BANK OF CANADA
BANQUE DU CANADA

Working Paper/Document de travail
2013-10

A New Linear Estimator for Gaussian Dynamic Term Structure Models

by Antonio Diez de los Rios

www.bank-banque-canada.ca

Bank of Canada Working Paper 2013-10

April 2013

A New Linear Estimator for Gaussian Dynamic Term Structure Models

by

Antonio Diez de los Rios

Financial Markets Department
Bank of Canada
Ottawa, Ontario, Canada K1A 0G9
diez@bankofcanada.ca

ISSN 1701-9397

Acknowledgements

I would like to thank Narayan Bulusu, Bruno Feunou, Sermin Gungor, Scott Hendry, Enrique Sentana, Jon Witmer and seminar participants at the Bank of Canada, CEMFI, and European Central Bank for their suggestions. Any remaining errors are my own.

Abstract

This paper proposes a novel regression-based approach to the estimation of Gaussian dynamic term structure models that avoids numerical optimization. This new estimator is an asymptotic least squares estimator defined by the no-arbitrage conditions upon which these models are built. We discuss some efficiency considerations of this estimator, and show that it is asymptotically equivalent to maximum likelihood estimation. Further, we note that our estimator remains easy-to-compute and asymptotically efficient in a variety of situations in which other recently proposed approaches lose their tractability. We provide an empirical application in the context of the Canadian bond market.

JEL classification: E43, C13, G12
Bank classification: Asset pricing; Econometric and statistical methods; Interest rates

Résumé

Un cadre de régression novateur permettant d'éviter l'optimisation numérique est proposé pour l'estimation de modèles dynamiques gaussiens de la structure par terme des taux d'intérêt. Ce nouvel estimateur est un estimateur des moindres carrés asymptotiques et est défini par les conditions d'absence d'arbitrage à la base de ces modèles. L'auteur analyse les caractéristiques d'efficience de son estimateur et montre que celui-ci est asymptotiquement équivalent à un estimateur du maximum de vraisemblance. De plus, il reste simple à calculer et asymptotiquement efficient dans un éventail de situations où d'autres approches récentes deviennent très difficiles à utiliser. L'auteur présente une application empirique de son cadre au cas du marché obligataire canadien.

Classification JEL : E43, C13, G12
Classification de la Banque : Évaluation des actifs; Méthodes économétriques et statistiques; Taux d'intérêt

1 Introduction

The maximum likelihood (ML) approach is considered as the most natural way to estimate Gaussian dynamic term structure models (GDTSMs), since they provide a complete characterization of the joint distribution of yields.[1] However, the solution of the optimization problem involving maximization of the density of the yields does not exist in closed form, except in very few specific cases. Consequently, researchers often have to rely on cumbersome optimization techniques to estimate the parameters of the model, facing diverse numerical issues that are usually magnified by (i) the large number of parameters describing the dynamics of the term structure of interest rates, (ii) the highly non-linear nature of the likelihood function, and/or (iii) the existence of multiple local optima (see, for example, the discussions in Duffee and Stanton, 2012; Hamilton and Wu, 2012).

Motivated by these numerical challenges, this paper considers a new linear regression approach to the estimation of GDTSMs that completely avoids numerical optimization methods. Specifically, our linear estimator is an asymptotic least squares (ALS) estimator that exploits three features that characterize this class of models. First, GDTSMs have a reduced-form representation whose parameters can be easily estimated via a set of ordinary least squares (OLS) regressions. Second, the no-arbitrage assumption upon which GDTSMs are built can be characterized as a set of implicit constraints between these reduced-form parameters and the parameters of interest. Third, this set of restrictions is linear in the parameters of interest. Consequently, we propose a two-step estimator. In the first step, estimates of the reduced-form parameters are obtained by OLS. In the second step, the parameters of the GDTSMs are inferred by forcing the no-arbitrage constraints, evaluated at the first-stage estimates of the reduced-form parameters, to be as close as possible to zero in the metric defined by a given weighting matrix. Note that, since the constraints are linear in the parameters of interest, the solution to the estimation problem in this second step is known in closed form. In fact, in its most basic form (i.e., using an identity weighting matrix), the estimates of the parameters of the GDTSMs resemble those obtained from an OLS cross-sectional regression involving the reduced-form parameter estimates (i.e., the estimated bond factor loadings). Moreover, our ALS estimator is consistent and asymptotically normally distributed.

As in the case of generalized method of moments (GMM) estimation, efficiency gains can be achieved by selecting an appropriate weighting matrix. As noted by Gourieroux, Monfort and Trognon (1982, 1985) (GMT hereafter), the optimal weighting matrix is equal to the inverse of the asymptotic covariance matrix of the set of implicit constraints

[1] See, for example, Chen and Scott (1993), Dai and Singleton (2002), Kim and Wright (2005), Kim and Orphanides (2005), Aït-Sahalia and Kimmel (2010), Christensen, Diebold and Rudebusch (2011) and Joslin, Singleton and Zhu (2011) for some term structure models estimated by ML.

between reduced-form and the parameters of interest. However, we show that such a matrix is singular in the context of GDTSM estimation and therefore the definition of an optimal ALS estimator in GMT breaks down. For this reason, we borrow from Peñaranda and Sentana (2012), who study the problem of obtaining an optimal GMM estimator when the asymptotic variance of the moment conditions is singular in the population, to extend the theory of optimal ALS estimation to cover the singular set-up.

We also discuss several extensions of our estimation method. First, we show how to estimate GDTSMs subject to certain equality constraints on the structural parameters. This includes the important case of exclusion (zero) restrictions on the parameters driving the prices of risk (see, e.g., Cochrane and Piazzesi, 2008), but also the case of more complicated non-linear restrictions. Second, we discuss how to estimate GDTSMs where some of the factors are unspanned, as in Joslin, Priebsch and Singleton (2012). Such factors are not related to the contemporaneous cross-section of interest rates, but they help forecast future excess returns on the bonds. Third, we show how to accommodate for higher-order dynamics in the parameterization of the distribution of yields under the physical measure (i.e., a VAR(p) model with $p > 1$), while preserving the parsimonious factor representation of yields, as in Joslin, Le and Singleton (2013). Fourth, we show how our framework can be adapted to handle autocorrelation in the measurement errors and/or overlapping in the dynamics under the physical measure. Fifth, we discuss how the choice of the bonds to be used in the estimation of GDTSMs has consequences on the properties of the GDTSM estimators. Sixth, we show how to compute small-sample standard errors using bootstrap methods and how to address some of the biases associated with the extreme persistence found in interest rates (see, e.g., Bauer, Rudebusch and Wu, 2012).

Recent approaches to the estimation of GDTSMs that have substantially lessened some of the numerical challenges faced by researchers include the maximum likelihood approach of Joslin, Singleton and Zhu (2011), the minimum-chi-square estimator of Hamilton and Wu (2012), and the regression-based approach of Adrian, Crump and Moench (2012). In particular, we show that the optimal ALS estimator of the parameters of a GDTSM is asymptotically equivalent to the ML estimator of Joslin, Singleton and Zhu (2011). We also show that both the Hamilton and Wu (2012) and Adrian, Crump and Moench (2012) estimators are asymptotic least squares estimators, where these estimators differ from ours either in the weighting matrix employed, the parameterization of the no-arbitrage conditions, the reduced-form model estimates, and/or the existence of restrictions on the parameters of interest. This unified framework allows us to conclude that our ALS estimator remains tractable and asymptotically efficient in a variety of situations in which the other approaches lose their tractability. Along these lines, we provide a Monte Carlo

study to confirm that the tractability of the ALS estimator does not come at the expense of efficiency losses or bad finite-sample properties.

For illustrative purposes, we estimate a three-factor model and decompose the Canadian ten-year zero-coupon bond yield into an expectations and term premium component. Our three-factor specification is designed to capture all the economically interesting variation in both the cross-section of interest rates and bond risk premia, and resembles the Cochrane and Piazzesi (2008) model of the U.S. yield curve. Specifically, we identify our first two factors with the first two principal components of the Canadian yield curve, while the third one is a return-forecasting factor similar in spirit to that presented in Cochrane and Piazzesi (2005). Moreover, the model is estimated subject to a variety of non-linear restrictions on the parameters of the model which, absent our proposed regression-based framework, would greatly complicate both the estimation and inference. In fact, we exploit the numerical tractability of our estimation method to compute bootstrap p-values that correct for the generated regressor problem inherent in the estimation of our model.

The structure of the paper is as follows. In section 2, we briefly describe the class of GDTSMs, introduce our new linear estimator and reinterpret this estimator within the ALS framework. In section 3, we briefly review the asymptotic properties of ALS estimators, and discuss how to conduct optimal ALS estimation in the singular set-up that characterizes GDTSMs. We discuss the advantages of our method with respect to recently suggested approaches to the estimation of GDTSMs in section 4. In section 5, we present a Monte Carlo exercise designed to assess the finite-sample properties of our new linear estimator. Section 6 discusses several extensions of our regression-based framework, and Section 7 contains our empirical results. Finally, we provide some concluding remarks and future lines of research in Section 8. Auxiliary results are gathered in the appendix.

2 Gaussian affine term structure models

2.1 General framework

We start by considering a $(M \times 1)$ vector of state variables (or pricing factors), $\mathbf{f}_t$, that describes the state of the economy. For the moment, we remain "agnostic" as to the nature of these pricing factors. The dynamic evolution of the state variables under the physical, or historical measure, $\mathbb{P}$, is given by a Gaussian VAR(1) process:

$$\mathbf{f}_{t+1} = \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{f}_t + \mathbf{v}_{t+1}, \tag{1}$$

where $\mathbf{v}_{t+1} \sim iid\ N(\mathbf{0}, \boldsymbol{\Sigma})$.

Let r_t be the continuously compounded one-period, or short-term, interest rate. The

short rate is related to the set of state variables through the following affine relation:

$$r_t = \delta_0 + \boldsymbol{\delta}_1' \mathbf{f}_t. \tag{2}$$

The model is completed by specifying the stochastic discount factor (SDF) to be exponentially affine in $\mathbf{f}_t$ (e.g., Ang and Piazzesi, 2003):

$$\xi_{t+1} = \exp\left(-r_t - \frac{1}{2}\boldsymbol{\lambda}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_t'\boldsymbol{\Sigma}^{-1}\mathbf{v}_{t+1}\right), \tag{3}$$

with prices of risk given by $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}_0 + \boldsymbol{\lambda}_1 \mathbf{f}_t$. This (strictly positive) SDF, ξ_{t+1}, can be used to price zero-coupon bonds using the following recursive relation:

$$P_{t,n} = E_t\left[\xi_{t+1} P_{t+1,n-1}\right], \tag{4}$$

where $P_{t,n}$ is the price of a zero-coupon bond of maturity n periods at time t. In particular, it is possible to show that solving equation (4) is equivalent to solving the following equation:

$$P_{t,n} = E_t^{\mathbb{Q}}\left[\exp\left(-\sum_{i=0}^{n-1} r_{t+i}\right)\right],$$

where $E_t^{\mathbb{Q}}$ denotes the expectation under the risk-neutral probability measure, $\mathbb{Q}$. Under the risk-neutral probability measure, the dynamics of the state vector $\mathbf{f}_t$ are characterized by the following VAR(1) process:

$$\mathbf{f}_{t+1} = \boldsymbol{\mu}^{\mathbb{Q}} + \boldsymbol{\Phi}^{\mathbb{Q}}\mathbf{f}_t + \mathbf{v}_{t+1}^{\mathbb{Q}}, \tag{5}$$

with $\boldsymbol{\mu}^{\mathbb{Q}} = \boldsymbol{\mu} - \boldsymbol{\lambda}_0$, $\boldsymbol{\Phi}^{\mathbb{Q}} = \boldsymbol{\Phi} - \boldsymbol{\lambda}_1$ and $\mathbf{v}_{t+1}^{\mathbb{Q}} \sim iid\ N(0, \boldsymbol{\Sigma})$.

Solving (4), we find that the continuously compounded yield on an n-period zero-coupon bond at time t, $y_{t,n} = -\frac{1}{n}\log P_{t,n}$, is given by

$$y_{t,n} = a_n + \mathbf{b}_n'\mathbf{f}_t, \tag{6}$$

where $a_n = -A_n/n$ and $\mathbf{b}_n = -\mathbf{B}_n/n$, and A_n and $\mathbf{B}_n$ satisfy the following set of recursive relations:

$$\mathbf{B}_n' = \mathbf{B}_{n-1}'\boldsymbol{\Phi}^{\mathbb{Q}} + \mathbf{B}_1', \tag{7}$$

$$A_n = A_{n-1} + \mathbf{B}_{n-1}'\boldsymbol{\mu}^{\mathbb{Q}} + \frac{1}{2}\mathbf{B}_{n-1}'\boldsymbol{\Sigma}\mathbf{B}_{n-1} + A_1, \tag{8}$$

for $n = 2, ..., N$ where is N is the largest maturity under consideration.

The recursion is started by exploiting the fact that the affine pricing relationship is trivially satisfied for one-period bonds ($n = 1$), which implies that $A_1 = -\delta_0$, and

$\mathbf{B}_1 = -\boldsymbol{\delta}_1$. Specifically, solving equations (7) and (8) forward, we obtain:

$$\mathbf{b}_n' = \frac{1}{n}\boldsymbol{\delta}_1' \sum_{j=0}^{n-1} \left(\boldsymbol{\Phi}^{\mathbb{Q}}\right)^j, \tag{9}$$

$$a_n = \delta_0 + \frac{1}{n}\sum_{j=1}^{n-1} j\mathbf{b}_j'\boldsymbol{\mu}^{\mathbb{Q}} - \frac{1}{2n}\sum_{j=1}^{n-1} j^2\mathbf{b}_j'\boldsymbol{\Sigma}\mathbf{b}_j. \tag{10}$$

2.2 A new linear estimator for GDTSMs

An important characteristic of the Gaussian affine bond pricing model above is that the pricing recursive relations in (7) and (8) are linear in $\boldsymbol{\Phi}^{\mathbb{Q}}$ and $\boldsymbol{\mu}^{\mathbb{Q}}$. Therefore, if the innovation covariance matrix $\boldsymbol{\Sigma}$ and the set of coefficients A_n and $\mathbf{B}_n$ were observed directly, one could easily estimate the risk-neutral parameters of the model using a set of (cross-sectional) OLS regressions. In such a case, the linear structure of the model would allow us to recover an estimate of $\boldsymbol{\Phi}^{\mathbb{Q}}$ from the (cross-sectional) OLS regression of $\left(\mathbf{B}_{n+1}' - \mathbf{B}_1'\right)$ on $\mathbf{B}_n'$:

$$\widehat{\boldsymbol{\Phi}}^{\mathbb{Q}} = \left(\sum_{n=1}^{N}\mathbf{B}_n\mathbf{B}_n'\right)^{-1}\left[\sum_{n=1}^{N}\mathbf{B}_n\left(\mathbf{B}_{n+1}' - \mathbf{B}_1'\right)\right], \tag{11}$$

while an estimate of $\boldsymbol{\mu}^{\mathbb{Q}}$ can be obtained from the regression of $(A_{n+1} - A_n - \frac{1}{2}\mathbf{B}_n'\boldsymbol{\Sigma}\mathbf{B}_n - A_1)$ on $\mathbf{B}_n'$:

$$\widehat{\boldsymbol{\mu}}^{\mathbb{Q}} = \left(\sum_{n=1}^{N}\mathbf{B}_n\mathbf{B}_n'\right)^{-1}\left[\sum_{n=1}^{N}\mathbf{B}_n(A_{n+1} - A_n - \frac{1}{2}\mathbf{B}_n'\boldsymbol{\Sigma}\mathbf{B}_n - A_1)\right]. \tag{12}$$

However, this linear estimator is infeasible because the innovation covariance matrix $\boldsymbol{\Sigma}$, and the set of coefficients A_n and $\mathbf{B}_n$ are, in practice, unknown. Nevertheless, consistent estimates of these objects are readily available from a reduced-form representation of the model (see Hamilton and Wu, 2012). We propose, instead, to replace the unknown objects in equations (11) and (12) above by consistent estimates obtained from the reduced-form representation of the model.

To obtain the reduced-form representation of the GDTSM, it is convenient to resort to the state-space representation of the observed variables (i.e., the bond yields) implied by the model. In a general state-space representation, there is a transition equation that describes the dynamic evolution of the state factors over time, and a measurement equation that relates the observed data to the state factor. In terms of our asset pricing model, the VAR dynamics in (1) can be interpreted as the transition equation, while the pricing relationship in (6) is the measurement equation. Let $y_{t,n}^o$ denote the observed yields, which we assume are subject to measurement error. Let $\mathbf{y}_t = [y_{t,1}, y_{t,2}, ..., y_{t,N}]'$ be the vector of model-implied yields that stack the affine mapping (6), and let $\mathbf{y}_t^o$ be the

equivalent vector of observed yields. Let $\boldsymbol{\eta}_t$ be a zero-mean measurement error that is *i.i.d.* across time and that has a covariance matrix $\boldsymbol{\Omega}$. Then, the asset pricing model, joint with our assumption on the measurement errors, implies that the vector $\mathbf{y}_t^o$ has the following state-space representation:

$$\mathbf{y}_t^o = \mathbf{a} + \mathbf{b}\mathbf{f}_t + \boldsymbol{\eta}_t, \tag{13}$$

$$\mathbf{f}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{f}_{t-1} + \mathbf{v}_t, \tag{14}$$

where the corresponding elements of $\mathbf{a}$ and $\mathbf{b}$ satisfy equations (9) and (10). In particular, note that $\mathbf{a} = \mathbf{a}(\boldsymbol{\mu}^{\mathbb{Q}}, \boldsymbol{\Phi}^{\mathbb{Q}}, \boldsymbol{\Sigma})$ and $\mathbf{b} = \mathbf{b}(\boldsymbol{\Phi}^{\mathbb{Q}})$ are non-linear functions of $\boldsymbol{\mu}^{\mathbb{Q}}$, $\boldsymbol{\Phi}^{\mathbb{Q}}$, and $\boldsymbol{\Sigma}$. For completeness, we assume that $E(\boldsymbol{\eta}_t \mathbf{v}_s) = \mathbf{0}$ for all t and s.

Once again, estimation of the reduced-form parameters in equations (13) and (14) could be greatly simplified if the bond factors, $\mathbf{f}_t$, were observed. Specifically, since the errors of the model are conditionally homoskedastic, the maximum likelihood estimates of the reduced-form parameters could be trivially obtained via a set of OLS regressions (see Sentana, 2002, and Hamilton and Wu, 2012): (i) the (cross-sectional) coefficients $\mathbf{a}$ and $\mathbf{b}$ could be estimated from the OLS regression of $\mathbf{y}_t^o$ on a constant and $\mathbf{f}_t$; (ii) the (time-series) coefficients $\boldsymbol{\mu}$ and $\boldsymbol{\Phi}$ could be estimated from the OLS regression of $\mathbf{f}_t$ on a constant and its lag.

In order to overcome this issue, we follow Joslin, Singleton and Zhu (2011) in working with bond state variables that are linear combinations (i.e., portfolios) of the yields themselves, $\mathbf{f}_t = \mathbf{P}'\mathbf{y}_t^o$, where $\mathbf{P}$ is a full-rank matrix of weights, and by further assuming that $\mathbf{f}_t$ is observed perfectly.[2] That is, $\mathbf{P}'(\mathbf{y}_t^o - \mathbf{y}_t) = \mathbf{P}'\boldsymbol{\eta}_t = 0 \ \forall t$.[3] This assumption allows us to factorize the joint likelihood function into the marginal component of $\mathbf{f}_t$ and the conditional components corresponding to all the individual yields. That is, this assumption makes $\mathbf{f}_t$ observable in practice and, as noted above, allows us to obtain maximum likelihood estimates of the reduced-form parameters by OLS. In practice, we choose $\mathbf{P}$ so that $\mathbf{f}_t$ are the first M principal components of the cross-section of yields. Finally, estimates of the prices of risk parameters $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$ can then be obtained as the difference between the parameters driving the physical and risk-neutral measures.[4]

[2] An implication of the affine pricing framework is that we are free to consider either latent variables or linear combination of yields as the set of factors, given that one can be understood as a rotation of the other (see Joslin, Singleton and Zhu, 2011, and Joslin, Le and Singleton, 2012).

[3] We follow Joslin, Singleton and Zhu (2011) in assuming that $\boldsymbol{\Omega} = \boldsymbol{\sigma}_\eta^2 \times (\mathbf{P}_\perp \mathbf{P}_\perp')$ where $\mathbf{P}_\perp'$ is a basis for the orthogonal component of the row span of $\mathbf{P}'$. This guarantees that $\mathbf{P}'\boldsymbol{\Omega}\mathbf{P} = \mathbf{0}$. In addition, Joslin, Singleton and Zhu (2011) note that one can concentrate $\boldsymbol{\sigma}_\eta^2$ from the likelihood function through $\widehat{\boldsymbol{\sigma}}_\eta^2 = \sum_{t=1}^{T}\sum_{n=1}^{N}(y_{t,n}^o - y_{t,n})^2/(T \times (N-M))$.

[4] The approach in this paper can be extended to the case of observable factors with measurement errors. In such a case, and given the dimensionality of the problem, one needs to estimate the reduced-form parameters using the computationally efficient techniques of Jungbacker and Koopman (2008). Still, it is important to recall that Joslin, Le and Singleton (2012) show that, in practice, the (fitting) gain

We end this section by providing a multi-step algorithm that summarizes the implementation of our new linear estimator for GDTSMs:

Step 1 Estimate the parameters of the reduced-form model by linear regressions:

(1a) Estimate the cross-sectional coefficients $\mathbf{a}$ and $\mathbf{b}$ in equation (13) from OLS regressions of the observed yields, $\mathbf{y}_t^o$, on a constant and the bond factors $\mathbf{f}_t$

(1b) Obtain the coefficients $\boldsymbol{\mu}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Sigma}$ driving the VAR dynamics in (1) by OLS.

Step 2 Recover the coefficients driving the risk-neutral dynamics of the factors using cross-sectional regressions:

(2a) Recover $\widehat{A}_n = -n\widehat{a}_n$ and $\widehat{\mathbf{B}}_n = -n\widehat{\mathbf{b}}_n$, where $\widehat{a}_n$, and $\widehat{\mathbf{b}}_n$ are the estimates of a and $\mathbf{b}$ obtained in step 1. Set $\widehat{\delta}_0 = \widehat{a}_1$ and $\widehat{\boldsymbol{\delta}}_1 = \widehat{\mathbf{b}}_1$

(2b) Run the cross-sectional regressions in (11) and (12) to obtain an estimate of $\boldsymbol{\mu}^{\mathbb{Q}}$ and $\boldsymbol{\Phi}^{\mathbb{Q}}$.

Step 3 Obtain an estimate of the prices of risk as a difference between the coefficients driving the dynamics under the physical and risk-neutral measures: $\widehat{\boldsymbol{\lambda}}_0 = \widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}^{\mathbb{Q}}$ and $\widehat{\boldsymbol{\lambda}}_1 = \widehat{\boldsymbol{\Phi}} - \widehat{\boldsymbol{\Phi}}^{\mathbb{Q}}$.

2.3 An asymptotic least squares interpretation

In this section, we provide an alternative interpretation of this estimator based on the asymptotic least squares framework of GMT. As noted by these authors, many empirical models, including the one presented in this paper, can be formalized as a set of relationships $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \mathbf{0}$ between the parameters of interest $\boldsymbol{\theta}$, and a set of auxiliary parameters $\boldsymbol{\pi}$, for which a consistent and asymptotically normal estimate $\widehat{\boldsymbol{\pi}}$ is available. This framework suggests estimating the structural parameters by trying to find the set of parameters $\boldsymbol{\theta}$ that makes $\mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta})$ as close as possible to zero, in the metric of a given weighting matrix $\mathbf{W}_T$.[5] For this reason, the ALS estimation framework is also known as a minimum distance estimation, and $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta})$ is sometimes referred to as a distance function.[6]

from assuming that all observable factors are subject to measurement errors is minimal when one uses the first M principal components of yields as factors.

[5] The ALS estimator is similar in spirit to Hansen's (1982) GMM, which is based on moment conditions $E\left[\mathbf{h}(\mathbf{z}_t, \boldsymbol{\theta})\right] = 0$, $t = 1, ..., T$, where $\mathbf{z}_t$ is a vector of data in the time t information set. These moments conditions are then averaged to obtain $T^{-1}\sum_{t=1}^{T}\mathbf{h}(\mathbf{z}_t, \boldsymbol{\theta})$, and an estimate of $\boldsymbol{\theta}$ is obtained by minimizing a quadratic form in these sample moment conditions, $\left[T^{-1}\sum_{t=1}^{T}\mathbf{h}(\mathbf{z}_t, \boldsymbol{\theta})\right]' \mathbf{W} \left[T^{-1}\sum_{t=1}^{T}\mathbf{h}(\mathbf{z}_t, \boldsymbol{\theta})\right]$. Notice that the main difference between the GMM and ALS frameworks is that, in the latter case, the distance function $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta})$ is not necessarily a linear function of sample moments.

[6] Specifically, (classical) minimum distance estimation refers to the case where the distance function has the form $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \boldsymbol{\pi} - \mathbf{p}(\boldsymbol{\theta})$ (see, e.g., Chamberlain, 1982).

The linear estimator above falls under the ALS estimation framework. Specifically, we start by noting that the vector of auxiliary parameters is given by $\boldsymbol{\pi} = (\boldsymbol{\pi}_1', \boldsymbol{\pi}_2', \boldsymbol{\pi}_3')'$ (i.e., the reduced-form parameters), where

$$\begin{aligned}\boldsymbol{\pi}_1 &= \left\{vec\left[(\mathbf{a}\ \mathbf{b})'\right]\right\}',\\ \boldsymbol{\pi}_2 &= \left\{vec\left[(\boldsymbol{\mu}\ \boldsymbol{\Phi})'\right]\right\}',\\ \boldsymbol{\pi}_3 &= \left[vech\left(\boldsymbol{\Sigma}^{1/2}\right)\right]'.\end{aligned}$$

In order to guarantee the positivity of the covariance matrix $\boldsymbol{\Sigma}$, we focus on its Cholesky decomposition, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2\prime}$ rather than on $\boldsymbol{\Sigma}$ itself. Thus, we have a total of $H = (N+M)\times(M+1) + M\times(M+1)/2$ auxiliary parameters.

Note that the maximum likelihood estimation of the reduced-form parameters coincides with OLS estimation equation-by-equation, and therefore there is a consistent and asymptotically normal estimate $\widehat{\boldsymbol{\pi}}$ available. Specifically, we have that

$$\sqrt{T}\left[\begin{pmatrix}\widehat{\boldsymbol{\pi}}_1\\ \widehat{\boldsymbol{\pi}}_2\\ \widehat{\boldsymbol{\pi}}_3\end{pmatrix} - \begin{pmatrix}\boldsymbol{\pi}_1^0\\ \boldsymbol{\pi}_2^0\\ \boldsymbol{\pi}_3^0\end{pmatrix}\right] \xrightarrow{d} N\left[\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{V}_{\pi_1} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{V}_{\pi_2} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{V}_{\pi_3}\end{pmatrix}\right],$$

$$\sqrt{T}\left(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^0\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{V}_{\pi}\right),$$

where $\mathbf{V}_{\pi_1} = \boldsymbol{\Omega}\otimes E(\mathbf{x}_t\mathbf{x}_t')^{-1}$, $\mathbf{V}_{\pi_2} = \boldsymbol{\Sigma}\otimes E(\mathbf{x}_t\mathbf{x}_t')^{-1}$, $\mathbf{V}_{\pi_3} = 2\mathbf{E}(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\mathbf{E}'$; with $\mathbf{x}_t = (1\ \mathbf{f}_t')'$, $\mathbf{E} = \left[\mathbf{L}_M(\mathbf{I}+\mathbf{K}_{MM})(\boldsymbol{\Sigma}^{1/2}\otimes\mathbf{I})\mathbf{L}_M'\right]^{-1}\mathbf{D}_M^+$, where $\mathbf{L}_M$ is an "elimination matrix" such that $vech\left(\boldsymbol{\Sigma}\right) = \mathbf{L}_M vec\left(\boldsymbol{\Sigma}\right)$, $\mathbf{K}_{MM}$ is a "commutation matrix" such that $\mathbf{K}_{MM}vec(\mathbf{F}) = vec(\mathbf{F}')$ for any $(M\times M)$ matrix $\mathbf{F}$, and $\mathbf{D}_M^+ = (\mathbf{D}_M'\mathbf{D}_M)^{-1}\mathbf{D}_M'$ where $\mathbf{D}_M$ is a "duplication matrix" satisfying $\mathbf{D}_M vech\left(\boldsymbol{\Sigma}\right) = vec\left(\boldsymbol{\Sigma}\right)$ (see Lütkepohl, 1989).

Next, we consider the pricing recursions in equations (7) and (8). By stacking these two sets of equations for all bond yields, we can express the restrictions implied by the no-arbitrage model in compact form as

$$\mathbf{G}(\boldsymbol{\pi},\boldsymbol{\theta})' = \mathbf{Y}(\boldsymbol{\pi}) - \mathbf{X}(\boldsymbol{\pi})\boldsymbol{\Theta}^{\mathbb{Q}\prime} = \mathbf{0}, \tag{15}$$

where

$$\mathbf{Y}(\boldsymbol{\pi}) = \begin{pmatrix} A_1 & \mathbf{B}_1'\\ A_2 - A_1 - \frac{1}{2}\mathbf{B}_1'\boldsymbol{\Sigma}\mathbf{B}_1 - A_1 & \mathbf{B}_2' - \mathbf{B}_1'\\ \vdots & \vdots\\ A_n - A_{n-1} - \frac{1}{2}\mathbf{B}_{n-1}'\boldsymbol{\Sigma}\mathbf{B}_{n-1} - A_1 & \mathbf{B}_n' - \mathbf{B}_1'\\ \vdots & \vdots\\ A_N - A_{N-1} - \frac{1}{2}\mathbf{B}_{N-1}'\boldsymbol{\Sigma}\mathbf{B}_{N-1} - A_1 & \mathbf{B}_N' - \mathbf{B}_1' \end{pmatrix}, \quad \mathbf{X}(\boldsymbol{\pi}) = \begin{pmatrix} -1 & 0\\ 0 & \mathbf{B}_1'\\ \vdots & \vdots\\ 0 & \mathbf{B}_{n-1}'\\ \vdots & \vdots\\ 0 & \mathbf{B}_{N-1}' \end{pmatrix},$$

and where $\boldsymbol{\Theta}^{\mathbb{Q}}$ is a matrix that collects the parameters driving the dynamics under the risk-neutral measure:

$$\boldsymbol{\Theta}^{\mathbb{Q}\prime} = \begin{pmatrix}\delta_0 & \boldsymbol{\delta}_1'\\ \boldsymbol{\mu}^{\mathbb{Q}} & \boldsymbol{\Phi}^{\mathbb{Q}}\end{pmatrix}.$$

In addition to considering $vec(\mathbf{\Theta}^{\mathbb{Q}})$, it is convenient to add the parameters describing the dynamics of the factors under the physical measure to the vector of structural parameters, $\boldsymbol{\theta}$, such that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \boldsymbol{\theta}_3')'$ where

$$\begin{aligned}\boldsymbol{\theta}_1 = vec(\mathbf{\Theta}^{\mathbb{Q}}) &= \left((\delta_0 \ \boldsymbol{\delta}_1'), \left\{ vec\left[\left(\boldsymbol{\mu}^{\mathbb{Q}} \ \mathbf{\Phi}^{\mathbb{Q}} \right)' \right] \right\}' \right)', \\ \boldsymbol{\theta}_2 &= \left\{ vec\left[(\boldsymbol{\mu} \ \mathbf{\Phi})' \right] \right\}', \\ \boldsymbol{\theta}_3 &= \left[vech\left(\mathbf{\Sigma}^{1/2} \right) \right]'.\end{aligned}$$

Thus, we have a total of $K = (2M+1) \times (M+1) + M \times (M+1)/2$ parameters of interest.

By vectorizing equation (15) and adding a set of identities that define these new elements of $\boldsymbol{\theta}$ to be equal to the corresponding elements of $\boldsymbol{\pi}$, we arrive at the following expression for $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta})$:

$$\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \boldsymbol{\gamma}(\boldsymbol{\pi}) - \mathbf{\Gamma}(\boldsymbol{\pi})\boldsymbol{\theta},$$

where $\boldsymbol{\gamma}(\boldsymbol{\pi}) = \left\{ vec\left[\mathbf{Y}(\boldsymbol{\pi})' \right]', \boldsymbol{\pi}_2', \boldsymbol{\pi}_3' \right\}$ and

$$\mathbf{\Gamma}(\boldsymbol{\pi}) = \begin{pmatrix} \mathbf{X}(\boldsymbol{\pi}) \otimes \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix}.$$

We note that the linear estimator proposed in section 2.2 above is numerically equivalent to the estimator that minimizes a quadratic form in the distance function, evaluated at the estimates of the reduced-form parameters, $\widehat{\boldsymbol{\pi}}$, where the weighting matrix has been chosen to be the identity matrix, $\mathbf{W}_T = \mathbf{I}$:

$$\widehat{\boldsymbol{\theta}}_{OLS} = \arg\min_{\boldsymbol{\theta}} T \left[\boldsymbol{\gamma}(\widehat{\boldsymbol{\pi}}) - \mathbf{\Gamma}(\widehat{\boldsymbol{\pi}})\boldsymbol{\theta} \right]' \left[\boldsymbol{\gamma}(\widehat{\boldsymbol{\pi}}) - \mathbf{\Gamma}(\widehat{\boldsymbol{\pi}})\boldsymbol{\theta} \right]. \tag{16}$$

More importantly, since the distance function is linear in $\boldsymbol{\theta}$, the solution to the minimization problem in equation (16), $\widehat{\boldsymbol{\theta}}_{OLS}$, is known in closed form. In particular, we have that $\mathbf{\Gamma}(\widehat{\boldsymbol{\pi}})' \left[\boldsymbol{\gamma}(\widehat{\boldsymbol{\pi}}) - \mathbf{\Gamma}(\widehat{\boldsymbol{\pi}})\widehat{\boldsymbol{\theta}}_{OLS} \right] = 0$, and therefore

$$\widehat{\boldsymbol{\theta}}_{OLS} = \left(\widehat{\mathbf{\Gamma}}'\widehat{\mathbf{\Gamma}} \right)^{-1} \left(\widehat{\mathbf{\Gamma}}'\widehat{\boldsymbol{\gamma}} \right), \tag{17}$$

where $\widehat{\boldsymbol{\gamma}} \equiv \boldsymbol{\gamma}(\widehat{\boldsymbol{\pi}})$ and $\widehat{\mathbf{\Gamma}} = \mathbf{\Gamma}(\widehat{\boldsymbol{\pi}})$.

2.4 Self-consistency and optimal ALS estimation

As in the case of GMM estimation, an identity weighting matrix is not necessarily optimal in this context and (asymptotic) efficiency gains can be achieved by selecting an appropriate weighting matrix. As noted by GMT, the optimal weighting matrix is equal to (an estimate of) the inverse of the asymptotic variance of the distance function. By

the standard delta method, this covariance matrix is related to the asymptotic covariance of the reduced-form coefficients $\mathbf{V}_\pi$ through the following relationship: $\mathbf{V}_g = \mathbf{G}_{\boldsymbol{\pi}} \mathbf{V}_{\boldsymbol{\pi}} \mathbf{G}'_{\boldsymbol{\pi}}$, where $\mathbf{G}_\pi(\boldsymbol{\pi}, \boldsymbol{\theta}) = \partial \mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta}) / \partial \boldsymbol{\pi}'$ is the Jacobian of the distance function with respect to the auxiliary parameters.

However, the matrix $\mathbf{V}_\pi$ has a reduced-rank structure in our set-up which, given that $\mathbf{G}_{\boldsymbol{\pi}}$ is a non-singular $H \times H$ matrix, implies a singularity in $\mathbf{V}_g$ as well. To understand why this happens, we need to discuss the concept of self-consistency of an affine term structure model. As noted by Cochrane and Piazzesi (2005), one has to guarantee that, when choosing state variables that are linear combinations (portfolios) of the yields, $\mathbf{f}_t = \mathbf{P}' \mathbf{y}_t^o$, the state variables that come out of the model need to be the same as the state variables that we started with. In other words, it is necessary to ensure that the pricing of portfolios of yields is also consistent with equation (6) such that $\mathbf{f}_t = \mathbf{P}' \mathbf{y}_t = \mathbf{P}' \mathbf{a}(\boldsymbol{\theta}) + \mathbf{P}' \mathbf{b}(\boldsymbol{\theta}) \mathbf{f}_t$. Thus, self-consistency of the model amounts to imposing the following set of constraints when estimating the model:

$$\mathbf{P}' \mathbf{a}(\boldsymbol{\theta}) = \mathbf{0}, \qquad \mathbf{P}' \mathbf{b}(\boldsymbol{\theta}) = \mathbf{I}. \tag{18}$$

While the OLS estimator in equation (17) does not necessarily satisfy these constraints, we find that, in practice, $\widehat{\boldsymbol{\theta}}_{OLS}$ delivers parameter estimates that almost satisfy such restrictions. Two features of our proposed estimation approach explain this result. First, given our assumption that $\mathbf{f}_t$ is observed perfectly, the OLS estimates of the reduced-form coefficients automatically satisfy such restrictions:

$$\mathbf{P}' \widehat{\mathbf{a}} = \mathbf{0}, \qquad \mathbf{P}' \widehat{\mathbf{b}} = \mathbf{I}, \tag{19}$$

and, second, our linear estimator tries to match those as closely as possible. The problem is that, since the reduced-form coefficients satisfy the self-consistency restrictions in (19), $\mathbf{V}_\pi$ has a reduced rank structure, which implies a singular $\mathbf{V}_g$. Hence, the definition provided by GMT of an optimal ALS estimator breaks down in our set-up.

In the next section, we briefly review the asymptotic distribution of ALS estimators and discuss some optimality considerations for these estimators when the covariance matrix of the distance function is singular. In particular, we adapt the work of Peñaranda and Sentana (2012) – who study the problem of obtaining an optimal GMM estimator when the asymptotic variance of the moment conditions is singular in the population – to the ALS framework. In particular, we show that imposing the self-consistency restrictions when estimating the model and, simultaneously, replacing the ordinary inverse of the covariance of the distance function by any of its generalized inverses delivers an optimal ALS estimator.

3 Asymptotic least squares estimation of GDTSMs

Prior to defining an optimal ALS estimator in a singular set-up, we introduce some formal notation and briefly review the estimation framework proposed by GMT.

3.1 Asymptotic distribution of ALS estimators

As already discussed in the previous section, let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset R^K$ be the vector of parameters of interest, and let $\boldsymbol{\pi} \in \boldsymbol{\Pi} \subset R^H$ be the vector of auxiliary parameters. Both sets of parameters are related through a system of implicit equations of the form $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \mathbf{0}$, where $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta})$ is a $(G \times 1)$ twice continuously differentiable distance function. Let the Jacobian of this distance function be denoted by

$$\mathbf{G}_{\pi}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \frac{\partial \mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta})}{\partial \boldsymbol{\pi}'}, \qquad \mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \frac{\partial \mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}.$$

Let $\mathbf{p}(\boldsymbol{\theta})$ be a function satisfying $\mathbf{g}\left[\mathbf{p}(\boldsymbol{\theta}), \boldsymbol{\theta}\right] = \mathbf{0}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, which implies that the system of implicit equations $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \mathbf{0}$ has a unique solution for $\boldsymbol{\pi}$ given $\boldsymbol{\theta}$, and let this solution be given by $\boldsymbol{\pi} = \mathbf{p}(\boldsymbol{\theta})$. For example, the function $\mathbf{p}(\boldsymbol{\theta})$ can be thought of as the set of auxiliary parameters implied by the set of parameters $\boldsymbol{\theta}$. In the case of the estimation of GDTSMs, we note that the set of implicit equations $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \mathbf{0}$ is related to the pricing recursions in equations (7) and (8); the set of implied auxiliary parameters can be found in equations (9) and (10); and the number of auxiliary parameters, H, is equal to the dimension of the distance function, G.

Let $\widehat{\boldsymbol{\pi}}$ denote a strongly consistent and asymptotically normal estimator of the auxiliary parameters, such that as $T \to \infty$, $\widehat{\boldsymbol{\pi}} \to \boldsymbol{\pi}^0 = \mathbf{p}(\boldsymbol{\theta}^0)$, P_{θ^0} almost surely; and $\sqrt{T}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^0) \xrightarrow{d} N\left[\mathbf{0}, \mathbf{V}_{\pi}(\boldsymbol{\theta}^0)\right]$, where T denotes the number of observations in the sample and $\boldsymbol{\theta}^0$ denotes the true value of the parameters of interest; i.e., $\mathbf{g}(\boldsymbol{\pi}^0, \boldsymbol{\theta}^0) = \mathbf{0}$.

GMT propose to minimize a quadratic form in the distance function evaluated at the estimates of the auxiliary parameters, $\widehat{\boldsymbol{\pi}}$:

$$\widehat{\boldsymbol{\theta}}_{ALS} = \arg\min_{\boldsymbol{\theta}} T\mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta})' \mathbf{W}_T \mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta}), \tag{20}$$

where $\mathbf{W}_T$ is a positive semi-definite weighting matrix that possibly depends on the observations. In other words, the ALS estimation principle consists of forcing the G implicit equations evaluated at $\widehat{\boldsymbol{\pi}}$ to be as close as possible to zero in the metric defined by $\mathbf{W}_T$. Further notice that, when the distance function is linear in the set of parameters of interest (as in the case of the estimation of GDTSMs), the solution to the optimization problem in (20) is known in closed form.

Now, let $\mathbf{W}_T$ converge P_{θ^0} almost surely to $\mathbf{W}$, a non-stochastic semi-definite weighting matrix of size G, and rank greater or equal than K. If the true values of the parameters

of interest and auxiliary parameters, $\boldsymbol{\theta}^0$ and $\boldsymbol{\pi}^0$, both belong to the interior of $\boldsymbol{\Theta}$ and $\boldsymbol{\Pi}$, respectively, and $\mathbf{G}_{\boldsymbol{\theta}}'\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}}$ evaluated at $\boldsymbol{\theta}^0$ and $\boldsymbol{\pi}^0$ is non-singular (which implies that the rank of $\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\pi}^0, \boldsymbol{\theta}^0) = K$ and that $K \leq G$), then (see GMT for the proof) $\widehat{\boldsymbol{\theta}}_{ALS}$ is strongly consistent for every choice of $\mathbf{W}_T$, and its asymptotic distribution is given by

$$\sqrt{T}(\widehat{\boldsymbol{\theta}}_{ALS} - \boldsymbol{\theta}^0) \xrightarrow{d} N\left[\mathbf{0}, (\mathbf{G}_{\boldsymbol{\theta}}'\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}})^{-1}\mathbf{G}_{\boldsymbol{\theta}}'\mathbf{W}\mathbf{G}_{\boldsymbol{\pi}}\mathbf{V}_{\pi}\mathbf{G}_{\pi}'\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}}(\mathbf{G}_{\boldsymbol{\theta}}'\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}})^{-1}\right], \tag{21}$$

where the various matrices in this equation are evaluated at $\boldsymbol{\theta}^0$ and $\boldsymbol{\pi}^0$.

As in the case of (overidentified) GMM estimation, it is possible to choose an "optimal" weighting matrix, in the sense that the difference between the asymptotic variance of the resulting ALS estimator and another ALS estimator based on any other quadratic form in the same distance function is positive definite. In particular, GMT show that when $\mathbf{G}_{\boldsymbol{\pi}}\mathbf{V}_{\pi}\mathbf{G}_{\pi}'$ and $\mathbf{G}_{\boldsymbol{\theta}}'(\mathbf{G}_{\boldsymbol{\pi}}\mathbf{V}_{\boldsymbol{\pi}}\mathbf{G}_{\boldsymbol{\pi}}')^{-1}\mathbf{G}_{\boldsymbol{\theta}}$ are non-singular when evaluated at $\boldsymbol{\theta}^0$ and $\boldsymbol{\pi}^0$ (which implies that the rank of $\mathbf{G}_{\boldsymbol{\pi}}(\boldsymbol{\pi}^0, \boldsymbol{\theta}^0) = G$ and that $G \leq H$), then an optimal ALS exists and corresponds to the choice of a weighting matrix $\mathbf{W}_T$ that converges to $\mathbf{W} = (\mathbf{G}_{\boldsymbol{\pi}}\mathbf{V}_{\pi}\mathbf{G}_{\pi}')^{-1}$. Note that, by the delta method, the optimal weighting matrix is simply the asymptotic covariance of the distance function evaluated at the estimates of the auxiliary parameters $\mathbf{V}_g(\boldsymbol{\theta}^0) = avar\left[\sqrt{T}\mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta}^0)\right] = \mathbf{G}_{\boldsymbol{\pi}}(\boldsymbol{\pi}^0, \boldsymbol{\theta}^0)\mathbf{V}_{\pi}(\boldsymbol{\theta}^0)\mathbf{G}_{\pi}'(\boldsymbol{\pi}^0, \boldsymbol{\theta}^0)$. However, using $\mathbf{W} = \mathbf{V}_g^{-1}$ in our set-up is not feasible because, in such a case, the weighting matrix needs to be evaluated at the true value $\boldsymbol{\theta}^0$ (which is unknown). Instead, as in the case of optimal GMM estimation, it is possible to replace $\mathbf{V}_g(\boldsymbol{\theta}^0)$ with an estimator $\widehat{\mathbf{V}}_g(\boldsymbol{\theta})$ evaluated at some initial consistent estimator of $\boldsymbol{\theta}^0$. As is standard in the literature, we will use a two-step ALS estimation in which we first obtain an estimate of $\boldsymbol{\theta}$ by using the identity matrix (OLS approach), and then evaluate the weighting matrix at this estimate:

$$\widehat{\mathbf{V}}_g = \mathbf{G}_{\pi}\left[\mathbf{p}(\widehat{\boldsymbol{\theta}}_{OLS}), \widehat{\boldsymbol{\theta}}_{OLS}\right]\widehat{\mathbf{V}}_{\pi}(\widehat{\boldsymbol{\theta}}_{OLS})\mathbf{G}_{\pi}\left[\mathbf{p}(\widehat{\boldsymbol{\theta}}_{OLS}), \widehat{\boldsymbol{\theta}}_{OLS}\right].$$

Further, Kodde, Palm and Pfann (1990) note that if (i) the system of relationships $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \mathbf{0}$ is complete,[7] and (ii) $\boldsymbol{\pi}$ is estimated by maximum likelihood (ML), or a method asymptotically equivalent to ML, then the optimal ALS estimator is asymptotically equivalent to the ML estimator of $\boldsymbol{\theta}$.

3.2 Optimal ALS estimation in a singular set-up

Unfortunately, $\mathbf{V}_{\pi}(\boldsymbol{\theta}^0)$ has a reduced-rank structure in our set-up, because the estimates of the reduced-form coefficients satisfy the self-consistency restrictions in equation (19). As noted above, this renders $\mathbf{V}_g(\boldsymbol{\theta}^0)$ singular as well, which in turn makes the definition provided by GMT of an optimal ALS estimator break down.

[7] The system $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \mathbf{0}$ is complete if the dimension of the set of reduced-form parameters, $\boldsymbol{\pi}$, is equal to the dimension of $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta})$; and the Jacobian $\mathbf{G}_{\pi}(\boldsymbol{\pi}, \boldsymbol{\theta})$ has full rank (i.e., rank$[\mathbf{G}_{\boldsymbol{\pi}}(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0)] = G$) when evaluated at the true value.

In order to provide an optimal ALS estimator when the asymptotic covariance of the distance function no longer has full rank, we borrow from Peñaranda and Sentana (2012), who study a conceptually similar problem to ours: obtaining an optimal GMM estimator when the asymptotic variance of the moment conditions is singular in the population. In the remainder of this section, we adapt their methodology to the case of ALS estimation and we refer the reader to their work for further details.

Specifically, we focus on singularities of the asymptotic covariance of the distance function that satisfy the following two assumptions:

Assumption 1. Let $\boldsymbol{\Xi}(\boldsymbol{\theta})$ denote a $G \times S$ matrix of continuously differentiable functions of $\boldsymbol{\theta}$, where $0 \leq S \leq K$. The subset of $\boldsymbol{\Theta}$ for which

$$\boldsymbol{\Xi}'(\boldsymbol{\theta}) \left[\sqrt{T}\mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta})\right] \xrightarrow{p} \mathbf{0}, \tag{22}$$

can be fully characterized by $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$, where $\mathbf{r}(\boldsymbol{\theta})$ is a $S \times 1$ known continuously differentiable transformation of $\boldsymbol{\theta}$.

Assumption 2. For $S > 0$, we have that $\mathbf{r}(\boldsymbol{\theta}^0) = \mathbf{0}$, $\text{rank}\big[\mathbf{V}_g(\boldsymbol{\theta}^0)\big] = G - S$, and $\text{rank}\big[\partial \mathbf{r}(\boldsymbol{\theta}^0)/\partial \boldsymbol{\theta}'\big] = S$.

These two assumptions are the ALS analogue to the assumptions in Peñaranda and Sentana (2012). The first assumption defines $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ as the implicit $K - S$-dimensional manifold in $\boldsymbol{\Theta}$ over which S linear combinations of $\sqrt{T}\mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta})$ converge in probability to zero. The second assumption ensures that the true values of $\boldsymbol{\theta}^0$ belong to that manifold. More importantly, these two assumptions cover the empirically relevant case of the estimation of GDTSMs within an ALS framework, where the singularity of the covariance matrix of the distance function is a consequence of the estimates of the reduced-form model satisfying the self-consistency restrictions.[8]

Similarly to Peñaranda and Sentana (2012), we show in Proposition 1 in Appendix A[9] that the optimal ALS estimator satisfies

$$\widehat{\boldsymbol{\theta}}_{OALS} = \arg\min_{\boldsymbol{\theta}} T\mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta})' \mathbf{V}_g^+(\boldsymbol{\theta}^0) \mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta}) \qquad \text{s.t. } \mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}. \tag{23}$$

[8] To be more specific, Assumption 1 covers the case where (i) the restrictions on $\boldsymbol{\theta}$ can be written as a set of restrictions on the auxiliary parameters implied by the model $\mathbf{p}(\boldsymbol{\theta})$, i.e. $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{r}\left[\mathbf{p}(\boldsymbol{\theta})\right] = \mathbf{0}$, (ii) $\widehat{\boldsymbol{\pi}}$ denote a strongly consistent and asymptotically normal estimator of the auxiliary parameters, such that as $T \to \infty$, $\widehat{\boldsymbol{\pi}} \to \boldsymbol{\pi}_0 = \mathbf{p}(\boldsymbol{\theta})$, P_θ almost surely for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and (iii) $\widehat{\boldsymbol{\pi}}$ satisfies that $\mathbf{r}(\widehat{\boldsymbol{\pi}}) = \mathbf{0}$. Under these three assumptions, we can expand both $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta})$ and $\mathbf{r}(\boldsymbol{\pi}) = \mathbf{0}$ around $\mathbf{p}(\boldsymbol{\theta})$:

$$\begin{aligned} \sqrt{T}\mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta}) &= \mathbf{G}_{\boldsymbol{\pi}}(\boldsymbol{\theta})\sqrt{T}\left[\widehat{\boldsymbol{\pi}} - \mathbf{p}(\boldsymbol{\theta})\right] + o_p(1), \\ \sqrt{T}\mathbf{r}(\widehat{\boldsymbol{\pi}}) &= \mathbf{R}_{\boldsymbol{\pi}}(\boldsymbol{\theta})\sqrt{T}\left[\widehat{\boldsymbol{\pi}} - \mathbf{p}(\boldsymbol{\theta})\right] + o_p(1), \end{aligned}$$

where $\mathbf{R}_{\boldsymbol{\pi}}(\boldsymbol{\theta}) = \partial \mathbf{r}\left[\mathbf{p}(\boldsymbol{\theta})\right]/\partial \boldsymbol{\theta}'$. Since $\mathbf{r}(\widehat{\boldsymbol{\pi}}) = \mathbf{0}$, the linear combinations of $\sqrt{T}\mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta})$ given by $\boldsymbol{\Xi}'(\boldsymbol{\theta}) = \mathbf{R}_{\boldsymbol{\pi}}(\boldsymbol{\theta})\left[\mathbf{G}_{\boldsymbol{\pi}}'(\boldsymbol{\theta})\mathbf{G}_{\boldsymbol{\pi}}(\boldsymbol{\theta})\right]^{-1}\mathbf{G}_{\boldsymbol{\pi}}'(\boldsymbol{\theta})$ converge in probability to zero.

[9] We note that this is simply the ALS analogue of Proposition 1 in Peñaranda and Sentana (2012).

That is, optimal ALS estimation and inference under the type of singularities characterized by Assumptions 1 and 2 requires both (i) imposing the parametric restriction $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ when estimating the model and, simultaneously (ii) replacing the ordinary inverse of $\mathbf{V}_g(\boldsymbol{\theta}^0)$ by any of its generalized inverses, $\mathbf{V}_g^+(\boldsymbol{\theta}^0)$. Moreover, as in the case of GMM, the optimized value of the ALS criterion function has an asymptotic χ^2 distribution with degrees of freedom equal to the number of overidentifying restrictions $(G - K)$.

An alternative way to view this estimator, as discussed in Peñaranda and Sentana (2012), is as follows. Let the spectral decomposition of $\mathbf{V}_g(\boldsymbol{\theta}^0)$ given above be written as

$$\mathbf{V}_g(\boldsymbol{\theta}^0) = \begin{pmatrix} \mathbf{T}_1 & \mathbf{T}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{T}_1' \\ \mathbf{T}_2' \end{pmatrix} = \mathbf{T}_1 \boldsymbol{\Lambda} \mathbf{T}_1',$$

where $\boldsymbol{\Lambda}$ is a $(G-S)\times(G-S)$ positive definite diagonal matrix and focus, for the moment, on the Moore-Penrose generalized inverse of $\mathbf{V}_g(\boldsymbol{\theta}^0)$, such that

$$\mathbf{V}_g^{MP+}(\boldsymbol{\theta}^0) = \mathbf{T}_1 \boldsymbol{\Lambda}^{-1} \mathbf{T}_1'.$$

Then, the optimal ALS estimator in this singular set-up is equivalent to the constrained ALS estimator that works with the reduced set of $K - S$ distance functions $\mathbf{T}_1' \mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta})$ and the restrictions $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$. In this way, note that the ALS estimator that uses the Moore-Penrose generalized inverse of $\mathbf{V}_g(\boldsymbol{\theta}^0)$ alone without the equality restrictions $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ will not likely be optimal, since it drops the S asymptotically degenerate, i.e. most informative, linear combinations of $\sqrt{T}\mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta})$. In fact, it might even be the case that $\boldsymbol{\theta}$ is not identified from the set of reduced implicit relations $\mathbf{T}_1' \mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta}) = \mathbf{0}$. This will occur, for example, if $K > G - S$.

Again the choice of the weighting matrix $\mathbf{W} = \mathbf{V}_g^+(\boldsymbol{\theta}^0)$ is not feasible because $\boldsymbol{\theta}^0$ is not known. In particular, we replace $\mathbf{V}_g^+(\boldsymbol{\theta}^0)$ by the generalized inverse of $\widehat{\mathbf{V}}_g(\boldsymbol{\theta})$ evaluated at some initial consistent estimator of $\boldsymbol{\theta}^0$ (e.g., using the identity matrix as the weighting matrix). We note that, since we focus on the empirical relevant case that $\widehat{\boldsymbol{\pi}}$ satisfies the parametric restriction $\mathbf{r}(\boldsymbol{\pi}) = \mathbf{0}$, $\widehat{\mathbf{V}}_g$ will, under standard regularity conditions, consistently estimate $\mathbf{V}_g^+(\boldsymbol{\theta}^0)$ given that $\text{rank}\left(\widehat{\mathbf{V}}_g\right) = G - S$.

Further, Proposition 2 in Appendix A presents the conditions under which optimal ALS estimation is asymptotically equivalent to ML when one considers the class of models with singularities in the asymptotic covariance matrix of the distance function covered in Assumptions 1 and 2. In particular, if (i) the restrictions on $\boldsymbol{\theta}$ can be written as a set of restrictions on the auxiliary parameters implied by the model $\mathbf{p}(\boldsymbol{\theta})$, i.e., $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{r}[\mathbf{p}(\boldsymbol{\theta})] = \mathbf{0}$; (ii) $\widehat{\boldsymbol{\pi}}$ has been estimated by a method that is asymptotically equivalent to ML, and satisfies $\mathbf{r}(\widehat{\boldsymbol{\pi}}) = \mathbf{0}$; and (iii) the system of implicit relationships $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \mathbf{0}$ is complete, then the optimal ALS estimator in equation (23) is asymptotically equivalent to the ML estimator of $\boldsymbol{\theta}$ that imposes $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$. Proposition 2 thus extends the results in Kodde, Palm and Pfann (1990) to the case of optimal ALS estimation in a singular set-up.

3.3 Asymptotic least squares estimation of GDTSMs

We now return to the case of the estimation of GDTSMs. In particular, from the results obtained in the previous section, we have that the optimal estimator of the parameters of interest of such a model is given by

$$\widehat{\boldsymbol{\theta}}_{CGLS} = \arg\min_{\boldsymbol{\theta}} T\left[\boldsymbol{\gamma}(\widehat{\boldsymbol{\pi}}) - \boldsymbol{\Gamma}(\widehat{\boldsymbol{\pi}})\boldsymbol{\theta}\right]' \widehat{\mathbf{V}}_g^{+} \left[\boldsymbol{\gamma}(\widehat{\boldsymbol{\pi}}) - \boldsymbol{\Gamma}(\widehat{\boldsymbol{\pi}})\boldsymbol{\theta}\right] \qquad \text{s.t. } \mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}, \tag{24}$$

where, by stacking and vectorizing (18), we have that

$$\mathbf{r}(\boldsymbol{\theta}) = \mathbf{r}\left[\mathbf{p}(\boldsymbol{\theta})\right] = vec\left(\mathbf{P}' \otimes \mathbf{I}\right) \mathbf{p}_1(\boldsymbol{\theta}) - \bar{\mathbf{r}}_1, \tag{25}$$

where $\mathbf{p}_1(\boldsymbol{\theta}) = vec\left\{\left[\mathbf{a}(\boldsymbol{\theta})\ \mathbf{b}(\boldsymbol{\theta})\right]'\right\}$, $\bar{\mathbf{r}}_1 = vec\left[(\mathbf{0}\ \mathbf{I})'\right]$, and the number of self-consistency restrictions is $S = M \times (M+1)$. Again, note that self-consistency of the model implies a set of restrictions on the auxiliary parameters implied by the model $\mathbf{p}(\boldsymbol{\theta})$, and that the OLS estimates of the reduced-form parameters already satisfy these self-consistency restrictions, $\mathbf{r}(\widehat{\boldsymbol{\pi}}) = \mathbf{0}$. We refer to this (optimal) estimator as the constrained generalized least squares (CGLS) estimator. More important, our estimating system $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \mathbf{0}$ is complete, which means that the CGLS estimator satisfies the conditions in Proposition 2 under which optimal ALS estimation is asymptotically equivalent to ML estimation.

Unfortunately, the solution to the optimal ALS (i.e., the CGLS) estimator in equation (24) is not known in closed form because $\mathbf{r}(\boldsymbol{\theta})$ is not linear in the set of parameters of interest, $\boldsymbol{\theta}$. Still, as noted by Newey and McFadden (1994) and Gourieroux and Monfort (1995) among others, estimating the model subject to a linearized version of the constraint delivers an estimator that is asymptotically equivalent to the one that uses the non-linear constraint. For this reason, we start by considering the (suboptimal) ALS estimator that uses a consistent estimate of the generalized inverse of $\mathbf{V}_g(\boldsymbol{\theta})$ as weighting matrix but that does not impose the restrictions $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$:

$$\widehat{\boldsymbol{\theta}}_{GLS} = \left(\widehat{\boldsymbol{\Gamma}}'\widehat{\mathbf{V}}_g^{+}\widehat{\boldsymbol{\Gamma}}\right)^{-1} \left(\widehat{\boldsymbol{\Gamma}}'\widehat{\mathbf{V}}_g^{+}\widehat{\boldsymbol{\gamma}}\right). \tag{26}$$

We will refer to this estimator as the generalized least squares (GLS) estimator. Then, the linearized constrained GLS estimator, $\widetilde{\boldsymbol{\theta}}_{CGLS}$ is defined as

$$\begin{aligned} \widetilde{\boldsymbol{\theta}}_{CGLS} = \arg\min_{\boldsymbol{\theta}} T\left[\boldsymbol{\gamma}(\widehat{\boldsymbol{\pi}}) - \boldsymbol{\Gamma}(\widehat{\boldsymbol{\pi}})\boldsymbol{\theta}\right]' \widehat{\mathbf{V}}_g^{+} \left[\boldsymbol{\gamma}(\widehat{\boldsymbol{\pi}}) - \boldsymbol{\Gamma}(\widehat{\boldsymbol{\pi}})\boldsymbol{\theta}\right], \\ \text{s.t. } \mathbf{r}(\widehat{\boldsymbol{\theta}}_{GLS}) = \widehat{\mathbf{R}}_{\theta}(\widehat{\boldsymbol{\theta}}_{GLS} - \boldsymbol{\theta}), \end{aligned} \tag{27}$$

where $\widehat{\mathbf{R}}_{\theta} = \frac{\partial \mathbf{r}(\widehat{\boldsymbol{\theta}}_{GLS})}{\partial \boldsymbol{\theta}'}$ and the constraint $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ has been linearized around the unconstrained (GLS) estimate of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}_{GLS}$. The main advantage of such linearization is that, since the objective function is quadratic and the restrictions are now linear in the parameters of interest, the solution of the estimation problem is known in closed form:

$$\widetilde{\boldsymbol{\theta}}_{CGLS} = \widehat{\boldsymbol{\theta}}_{GLS} - \left(\widehat{\boldsymbol{\Gamma}}'\widehat{\mathbf{V}}_g^{+}\widehat{\boldsymbol{\Gamma}}\right)^{-1} \widehat{\mathbf{R}}_{\theta}' \left[\widehat{\mathbf{R}}_{\theta} \left(\widehat{\boldsymbol{\Gamma}}'\widehat{\mathbf{V}}_g^{+}\widehat{\boldsymbol{\Gamma}}\right)^{-1} \widehat{\mathbf{R}}_{\theta}'\right]^{-1} \mathbf{r}(\widehat{\boldsymbol{\theta}}_{GLS}). \tag{28}$$

In particular, this linearized estimator corrects the unconstrained GLS estimates with an additive term that is a function of the distance $\mathbf{r}(\widehat{\boldsymbol{\theta}}_{GLS})$ by which the contraints are not satisfied by $\widehat{\boldsymbol{\theta}}_{GLS}$ (see chapter 10 in Gourieroux and Monfort, 1995).

However, $\widetilde{\boldsymbol{\theta}}_{CGLS}$ still does not satisfy the constraints $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ exactly, even though $\widetilde{\boldsymbol{\theta}}_{CGLS}$ is asymptotically equivalent to the estimator that uses the non-linear constraint. This is why we follow Bekaert and Hodrick (2001) in iterating equation (28) when constructing our constrained estimates. Specifically, we start by obtaining a first restricted estimate of $\boldsymbol{\theta}$ using equation (28) and linearizing the constraint $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ around $\widehat{\boldsymbol{\theta}}_{GLS}$. Denote this first restricted estimate $\widetilde{\boldsymbol{\theta}}^{(1)}_{CGLS}$. Then, we substitute the initial unconstrained estimate, $\widehat{\boldsymbol{\theta}}_{GLS}$, in (28) $\widetilde{\boldsymbol{\theta}}^{(1)}_{CGLS}$ to obtain a second restricted estimate of $\boldsymbol{\theta}$. Denote this second restricted estimate by $\widetilde{\boldsymbol{\theta}}^{(2)}_{CGLS}$. We repeat this process until the resulting constrained estimate satisfies the self-consistency restrictions, $\mathbf{r}(\widetilde{\boldsymbol{\theta}}^{(n)}_{CGLS}) = \mathbf{0}$ within a given tolerance. In practice, only a few iterations of equation (28) are required.

Finally, we note that one has to be careful not to delete any of the identity equations in $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \mathbf{0}$ that define $\boldsymbol{\theta}_2 = \boldsymbol{\pi}_2$ and $\boldsymbol{\theta}_3 = \boldsymbol{\pi}_3$ when computing the generalized inverse of $\widehat{\mathbf{V}}_g$. In particular, $\mathbf{g}_1(\boldsymbol{\pi}, \boldsymbol{\theta}) = vec[\mathbf{G}(\boldsymbol{\pi}, \boldsymbol{\theta})]$ with $\mathbf{G}(\boldsymbol{\pi}, \boldsymbol{\theta})$ in equation (15) does not identify $\boldsymbol{\theta}_2$, and only weakly identifies $\boldsymbol{\theta}_3$ (i.e., the innovation parameters of the VAR dynamics under $\mathbb{P}$ only appear in the pricing equations through a (small) Jensen's inequality term). Thus, eliminating any of the identities in $\mathbf{g}_2(\boldsymbol{\pi}, \boldsymbol{\theta}) = \boldsymbol{\pi}_2 - \boldsymbol{\theta}_2$ or $\mathbf{g}_3(\boldsymbol{\pi}, \boldsymbol{\theta}) = \boldsymbol{\pi}_3 - \boldsymbol{\theta}_3$, might lead to numerical instabilities of the CGLS estimates. We solve this problem by orthogonalizing $\mathbf{g}_1(\boldsymbol{\pi}, \boldsymbol{\theta})$ with respect to $\mathbf{g}_3(\boldsymbol{\pi}, \boldsymbol{\theta})$, and computing the generalized inverse of the residual. Undoing this transformation, we have the following generalized inverse of $\mathbf{V}_g$:

$$\mathbf{V}_g^+ = \begin{pmatrix} \mathbf{V}_{g_{1|3}}^+ & \mathbf{0} & -\mathbf{V}_{g_{1|3}}^+ \mathbf{V}_{g_{13}} \mathbf{V}_{g_{33}}^{-1} \\ 0 & \mathbf{V}_{g_{22}}^{-1} & \mathbf{0} \\ -\mathbf{V}_{g_{33}}^{-1} \mathbf{V}_{g_{31}} \mathbf{V}_{g_{1|3}}^+ & \mathbf{0} & \mathbf{V}_{g_{33}}^{-1} \end{pmatrix},$$

where $\mathbf{V}_{g_{1|3}}^+$ is a generalized inverse of $\mathbf{V}_{g_{1|3}} = \mathbf{V}_{g_{11}} - \mathbf{V}_{g_{13}} \mathbf{V}_{g_{33}}^{-1} \mathbf{V}_{g_{31}}$.[10]

4 Discussion of related literature

In this section, we compare our linear regression estimator to three recent approaches to the estimation of GDTSMs: the maximum likelihood approach of Joslin, Singleton and Zhu (2011), the minimum-chi-square estimator of Hamilton and Wu (2012), and the regression-based approach of Adrian, Crump and Moench (2012).

[10] After some tedious but straightforward algebra, it is easy to show that (i) $\mathbf{V}_g \mathbf{V}_g^+ \mathbf{V}_g = \mathbf{V}_g$ and (ii) $\mathbf{V}_g^+ \mathbf{V}_g \mathbf{V}_g^+ = \mathbf{V}_g^+$. However, $\mathbf{V}_g^+$ does not satisfy (iii) $\mathbf{V}_g \mathbf{V}_g^+ = (\mathbf{V}_g \mathbf{V}_g^+)'$ nor (iv) $\mathbf{V}_g^+ \mathbf{V}_g = (\mathbf{V}_g^+ \mathbf{V}_g)'$. Thus $\mathbf{V}_g^+$ is a generalized inverse of $\mathbf{V}_g$, but it is not the Moore-Penrose generalized inverse of $\mathbf{V}_g$.

4.1 Joslin, Singleton and Zhu (2011)

The maximum likelihood (ML) approach has traditionally been considered a natural way to estimate GDTSMs given that, once one has specified the distribution of the pricing errors, these models provide a complete characterization of the joint distribution of yields. However, the solution of the optimization problem involving the maximization of the density of the yields does not exist in closed form, except in very few specific cases. Consequently, researchers often have to rely on cumbersome optimization techniques to estimate the parameters of the model, facing diverse numerical issues that are usually magnified by (i) the large number of parameters describing the dynamics of the term structure of interest rates, (ii) the highly non-linear nature of the likelihood function, and/or (iii) the existence of multiple local optima (see, for example, the discussions in Duffee and Stanton, 2012; Hamilton and Wu, 2012).

In a recent paper, Joslin, Singleton and Zhu (2011) (JSZ) propose a new canonical representation of GDTSMs that has substantially lessened many of these numerical challenges faced when estimating GDTSMs by ML. They note that by focusing on "bond" state variables that are linear combinations (i.e., portfolios) of the yields themselves, it is possible to represent the model in a way such that there is a separation between the parameters driving the state variables under the physical measure, $\mathbb{P}$, and those in the risk-neutral distribution, $\mathbb{Q}$. Such separation can be exploited in order to simplify the estimation of the model. In particular, JSZ show that the generic representation of a GDTSM in equations (1), (2) and (5) is observationally equivalent to a canonical model with $r_t = r_\infty^{\mathbb{Q}} + \mathbf{1}_K' \mathbf{z}_t$,

$$\mathbf{z}_{t+1} = \mathbf{\Psi}^{\mathbb{Q}} \mathbf{z}_t + \mathbf{u}_{t+1}^{\mathbb{Q}},$$

where the state variables $\mathbf{z}_t$ are latent, $\mathbf{u}_t^{\mathbb{Q}} \sim iid\ N(0, \mathbf{\Sigma}_z)$, $\mathbf{1}_K$ is a K-dimensional vector of ones, the matrix $\mathbf{\Psi}^{\mathbb{Q}}$ is in ordered real Jordan form with relevant elements (i.e., eigenvalues) collected in the vector $\boldsymbol{\psi}$, and $\mathbf{z}_t$ follows an unrestricted VAR(1) process under the historical measure, $\mathbb{P}$.[11]

By further realizing that $\mathbf{f}_t = \mathbf{P}'\mathbf{y}_t = \mathbf{P}'(\mathbf{a}_z + \mathbf{b}_z \mathbf{z}_t)$ where $\mathbf{a}_z$, $\mathbf{b}_z$ are the constant and factor loadings implied by the JSZ canonical model bond pricing, and using results on invariant transformations of affine term structure models (see Dai and Singleton, 2000), JSZ show that a self-consistent model that uses state variables that are linear combinations

[11] This canonical model is only valid under the assumption of stationarity under $\mathbb{Q}$. While JSZ provide a second canonical model that is valid when $\mathbf{\Phi}^{\mathbb{Q}}$ has a unit root, we still prefer to focus on the parameter $r_\infty^{\mathbb{Q}}$ given its natural economic interpretation and the fact that, in the empirical illustration, we find the largest eigenvalue of $\mathbf{\Phi}^{\mathbb{Q}}$ is less than one.

of yields, $\mathbf{f}_t = \mathbf{P}'\mathbf{y}_t$, must satisfy

$$\begin{aligned} \delta_0 &= r_\infty^{\mathbb{Q}} - \boldsymbol{\delta}'\mathbf{c}, \qquad \boldsymbol{\delta}' = \mathbf{1}_K'\mathbf{D}^{-1}, \\ \boldsymbol{\mu}^{\mathbb{Q}} &= (\mathbf{I} - \boldsymbol{\Phi}^{\mathbb{Q}})\mathbf{c}, \qquad \boldsymbol{\Phi}^{\mathbb{Q}} = \mathbf{D}\boldsymbol{\Psi}^{\mathbb{Q}}\mathbf{D}^{-1}, \end{aligned} \tag{29}$$

where $\mathbf{c} = \mathbf{P}'\mathbf{a}_z$, and $\mathbf{D} = \mathbf{P}'\mathbf{b}_z$. As a result, the risk-neutral dynamics of the yield curve (and therefore, the cross-section of interest rates) is entirely determined by (a) $r_\infty^{\mathbb{Q}}$, the long-run mean of the short rate under $\mathbb{Q}$; (b) $\boldsymbol{\psi}$, the speed of mean reversion of the state variables under $\mathbb{Q}$; and (c) $\boldsymbol{\Sigma}$, the covariance matrix of the innovations from the VAR. On the other hand, the VAR dynamics under $\mathbb{P}$ remain unrestricted.

Given this separation between risk-neutral and physical dynamics, and given the fact that the VAR dynamics remain unrestricted, JSZ propose the following two-step estimator. In the first step, they estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Phi}$ by OLS given that, since the VAR dynamics are unrestricted, OLS recovers the estimates of the conditional mean (Zellner, 1962). In the second step, they estimate the remaining parameters of the model ($r_\infty^{\mathbb{Q}}$, $\boldsymbol{\psi}$, $\boldsymbol{\Sigma}$) via numerical maximization of the likelihood function taking as given the $\mathbb{P}$-dynamics estimates obtained in the first step. Consequently, JSZ report improved convergence and speed of maximum likelihood estimation over other canonical representations.

We note that it is possible to recover the coefficients of the JSZ canonical representation,

$$\boldsymbol{\varphi} = \left(\left(r_\infty^{\mathbb{Q}},\ \boldsymbol{\psi}'\right)', \left\{ vec\left[(\boldsymbol{\mu}\ \boldsymbol{\Phi})' \right] \right\}', \left[vech\left(\boldsymbol{\Sigma}^{1/2} \right) \right]' \right)', \tag{30}$$

using our linear estimation framework. In order to do so, one would start by estimating the model by either OLS or, preferably, CGLS. Second, note from equation (29) that $\boldsymbol{\Psi}^{\mathbb{Q}}$ is related to the Jordan decomposition of $\boldsymbol{\Phi}^{\mathbb{Q}}$. Therefore, an estimate of $\boldsymbol{\Psi}^{\mathbb{Q}}$ can be obtained by finding the real Jordan order form of $\widehat{\boldsymbol{\Phi}}^{\mathbb{Q}}$. In particular, when the eigenvalues in $\boldsymbol{\Psi}^{\mathbb{Q}}$ are real and distinct, $\widehat{\boldsymbol{\psi}}^{\mathbb{Q}}$ can be obtained by a simple spectral decomposition of $\widehat{\boldsymbol{\Phi}}^{\mathbb{Q}} = \widehat{\mathbf{D}} diag(\widehat{\boldsymbol{\psi}}^{\mathbb{Q}})\widehat{\mathbf{D}}^{-1}$. Third, an estimate of the long-run mean of the short rate under $\mathbb{Q}$ can be obtained from $\widehat{r}_\infty^{\mathbb{Q}} = \widehat{\delta}_0 + \widehat{\boldsymbol{\delta}}'(I - \widehat{\boldsymbol{\Phi}}^{\mathbb{Q}})^{-1}\widehat{\boldsymbol{\mu}}^{\mathbb{Q}}$. Fourth, given the structure of the optimization problems in (16) and (27), the estimates of the $\mathbb{P}$-dynamics parameters of the state variables implied by our linear framework also coincide with the OLS estimates of the VAR model in equation (1). Finally, standard errors for the coefficients of the JSZ canonical representation can be obtained using the Delta method and the results in Magnus (1985) regarding differentiation of eigenvalues and eigenvectors.

We see three main advantages of our linear regression approach when compared to JSZ. First, while less computationally demanding, estimates of the JSZ normalization parameters obtained from the self-consistent GLS estimates of the risk-neutral dynamics of the bond factors are asymptotically equivalent to those obtained using the JSZ ML approach (see Proposition 2 in Appendix A).

Second, we note that by imposing the self-consistency restrictions using equation (29) and focusing on their canonical representation of a GDTSM, JSZ are effectively reparameterizing the model in terms of $K - S$ free parameters as in the proofs of Proposition 1 and 2 in the appendix. Yet, their normalization requires the analysis of several different subcases depending on whether all the eigenvalues $\mathbf{\Psi}^{\mathbb{Q}}$ are real and distinct, there are repeated eigenvalues or such eigenvuales are complex. In fact, most researchers only analyze the case of real and distinct eigenvalues (i.e., Duffee, 2011; Bauer, Rudebusch and Wu, 2012; Joslin, Priebsch and Singleton, 2012). On the other hand, one does not need to a priori determine whether the eigenvalues are real and distinct when estimating the model using our linear regression approach given that our method will, in practice, numerically determine which subcase is most empirically relevant.

Third, the improved convergence of the optimization algorithm reported by JSZ relies on the separation between the parameters driving the state variables under the physical measure, $\mathbb{P}$, and those in the risk-neutral distribution, $\mathbb{Q}$. However, when there are restrictions across the parameters of the $\mathbb{P}$ and $\mathbb{Q}$ distributions (i.e., exclusion restrictions on the price of risk parameters), such a separation breaks down, and the ML estimates of the $\mathbb{P}$ parameters are no longer recovered by the OLS estimates of (1). Instead, one has to maximize the likelihood function with respect to all of the parameters of the model in a single step, thus losing the main computational advantage of the JSZ approach. In contrast, we show below (see section 6.1) that our linear approach remains computationally tractable even when one considers the estimation of GDTSMs subject to equality constraints on the structural parameters.

4.2 Hamilton and Wu (2012)

An alternative estimation method to ML estimation is the minimum-chi-square estimation proposed by Hamilton and Wu (2012) (HW). These authors propose to minimize the value of a Wald test statistic for the null hypothesis that the restrictions implied by the no-arbitrage model are consistent with the data. That is, they propose to minimize a quadratic form in the difference between the estimated reduced-form parameters and the reduced-form coefficients implied by the no-arbitrage model:

$$\widehat{\boldsymbol{\varphi}}_{HW} = \arg\min_{\boldsymbol{\varphi}} T \left[\widehat{\boldsymbol{\pi}}_{HW} - \mathbf{p}_{HW}(\boldsymbol{\varphi})\right]' \widehat{\mathbf{V}}_{\pi_{HW}}^{-1} \left[\widehat{\boldsymbol{\pi}}_{HW} - \mathbf{p}_{HW}(\boldsymbol{\varphi})\right], \tag{31}$$

where $\boldsymbol{\varphi}$ is the vector of canonical parameters in equation (30), and where the reduced-form and implied coefficients, $\widehat{\boldsymbol{\pi}}_{HW}$ and $\mathbf{p}_{HW}(\boldsymbol{\varphi})$ respectively, have the subscript HW as a reminder that the set of reduced-form estimates $\widehat{\boldsymbol{\pi}}_{HW}$ used in this estimation method are not necessarily the same as the ones employed when computing the linear estimator discussed in section 2.3. For example, the set of bonds used when estimating the model

using the Hamilton and Wu (2012) approach does not need to be the same as the set employed when using our regression-based methods (we return to this point below).

It is thus straightforward to realize that the HW minimum-chi-square estimator falls within the ALS framework as well. In their case, the distance function is linear in the set of reduced-form coefficients, and the optimal weighting matrix is given by the inverse of the variance of the reduced-form parameters which, if $\boldsymbol{\pi}_{HW}$ is estimated by maximum likelihood, coincides with the (reduced-form) information matrix. We note that, in fact, $\boldsymbol{\pi}_{HW}$ can be viewed as a reparameterization of $\boldsymbol{\pi}$ above in terms of $G-S$ free reduced-form coefficients after imposing the self-consistency restrictions as in the proof of Proposition 2 in Appendix A. Consequently, $\widehat{\mathbf{V}}_{\pi_{HW}}$ is, in general, invertible and $\widehat{\boldsymbol{\varphi}}_{HW}$ thus optimal. Further, HW show that their estimation approach is asymptotically equivalent to ML estimation of $\boldsymbol{\varphi}$, which is a consequence of satisfying the Kodde, Palm and Pfann (1990) conditions under which an ALS estimator is asymptotically equivalent to the ML estimator (see section 3.1).

We note that the CGLS estimator in equation (27) can also be interpreted as minimizing the value of a Wald test statistic for the null hypothesis that the restrictions implied by the no-arbitrage model are consistent with the data (within the set of $\boldsymbol{\theta}$'s that imply a self-consistent model). In particular, the criterion function of our optimal linear estimator resembles a Wald statistic for the null hypothesis that $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \boldsymbol{\gamma}(\boldsymbol{\pi}) - \boldsymbol{\Gamma}(\boldsymbol{\pi})\boldsymbol{\theta} = \mathbf{0}$ which is, in fact, a reparameterization of the null hypothesis considered in HW. Still, it is worth pointing out, again, that the main advantage of our parameterization is that the distance function is linear in $\boldsymbol{\theta}$ and, thus, the solution to the minimization of the chi-square criterion function can be obtained in closed form.

The advantages of our method with respect to the HW approach are similar to the advantages with respect to the ML estimator described in the previous section. For example, the computational tractability of the HW method is lost when there are restrictions across the parameters of the $\mathbb{P}$ and $\mathbb{Q}$ distributions, as one has to minimize (31) directly. Moreover, most of the numerical advantages of the HW approach with respect to direct maximization of the likelihood of the model come from considering exactly identified (unrestricted) models (i.e., models where the number of linear combinations of yields used in the estimation is equal to $N = M + 1$). That is, one has to discard a lot of the information contained in the term structure of interest rates and, by thus reducing the number of bonds used in the estimation of the model, it is possible to incur potentially large efficiency losses. We will illustrate this point below in a Monte Carlo experiment.

4.3 Adrian, Crump and Moench (2012)

In a recent paper, Adrian, Crump and Moench (2012) (ACM) propose a regression-based approach to the estimation of GDTSMs that, as with our ALS approach, completely avoids numerical optimization. Using observable pricing factors, (i.e., principal components of yields), and focusing on bond excess holding period returns, they develop a four-step OLS estimation method. In their first step, they estimate the VAR(1) process in equation (1) to obtain $\boldsymbol{\mu}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$, and decompose pricing factors into predictable components and innovations. In the second step, they estimate the exposures of bond returns to (a) lagged levels of pricing factors $\mathbf{f}_{t-1}$, and (b) contemporaneous pricing factor innovations $\widehat{\mathbf{v}}_t$. In the third step, they estimate the market prices of risk parameters, $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$, from a cross-sectional regression of exposures to contemporaneous pricing factors (i.e., a la Fama-MacBeth). Lastly, they recover parameters of the short rate, δ_0 and $\boldsymbol{\delta}_1$, by regressing the short rate on the pricing factors.

We note that the ACM differs from our method (and JSZ and HW for that matter) in the distributional assumption of the pricing errors. In particular, ACM assume uncorrelated pricing errors on excess returns, instead of uncorrelated pricing errors on the yields. As shown in their paper, when yield pricing errors are uncorrelated, the model has the undesirable feature of generating return pricing errors that are cross-sectionally and serially correlated, thus implying the existence of bond return predictability that is not captured by the pricing factors. While (for space considerations and given that the assumption of uncorrelated pricing errors on excess returns is not standard in the literature) we do not explore the ACM assumption on the pricing errors in this paper, we note that it is possible to handle autocorrelated yield errors within our framework given that OLS estimates of the reduced-form parameters remain consistent and asymptotically normal under such an assumption. In particular, one would simply need to estimate the covariance of the reduced-form parameters in (21) using a method that is robust to autocorrelation (e.g., using Newey and West, 1987).

More importantly, we show in Appendix B how to reinterpret the ACM estimator within the ALS framework, thus easing the comparison with our proposed linear estimators. By doing so, we find three main advantages of our linear regression approach with respect to ACM. First, we show how to impose self-consistency of the model and how to obtain the parameters of the JSZ normalization (features absent in the ACM framework). Second, our regression approach is likely to provide asymptotic efficiency gains with respect to ACM, given that they use an identity weighting matrix and do not impose self-consistency of the model. Finally, we show in the appendix that the system of implicit relationships that defines the ACM estimator is not complete (the number of reduced-form parameters is larger than the dimension of the distance function). This implies

that their approach will not be equivalent to maximum likelihood estimation, even if the self-consistency restrictions were imposed and an optimal weighting matrix was chosen.[12]

5 Monte Carlo simulations

In this section, we carry out a Monte Carlo study to assess the finite-sample properties of the proposed ALS estimators of the parameters of GDTSMs. In addition, we also compare our proposed OLS and CGLS estimators to two of the main approaches to the estimation of GDTSMs described above: the maximum likelihood estimator of Joslin, Singleton and Zhu (2011) and the minimum chi-square estimator of Hamilton and Wu (2012). We leave for further research a comparison with the linear estimator of Adrian, Crump and Moench (2012), because parameter estimates obtained using this approach are not directly comparable to the JSZ, HW and ALS estimators given the different distributional assumption on the measurement errors.

5.1 Design

We simulate 10,000 samples of 25 years of quarterly interest rates ($T = 100$), ranging from one quarter to 15 years ($N = 60$), from a one-factor model using equations (1), (2) and (3) above. We focus on a one-factor model for its simplicity. For example, note that in the one-factor case there is no need to consider complex or repeated eigenvalues when estimating the model subject to the JSZ normalization. Still, we can illustrate most of the properties of our estimators within this simple framework.

To ensure that self-consistency is satisfied in the simulated data, we focus on the JSZ canonical representation of a GDTSM. That is, we use δ_0, $\boldsymbol{\delta}$, $\boldsymbol{\mu}^{\mathbb{Q}}$ and $\boldsymbol{\Phi}^{\mathbb{Q}}$ defined in equation (29), and the following parameter values: $r_\infty^{\mathbb{Q}} = 0.03$, $\boldsymbol{\Psi}^{\mathbb{Q}} = 0.975$, $\boldsymbol{\mu} = 0.0015$, $\boldsymbol{\Phi} = 0.9$, $\boldsymbol{\Sigma} = 0.003$, and $\sigma_\eta = 0.0015$. Such parameter values are chosen to match the empirical characteristics of our data set. In order to capture the level factor that characterizes yield curve data, we choose $\mathbf{f}_t$ to be equal to the one-year yield, which is assumed to be observed without measurement error. That is, we take $\mathbf{P} = \mathbf{e}_4$ where $\mathbf{e}_j$ is a $N \times 1$ vector with a one in the jth position and zeroes in the other. Finally, we draw starting values for the one-year yield (i.e., the factor) from its stationary distribution.

5.2 Results of the simulations

Table 1 reports the results of our Monte Carlo exercise. In this table, OLS and CGLS refer to the linear ALS estimators defined in equations (16) and (27), respectively; HW-ei and

[12] We also note that the use of the ALS framework greatly simplifies obtaining the asymptotic distribution of their linear estimator.

HW-oi refer to the exactly identified and overidentified minimum-chi-square estimators in HW; and, finally, ML refers to the maximum likelihood estimator proposed in JSZ.[13] In the exactly identified case, the HW estimator uses only two points of the yield curve. In particular, we choose to match both the one- and ten-year yields. In addition, as is customary in the literature, we use only a selected set of bond yields when computing the HW and JSZ estimators. In particular, we use yields of maturities two, four, eight, twelve, twenty, forty and sixty quarters when estimating the model using these two methods.[14] On the other hand, we note that our linear estimator uses all the information available in the term structure of interest rates. Therefore, in order to provide a relevant bechmark for our linear framework and to gauge the efficiency loss from discarding data, as traditionally done in the literature, we also compute the ML estimates that use the full span of maturities (ML-all).

The results reported in Table 1 are as follows: *Mean,* the mean (across Monte Carlo replications) of the estimate; *Std,* the standard deviation of the estimate; *EStd,* the sample mean of the estimated asymptotic standard error; *RMSE*, the root-mean-squared error of the estimate; and *CINT-95,* the proportion of times that the true parameter value lies within the 95 per cent asymptotic confidence interval.

Our simulations show that, in terms of bias, the CGLS approach accurately estimates the parameters describing the dynamic evolution of the factor under the risk-neutral measure, $\mathbb{Q}$. On the other hand, the OLS estimates of the canonical parameters seem to be estimated subject to a small downward bias. The bias in the estimated $\mathbb{P}$-parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\Phi}$, is sizable, which is consistent with Bauer, Rudebusch and Wu (2012) and their study of the properties of the HW and JSZ estimators. We note that this is not a problem exclusive to our methodology. In fact, the six estimation methods considered in this Monte Carlo study recover exactly the same OLS estimates of the parameters driving the $\mathbb{P}$-dynamics of the factors. In consequence, our linear estimators will also suffer from the well-known problem that OLS estimates of autoregressive parameters tend to underestimate the persistence of the system in finite samples.

Choosing an optimal weighting matrix and imposing the self-consistency restrictions clearly matters. First, the CGLS estimator has lower variability than the OLS estimator. For example, the standard error of the CGLS estimate of $\boldsymbol{\Phi}^{\mathbb{Q}}$ is larger than the one corresponding to the OLS estimate by a factor of five. Second, the coverage rate of the CGLS estimator is very close to the 95 per cent nominal rate. In contrast, the fact that the OLS estimated standard errors slightly understate the true variability of the estimate,

[13] We use the estimates of the canonical parameters from our OLS approach to initialize the optimization algorithm.

[14] This is the same set of maturities considered by JSZ, augmented by the sixty-quarter (fifteen-year) bond yield.

combined with the slight bias in this estimator, results in non-trivial differences between the empirical coverage rate and the nominal rate of 95 per cent.

Similarly, discarding bonds when estimating the model has an important effect on the efficiency of the estimator. For example, the standard error of the HW-ei (HW-oi) estimator of $r_\infty^{\mathbb{Q}}$ more than triples (doubles) the standard error of the corresponding CGLS (or ML-full) estimate. A similar pattern can be observed when considering the ML estimator of the GDTSM parameters. In fact, the loss of efficiency incurred from focusing on the HW exactly identified estimator is similar to the loss of efficiency incurred from using an identity matrix (versus using the optimal estimator) within our linear framework. On the other hand, and as predicted by asymptotic theory, the properties of our CGLS estimator are almost identical to the properties of the ML estimator that uses the full set of bonds: a result that we find particularly reassuring.

Finally, we briefly analyze the finite-sample properties of the ALS overidentification test. Specifically, we have that the optimized value of the ALS criterion function has an asymptotic χ^2 distribution with degrees of freedom equal to the number of overidentifying restrictions $(G - K)$. We find that this test has a slight tendency to under-reject. In particular, we find that, in our simulations, the empirical rejection rate of a 5 per cent (10 per cent) confidence-level overidentification test is 2.8 per cent (6.9 per cent).

6 Extensions

6.1 Estimation subject to equality constraints

Several recent studies in this literature have considered estimation of GDTSMs subject to certain equality constraints on the structural parameters, including the case where some of the elements of the prices of risk are set to zero (see, e.g., Cochrane and Piazzesi, 2008; Bauer, 2011; Bauer and Diez de los Rios, 2012; Joslin, Priebsch and Singleton, 2012). There are two main reasons to impose such restrictions on the prices of risk. The first concerns the trade-off between model misspecification and sampling uncertainty. As noted by Cochrane and Piazzesi (2008), the risk-neutral distribution can provide a lot of information about the time-series dynamics of the yields. For example, if the price of risk were zero (i.e., agents were risk-neutral), both physical and risk-neutral dynamics would coincide and we could obtain estimates of the parameters driving the time-series process of yields exclusively from the cross-section of interest rates. Since the risk-neutral dynamics can be measured with great precision, one could reduce the sampling uncertainty by following this approach. On the other hand, when the prices of risk are completely unrestricted, no-arbitrage restrictions are irrelevant for the conditional distribution of yields under the physical measure, and thus the cross-section of bond yields does not

contain any information about the time-series properties of interest rates (see JSZ). In this case, notice above that the estimates of the physical dynamic parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\Phi}$, coincide with the OLS estimates of an unrestricted VAR(1) process for $\mathbf{f}_t$. Imposing restrictions on the prices of risk can thus be understood as a trade-off between these two extreme cases.

The second reason concerns the estimated persistence of the data. When the prices of risk are completely unrestricted, the largest eigenvalue of the physical measure $\boldsymbol{\Phi}$ estimated from the VAR(1) representation in equation (1) is usually less than 1.00, with the result that expected future bond yields beyond ten years are almost constant.[15] However, the existence of a level factor in the cross-section of interest rates implies a very persistent process for bond yields under the risk-neutral measure. The largest eigenvalue of $\boldsymbol{\Phi}^{\mathbb{Q}}$ thus tends to be close to or equal to one. By imposing restrictions on the prices of risk, we will be effectively pulling the largest eigenvalue of $\boldsymbol{\Phi}$ closer to that of $\boldsymbol{\Phi}^{\mathbb{Q}}$ so that the physical time-series can inherit more of the high persistence that exists under the risk-neutral measure. In fact, motivated by this persistence issue, Joslin, Priebsch and Singleton (2012) directly force the largest eigenvalue of $\boldsymbol{\Phi}$ to be equal to the largest eigenvalue of $\boldsymbol{\Phi}^{\mathbb{Q}}$.

By having already discussed how to estimate GDTSMs subject to self-consistency restrictions, it is straightforward to see that estimation subject to (additional) equality constraints can naturally be handled in our set-up. In the case of optimal ALS estimation, for example, one just needs to (i) add a new set of restrictions to the estimation problem in (27) and (ii) use the iterative procedure described in section 3.3.

6.2 Unspanned risks

An important development in this literature is the role of unspanned variables.[16] A variable is unspanned if its value is not related to the contemporaneous cross-section of interest rates but it does help forecast both future excess returns on the bonds (i.e., term structure risk premia) and future interest rates. That is, a variable is unspanned if its bond yield factor loadings are equal to zero, yet it helps in explaining the dynamics of interest rates.

Such unspanned variables can be accommodated in our framework in the following way. Specifically, let the pricing factors $\mathbf{f}_t = (\mathbf{f}_{1t}', \mathbf{f}_{2t}')'$ be partitioned into spanned factors

[15] Problems with measuring the persistence of the term structure physical dynamics given the short data samples available have been noted by Ball and Torous (1996), Bekaert, Hodrick and Marshall (1997), Kim and Orphanides (2005), Cochrane and Piazzesi (2008), Bauer (2011), Bauer, Rudebusch and Wu (2012), Duffee and Stanton (2012), and Joslin, Priebsch and Singleton (2012).

[16] See, for example, Cochrane and Piazzesi (2005, 2008), Kim (2007), Cooper and Priestly (2008), Ludvigson and Ng (2009), Orphanides and Wei (2010), Duffee (2011), Chernov and Mueller (2012), Joslin, Priebsch and Singleton (2012).

($\mathbf{f}_{1t}$) and unspanned factors ($\mathbf{f}_{2t}$), and let $\boldsymbol{\delta}$, $\boldsymbol{\mu}^{\mathbb{Q}}$ and $\boldsymbol{\Phi}^{\mathbb{Q}}$ be partitioned accordingly. If (i) the short rates in each country are affine functions of $\mathbf{f}_{1t}$ only ($\boldsymbol{\delta}_2 = \mathbf{0}$) and (ii) we set the right, upper block of the autocorrelation matrix $\boldsymbol{\Phi}^{\mathbb{Q}}$ to zero ($\boldsymbol{\Phi}^{\mathbb{Q}}_{12} = \mathbf{0}$), then $\mathbf{f}_{2t}$ will be unspanned by the cross-section of interest rates. Absent these two assumptions, no-arbitrage pricing would imply that bond yields would be affine functions of all $\mathbf{f}_t$ (cf equation (7) above). While the no-spanning assumptions imply that it is not possible to identify $\boldsymbol{\mu}^{\mathbb{Q}}_2$, $\boldsymbol{\Phi}^{\mathbb{Q}}_{21}$ nor $\boldsymbol{\Phi}^{\mathbb{Q}}_{22}$, given that they affect neither the prices of the bonds nor their risk premia (see JPS for additional details), the linear estimator in section 2.2 (as well as its optimal implementation), can still be easily adapted to recover estimates of $\boldsymbol{\delta}_1$, $\boldsymbol{\mu}^{\mathbb{Q}}_1$ and $\boldsymbol{\Phi}^{\mathbb{Q}}_{11}$ from the coefficients of a cross-sectional regression of yields on $\mathbf{f}_{1t}$.

Our estimation method reveals, however, that caution should be exercised when selecting the factors that explain the cross-section of interest rates. As noted in section 3.1, inference is based on the assumption that $\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\pi}^0, \boldsymbol{\theta}^0) = -\boldsymbol{\Gamma}(\boldsymbol{\pi}^0)$ has a full rank structure. If we wrongly assume that a factor is spanned when, in fact, it is not, we have that the estimates of its factor loadings converge to a vector of zeroes. Consequently, $rank\left[\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\pi}^0, \boldsymbol{\theta}^0)\right] < K$ and the asymptotic approximations to the distribution of the ALS estimator (both OLS and GLS-type implementation) become non-standard. In fact, standard inference still breaks down when the $\mathbf{b}$'s are close to zero (i.e., $\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\pi}^0, \boldsymbol{\theta}^0)$ is almost of reduced rank).[17] This can occur, for example, when there are "hidden" factors in the sense of Duffee (2011): the factor j's loadings are in the neighborhood of zero, so their tiny contemporaneous effect on yields can be lost in the noise that contaminates observed yields. Further, since (the optimal implementation of) our method delivers an estimator that is equivalent to the ML estimator, we suspect that ML estimates of the parameters of the GDTSM might also be subject to weak identification concerns. This situation mirrors the problems that plague the statistical inference in linear factor models when the betas are close to or equal to zero, or the matrix of betas has a near-reduced rank (see Kan and Zhang, 1999; Kleibergen, 2009; Beaulieu, Dufour and Khalaf, 2012).

Since it is well documented that three principal components (i.e., level, slope and curvature) are sufficient to explain at least 99 per cent of the variation in yields (Litterman and Scheinkman, 1991), one might expect any other variable added to the regression of yields on these three principal components to have very little (additional) explanatory power for the cross-section of interest rates. This suggests that any variable beyond the first three PCs should be modelled as unspanned, in order to avoid non-standard asymptotics of the

[17] As noted by Magnusson and Mavroeidis (2010), expanding $\mathbf{G}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)$ around the true value $\boldsymbol{\pi}_0$, we have that $\mathbf{G}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)$ is approximately equal to $\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) + \sqrt{T}\boldsymbol{\Psi}_G$, where $\boldsymbol{\Psi}_G$ is a normally distributed matrix. When $\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0)$ has full rank, $\mathbf{G}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)$ is approximately constant because $\sqrt{T}\boldsymbol{\Psi}_G$ tends to vanish. Hence, $\mathbf{G}'_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)\mathbf{V}_g^{-1}(\boldsymbol{\theta}_0)\mathbf{G}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)$ converges to a non-random invertible matrix. However, such an approximation fails when the matrix $\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0)$ is small compared to $\sqrt{T}\boldsymbol{\Psi}_G$.

parameter estimates. An alternative interesting avenue, left for further research, would be the use of identification-robust methods in the estimation of term structure models. For example, Magnusson (2010) presents ALS versions of the identification-robust tests proposed by Stock and Wright (2000) and Kleibergen (2005) that could be used in our set-up.

6.3 GDTSMs with lags

In a recent paper, Joslin, Le and Singleton (2013) extend the family of GDTSMs to accommodate higher-order dynamics (i.e., beyond the VAR(1) model in equation (1)) in the parameterization of the distribution of yields under $\mathbb{P}$, while preserving the parsimonious factor representation of yields. These authors assume that the factors $\mathbf{f}_t$ follow a VAR(p) under the physical distribution, and a VAR(1) under the risk-neutral measure. Since this can be achieved by assuming that the lags of $\mathbf{f}_t$ are, in essence, unspanned from the cross-section of interest rates, our linear estimator can still be used to estimate this new class of GDTSMs with lags.

6.4 Autocorrelation of the residuals

Our framework can be easily adapted to handle autocorrelation in the measurement errors and/or overlapping in the dynamics under the physical measure. For example, both Cochrane and Piazzesi (2008), and Bauer and Diez de los Rios (2012) focus on annual dynamics of yields estimated using monthly data, which induces a moving average structure on the residuals of the VAR dynamics in equation (1). Similarly, and as noted above, ACM show that uncorrelated pricing errors on excess returns deliver autocorrelated pricing errors on yields. As a difference with maximum likelihood estimation, our framework can naturally handle the presence of autocorrelation in the residuals as long as we estimate the covariance of the reduced-form parameters using a method that is robust to autocorrelation, i.e., using Newey and West (1987).

6.5 Temporal aggregation

Interest rates evolve on a much finer time scale than the frequency of observations typically employed by empirical researchers. While the sampling frequency is often given because collecting data is very expensive in terms of time and money (e.g., output or labor force statistics), this is no longer the case for financial prices. In fact, for interest rates (i.e., bond prices), currently the sampling frequency is, to a large extent, chosen by the researchers. Yet, in the context of the estimation of time-series models, Marcellino (1999) and Diez de los Rios and Sentana (2011), among others, show that this choice has an impact on the

properties of the estimators/tests considered.

In our case, researchers also need to choose which bonds to use in the estimation of GDTSMs. Under our estimation framework, we have that (i) the a_n and $\mathbf{b}_n$'s can be considered as time-series processes indexed by maturity, and (ii) we need the full set of maturities for $n = 1$ to N. This implies that this second choice faced by the researcher is essentially equivalent to choosing a "cross-sectional" or risk-neutral frequency of observation. Thus, paralleling the case of time-aggregation under $\mathbb{P}$, we note that this second choice also has consequences on the properties of the GDTSM estimators.

A first example of how this choice might affect the statistical properties of the GDTSM estimators is related to the efficiency of the parameter estimates. In particular, we note that using the full spectrum of maturities (as prescribed by our methodology) versus using only a sparse selection of yields (as usually done in the literature), or increasing the number of bonds in the estimation, might deliver efficiency gains. To see why this is the case, we can partition $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta}) = [\mathbf{g}_1'(\boldsymbol{\pi}, \boldsymbol{\theta})\ \mathbf{g}_2'(\boldsymbol{\pi}, \boldsymbol{\theta})]' = \mathbf{0}$, and note that, in general, $\mathbf{V}_g(\boldsymbol{\theta}^0) = avar\left[\sqrt{T}\mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta}^0)\right]$ and $\mathbf{V}_{g_1}(\boldsymbol{\theta}^0) = avar\left[\sqrt{T}\mathbf{g}_1(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta}^0)\right]$ are not necessarily the same. Hence, the estimator that uses only a subset of the number of bonds (and therefore, just a subset of the distance functions) and a weighting matrix given by $\mathbf{V}_{g_1}^+$ will not be optimal in this set-up, even if it imposes the self-consistency restrictions.

A second example is related to the identification of the parameters driving the risk-neutral dynamics. Iterating the pricing equations for bond loadings in (7), we have that

$$\mathbf{B}_{n+j}' = \mathbf{B}_n' \left(\boldsymbol{\Phi}^{\mathbb{Q}}\right)^j + \mathbf{B}_j' \qquad j \geq 1. \tag{32}$$

Now assume that, from the N bonds in the original sample, only the jth maturities are retained, $\left\{y_t^{(nj)}\right\}_{n=1}^{N}$, where j is the frequency of (cross-sectional) aggregation. Equation (32) implies that, if we were to estimate this model using this reduced set of bonds, we could only identify the matrix $\digamma^{\mathbb{Q}} = \left(\boldsymbol{\Phi}^{\mathbb{Q}}\right)^j$. However, when $\boldsymbol{\Phi}^{\mathbb{Q}}$ has complex eigenvalues, there are several matrices $\boldsymbol{\Phi}^{\mathbb{Q}}$ that deliver the same $\digamma^{\mathbb{Q}}$ (i.e., there is no bijection between $\boldsymbol{\Phi}^{\mathbb{Q}}$ and $\digamma^{\mathbb{Q}}$).[18] This problem, known as aliasing (see, e.g., Phillips, 1973; Hansen and Sargent, 1981; Bergstrom, 1984), implies that it is not possible to distinguish between parameter structures generating oscillations under $\mathbb{Q}$ at frequencies higher than the interval chosen for the maturities of bonds (i.e., the "cross-sectional" frequency). For example, it will not be possible to identify the parameters driving the $\mathbb{Q}$-dynamics at the monthly frequency if we only have interest rates with quarterly maturities. We leave the study of

[18] To see this point, let $\psi = \rho(\cos\omega \pm i\sin\omega)$ denote a complex eigenvalue of $\boldsymbol{\Phi}^Q$ and its conjugate. By De Moivre's theorem, we have that $\psi^n = \rho^n(\cos n\omega \pm i\sin n\omega)$. Now consider the following complex number in polar form $\widetilde{\psi} = \rho(\cos\widetilde{\omega} \pm i\sin\widetilde{\omega})$ with $\widetilde{\omega} = \omega + 2\pi/n$. Using standard trigonometry results, we have that $\psi^n = \widetilde{\psi}^n$. Thus, two matrices $\boldsymbol{\Phi}^{\mathbb{Q}}$ and $\widetilde{\boldsymbol{\Phi}}^{\mathbb{Q}}$ with the same eigenvectors, and eigenvalues given by ψ and $\widetilde{\psi}$, respectively, will deliver the same $\digamma^{\mathbb{Q}}$.

the econometric issues generated by the temporal aggregation and aliasing problem under the risk-neutral measure for further research, since it is beyond the scope of this paper.

6.6 Small-sample standard errors and bias corrections

Given the computational simplicity of our new linear estimation method, (small-sample) standard errors can be computed using a parametric bootstrap similar to those in JSZ and Hamilton and Wu (2012). The proposed method is as follows:

Step 1 Initialize the artificial sample j of bond factors at their value on the first date from the original sample: $\mathbf{f}_1^{(j)} = \mathbf{P}'\mathbf{y}_1^o$.[19]

Step 2 Generate a sequence $\left\{\mathbf{v}_t^{(j)}\right\}_{t=2}^{T}$ of $N(0, \widehat{\boldsymbol{\Sigma}})$ variables, and then recursively generate a path of the bond factors $\mathbf{f}_t^{(j)} = \widehat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\Phi}}\mathbf{f}_{t-1}^{(j)} + \mathbf{v}_t^{(j)}$ for $t = 2, ..., T$.

Step 3 Generate a sequence of $\left\{\boldsymbol{\eta}_t^{(j)}\right\}_{t=1}^{T}$ of $N(0, \widehat{\boldsymbol{\Omega}})$ variables, and then generate a path of the term structure for the original sample size, T, as $\mathbf{y}_t^{(j)} = \mathbf{a}(\widehat{\boldsymbol{\mu}}^{\mathbb{Q}}, \widehat{\boldsymbol{\Phi}}^{\mathbb{Q}}, \widehat{\boldsymbol{\Sigma}}) + \mathbf{b}(\widehat{\boldsymbol{\Phi}}^{\mathbb{Q}})\mathbf{f}_t^{(j)} + \boldsymbol{\eta}_t^{(j)}$ for $t = 1, ..., T$.

Step 4 Compute an estimate of $\boldsymbol{\theta}$ for the artificial sample j using the linear method described above.

These four steps are repeated J times and small-sample standard errors for the parameters of the model can be computed as the standard deviation of the artificial sequence of bootstrap parameter estimates.

Additionally, since the bond factors are linear combinations of yields, the OLS estimates of the VAR dynamics are likely to be subject to the small-sample biases associated with the extreme persistence found in interest rates (see, i.e., Beckaert, Hodrick and Marshall, 1997; Bauer, Rudebusch and Wu, 2012). Such a problem can be easily dealt with by adapting step 2 of our bootstrap. For example, one could correct these OLS estimates using the analytical formula of Pope (1990), using the bootstrap-after-bootstrap method of Kilian (1998) or using the indirect inference estimator of Bauer, Rudebusch and Wu (2012).

7 Decomposing Canadian yields

In this section, we use the iterative procedure outlined in section 3.3 to estimate a three-factor model and decompose the Canadian ten-year zero-coupon bond yield into an expectations and term premium component. This three-factor specification is designed to

[19] Consistent with the modelling approach of JSZ, one needs to assume that the matrix of "portfolio weights," $\mathbf{P}$, is known and, therefore, remains fixed across bootstrap replications.

capture all the economically interesting variation in both the cross-section of interest rates and bond risk premia, and resembles the Cochrane and Piazzesi (2008) model of the U.S. yield curve.

Our data set consists of end-of-quarter observations over the period March 1986 (1986Q1) to June 2012 (2012Q2) of the term structure of Canadian zero-coupon bond yields obtained from the Bank of Canada website.[20] We consider the full spectrum of maturities from one quarter to fifteen years. By focusing on the Canadian bond market, we expect to alleviate some data-snooping concerns related to the fact that most of the research on yield curve modelling and bond premia focuses on U.S. data.

In order to capture the cross-sectional variation of bond yields, we identify our first two factors with the first two principal components of the term structure of Canadian interest rates. These two factors explain 99.8 per cent of the variation of yields, and have the traditional interpretation of level and slope (Litterman and Scheinkman, 1991). Curvature (i.e., the third principal component) explains only 0.15 per cent of the variability on bond yields. Thus, motivated by parsimony and the discussion on weak identification in section 6.2, we drop the curvature from our study.

Our third factor, on the other hand, builds on the recent evidence documenting the existence of unspanned or nearly spanned factors that are not related to the contemporaneous cross-section of interest rates, but that do help forecast both future excess returns on the bonds (i.e., term structure risk premia) and future interest rates (see Cochrane and Piazzesi 2008, and Duffee, 2011). In particular, we include a return-forecasting factor that is similar in spirit to the one presented in Cochrane and Piazzesi (2005), and that captures all of the economically interesting variation in one-year excess returns for Canadian bonds of all maturities. While this factor can be written as a linear combination of yields and it is fully spanned by bond yields, we find that it has very little (if any) explanatory power for the cross-section of interest rates once level and slope are included in the set of factors. Thus, once again motivated by the discussion on weak identification in section 6.2, we treat the Canadian return-forecasting factor as fully unspanned for the purposes of estimation.

Before turning to the estimates of the model and the decomposition of the Canadian bond yields, we first motivate the choice of such a Cochrane-Piazzesi-like factor for the Canadian yield curve.

[20] Canadian zero-coupon yields are constructed using an exponential spline model (see Bolder, Johnson and Metzler, 2004) for details. The data can be obtained from the Bank of Canada website at http://www.bankofcanada.ca/rates/interest-rates/bond-yield-curves/.

7.1 A return-forecasting factor for Canada

Cochrane and Piazzesi (2005) (CP) show that (i) a linear combination of forward rates predicts annual bond excess holding period returns with R^2 values as high as 0.44, (ii) this single factor has a tent-shaped structure, (iii) this factor captures all of the economically interesting variation in one-year excess returns for bonds of all maturities.

In order to investigate the existence of a similar factor in the Canadian term structure of interest rates, we start by regressing the average (across maturity) annual excess returns at time $t+4$ on forward rates at time t:

$$\frac{1}{14}\sum_{n=2}^{15} rx_{t\to t+4,n} = \gamma_0 + \boldsymbol{\gamma}'\mathbf{g}_t + \epsilon_{t+4}, \tag{33}$$

where $rx_{t\to t+4,n} \equiv \log\left(P_{t+4,n-4}/P_{t,n}\right) - y_{t,4}$ is the annual bond excess holding period returns, and $\mathbf{g}_t$ is a vector of log forward (annual) interest rates, $g_t^{(n\to n+4)} = p_{t,n} - p_{t,n+4}$.[21] Given the overlapping nature of the regression equation (33), we follow CP in computing Newey and West (1987) standard errors with six lags.[22]

The first row in Table 2 reports the estimated values of γ_0 and $\boldsymbol{\gamma}$ for the original choice of five forwards in CP ($g_t^{(n\to n+4)}$ for $n = 0, 4, 8, 12, 16$). While the predictability is weaker than in the original CP paper (i.e., $R^2 = 0.20$ versus 0.44), the Wald test for the hypothesis that $\boldsymbol{\gamma} = \mathbf{0}$ cannot be rejected at standard confidence levels. However, the regression coefficients present an M shape that is suggestive of multicollinearity. For this reason, we follow Sekkel (2011), who tests the robustness of the CP factor across several international markets, in using only the one-, three-, and five-year forwards ($g_t^{(n\to n+4)}$ for $n = 0, 8, 16$) when estimating equation (33). Such results are reported in the second row of Table 2. In this case, the M pattern in the estimated coefficients disappears, and we recover a tent-shaped forecasting factor. Yet, the R^2 decreases to 0.17 given the loss of information from reducing the number of forecasting instruments.

Note that neither the CP nor the Sekkel (2011) specifications incorporate the information that long-dated forwards potentially contain. However, rather than using the full set of forward rates, which could lead again to potential collinearity issues, we focus on one specification of equation (33) that has only five forward rates as regressors. In particular, we use the one-, two-, five-, ten- and fifteen-year forward rates ($g_t^{(n\to n+4)}$ for $n = 0, 4, 16, 36, 56$) as our set of regressors. By doing so, the R^2 increases to 0.46 and the estimated regression coefficients have the desired tent-shaped structure.[23] Thus, we

[21] In particular, $g_t^{(n\to n+4)}$ is the interest rate at time t for loans between time $t+n$ and $t+n+4$.

[22] The results remain the same when we use Hansen and Hodrick (1980) standard errors with three lags (i.e., the order of the MA process of the error term ϵ_{t+4} induced by the overlapping problem).

[23] Following CP, we conduct two robustness exercises. First, in order to address the concern that forward rates on the right-hand side show secular decline over the sample studied, we analyze a specification that

conclude that the information contained in long-dated forwards seems to be important for explaining time variation in Canadian bond premia.

We also verify that this new version of the CP factor captures all of the economically interesting variation in one-year excess returns for Canadian bonds. We do so by following CP once more in comparing the R^2 of an unrestricted regression of individual bond excess holding period returns with the R^2 of the predictive regression that imposes the one-factor structure in expected excess returns. While unreported for the sake of space, we find that the R^2's from both regressions are essentially the same, which indicates that the single-factor structure in expected returns does little damage to its forecast ability. In other words, this Canadian CP factor captures all of the economically interesting variations in one-year excess returns for bonds of all maturities. For these reasons, we use this version of the return-forecasting factor in the remainder of the paper.

7.2 Bias corrections and parameter restrictions

When the prices of risk are completely unrestricted, the estimates of $\mathbb{P}$-parameters coincide with the OLS estimates of an unrestricted VAR(1) process for $\mathbf{f}_t$ and, therefore, suffer from the well-known problem that OLS estimates of autoregressive parameters tend to underestimate the persistence of the system in finite samples. Consequently, the largest eigenvalue of $\mathbf{\Phi}$ estimated from the VAR(1) representation under $\mathbb{P}$ in equation (1) is usually less than 1.00, with the result that expected future bond yields beyond ten years are almost constant.

We tackle this persistence bias in two ways. First, we follow Bauer, Rudebusch and Wu (2012) in replacing the reduced-form OLS estimates of the VAR(1) equation in (1) with bias-corrected estimates. Specifically, we use the analytical approximation for the mean bias in VARs presented in Pope (1990) with the adjustment suggested by Kilian (1998), in order to guarantee that the bias-corrected estimates are stationary. Second, we follow Cochrane and Piazzesi (2008) in forcing one-year expected excess returns to have a single-factor structure, so that time variation in bond premia is driven (solely) by the return-forecasting factor. Thus, by pulling $\mathbf{\Phi}$ to be close to $\mathbf{\Phi}^{\mathbb{Q}}$, we expect the dynamics under $\mathbb{P}$ to inherit more of the high persistence that characterizes the $\mathbb{Q}$-measure.

We now analyze the Cochrane-Piazzesi restrictions in detail. We note that the m-period excess return for holding an n-period zero-coupon bond is given by

$$E_t rx_{t\to t+m}^{(n)} \equiv JIT + \mathbf{B}'_{n-4}\left[\boldsymbol{\lambda}_0^{(m)} + \boldsymbol{\lambda}_1^{(m)}\mathbf{f}_t\right], \tag{34}$$

uses spread information, $g_t^{(n\to n+4)} - r_t$. Second, in order to address the concern that the price at t is common to both the left- and right-hand sides of the regression, we run the regression in equation (33) using forecasting instruments measured at time $t-1$. In both cases, the forecasting power and the tent-shaped pattern are preserved.

where JIT is a (constant) Jensen's inequality term and

$$\boldsymbol{\lambda}_0^{(m)} = \sum_{j=0}^{m-1} \left[\boldsymbol{\Phi}^j \boldsymbol{\mu} - \left(\boldsymbol{\Phi}^{\mathbb{Q}} \right)^j \boldsymbol{\mu}^{\mathbb{Q}} \right], \tag{35}$$

$$\boldsymbol{\lambda}_1^{(m)} = \boldsymbol{\Phi}^j - \left(\boldsymbol{\Phi}^{\mathbb{Q}} \right)^j, \tag{36}$$

where it is easy to see that $\boldsymbol{\lambda}_0^{(m)} = \boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1^{(m)} = \boldsymbol{\lambda}_1$ for $m = 1$. Thus, the risk premia on holding a bond for a year is linear in the factors, $\mathbf{f}_t$, and have three terms: (i) a Jensen's inequality term; (ii) a constant risk premium related to $\boldsymbol{\lambda}_0^{(m)}$; and, (iii) a time-varying risk-premium component where time variation is governed by the parameters in matrix $\boldsymbol{\lambda}_1^{(m)}$. Further, when agents are risk-neutral (i.e., $\boldsymbol{\mu} = \boldsymbol{\mu}^{\mathbb{Q}}$ and $\boldsymbol{\Phi} = \boldsymbol{\Phi}^{\mathbb{Q}}$), we have that $\boldsymbol{\lambda}_0^{(m)}$ and $\boldsymbol{\lambda}_1^{(m)}$ are equal to zero for any holding period, m. Consequently, $\boldsymbol{\lambda}_0^{(m)}$ and $\boldsymbol{\lambda}_1^{(m)}$ play the role of the price of risk parameters when focusing on a holding period $m > 1$.

Note that the single-factor structure of annual bond premia implies that the return-forecasting factor is the only variable driving $E_t rx_{t \to t+4}^{(n)}$. This can be achieved by means of exclusion restrictions on the matrix $\boldsymbol{\lambda}_1^{(4)}$. For example, if we assume that $\mathbf{f}_t = (pc1_t, pc2_t, x_t)'$, where pcj_t is the jth principal component of yields and x_t is the return-forecasting factor, the single-factor expected return model requires the first two columns of $\boldsymbol{\lambda}_1^{(4)}$ to be equal to zero. In fact, we follow Cochrane and Piazzesi (2008) in going a step further and assuming that only the level risk is priced. This imposes the additional restriction that the last two elements of $\boldsymbol{\lambda}_0^{(4)}$ and the last two rows of $\boldsymbol{\lambda}_1^{(4)}$ are equal to zero. That is, we have that the multi-period prices of risk in (35) and (36) take the following form:

$$\boldsymbol{\lambda}_0^{(4)} = \begin{pmatrix} \boldsymbol{\lambda}_{01}^{(4)} \\ 0 \\ 0 \end{pmatrix}, \qquad \boldsymbol{\lambda}_1^{(4)} = \begin{pmatrix} 0 & 0 & \boldsymbol{\lambda}_{13}^{(4)} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Note that both the eigenvalue and exclusion restrictions on $\boldsymbol{\lambda}_0^{(4)}$ and $\boldsymbol{\lambda}_1^{(4)}$ are non-linear with respect to the parameters of the GDTSMs. Consequently, we will use the iterative procedure described in section 3.3 to obtain a set of parameter estimates that satisfies the self-consistency, and Cochrane-Piazzesi restrictions at the same time. We show below that, in fact, these restrictions cannot be rejected by the data.

Finally, we note that, while similar in spirit to the model developed by Cochrane and Piazzesi (2008) for the U.S. term structure, our approach differs from theirs in several aspects. First, they focus directly on the annual dynamics of the factor, and therefore their exclusion restrictions on the prices of risk can be viewed as linear restrictions on the underlying parameters of the model. Second, we do not include curvature in our study due to the concerns of weak identification of the parameters driving the $\mathbb{Q}$-dynamics of such a factor. Third, we exploit the information contained in long-dated futures when

computing our return-forecasting factor. Fourth, we estimate the model subject to not only restrictions on the prices of risk, but also to self-consistency restrictions.

7.3 Parameter estimates

Table 3 reports the parameter estimates of our three-factor GDTSM and, specifically, panel a presents the estimates of the risk-neutral parameters under the JSZ normalization. The largest eigenvalue of $\mathbf{\Phi}^{\mathbb{Q}}$ is almost equal to one (0.9987), a feature needed to explain the existence of the level factor in interest rates. Specifically, long rates are essentially expected future short-term interest rates under the risk-neutral measure corrected by a Jensen's inequality term. Hence, the very persistent dynamics under $\mathbb{Q}$ imply that shocks to the first principal component raise expected future rates in parallel, making sense of the level factor of interest rates (see Cochrane and Piazzesi, 2008). On the other hand, this extreme persistence makes inference about the risk-neutral long-run mean of the short rate very difficult (see Hamilton and Wu, 2012), which is reflected in an estimated $r_{\infty}^{\mathbb{Q}}$ that is unusually high.

Since the cross-section of bond yields is determined by the $\mathbb{Q}$-parameters, Figure 1 presents both the estimated bond yield loadings implied by the affine term structure model and the regression coefficients that one would obtain from projecting bond yields on the first two principal components (i.e., the loadings from a principal-components analysis). The latter coefficients thus provide a natural benchmark to compare the pricing errors implied by our no-arbitrage model. Figure 1 shows that our term structure model is flexible enough to replicate the level and slope shapes of the loadings on individual bond yields obtained from a principal-components analysis. We quantify this remark by looking at the fit of the model. Specifically, the root mean squared pricing error (RMSPE) of the model is 11.81 basis points (bps), which is only marginally worse than the 11.64 bps RMSPE of the reduced-form model. This is confirmed when we look at mean absolute pricing errors (MAPEs). In this case, we have 8.20 versus 7.98 bps. Hence, the loss from imposing the no-arbitrage conditions is minimal: the difference in pricing errors is less than one basis point. In fact, since the risk-neutral measure parameters are pinned down by the cross-section of interest rates, this accuracy in fitting bond yields translates into tight standard errors around these estimates.[24]

Panel b of Table 3, in which we present the estimates of the parameters driving the physical dynamics of the factors, shows that the level factor is very persistent, with a

[24]For completeness, we also note that the OLS estimates of the model imply RMSPE and MAPE of 12.20 and 8.50 bps, respectively. While the model fit is worse than under CGLS estimation, we note that the loss from using the simpler estimator is minimal. This result is even more compelling once we recall that the OLS estimates do not impose self-consistency. Still, for illustrative reasons, we continue using the CGLS estimator in the rest of the paper.

0.98 coefficient, while the slope coefficient decays somewhat faster, with a 0.90 coefficient. Lastly, the return-forecasting factor is the least persistent factor, with a 0.75 coefficient.

The estimated dynamics seem to be consistent with those found by Cochrane and Piazzesi (2008) for the United States. First, a change in the level factor does not seem to have an impact on anything else: the effect of a change in the level factor on the slope is 0.0074, while the effect on the return-forecasting factor is -0.0001. Second, the return-forecasting factor seems to set off a small effect on the level but not on the slope. Lastly, a movement in the slope affects all three factors. For example, the coefficient that measures the effect of a change in the slope on the return-forecasting factor is 0.36. This implies that, even if current expected returns are zero today, one would forecast high future returns when the term structure is upward sloping. We have to be careful in interpreting this result, since this coefficient is measured with great imprecision.

Also, it is interesting to analyze the effects of the restrictions on the prices of risk and the bias correction of the reduced-form parameters on the estimated persistence of this system. In particular, we find that the estimated persistence of a model with unrestricted VAR(1) dynamics under $\mathbb{P}$ is 0.9897. Imposing the CP-like restrictions, on the other hand, increases the estimated persistence only marginally (0.9905). Finally, a combined approach of the price of risk restrictions and bias corrections pushes the persistence up to 0.9938.

We can use the fact that the minimized value of the ALS criterion function has an asymptotic χ^2 distribution to test the validity of the model. Specifically, we have that the dimensionality of the distance function is 198, the number of parameters of interest is 27 and there are six additional restrictions imposed by the CP single-factor structure of excess returns. This leaves 165 degrees of freedom. The 1 per cent (5 per cent) critical value for a $\chi^2(165)$ is 210.17 (195.97), while the minimized value of the ALS criterion is 168.76. Hence, there is no evidence that the restrictions imposed by the model are inconsistent with the data.

We also test the validity of the CP restrictions using an ALS-based distance metric test. In particular, we calculate the difference between the ALS criterion function evaluated at the estimate that imposes the self-consistency and CP restrictions, and the ALS criterion function evaluated at an estimate that only imposes the self-consistency restrictions. We note that both estimates have been computed using the same weighting matrix. This difference has an asymptotic χ^2 distribution with six degrees of freedom (i.e., the number of restrictions). The 1 per cent (5 per cent) critical value for a $\chi^2(6)$ is 16.81 (12.59), while the difference between the ALS criteria is 1.26. Thus, we do not find evidence against the CP restrictions.

7.4 Prices of risk

Estimates of the prices of risk subject to the CP restrictions are shown in Table 4. Panel a focuses on the (quarterly) price of risk coefficients. We find that both the level and slope factors are priced, and that their prices of risk are solely driven by the return-forecasting factor. Still, it is interesting to note that the coefficient on the price of slope risk is very small, which implies that compensation for level risk is the dominant factor in quarterly expected excess returns. Panel b reports the estimated coefficients driving one-year bond premia. Given our set of restrictions, there are only two free parameters: the constant and the coefficient on the return-forecasting factor. We find that they are both significantly different from zero.

A concern regarding the validity of these results relates to the fact that the return-forecasting factor is a generated regressor. In particular, since our standard errors do not account for this problem, it can be the case that our tests are misleading. For this reason, we exploit the numerical tractability of our estimation method to compute bootstrap p-values that correct for the generated regressor problem. In particular, we first estimate the model subject to risk-neutrality (i.e., $\boldsymbol{\mu} = \boldsymbol{\mu}^{\mathbb{Q}}$ and $\boldsymbol{\Phi} = \boldsymbol{\Phi}^{\mathbb{Q}}$) and the self-consistency restrictions using the methods described in section 3.3. Second, using these estimates and the methods described in section 6.6, we generate an artificial sample of yields that satisfies the null that the prices of risk are all zero. Third, using this time-series of pseudo yields, we recompute the return-forecasting factor, estimate the model subject to the Cochrane-Piazzesi and self-consistency restrictions, and save the corresponding t-stats for the null hypothesis that the prices of risk are zero. Finally, we repeat these steps to compute $J = 2000$ bootstrap replications and build the distribution of our test statistics.

We report bootstrap p-values in curly brackets in Table 4. We note that, while the bootstrap p-values are slightly higher than asymptotic p-values (reported in square brackets), we still find that the coefficients on the return-forecasting factor are significantly different from zero.

7.5 Decomposing the Canadian yield curve

In this section, we use the parameter estimates of our three-factor GDTSM to decompose long-term interest rates into expectations of future short-term rates and term premia. In particular,

$$y_{t,n} = \frac{1}{n}\sum_{h=1}^{n} E_t r_{t+h-1} + tp_{t,n}. \tag{37}$$

That is, the n-period interest rate at time t, $y_{t,n}$, is equal to the average path of the short-term rate over the following n periods and a risk-premium component, $tp_{t,n}$, usually called

the term premium. This term premium is the expected return from holding an n-period bond to maturity while financing this investment by selling a sequence of one-year bonds.

We start by analyzing the model's implications for expected future short-term interest rates, the first term in (37). Figure 2 plots the current one-quarter yields, $y_{t,1} = r_t$, and the expected average short-term rate over the next 10 years (60 quarters), $\frac{1}{n}\sum_{h=1}^{n} E_t r_{t+h-1}$, generated by our restricted and bias-corrected three-factor model and, for comparison purposes, a fully unrestricted three-factor model. Both the one-quarter yield and the expectations component generated by these two models tend to drift downwards from the mid-1980s until the end of the sample. In fact, the bias correction on the parameters driving the physical dynamics under $\mathbb{P}$ and the Cochrane-Piazzesi restrictions only seem to induce an almost parallel shift upwards on the projected path of the short rate.

We now analyze the second term in the decomposition in equation (37). Figure 3 plots the term premium implied by the restricted three-factor model. We find that the estimated term premium is countercyclical, increasing rapidly during the two Canadian recessions in our sample, and seems to be driven by both Canadian and global factors. For example, the term premium spikes several times during the mid-1990s, reflecting Canada's loss of its AAA credit-rating status after Moody's downgrade of its sovereign credit rating from Aaa to Aa1 in June 1994, and the uncertainty about the outcome of the fiscal reforms implemented in Canada at the time.[25]

Interestingly, and as recently noted by Bauer and Diez de los Rios (2012), the Canadian term premium also reacts to global events. For example, we observe a surge of the term premium during the U.S. recession that followed the burst of the dotcom bubble in 2000-01, even though output in Canada did not suffer as much. More recently, the Canadian term premium spikes in the last part of the sample due to the European sovereign crisis and the 2011 Japanese earthquake.

8 Final remarks

In this paper, we consider a new linear regression approach to the estimation of GDTSMs that completely avoids numerical optimization methods. Specifically, our linear estimator is an asymptotic least squares estimator that exploits three features that characterize this class of models. First, GDTSMs have a reduced-form representation whose parameters can be easily estimated via OLS regressions. Second, the no-arbitrage assumption upon which GDTSMs are built can be characterized as a set of implicit constraints between these reduced-form parameters and the parameters of interest. Third, this set of restrictions is

[25] On the other hand, the impact of the S&P downgrade by one notch from AAA in October 1992 was small.

linear in the parameters of interest. Consequently, we use the asymptotic least squares estimation principle and infer the parameters of the term structure model by forcing the no-arbitrage constraints, evaluated at the OLS estimates of the reduced-form parameters, to be as close as possible to zero.

In addition, we discuss the advantages of our method with respect to recently suggested approaches to the estimation of GDTSMs. In particular, we find that our asymptotic least squares estimator remains tractable and asymptotically efficient in a variety of situations (i.e., estimation subject to equality constraints) in which the other approaches lose their tractability. Furthermore, we provide a Monte Carlo study to confirm that the tractability of the ALS estimator does not come at the expense of efficiency losses or bad finite-sample properties.

Finally, we use our estimation method to decompose the Canadian ten-year zero-coupon bond yield into an expectations and term premium component. In particular, we use a three-factor specification that is designed to capture all the economically interesting variations in both the cross-section of interest rates and bond risk premia. Moreover, we exploit the numerical tractability of our estimation method to compute bootstrap p-values that correct for the generated regressor problem that is inherent in the estimation of our model.

Our methodology suggests that caution should be exercised when selecting the factors driving the cross-section of interest rates. In particular, our method reveals that when the factor loadings are close to or equal to zero, or the matrix of loadings has a near-reduced rank structure, the asymptotic approximations to the distribution of our asymptotic least squares estimators become non-standard. Further, since (the optimal implementation of) our method delivers an estimator that is equivalent to maximum likelihood estimation, we suspect that maximum likelihood estimates of the parameters of the GDTSM might also be subject to weak identification concerns. Thus, an alternative interesting avenue for further research would be the use of identification-robust methods in the estimation of term structure models.

Another area that deserves further investigation is the application of the asymptotic least squares principle to non-Gaussian term structure models that allow for stochastic volatility given that, in such a case, OLS estimates of the reduced-form parameters remain consistent and asymptotically normal.

References

[1] Adrian, T., R.K. Crump and E. Moench (2012): "Pricing the term structure with linear regressions," *Journal of Financial Economics* (forthcoming).

[2] Aït-Sahalia, Y. and R.L. Kimmel (2010): "Estimating affine multifactor term structure models using closed-form likelihood expansions," *Journal of Financial Economics*, 98, 113–144.

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

[4] Ball, C.A. and W.N. Torous (1996): "Unit roots and the estimation of interest rate dynamics", *Journal of Empirical Finance* 3, 215-238.

[5] Bauer, G.H. and A. Diez de los Rios (2012): "An international dynamic term structure model with economic restrictions and unspanned risks," Bank of Canada Working Paper No. 2012-5.

[6] Bauer, M.D. (2011): "Term premia and the news," Federal Reserve Bank of San Francisco Working Paper No. 2011-03.

[7] Bauer, M.D., G.D. Rudebusch and C. Wu (2012): "Correcting estimation bias in dynamic term structure models," *Journal of Business and Economic Statistics*, 30, 454-467.

[8] Beaulieu, M.C., J.M. Dufour and L. Khalaf (2012): "Identification-robust estimation and testing of the zero-beta CAPM," *Review of Economic Studies* (forthcoming).

[9] Bekaert, G. and R.J. Hodrick (2001): "Expectations hypotheses tests," *Journal of Finance*, 56, 4, 1357-1393.

[10] Bekaert, G., R.J. Hodrick and D. Marshall (1997): "On biases in tests of the expectation hypothesis of the term structure of interest rates," *Journal of Financial Economics* 44, 309-348.

[11] Bergstrom, A. R. (1984): "Continuous time stochastic models and issues of aggregation over time," in Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics: Vol. 2*, Elsevier Science Press (Amsterdam), 1145-1212.

[12] Bolder, D.J., G. Johnson and A. Metzler (2004): "An empirical analysis of the Canadian term structure of zero-coupon interest rates," Bank of Canada Working Paper No. 2004-48.

[13] Chamberlain, G. (1982): "Multivariate regression models for panel data," *Journal of Econometrics*, 18, 5-46.

[14] Chen, R.R. and L. Scott (1993): "Maximum likelihood estimation for a multifactor equilibrium model of the term structure of interest rates," *Journal of Fixed Income,* 3, 14-31.

[15] Chernov, M. and P. Mueller (2012): "The term structure of inflation expectations," *Journal of Financial Economics*, 106, 367–394.

[16] Christensen, J.H.E., F.X. Diebold and G.D. Rudebusch (2011): "The affine arbitrage-free class of Nelson-Siegel term structure models," *Journal of Econometrics*, 164, 4-20.

[17] Cochrane, J. and M. Piazzesi (2005): "Bond risk premia," *American Economic Review*, 95, 138-60.

[18] Cochrane, J. and M. Piazzesi (2008): "Decomposing the yield curve," mimeo, University of Chicago.

[19] Cooper, I. and R. Priestly (2008): "Time-varying risk premiums and the output gap," *Review of Financial Studies* 22, 2801-2833.

[20] Dai, Q. and K.J. Singleton (2000): "Specification analysis of affine term structure models," *Journal of Finance*, 55, 1943-1978.

[21] Dai, Q. and K.J. Singleton (2002): "Expectations puzzles, time-varying risk premia, and affine models of the term structure," *Journal of Financial Economics*, 63, 415-411.

[22] Diez de los Rios, A. and E. Sentana (2011): "Testing uncovered interest parity: a continuous-time approach," *International Economic Review*, 52, 1215-1251.

[23] Duffee, G.R. (2011): "Information in (and not in) the term structure," *Review of Financial Studies* 24, 2895-2934.

[24] Duffee, G.R. and R. Stanton (2012): "Estimation of dynamic term structure models," *Quarterly Journal of Finance,* 2, 1-51.

[25] Gourieroux, C. and A. Monfort (1995): *Statistics and econometric models*, Cambridge University Press (Cambridge).

[26] Gourieroux, C., A. Monfort and A. Trognon (1982): "Nonlinear asymptotic least squares," INSEE Document de travail no. 8207.

[27] Gourieroux, C., A. Monfort and A. Trognon (1985): "Moindres carres asymptotiques," Annales de l'INSEE 58, 91-122.

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

[29] Hansen, L.P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029-1054.

[30] Hansen, L.P. and R.J. Hodrick (1980): "Forward exchange rates as optimal predictors of future spot rates, an econometric analysis," *Journal of Political Economy,* 88, 829-53.

[31] Hansen, L.P. and T.J. Sargent (1981): "Identification of continuous-time rational expectations models from discrete-time data," in L.P. Hansen and T.J. Sargent (eds.), *Rational Expectations Econometrics*, Westview Press (Boulder), 219-236.

[32] Joslin, S., A. Le and K.J. Singleton (2012): "Why Gaussian macro-finance term structure models are (nearly) unconstrained factor-VARs," *Journal of Financial Economics* (forthcoming).

[33] Joslin, S., A. Le and K.J. Singleton (2013): "Gaussian macro-finance term structure models with lags," *Journal of Financial Econometrics* (forthcoming).

[34] Joslin, S., M. Priebsch and K.J. Singleton (2012): "Risk premiums in dynamic term structure models with unspanned macro risks," Stanford University, mimeo.

[35] Joslin, S., K.J. Singleton and H. Zhu (2011): "A new perspective on Gaussian DTSMs," *Review of Financial Studies,* 24, 926-970.

[36] Jungbacker, B. and S.J. Koopman (2008): "Likelihood-based analysis for dynamic factor models," Tinbergen Institute Discussion Paper No. TI 2008-0007/4.

[37] Kan, R. and C. Zhang (1999): "Two-pass tests of asset pricing models with useless factors," *Journal of Finance*, 54, 204-235.

[38] Kilian, L. (1998): "Small-sample confidence intervals for impulse-response functions," *Review of Economics and Statistics*, 80, 218-230.

[39] Kim, D. (2007): "Challenges in macro-finance modeling," BIS Working Papers No. 240.

[40] Kim, D. and A. Orphanides (2005): "Term structure estimation with survey data on interest rate forecasts," Federal Reserve Board, mimeo.

[41] Kim, D. and J.H. Wright (2005): "An arbitrage-free three-factor term structure model and the recent behavior of long-term yields and distant-horizon forward rates," Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series No. 2005-33.

[42] Kleibergen, F. (2005): "Testing parameters in GMM without assuming that they are identified," *Econometrica*, 73, 1103-23.

[43] Kleibergen, F. (2009): "Tests of risk premia in linear factor models," *Journal of Econometrics*, 149, 149-173.

[44] Kodde, D.A., F.C. Palm and G.A. Pfann (1990): "Asymptotic least-squares estimation efficiency considerations and applications," *Journal of Applied Econometrics*, 5, 229-243.

[45] Litterman, R. and J.A. Scheinkman (1991): "Common factors affecting bond returns," *Journal of Fixed Income*, June, 54-61.

[46] Ludvigson, S.C. and S. Ng. (2009): "Macro factors in bond risk premia," *Review of Financial Studies*, 22, 5027-5067.

[47] Lütkepohl, H. (1989): "A note on the asymptotic distribution of impulse response functions of estimated VAR models with orthogonal residuals," *Journal of Econometrics*, 42, 371-376.

[48] Magnus, J. (1985): "On differentiating eigenvalues and eigenvectors," *Econometric Theory*, 1, 179-191.

[49] Magnusson, L.M. (2010): "Inference in limited dependent variable models robust to weak identification," *Econometrics Journal,* 13, S56-S79.

[50] Magnusson, L.M. and S. Mavroeidis (2010): "Identification-robust minimum distance estimation of the new Keynesian Phillips curve," *Journal of Money, Credit and Banking*, 42, 465-481.

[51] Marcellino, M. (1999): "Some consequences of temporal aggregation in empirical analysis," *Journal of Business and Economic Statistics*, 17, 129-36.

[52] Newey, W.K. and D.L. McFadden (1994): "Large sample estimation and hypothesis testing," in R.F. Engle and D.L. McFadden (eds), *Handbook of Econometrics: Vol. 4*, Elsevier Science Press (Amsterdam), 2111-2245.

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

[54] Orphanides, A. and M. Wei (2010): "Evolving macroeconomic perceptions and the term structure of interest rates," Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series No. 2010-01.

[55] Peñaranda, F. and E. Sentana (2012): "Spanning tests in return and stochastic discount factor mean-variance frontiers: a unifying approach," *Journal of Econometrics,* 170, 303-324.

[56] Phillips, P.C.B. (1973): "The problem of identification in finite parameter continuous time models," *Journal of Econometrics*, 1, 351–262.

[57] Pope, A. L. (1990): "Biases of estimators in multivariate non-Gaussian autoregressions," *Journal of Time Series Analysis*, 11, 249-258.

[58] Sekkel, R. (2011): "International evidence on bond risk premia," *Journal of Banking and Finance*, 35, 174-181.

[59] Sentana, E. (2002): "Did the EMS reduce the cost of capital?," *Economic Journal*, 112, 786-809.

[60] Stock, J.H. and J.H. Wright (2000): "GMM with weak identification," *Econometrica*, 68, 1055-96.

[61] Zellner, A. (1962): "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias," *Journal of the American Statistical Association*, 57, 348-68.

Table 1
Finite-sample properties of GDTSMs estimators

		$100 \times r_\infty^{\mathbb{Q}}$	$\boldsymbol{\Psi}^{\mathbb{Q}}$	$100 \times \boldsymbol{\mu}$	$\boldsymbol{\Phi}$	$100 \times \boldsymbol{\Sigma}$
	True values	3.00	0.9750	0.15	0.900	0.300
OLS	Mean	2.91	0.9734	0.21	0.860	0.296
	Std	0.14	0.0022	0.09	0.056	0.021
	EStd	0.13	0.0021	0.08	0.050	0.021
	RMSE	0.17	0.0027	0.11	0.070	0.022
	CINT-95	82.5%	88.8%	91.1%	90.7%	93.2%
CGLS	Mean	2.99	0.9750	0.21	0.860	0.297
	Std	0.04	0.0004	0.09	0.056	0.013
	EStd	0.04	0.0004	0.08	0.050	0.013
	RMSE	0.04	0.0004	0.11	0.070	0.013
	CINT-95	93.8%	95.6%	91.1%	90.7%	94.9%
HW-ei	Mean	3.00	0.9749	0.21	0.860	0.296
	Std	0.15	0.0026	0.09	0.056	0.021
	EStd	0.15	0.0025	0.08	0.050	0.021
	RMSE	0.15	0.0026	0.11	0.070	0.021
	CINT-95	94.5%	94.8%	91.1%	90.7%	93.2%
HW-oi	Mean	3.00	0.9750	0.21	0.863	0.297
	Std	0.09	0.0014	0.09	0.056	0.019
	EStd	0.09	0.0013	0.08	0.050	0.019
	RMSE	0.09	0.0014	0.11	0.070	0.019
	CINT-95	93.9%	95.3%	91.1%	90.7%	93.1%
ML	Mean	3.00	0.9750	0.21	0.860	0.297
	Std	0.09	0.0014	0.09	0.056	0.019
	EStd	0.09	0.0014	0.08	0.050	0.019
	RMSE	0.09	0.0014	0.11	0.070	0.019
	CINT-95	94.1%	95.4%	91.1%	90.7%	93.2%
ML-all	Mean	3.00	0.9750	0.21	0.863	0.299
	Std	0.04	0.0004	0.09	0.053	0.013
	EStd	0.04	0.0004	0.08	0.050	0.012
	RMSE	0.04	0.0004	0.11	0.065	0.013
	CINT-95	94.6%	95.1%	91.1%	90.7%	94.5%

Table 2
Cochrane and Piazzesi regressions

	constant	$g_t^{(0\to4)}$	$g_t^{(4\to8)}$	$g_t^{(8\to12)}$	$g_t^{(12\to16)}$	$g_t^{(16\to20)}$	$g_t^{(36\to40)}$	$g_t^{(56\to60)}$	R^2	Wald
Original CP (2005)	-0.90	-5.00	10.70	-21.72	26.65	-10.28			0.20	25.34
	(3.00)	(1.42)	(5.21)	(10.26)	(11.03)	(5.16)				[<0.001]
Sekkel (2011)	-1.22	-2.78		3.00		0.19			0.17	10.22
	(2.91)	(0.92)		(2.55)		(2.05)				[0.02]
Long-dated forwards	-0.25	-2.95	0.79			5.94	2.55	-5.48	0.46	82.92
	(2.97)	(1.17)	(1.90)			(1.64)	(0.87)	(1.12)		[<0.001]

Note: Data are sampled quarterly from 1986Q1 to 2012Q2. Newey and West (1987) asymptotic standard errors are given in parentheses.

Table 3
Parameter estimates

Panel a: Risk-neutral parameters

	$100 \times r_\infty^{\mathbb{Q}}$	ψ_1	ψ_2
JSZ coefficient	10.465	0.9989	0.8809
	(1.730)	(0.0002)	(0.0016)
	[<0.001]	[<0.001]	[<0.001]

Panel b: Physical parameters

		$\boldsymbol{\Phi}$		
	$\boldsymbol{\mu}$	level	slope	CP
level	0.0358	0.9846	0.2371	-0.0760
	(0.1351)	(0.0046)	(0.0449)	(0.0222)
	[0.791]	[<0.001]	[<0.001]	[0.001]
slope	0.0623	0.0074	0.8941	0.0011
	(0.0046)	(0.0002)	(0.0019)	(0.0004)
	[<0.001]	[<0.001]	[<0.001]	[0.003]
CP	0.0323	-0.0001	0.3581	0.7577
	(0.8287)	(0.0458)	(0.3982)	(0.0636)
	[0.969]	[0.998]	[0.369]	[<0.001]

Panel c: Innovation variance parameters

	$\boldsymbol{\Sigma}^{1/2}$		
	level	slope	CP
level	0.9436	0	0
	(0.0816)	-	-
	[<0.001]	-	-
slope	-0.1073	1.0575	0
	(0.0796)	(0.1050)	-
	[0.178]	[<0.001]	-
CP	0.2664	2.1517	1.8002
	(0.1778)	(0.3362)	(0.3101)
	[0.134]	[<0.001]	[<0.001]

Note: Data are sampled quarterly from 1986Q1 to 2012Q2. Asymptotic standard errors are given in parentheses, and asymptotic p-values in square brackets.

Table 4
Price of risk estimates

Panel a: Quarterly bond premia

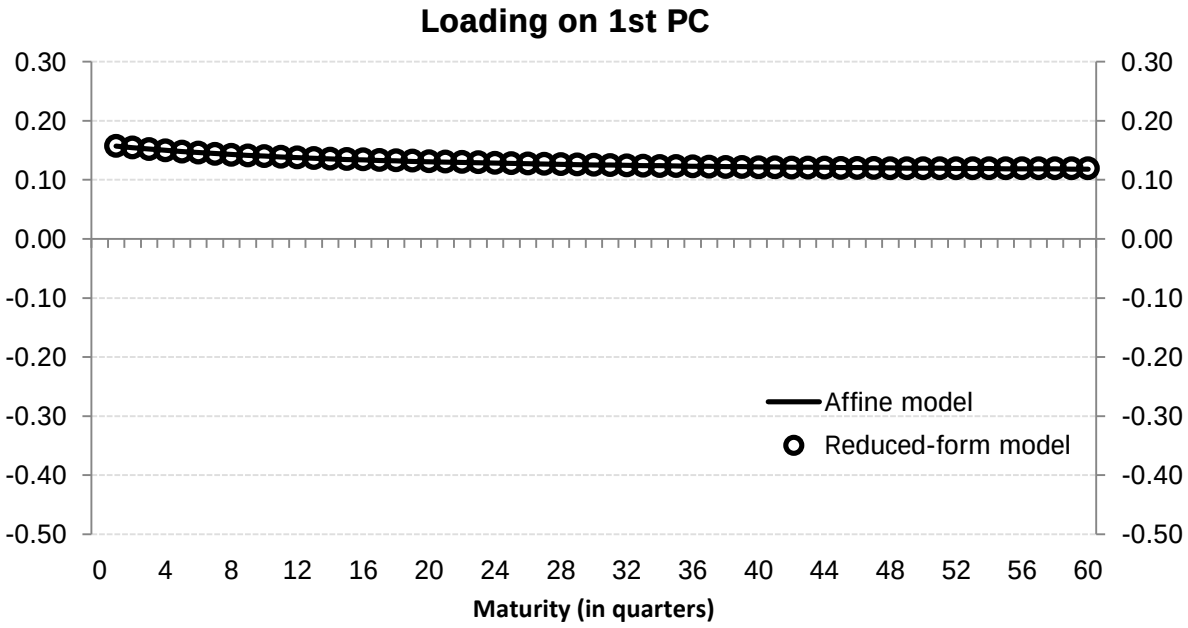
		$\boldsymbol{\lambda}_1$		
	$\boldsymbol{\lambda}_0$	level	slope	CP
level	0.0582	-0.0002	0.03823	-0.0760
	(0.1351)	(0.0046)	(0.0449)	(0.0222)
	[0.667]	[0.960]	[0.395]	[<0.001]
	{0.823}	{0.973}	{0.526}	{0.011}
slope	-0.0007	0.0001	-0.0009	0.0011
	(0.0023)	(0.0001)	(0.0011)	(0.0004)
	[0.764]	[0.939]	[0.404]	[0.003]
	{0.865}	{0.957}	{0.534}	{0.018}

Panel b: Annual bond premia

		$\boldsymbol{\lambda}_1^{(4)}$		
	$\boldsymbol{\lambda}_0^{(4)}$	level	slope	CP
level	0.2216	0	0	-0.2037
	(0.3823)	-	-	(0.0516)
	[0.562]	-	-	[<0.001]
	{0.816}	-	-	{0.008}
slope	0	0	0	0
	-	-	-	-
	-	-	-	-
	-	-	-	-

Note: Data are sampled quarterly from 1986Q1 to 2012Q2. Asymptotic standard errors are given in parentheses, asymptotic p-values in square brackets and bootstrap p-values in curly brackets.

Figure 1: Bond factor loadings: affine term structure versus OLS estimates



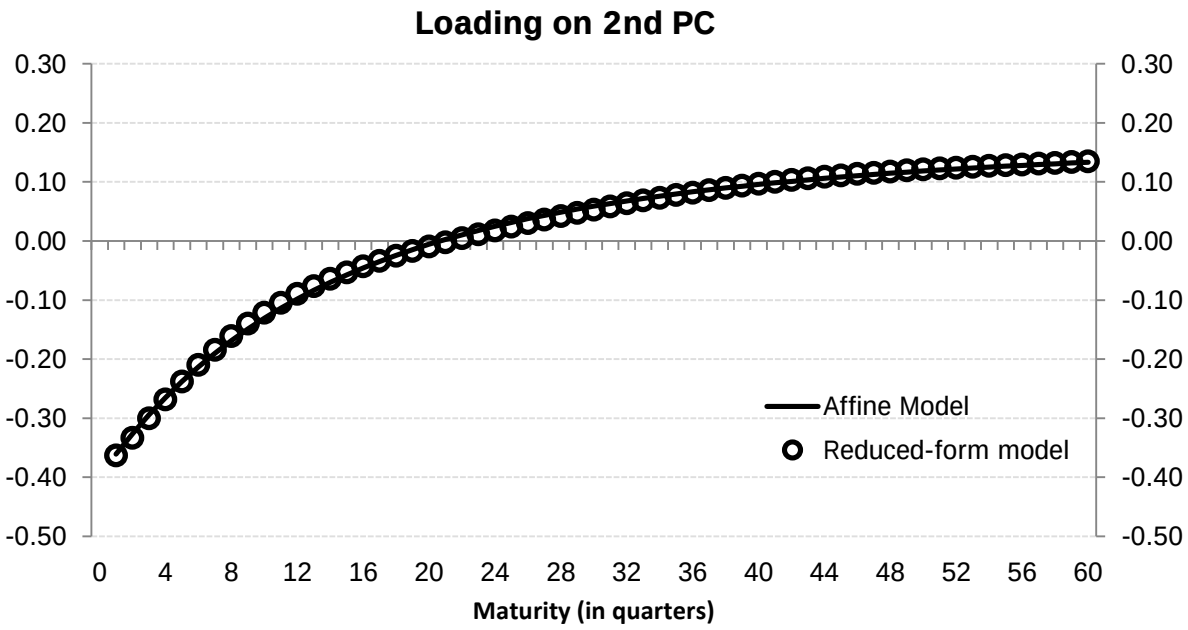


Figure 2: Expected average path of the Canadian one-quarter yield over the next ten years

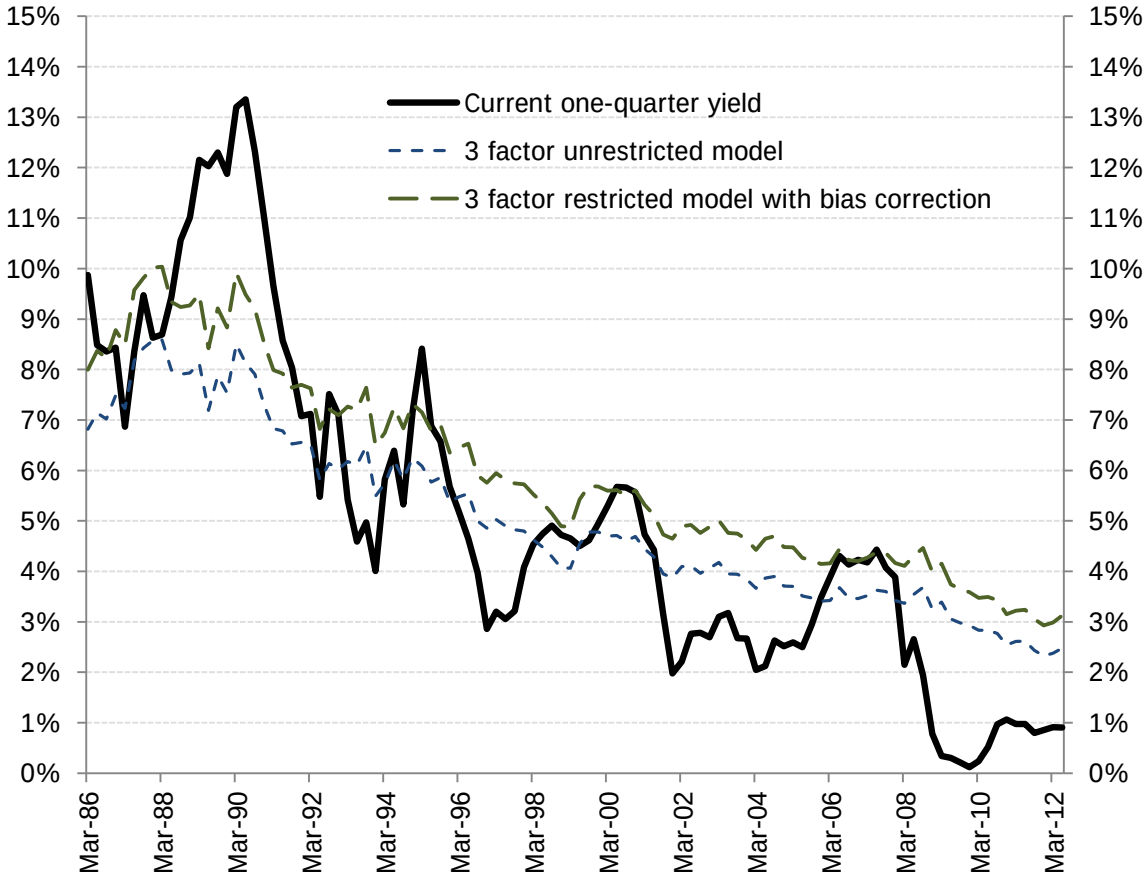
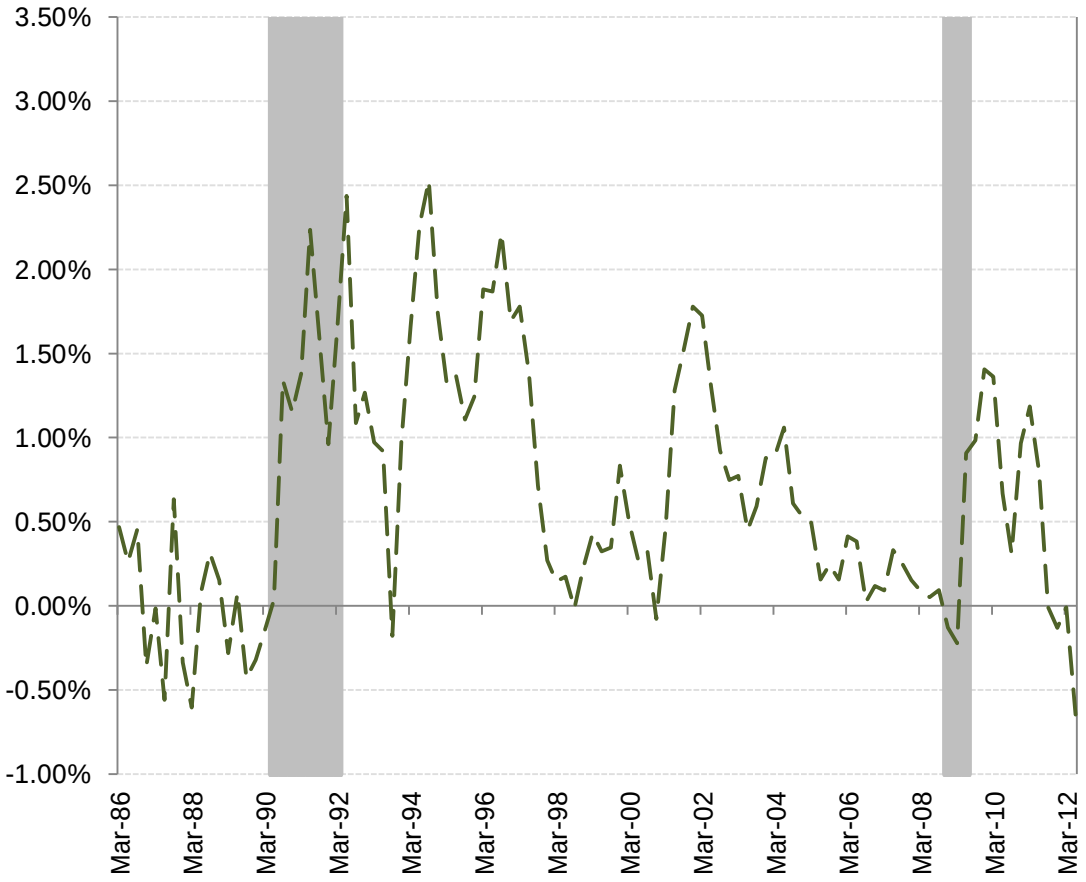


Figure 3: Term premium on Canadian ten-year yields



Appendix

A Propositions

Proposition 1. Let

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} T\mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta})'\mathbf{V}_g^+(\boldsymbol{\theta}^0)\mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta}) \qquad \text{s.t. } \mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$$

denote the optimal ALS estimator of the $(K \times 1)$ vector of unknown parameter $\boldsymbol{\theta}$ defined by the $(G \times 1)$ system of implicit equations $\mathbf{g}(\widehat{\boldsymbol{\pi}}, \boldsymbol{\theta}) = \mathbf{0}$ where $\widehat{\boldsymbol{\pi}}$ denote a strongly consistent and asymptotically normal estimator of the auxiliary parameters $\boldsymbol{\pi}$. Under the usual regularity conditions, together with assumptions 1 and 2 in the main text,

(i) $\widehat{\boldsymbol{\theta}}$ is identified along the manifold $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$

(ii) $\widehat{\boldsymbol{\theta}}$ is asymptotically efficient, in the sense that the difference between the asymptotic covariance matrix of this estimator and any other ALS estimator based on the same set of implicit equations and the same consistent and asymptotically normal estimator of the auxiliary parameters is negative semidefinite regardless of the choice of the weighting matrix $\mathbf{W}_T$ or whether the equality restrictions $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ are imposed.

(iii) the minimized value of the ALS criterion function, $T\mathbf{g}(\widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}})'\mathbf{V}_g^+(\boldsymbol{\theta}^0)\mathbf{g}(\widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}})$, has an asymptotic χ^2 distribution with degrees of freedom equal to $G - K$.

Proof. This proof closely follows Peñaranda and Sentana (2012), where further details can be found.

Let the spectral decomposition of $\mathbf{V}_g(\boldsymbol{\theta}^0)$ be given by

$$\mathbf{V}_g(\boldsymbol{\theta}^0) = \begin{pmatrix} \mathbf{T}_1 & \mathbf{T}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{T}_1' \\ \mathbf{T}_2' \end{pmatrix} = \mathbf{T}_1\boldsymbol{\Lambda}\mathbf{T}_1',$$

where $\boldsymbol{\Lambda}$ is a $(G - S) \times (G - S)$ positive definite diagonal matrix; and, without loss of generality, let $\mathbf{V}_g^+(\boldsymbol{\theta}^0)$ be the Moore-Penrose[26] generalized inverse of $\mathbf{V}_g(\boldsymbol{\theta}^0)$:

$$\mathbf{V}_g^+(\boldsymbol{\theta}^0) = \mathbf{T}_1\boldsymbol{\Lambda}^{-1}\mathbf{T}_1'.$$

In order to simplify the notation, it is convenient to reparameterize the parameter space into the alternative K parameters $\boldsymbol{\alpha}$ $(S \times 1)$ and $\boldsymbol{\beta}$ $((K - S) \times 1)$ such that

$$\mathbf{R}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\alpha}' & \boldsymbol{\beta}' \end{pmatrix}',$$

[26] As noted by Peñaranda and Sentana (2012), it is possible to show that the results in this proposition hold for any generalized inverse of $\mathbf{V}_g(\boldsymbol{\theta}^0)$. While a similar argument would apply here, we focus on the Moore-Penrose generalized inverse for simplicity.

where the first S elements of $\mathbf{R}(\boldsymbol{\theta})$ are such $\boldsymbol{\alpha} = \mathbf{r}(\boldsymbol{\theta})$. In particular, we can choose $\mathbf{R}(\boldsymbol{\theta})$ to be a regular transformation of $\boldsymbol{\theta}$ on an open neighbourhood of $\boldsymbol{\theta}^0$. Further, let $\mathbf{q}[\mathbf{R}(\boldsymbol{\theta})] = \boldsymbol{\theta}$ be the corresponding inverse transformation of $\mathbf{R}(\boldsymbol{\theta})$ that recovers $\boldsymbol{\theta}$ back. Let the Jacobians of the inverse transformation be given by

$$\mathbf{Q}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{\partial \mathbf{q}(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial(\boldsymbol{\alpha}',\boldsymbol{\beta}')} = \left[\begin{array}{cc} \mathbf{Q}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha},\boldsymbol{\beta}) & \mathbf{Q}_{\boldsymbol{\beta}}(\boldsymbol{\alpha},\boldsymbol{\beta}) \end{array}\right].$$

This transformation allows us to impose the parametric restrictions $\mathbf{r}(\boldsymbol{\theta}) = \boldsymbol{\alpha} = \mathbf{0}$ by simply working with the smaller set of parameters $\boldsymbol{\beta}$ and the distance functions $\mathbf{g}[\widehat{\boldsymbol{\pi}}, \mathbf{q}(\mathbf{0},\boldsymbol{\beta})]$. Thus the optimal ALS estimator can be defined as $\widehat{\boldsymbol{\theta}} = \mathbf{q}(\mathbf{0},\widehat{\boldsymbol{\beta}})$ where

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} T\mathbf{g}[\widehat{\boldsymbol{\pi}}, \mathbf{q}(\mathbf{0},\boldsymbol{\beta})]' \mathbf{V}_g^{+}(\boldsymbol{\theta}^0)\mathbf{g}[\widehat{\boldsymbol{\pi}}, \mathbf{q}(\mathbf{0},\boldsymbol{\beta})].$$

(i) Since $(\mathbf{T}_1, \mathbf{T}_2)$ is an orthogonal matrix, and the $rank[\mathbf{Q}(\boldsymbol{\alpha},\boldsymbol{\beta})] = K$ given that $\mathbf{R}(\boldsymbol{\theta})$ is a regular transformation of $\boldsymbol{\theta}$ on open neighbourhood of $\boldsymbol{\theta}^0$, we have by the inverse function theorem that

$$rank\left[\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\pi}^0,\boldsymbol{\theta}^0)\right] = rank\left[\begin{array}{cc} \mathbf{T}_1'\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\pi}^0,\boldsymbol{\theta}^0)\mathbf{Q}_{\boldsymbol{\alpha}}(\mathbf{0},\boldsymbol{\beta}) & \mathbf{T}_1'\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\pi}^0,\boldsymbol{\theta}^0)\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}) \\ \mathbf{T}_2'\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\pi}^0,\boldsymbol{\theta}^0)\mathbf{Q}_{\boldsymbol{\alpha}}(\mathbf{0},\boldsymbol{\beta}) & \mathbf{T}_2'\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\pi}^0,\boldsymbol{\theta}^0)\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}) \end{array}\right] = K. \tag{38}$$

Note now that Assumptions 1 and 2 imply that $\boldsymbol{\Xi}'[\mathbf{l}(\mathbf{0},\boldsymbol{\beta})]\sqrt{T}\mathbf{g}[\widehat{\boldsymbol{\pi}}, \mathbf{q}(\mathbf{0},\boldsymbol{\beta})] \xrightarrow{p} \mathbf{0}$ for all $\boldsymbol{\beta}$ in the neighbourhood. So, by differentiating this random process with respect to $\boldsymbol{\beta}$ and evaluating the derivatives at the true value $\boldsymbol{\beta}^0$ we have, by the continuous mapping theorem, that

$$\sqrt{T}\left\{\mathbf{g}\left[\widehat{\boldsymbol{\pi}}, \mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right] \otimes \mathbf{I}_S\right\} \frac{\partial vec\left\{\boldsymbol{\Xi}'\left[\mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right]\right\}}{\partial \boldsymbol{\beta}'} + \boldsymbol{\Xi}'\left[\mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right]\sqrt{T}\frac{\partial \sqrt{T}\mathbf{g}\left[\widehat{\boldsymbol{\pi}}, \mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right]}{\partial \boldsymbol{\beta}'} \xrightarrow{p} \mathbf{0}$$

$$\left\{\mathbf{g}\left[\widehat{\boldsymbol{\pi}}, \mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right] \otimes \mathbf{I}_S\right\} \frac{\partial vec\left\{\boldsymbol{\Xi}'\left[\mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right]\right\}}{\partial \boldsymbol{\beta}'} + \boldsymbol{\Xi}'\left[\mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right]\frac{\partial \sqrt{T}\mathbf{g}\left[\widehat{\boldsymbol{\pi}}, \mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right]}{\partial \boldsymbol{\beta}'} \xrightarrow{p} \mathbf{0},$$

since $1/\sqrt{T} \xrightarrow{p} \mathbf{0}$.

Using the chain rule, the previous expression can be written as

$$\left\{\mathbf{g}\left[\widehat{\boldsymbol{\pi}}, \mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right] \otimes \mathbf{I}_S\right\} \frac{\partial vec\left\{\boldsymbol{\Xi}'\left[\mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right]\right\}}{\partial \boldsymbol{\theta}'}\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0) + \boldsymbol{\Xi}'\left[\mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right]\mathbf{G}_{\boldsymbol{\theta}}\left[\widehat{\boldsymbol{\pi}}, \mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right]\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0),$$

which implies that

$$\boldsymbol{\Xi}'\left[\mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right]\mathbf{G}_{\boldsymbol{\theta}}\left[\mathbf{l}(\mathbf{0},\boldsymbol{\beta}^0)\right]\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0) = \mathbf{0}$$

with $\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathbf{G}_{\boldsymbol{\theta}}[\mathbf{p}(\boldsymbol{\theta}),\boldsymbol{\theta}]$ and where we have used that $\mathbf{g}\left[\widehat{\boldsymbol{\pi}}, \mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right] \xrightarrow{p} \mathbf{g}\left[\boldsymbol{\pi}^0, \mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right] = \mathbf{g}(\boldsymbol{\pi}^0,\boldsymbol{\theta}^0) = \mathbf{0}$, and that $\mathbf{G}_{\boldsymbol{\theta}}\left[\widehat{\boldsymbol{\pi}}, \mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right] \xrightarrow{p} \mathbf{G}_{\boldsymbol{\theta}}\left[\mathbf{p}(\boldsymbol{\theta}^0), \mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right] = \mathbf{G}_{\boldsymbol{\theta}}\left[\mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right]$.

Finally, note that since $\mathbf{T}_2'\mathbf{V}_g(\boldsymbol{\theta}^0) = 0$, then $\mathbf{T}_2$ must be a full-column rank linear transformation of $\boldsymbol{\Xi}(\boldsymbol{\theta})$. Therefore, it has to be that

$$\mathbf{T}_2'\mathbf{G}_{\boldsymbol{\theta}}\left[\mathbf{q}(\mathbf{0},\boldsymbol{\beta}^0)\right]\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0) = \mathbf{0},$$

which implies that $rank\left[\mathbf{Q}_1'\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\pi}^0,\boldsymbol{\theta}^0)\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta})\right] = K - S$ for (38) to be true. Thus, after imposing that $\boldsymbol{\alpha} = \mathbf{0}$, the reduced system of distance functions $\mathbf{Q}_1'\mathbf{g}\left[\widehat{\boldsymbol{\pi}}, \mathbf{q}(\mathbf{0},\boldsymbol{\beta})\right]$ will first-order identify $\boldsymbol{\beta}$ at $\boldsymbol{\beta}^0$.

(ii) Since the transformation from $\boldsymbol{\theta}$ to $(\boldsymbol{\alpha},\boldsymbol{\beta})$ is regular on an open neighbourhood of $\boldsymbol{\theta}^0$, a first-order expansion system of distance functions delivers:

$$\begin{aligned}\sqrt{T}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) &= -\left[\mathbf{Q}_{\boldsymbol{\beta}}'(\mathbf{0},\boldsymbol{\beta}^0)\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{V}_g^+(\boldsymbol{\theta}^0)\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)\right]^{-1} \\ &\quad \times\mathbf{Q}_{\boldsymbol{\beta}}'(\mathbf{0},\boldsymbol{\beta}^0)\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{V}_g^+(\boldsymbol{\theta}^0)\sqrt{T}\mathbf{g}(\widehat{\boldsymbol{\pi}},\boldsymbol{\theta}^0) + o_p(1). \qquad (39)\end{aligned}$$

Therefore,

$$\sqrt{T}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \xrightarrow{d} N\left[\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}}\right],$$

where

$$\mathbf{V}_{\boldsymbol{\beta}} = \left[\mathbf{Q}_{\boldsymbol{\beta}}'(\mathbf{0},\boldsymbol{\beta}^0)\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{V}_g^+(\boldsymbol{\theta}^0)\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)\right]^{-1}. \qquad (40)$$

In addition, note that since the optimal ALS estimator is given by $\widehat{\boldsymbol{\theta}} = \mathbf{q}(\mathbf{0},\widehat{\boldsymbol{\beta}})$, we can use the Delta method to compute its asymptotic distribution:

$$\sqrt{T}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{d} N\left[\mathbf{0}, \mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)\mathbf{V}_{\boldsymbol{\beta}}\mathbf{Q}_{\boldsymbol{\beta}}'(\mathbf{0},\boldsymbol{\beta}^0)\right]. \qquad (41)$$

We now compare the asymptotic covariance matrix of this optimal estimator with the ALS estimator that uses $\mathbf{W}$ as a weighting matrix and does not impose the restrictions $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$. In particular, the asymptotic covariance matrix of such an estimator is given by

$$\left[\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\right]^{-1}\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{V}_{\mathbf{g}}(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\left[\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\right]^{-1}.$$

Therefore, for $\widehat{\boldsymbol{\theta}}$ to be optimal, we need

$$\begin{aligned}&\left[\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\right]^{-1}\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{V}_{\mathbf{g}}(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\left[\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\right]^{-1} \\ &-\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)\mathbf{V}_{\boldsymbol{\beta}}\mathbf{Q}_{\boldsymbol{\beta}}'(\mathbf{0},\boldsymbol{\beta}^0)\end{aligned}$$

to be positive semidefinite, which in turn requires

$$\begin{aligned}&\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{V}_{\mathbf{g}}(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0) \\ &-\left[\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\right]\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)\mathbf{V}_{\boldsymbol{\beta}}\mathbf{Q}_{\boldsymbol{\beta}}'(\mathbf{0},\boldsymbol{\beta}^0)\left[\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\right]\end{aligned}$$

to be positive semidefinite as well.

It can be shown that this is the case given that this matrix is the asymptotic residual variance of the limiting least squares projection of $\sqrt{T}\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{g}\left[\widehat{\boldsymbol{\pi}},\boldsymbol{\theta}^0\right]$ on $\sqrt{T}\mathbf{Q}_{\boldsymbol{\beta}}'(\mathbf{0},\boldsymbol{\beta}^0)\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{V}_g^+(\boldsymbol{\theta}^0)\mathbf{g}(\widehat{\boldsymbol{\pi}},\boldsymbol{\theta})$. In particular:

$$\begin{aligned}\lim_{T\longrightarrow\infty} Var\left[\begin{array}{c}\sqrt{T}\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{g}\left[\widehat{\boldsymbol{\pi}},\boldsymbol{\theta}^0\right] \\ \sqrt{T}\mathbf{Q}_{\boldsymbol{\beta}}'(\mathbf{0},\boldsymbol{\beta}^0)\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{V}_g^+(\boldsymbol{\theta}^0)\mathbf{g}(\widehat{\boldsymbol{\pi}},\boldsymbol{\theta})\end{array}\right] = \\ \left[\begin{array}{cc}\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{V}_{\mathbf{g}}(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0) & \mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0) \\ \mathbf{Q}_{\boldsymbol{\beta}}'(\mathbf{0},\boldsymbol{\beta}^0)\mathbf{G}_{\boldsymbol{\theta}}'(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0) & \mathbf{V}_{\boldsymbol{\beta}}^{-1}\end{array}\right].\end{aligned}$$

Alternatively, we can consider the variance of a third ALS estimator that uses $\mathbf{W}$ as weighting matrix but imposes the restrictions $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$:

$$\begin{aligned}&\left[\mathbf{Q}'_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)\mathbf{G}'_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{G}'_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)\right]^{-1}\\ &\times\mathbf{Q}'_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)\mathbf{G}'_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{V}_g(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)\\ &\times\left[\mathbf{Q}'_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)\mathbf{G}'_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{W}\mathbf{G}'_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)\right]^{-1},\end{aligned}$$

and the variance of a fourth estimator that uses the generalized inverse of $\mathbf{V}_g(\boldsymbol{\theta}^0)$ as a weighting matrix but does not impose $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$:

$$\left[\mathbf{G}'_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{V}_g^+(\boldsymbol{\theta}^0)\mathbf{G}'_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\right]^{-1}\mathbf{G}'_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{V}_g^+(\boldsymbol{\theta}^0)\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\left[\mathbf{G}'_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{V}_g^+(\boldsymbol{\theta}^0)\mathbf{G}'_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\right]^{-1} = \left[\mathbf{G}'_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{V}_g^+(\boldsymbol{\theta}^0)\mathbf{G}'_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\right]^{-1}.$$

Again, it is possible to prove that the difference between any of these two matrices and $\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)\mathbf{V}_{\boldsymbol{\beta}}\mathbf{Q}'_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)$ is positive semidefinite.

(iii) Using a Taylor expansion of $\sqrt{T}\mathbf{g}\left[\widehat{\boldsymbol{\pi}},\mathbf{q}(\mathbf{0},\widehat{\boldsymbol{\beta}})\right]$ and equation (39), we have that

$$\begin{aligned}&\sqrt{T}\mathbf{g}\left[\widehat{\boldsymbol{\pi}},\mathbf{q}(\mathbf{0},\widehat{\boldsymbol{\beta}})\right]\\ &=\sqrt{T}\mathbf{g}(\widehat{\boldsymbol{\pi}},\boldsymbol{\theta}^0)+\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)\sqrt{T}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)+o_p(1)\\ &=\left[\mathbf{I}-\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)\mathbf{V}_{\boldsymbol{\beta}}\mathbf{Q}'_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)\mathbf{G}'_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{T}_1\boldsymbol{\Lambda}^{-1}\mathbf{T}'_1\right]\sqrt{T}\mathbf{g}(\widehat{\boldsymbol{\pi}},\boldsymbol{\theta}^0)+o_p(1),\end{aligned}$$

and rearranging the previous expression as

$$\begin{aligned}&\sqrt{T}\mathbf{g}\left[\widehat{\boldsymbol{\pi}},\mathbf{q}(\mathbf{0},\widehat{\boldsymbol{\beta}})\right]\\ =\;&\mathbf{T}_1\boldsymbol{\Lambda}^{1/2}\left[\mathbf{I}_{G-S}-\mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\right]\sqrt{T}\boldsymbol{\Lambda}^{-1/2}\mathbf{T}'_1\mathbf{g}(\widehat{\boldsymbol{\pi}},\boldsymbol{\theta}^0)+o_p(1),\end{aligned}$$

where $\mathbf{H} = \boldsymbol{\Lambda}^{-1/2}\mathbf{T}'_1\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)\mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0},\boldsymbol{\beta}^0)$. Therefore, the criterion function evaluated at the optimal ALS estimator is

$$T\mathbf{g}\left[\widehat{\boldsymbol{\pi}},\mathbf{q}(\mathbf{0},\widehat{\boldsymbol{\beta}})\right]'\mathbf{V}_g^+(\boldsymbol{\theta}^0)\mathbf{g}\left[\widehat{\boldsymbol{\pi}},\mathbf{q}(\mathbf{0},\widehat{\boldsymbol{\beta}})\right] = \widehat{\mathbf{z}}'\left[\mathbf{I}_{G-S}-\mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\right]\widehat{\mathbf{z}}+o_p(1),$$

where $\widehat{\mathbf{z}} = \boldsymbol{\Lambda}^{-1/2}\mathbf{T}'_1\mathbf{g}(\widehat{\boldsymbol{\pi}},\boldsymbol{\theta}^0)$ is asymptotically distributed as a standard multivariate normal, which implies that the criterion function converges to a chi-square distribution with $G-K$ degrees of freedom, given that the matrix $\left[\mathbf{I}_{G-S}-\mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\right]$ is idempotent with rank $(G-S)-(K-S) = G-K$.

Proposition 2. Assume, without loss of generality, that the function that implicitly defines the $K-S$-dimensional manifold in $\boldsymbol{\Theta}$ over which S linear combinations of $\sqrt{T}\mathbf{g}(\widehat{\boldsymbol{\pi}},\boldsymbol{\theta})$ converge in probability to zero (Assumption 1) can be written as $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{r}\left[\mathbf{p}(\boldsymbol{\theta})\right] = \mathbf{0}$. Let $\widehat{\boldsymbol{\pi}}$ be an estimator that is asymptotically equivalent to maximum likelihood and that satisfies $\mathbf{r}(\widehat{\boldsymbol{\pi}}) = \mathbf{0}$, and let the system of implicit relationships $\mathbf{g}(\boldsymbol{\pi},\boldsymbol{\theta}) = \mathbf{0}$ be complete (i.e., $G = H$ and $\mathbf{G}_{\boldsymbol{\pi}}$has full rank). Then, the optimal ALS

estimator that uses a generalized inverse of $\mathbf{V}_g(\boldsymbol{\theta}^0)$ as the weighting matrix and that, simultaneously, imposes the restriction $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{r}\left[\mathbf{p}(\boldsymbol{\theta})\right] = \mathbf{0}$ is asymptotically equivalent to the ML estimator that imposes that restriction.

Proof. As in the proof of Proposition 1, we will work with the alternative set of K structural parameters $\boldsymbol{\alpha}$ $(S \times 1)$ and $\boldsymbol{\beta}$ $((K-S) \times 1)$ such that

$$\mathbf{R}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\alpha}' & \boldsymbol{\beta}' \end{pmatrix}',$$

where the first S elements of $\mathbf{R}(\boldsymbol{\theta})$ are such that $\boldsymbol{\alpha} = \mathbf{r}(\boldsymbol{\theta})$. Again, let $\mathbf{q}\left[\mathbf{R}(\boldsymbol{\theta})\right] = \boldsymbol{\theta}$ be the inverse transformation of $\mathbf{R}(\boldsymbol{\theta})$ that recovers $\boldsymbol{\theta}$ back, and let its Jacobians be denoted by $\mathbf{Q}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \partial \mathbf{q}(\boldsymbol{\alpha}, \boldsymbol{\beta}) / \partial(\boldsymbol{\alpha}', \boldsymbol{\beta}')$. As noted earlier, this (regular) transformation allows us to impose the parametric restriction $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ by simply setting $\boldsymbol{\alpha} = \mathbf{0}$. In particular, the asymptotic distribution of the ML estimate of $\boldsymbol{\beta}$ subject to the restriction that $\boldsymbol{\alpha} = \mathbf{0}$ is given by

$$\sqrt{T}(\widehat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}^0) \xrightarrow{d} N\left[\mathbf{0}, \boldsymbol{\Upsilon}_{\beta\beta}^{-1}(\mathbf{0}, \boldsymbol{\beta}^0)\right],$$

where $\boldsymbol{\Upsilon}_{\beta\beta}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\frac{1}{T} E\left[\frac{\partial^2 \log L(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right]$ is the relevant block of the information matrix. Similarly, since the ML estimator of $\boldsymbol{\theta}$ that imposes the restriction $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ is given by $\widehat{\boldsymbol{\theta}}_{ML} = \mathbf{q}(\mathbf{0}, \widehat{\boldsymbol{\beta}}_{ML})$, we can use the Delta method to compute its asymptotic distribution:

$$\sqrt{T}(\widehat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}^0) \xrightarrow{d} N\left[\mathbf{0}, \mathbf{Q}_{\boldsymbol{\beta}}(\mathbf{0}, \boldsymbol{\beta}^0) \boldsymbol{\Upsilon}_{\beta\beta}^{-1}(\mathbf{0}, \boldsymbol{\beta}^0) \mathbf{Q}_{\boldsymbol{\beta}}'(\mathbf{0}, \boldsymbol{\beta}^0)\right].$$

In particular, the optimal ALS estimate of $\boldsymbol{\theta}$ will be asymptotically equivalent to ML if they have the same asymptotic variance. Comparing this expression with equation (41), it is straightforward to see that this will only occur when $\mathbf{V}_{\boldsymbol{\beta}} = \boldsymbol{\Upsilon}_{\beta\beta}^{-1}$.

In order to prove this result, we will work on an alternative set of G auxiliary parameters $\boldsymbol{\delta}$ $(S \times 1)$ and $\boldsymbol{\gamma}$ $((G-S) \times 1)$ such that

$$\mathbf{M}\left[\mathbf{p}(\boldsymbol{\theta})\right] = \begin{bmatrix} \boldsymbol{\delta}(\boldsymbol{\theta})' & \boldsymbol{\gamma}(\boldsymbol{\theta})' \end{bmatrix}',$$

where the first S elements of $\mathbf{M}(\boldsymbol{\pi})$ are such that $\boldsymbol{\delta} = \mathbf{r}(\boldsymbol{\pi})$. Let $\mathbf{l}\left[\mathbf{M}(\boldsymbol{\pi})\right] = \boldsymbol{\pi}$ be the corresponding inverse transformation of $\mathbf{M}(\boldsymbol{\pi})$ that recovers $\boldsymbol{\pi}$ back. Let the Jacobians of the inverse transformation be given by

$$\mathbf{L}(\boldsymbol{\delta}, \boldsymbol{\gamma}) = \frac{\partial \mathbf{l}(\boldsymbol{\delta}, \boldsymbol{\gamma})}{\partial(\boldsymbol{\delta}', \boldsymbol{\gamma}')} = \begin{bmatrix} \mathbf{L}_{\boldsymbol{\delta}}(\boldsymbol{\delta}, \boldsymbol{\gamma}) & \mathbf{L}_{\boldsymbol{\gamma}}(\boldsymbol{\delta}, \boldsymbol{\gamma}) \end{bmatrix}.$$

Note that this second (regular) transformation of the auxiliary parameters allows us to impose the parametric restriction $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ on both the estimation of the auxiliary and parameters of interest. Specifically, we have that $\boldsymbol{\delta}(\boldsymbol{\theta}) = \mathbf{r}\left[\mathbf{q}(\mathbf{0}, \boldsymbol{\beta})\right] = \mathbf{0}$ for all $\boldsymbol{\beta}$. Further, the asymptotic distribution of the ML estimate of $\boldsymbol{\gamma}$ subject to the restriction that $\boldsymbol{\delta} = \mathbf{0}$ is given by

$$\sqrt{T}(\widehat{\boldsymbol{\gamma}}_{ML} - \boldsymbol{\gamma}^0) \xrightarrow{d} N\left[\mathbf{0}, \boldsymbol{\Upsilon}_{\gamma\gamma}^{-1}(\mathbf{0}, \boldsymbol{\gamma}_0)\right],$$

where $\boldsymbol{\Upsilon}_{\gamma\gamma}(\boldsymbol{\delta},\boldsymbol{\gamma}) = -\frac{1}{T}E\left[\frac{\partial^2 \log L(\boldsymbol{\delta},\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\right]$ is the relevant block of the information matrix. Note that $\boldsymbol{\Upsilon}_{\beta\beta} = \frac{\partial\boldsymbol{\beta}'}{\partial\boldsymbol{\gamma}}\boldsymbol{\Upsilon}_{\gamma\gamma}\frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{\gamma}'}$.

Moreover, since the ML estimator of $\boldsymbol{\pi}$ that imposes the restriction $\mathbf{r}(\boldsymbol{\pi})$ is given by $\widehat{\boldsymbol{\pi}}_{ML} = \mathbf{l}(\mathbf{0},\widehat{\boldsymbol{\gamma}}_{ML})$, we can use the Delta method to compute its asymptotic distribution:

$$\sqrt{T}(\widehat{\boldsymbol{\pi}}_{ML} - \boldsymbol{\pi}^0) \xrightarrow{d} N\left[\mathbf{0}, \mathbf{L}_{\boldsymbol{\gamma}}(\mathbf{0},\boldsymbol{\gamma}^0)\boldsymbol{\Upsilon}^{-\mathbf{1}}_{\boldsymbol{\gamma\gamma}}(\mathbf{0},\boldsymbol{\gamma}^0)\mathbf{L}'_{\boldsymbol{\gamma}}(\mathbf{0},\boldsymbol{\gamma}^0)\right]. \tag{42}$$

Finally, note that, since the system is complete, and the fact that both $\mathbf{R}(\cdot)$ and $\mathbf{M}(\cdot)$ are regular imply that $\mathbf{Q}(\boldsymbol{\alpha},\boldsymbol{\beta})$ and $\mathbf{L}(\boldsymbol{\delta},\boldsymbol{\gamma})$ have full rank, we can write that

$$\frac{\partial\boldsymbol{\pi}}{\partial\theta'} = \frac{\partial\boldsymbol{\pi}}{\begin{pmatrix}\boldsymbol{\delta}' & \boldsymbol{\gamma}'\end{pmatrix}} \times \frac{\partial\begin{pmatrix}\boldsymbol{\delta}\\ \boldsymbol{\gamma}\end{pmatrix}}{\begin{pmatrix}\boldsymbol{\alpha}' & \boldsymbol{\beta}'\end{pmatrix}} \times \frac{\partial\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix}}{\partial\theta'}$$

$$-\mathbf{G}_{\boldsymbol{\pi}}^{-1}\mathbf{G}_{\boldsymbol{\theta}} = \mathbf{L}(\boldsymbol{\delta},\boldsymbol{\gamma})\begin{pmatrix}\frac{\partial\boldsymbol{\delta}}{\partial\boldsymbol{\alpha}'} & \frac{\partial\boldsymbol{\delta}}{\partial\boldsymbol{\beta}'}\\ \frac{\partial\boldsymbol{\gamma}}{\partial\boldsymbol{\alpha}'} & \frac{\partial\boldsymbol{\gamma}}{\partial\boldsymbol{\beta}'}\end{pmatrix}\mathbf{Q}^{-1}(\boldsymbol{\alpha},\boldsymbol{\beta})$$

$$-\mathbf{G}_{\boldsymbol{\pi}}^{-1}\mathbf{G}_{\boldsymbol{\theta}}\begin{pmatrix}\mathbf{Q}_{\boldsymbol{\alpha}} & \mathbf{Q}_{\beta}\end{pmatrix} = \begin{pmatrix}\mathbf{L}_{\boldsymbol{\delta}} & \mathbf{L}_{\boldsymbol{\gamma}}\end{pmatrix}\begin{pmatrix}\frac{\partial\boldsymbol{\delta}}{\partial\boldsymbol{\alpha}'} & \frac{\partial\boldsymbol{\delta}}{\partial\boldsymbol{\beta}'}\\ \frac{\partial\boldsymbol{\gamma}}{\partial\boldsymbol{\alpha}'} & \frac{\partial\boldsymbol{\gamma}}{\partial\boldsymbol{\beta}'}\end{pmatrix},$$

which, since $\boldsymbol{\delta}(\boldsymbol{\theta}) = \mathbf{r}\left[\mathbf{q}(\mathbf{0},\boldsymbol{\beta})\right] = \mathbf{0}$ for all $\boldsymbol{\beta}$ implies that $\partial\boldsymbol{\delta}/\partial\boldsymbol{\beta}' = 0$, we have that

$$-\mathbf{G}_{\boldsymbol{\theta}}\mathbf{Q}_{\beta} = \mathbf{G}_{\boldsymbol{\pi}}\mathbf{L}_{\boldsymbol{\gamma}}\frac{\partial\boldsymbol{\gamma}}{\partial\boldsymbol{\beta}'}. \tag{43}$$

Substituting equations (42) and (43) evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ in the expression for $\mathbf{V}_{\boldsymbol{\beta}}$ in (40) we have that

$$\mathbf{V}_{\boldsymbol{\beta}}^{-1} = \frac{\partial\boldsymbol{\gamma}'}{\partial\boldsymbol{\beta}}\left\{\mathbf{L}'_{\boldsymbol{\gamma}}(\mathbf{0},\boldsymbol{\gamma}^0)\mathbf{G}'_{\boldsymbol{\pi}}(\boldsymbol{\theta}^0)\left[\mathbf{G}_{\boldsymbol{\pi}}(\boldsymbol{\theta}^0)\mathbf{L}_{\boldsymbol{\gamma}}(\mathbf{0},\boldsymbol{\gamma}^0)\boldsymbol{\Upsilon}^{-1}_{\gamma\gamma}(\mathbf{0},\boldsymbol{\gamma}^0)\mathbf{L}'_{\boldsymbol{\gamma}}(\mathbf{0},\boldsymbol{\gamma}^0)\mathbf{G}_{\boldsymbol{\pi}}(\boldsymbol{\theta}^0)\right]^{+}\right.$$
$$\left.\times\mathbf{G}_{\boldsymbol{\pi}}(\boldsymbol{\theta}^0)\mathbf{L}_{\boldsymbol{\gamma}}(\mathbf{0},\boldsymbol{\gamma}^0)\right\}\frac{\partial\boldsymbol{\gamma}}{\partial\boldsymbol{\beta}'}.$$

Let $\mathbf{D}$ be the term inside the curly brackets. Premultiplying $\mathbf{D}$ by $\mathbf{G}_{\boldsymbol{\pi}}(\boldsymbol{\theta}^0)\mathbf{L}_{\boldsymbol{\gamma}}(\mathbf{0},\boldsymbol{\gamma}^0)\boldsymbol{\Upsilon}^{-1}_{\gamma\gamma}(\mathbf{0},\boldsymbol{\gamma}^0)$, and postmultiplying it by $\boldsymbol{\Upsilon}^{-1}_{\gamma\gamma}(\mathbf{0},\boldsymbol{\gamma}^0)\mathbf{L}'_{\boldsymbol{\gamma}}(\mathbf{0},\boldsymbol{\gamma}^0)\mathbf{G}_{\boldsymbol{\pi}}(\boldsymbol{\theta}^0)$, we find that

$$\mathbf{G}_{\boldsymbol{\pi}}(\boldsymbol{\theta}^0)\mathbf{L}_{\boldsymbol{\gamma}}(\mathbf{0},\boldsymbol{\gamma}^0)\boldsymbol{\Upsilon}^{-1}_{\gamma\gamma}(\mathbf{0},\boldsymbol{\gamma}^0)\mathbf{D}\boldsymbol{\Upsilon}^{-1}_{\gamma\gamma}(\mathbf{0},\boldsymbol{\gamma}^0)\mathbf{L}'_{\boldsymbol{\gamma}}(\mathbf{0},\boldsymbol{\gamma}^0)\mathbf{G}_{\boldsymbol{\pi}}(\boldsymbol{\theta}^0) =$$
$$\mathbf{G}_{\boldsymbol{\pi}}(\boldsymbol{\theta}^0)\mathbf{L}_{\boldsymbol{\gamma}}(\mathbf{0},\boldsymbol{\gamma}^0)\boldsymbol{\Upsilon}^{-1}_{\gamma\gamma}(\mathbf{0},\boldsymbol{\gamma}^0)\mathbf{L}'_{\boldsymbol{\gamma}}(\mathbf{0},\boldsymbol{\gamma}^0)\mathbf{G}_{\boldsymbol{\pi}}(\boldsymbol{\theta}^0),$$

where we have used the fact that a generalized inverse must satisfy $\mathbf{W}\mathbf{W}^{+}\mathbf{W} = \mathbf{W}$. Thus, $\mathbf{D} = \boldsymbol{\Upsilon}_{\boldsymbol{\gamma\gamma}}$ for the last equation to be true. This implies that,

$$\mathbf{V}_{\boldsymbol{\beta}} = \left(\frac{\partial\boldsymbol{\gamma}}{\partial\boldsymbol{\beta}'}\boldsymbol{\Upsilon}_{\gamma\gamma}\frac{\partial\boldsymbol{\gamma}'}{\partial\boldsymbol{\beta}}\right)^{-1} = \boldsymbol{\Upsilon}^{-1}_{\beta\beta}.$$

Therefore, the optimal ALS estimator that uses a generalized inverse of $\mathbf{V}_g(\boldsymbol{\theta}^0)$ as the weighting matrix and that, simultaneously, imposes the restriction $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{r}\left[\mathbf{p}(\boldsymbol{\theta})\right] = \mathbf{0}$ is asymptotically equivalent to the ML estimator that imposes that restriction.

B Adrian, Crump and Moench (2012) from an ALS perspective

Adrian, Crump and Moench (2012) focus on bond excess holding period returns rather than on yields themselves. In particular, the one-period excess return on a bond of maturity n is the gain from buying an n-period bond and selling it one year later, financing the position at the short rate:

$$rx_{t+1,n} \equiv \log\left(\frac{P_{t+1,n-1}}{P_{t,n}}\right) - r_t = ny_{t,n} - (n-1)y_{t+1,n-1} - r_t. \tag{44}$$

Substituting (1), (6), (7) and (8) into this last expression, we can show that the one-period excess return for holding an n-period zero-coupon bond is given by

$$rx_{t+1,n} = -\frac{1}{2}\mathbf{B}'_{n-1}\boldsymbol{\Sigma}\mathbf{B}_{n-1} + \mathbf{B}'_{n-1}(\boldsymbol{\lambda}_0 + \boldsymbol{\lambda}_1\mathbf{f}_t) + \mathbf{B}'_{n-1}\mathbf{v}_{t+1},$$

and, adding an *iid* pricing error $\varepsilon_{t,n}$ to this last equation, we have that

$$rx_{t+1,n} = D_n + \mathbf{E}'_n\mathbf{f}_t + \mathbf{F}'_n\mathbf{v}_{t+1} + \varepsilon_{t,n}, \tag{45}$$

where

$$D_n = \frac{1}{2}\mathbf{F}'_n\boldsymbol{\Sigma}\mathbf{F}_n + \mathbf{F}'_n\boldsymbol{\lambda}_0, \tag{46}$$

$$\mathbf{E}'_n = \mathbf{F}'_n\boldsymbol{\lambda}_1, \tag{47}$$

for $n = 2, ..., N$.

By a similar argument to the one proposed in section 2.2, we have that if the innovation covariance matrix $\boldsymbol{\Sigma}$ and the set of coefficients D_n, $\mathbf{E}_n$ and $\mathbf{F}_n$'s were observed directly, one could easily estimate the price of risk parameters of the model using a set of (cross-sectional) OLS regressions. In particular, one could recover an estimate of $\boldsymbol{\lambda}_1$ as

$$\widehat{\boldsymbol{\lambda}}_1 = \left(\sum_{n=2}^{N}\mathbf{F}_n\mathbf{F}'_n\right)^{-1}\left(\sum_{n=2}^{N}\mathbf{F}_n\mathbf{E}'_n\right), \tag{48}$$

while an estimate of $\boldsymbol{\lambda}_0$ from

$$\widehat{\boldsymbol{\lambda}}_0 = \left(\sum_{n=2}^{N}\mathbf{F}_n\mathbf{F}'_n\right)^{-1}\left[\sum_{n=1}^{N}\mathbf{F}_n(D_n - \frac{1}{2}\mathbf{F}'_n\boldsymbol{\Sigma}\mathbf{F}_n)\right]. \tag{49}$$

However, this estimator is (again) infeasible, because the innovation covariance matrix $\boldsymbol{\Sigma}$, and the set of coefficients D_n, $\mathbf{E}_n$ and $\mathbf{F}_n$'s are, in practice, unknown. Instead, we could follow the same principles used to develop the linear estimator in section 2.2, and replace these unknown quantities by some consistent estimates. Specifically, we could first estimate the VAR(1) process in equation (1) to obtain $\widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{\Phi}}$ and $\widehat{\boldsymbol{\Sigma}}$, as well as an estimate

of the innovation, $\widehat{\mathbf{v}}_{t+1}$. Second, we could use equation (44) to run a regression of $rx_{t,n}$ on a constant, the lagged pricing factors, $\mathbf{f}_{t-1}$, and the contemporaneous pricing factor innovations $\widehat{\mathbf{v}}_t$ to obtain a set of estimates of D_n, $\mathbf{E}_n$ and $\mathbf{F}_n$'s for $n = 2, ..., N$. Finally, we could recover an estimate for the market prices of risk parameters, $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$ from the cross-sectional regressions in (48) and (49) by simply replacing the unknown objects in these equations by the consistent estimates obtained in the previous step. In fact, such an approach is exactly the three-step linear regression method proposed by Adrian, Crump and Moench (2012). In addition, the short-rate parameters, δ_0 and $\boldsymbol{\delta}_1$, can be obtained by running an OLS regression of one-period yield on the pricing factors.

Along the same lines of section 2.3, we can now interpret the ACM estimator within the ALS framework. In particular, we have that the vector of reduced-form parameters is given by

$$\boldsymbol{\pi}_{ACM} = \left((\delta_0 \ \boldsymbol{\delta}_1'), \left\{ vec\left[(\mathbf{D} \ \mathbf{E} \ \mathbf{F})' \right] \right\}', \left\{ vec\left[(\boldsymbol{\mu} \ \boldsymbol{\Phi})' \right] \right\}', \left[vech\left(\boldsymbol{\Sigma}^{1/2} \right) \right]' \right)',$$

where $\mathbf{D}$ is a vector that stacks the corresponding elements of D_n, and $\mathbf{E}$ and $\mathbf{F}$ are matrices that stack the corresponding elements of $\mathbf{E}_n'$ and $\mathbf{F}_n'$.

On the other hand, using that $\boldsymbol{\mu}^{\mathbb{Q}} = \boldsymbol{\mu} - \boldsymbol{\lambda}_0$ and $\boldsymbol{\Phi}^{\mathbb{Q}} = \boldsymbol{\Phi} - \boldsymbol{\lambda}_1$ with equations (46) and (47) and stacking, it is possible to express the restrictions implied by the no-arbitrage model as

$$\mathbf{H}(\boldsymbol{\pi}_{ACM}, \boldsymbol{\theta})' = \mathbf{Y}_{ACM}(\boldsymbol{\pi}_{ACM}) - \mathbf{X}_{ACM}(\boldsymbol{\pi}_{ACM})\boldsymbol{\Upsilon}' = \mathbf{0}, \tag{50}$$

where

$$\mathbf{Y}_{ACM}(\boldsymbol{\pi}_{ACM}) = \begin{pmatrix} \delta_0 & \boldsymbol{\delta}_1' \\ D_2 - \frac{1}{2}\mathbf{F}_2'\boldsymbol{\Sigma}\mathbf{F}_2 & \mathbf{E}_2' \\ \vdots & \vdots \\ D_N - \frac{1}{2}\mathbf{F}_N'\boldsymbol{\Sigma}\mathbf{F}_N & \mathbf{E}_N' \\ \boldsymbol{\mu} & \boldsymbol{\Phi} \end{pmatrix}, \quad \mathbf{X}_{ACM}(\boldsymbol{\pi}_{ACM}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -\mathbf{F}_2' & \mathbf{F}_2' \\ \vdots & \vdots & \vdots \\ 0 & -\mathbf{F}_N' & \mathbf{F}_N \\ 0 & 0 & \mathbf{I} \end{pmatrix},$$

and where $\boldsymbol{\Upsilon}$ satisfies

$$\boldsymbol{\Upsilon}' = \begin{pmatrix} \delta_0 & \boldsymbol{\delta}_1' \\ \boldsymbol{\mu}^Q & \boldsymbol{\Phi}^Q \\ \boldsymbol{\mu} & \boldsymbol{\Phi} \end{pmatrix}.$$

Note that $\boldsymbol{\theta} = \left\{ vec(\boldsymbol{\Upsilon}'), \left[vech\left(\boldsymbol{\Sigma}^{1/2} \right) \right]' \right\}$, so by vectorizing equation (50) and adding a set of identities, it is possible to arrive at the following distance function for the case of ACM estimation:

$$\mathbf{h}(\boldsymbol{\pi}_{ACM}, \boldsymbol{\theta}) = vec\left[\mathbf{H}(\boldsymbol{\pi}_{ACM}, \boldsymbol{\theta}) \right] = \boldsymbol{\gamma}_{ACM}(\boldsymbol{\pi}_{ACM}) - \boldsymbol{\Gamma}_{ACM}(\boldsymbol{\pi}_{ACM})\boldsymbol{\theta},$$

where

$$\boldsymbol{\gamma}_{ACM}(\boldsymbol{\pi}_{ACM}) = \left[\mathbf{Y}_{ACM}(\boldsymbol{\pi}_{ACM}), vech\left(\boldsymbol{\Sigma}^{1/2} \right) \right]',$$
$$\boldsymbol{\Gamma}_{ACM}(\boldsymbol{\pi}_{ACM}) = \begin{pmatrix} \mathbf{X}_{ACM}(\boldsymbol{\pi}_{ACM}) \otimes \mathbf{I} & 0 \\ 0 & \mathbf{I} \end{pmatrix}.$$

With this notation, the ACM estimator is equivalent to the estimator that minimizes a quadratic form in the distance function $\mathbf{h}(\boldsymbol{\pi}_{ACM}, \boldsymbol{\theta})$, evaluated at the estimates of the reduced-form parameters, $\widehat{\boldsymbol{\pi}}_{ACM}$, where the weighting matrix has been chosen to be the identity matrix, $\mathbf{W}_T = \mathbf{I}$. Therefore, it is possible to achieve efficiency gains by selecting an appropriate weighting matrix and imposing the self-consistency of the model. However, note that that system of implicit relationships $\mathbf{h}(\boldsymbol{\pi}_{ACM}, \boldsymbol{\theta})$ is not complete (the number of reduced-form parameters is larger than the dimension of the distance function), so, even if we were using an optimal weighting matrix and impose self-consistency, the ACM approach would still not be asymptotically equivalent to maximum likelihood estimation.