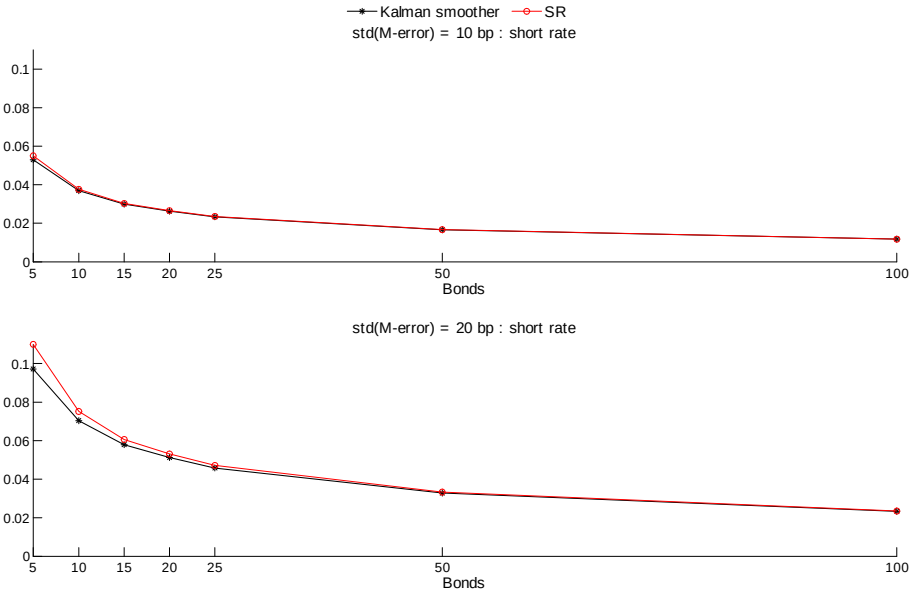

Figure 16: CIR model: Results for the estimated parameters

These results are calculated based on 1,000 repetitions of a sample of 480 observations and with measurement errors having a standard deviation of 10 basis points along the yield curve. Black lines with stars refer to the QML estimates, red lines with circles to the estimates from the three-step SR approach for the homoskedastic case, and blue lines with squares to the first-step SR approach for the homoskedastic case.