A New Perspective on Gaussian Dynamic Term Structure Models

Scott Joslin[*], Kenneth J. Singleton[†], and Haoxiang Zhu[‡]

First version: November 9, 2008
This version: October 17, 2010
Forthcoming *Review of Financial Studies*

Abstract

In any canonical Gaussian dynamic term structure model (*GDTSM*), the conditional forecasts of the pricing factors are invariant to the imposition of no-arbitrage restrictions. This invariance is maintained even in the presence of a variety of restrictions on the factor structure of bond yields. To establish these results, we develop a novel canonical *GDTSM* in which the pricing factors are observable portfolios of yields. For our normalization, standard maximum likelihood algorithms converge to the global optimum almost instantaneously. We present empirical estimates and out-of-sample forecasts for several *GDTSM*s using data on U.S. Treasury bond yields.

[*]Sloan School of Management, MIT, sjoslin@mit.edu
[†]Graduate School of Business, Stanford University and NBER, singleton_ken@gsb.stanford.edu
[‡]Graduate School of Business, Stanford University, haoxiang.zhu@stanford.edu.

Dynamic models of the term structure often posit a linear factor structure for a collection of yields, with these yields related to underlying factors $\mathcal{P}$ through a no-arbitrage relationship. Does the imposition of no-arbitrage in a Gaussian dynamic term structure model (*GDTSM*) improve the out-of-sample forecasts of yields relative to those from the unconstrained factor model, or sharpen model-implied estimates of expected excess returns? In practice, the answers to these questions are obscured by the imposition of over-identifying restrictions on the risk-neutral ($\mathbb{Q}$) or historical ($\mathbb{P}$) distributions of the risk factors, or on their market prices of risk, in addition to the cross-maturity restrictions implied by no-arbitrage.[1]

We show that, *within any canonical* GDTSM *and for any sample of bond yields, imposing no-arbitrage does not affect the conditional* $\mathbb{P}$*-expectation of* $\mathcal{P}$*,* $E^{\mathbb{P}}[\mathcal{P}_t|\mathcal{P}_{t-1}]$. *GDTSM*-implied forecasts of $\mathcal{P}$ are thus identical to those from the unrestricted vector-autoregressive (VAR) model for $\mathcal{P}$. To establish these results, we develop an all-encompassing canonical model in which the pricing factors $\mathcal{P}$ are linear combinations of the collection of yields y, (such as the first N principal components (PCs))[2] and in which these "yield factors" follow an unrestricted VAR. Within our canonical *GDTSM*, as long as $\mathcal{P}$ is measured without error, unconstrained ordinary least squares (OLS) gives the maximum likelihood (ML) estimates of $E^{\mathbb{P}}[\mathcal{P}_t|\mathcal{P}_{t-1}]$. Therefore, enforcing no-arbitrage has no effect on out-of-sample forecasts of $\mathcal{P}$. This result holds for *any* other canonical *GDTSM*, owing to observational equivalence (Dai and Singleton (2000)) and, as such, is a generic feature of *GDTSM*s.

Heuristically, under the assumption that the yield factors $\mathcal{P}$ are observed without error, these propositions follow from the factorization of the conditional density of y into the product of the conditional $\mathbb{P}$-density of $\mathcal{P}$ times the conditional density of measurement errors.[3] The density of $\mathcal{P}$ is determined by parameters controlling its conditional mean and its innovation covariance matrix. The measurement error density is determined by the "no arbitrage" cross-sectional relationship amongst the yields. We show that *GDTSM*s can be parameterized so that the parameters governing the $\mathbb{P}$-forecasts of $\mathcal{P}$ do not appear in the measurement-error density. Given this separation, the only link between the conditional $\mathbb{P}$-density and the measurement density is the covariance of the innovations. However, a classic result of Zellner (1962) implies that the ML estimates of $E^{\mathbb{P}}[\mathcal{P}_t|\mathcal{P}_{t-1}]$ are independent

[1] Recent studies that explore the forecasting performance of *GDTSM*s include Duffee (2002), Ang and Piazzesi (2003), Christensen, Diebold, and Rudebusch (2007), Chernov and Mueller (2008), and Jardet, Monfort, and Pegoraro (2009), among many others.

[2] Although standard formulations of affine term structure models use latent (unobservable) risk factors (e.g., Dai and Singleton (2000), Duffee (2002)), by Duffie and Kan (1996) we are free to normalize a model so that the factors are portfolios of yields on bonds and we choose PCs.

[3] See, for example, Chen and Scott (1993) and Pearson and Sun (1994).

of this covariance. Consequently, OLS recovers the ML estimates of $E^{\mathbb{P}}[\mathcal{P}_t|\mathcal{P}_{t-1}]$ and the no-arbitrage restriction is irrelevant for the conditional $\mathbb{P}$-forecast of $\mathcal{P}$.

Key to seeing this irrelevance is our choice of canonical form.[4] For any N-factor model with portfolios of yields $\mathcal{P}$ as factors, bond prices depend on the $N(N+1)$ parameters governing the risk-neutral conditional mean of $\mathcal{P}$ and the $(N+1)$ parameters linking the short rate to $\mathcal{P}$, for a total of $(N+1)^2$ parameters. Not all of these parameters are free, however, because internal consistency requires that the model-implied yields reproduce the yield-factors $\mathcal{P}$. We show that, given the N yield factors, the entire time-t yield curve can be constructed by specifying: (a) $r_\infty^{\mathbb{Q}}$, the long-run mean of the short rate under $\mathbb{Q}$; (b) $\lambda^{\mathbb{Q}}$, the speeds of mean reversion of the yield-factors under $\mathbb{Q}$; and (c) $\Sigma_{\mathcal{P}}$, the conditional covariance matrix of yields factors from the VAR. That is, given $\Sigma_{\mathcal{P}}$, the entire cross-section of bond yields in an N-factor $GDTSM$ is fully determined by only the $N+1$ parameters $r_\infty^{\mathbb{Q}}$ and $\lambda^{\mathbb{Q}}$. Moreover, $(r_\infty^{\mathbb{Q}}, \lambda^{\mathbb{Q}}, \Sigma_{\mathcal{P}})$ can be efficiently estimated independently of the $\mathbb{P}$-conditional mean of $\mathcal{P}_t$, rendering no-arbitrage irrelevant for forecasting $\mathcal{P}$.

With these results in place, we proceed to show that the conditional forecast $E^{\mathbb{P}}[\mathcal{P}_t|\mathcal{P}_{t-1}]$ from a no-arbitrage $GDTSM$ remains identical to its counterpart from an unrestricted VAR even in the presence of a large class of over-identifying restrictions on the factor structure of y. In particular, *regardless of the constraints imposed on the risk-neutral distribution of the yield-factors $\mathcal{P}$, the* GDTSM- *and VAR-implied forecasts of these factors are identical.* Put differently, OLS recovers the conditional forecasts of the yield factors even in the the presence of further cross-sectional restrictions on the shape of the yield curve beyond no arbitrage.

When does the structure of a $GDTSM$ improve out-of-sample forecasts of $\mathcal{P}$? We show that if constraints are imposed directly on the $\mathbb{P}$ distribution of $\mathcal{P}$ within a no-arbitrage $GDTSM$, then the ML estimate of $E^{\mathbb{P}}[\mathcal{P}_t|\mathcal{P}_{t-1}]$ is more efficient than its OLS counterpart from a VAR. Thus our theoretical results, as well as subsequent empirical illustrations, show that gains from forecasting using a $GDTSM$, if any, must come from auxiliary constraints on the $\mathbb{P}$ distribution of $\mathcal{P}$, and not from the no-arbitrage restriction *per se*.[5]

An important example of such auxiliary constraints is the number of risk factors that determine risk premiums. Motivated by the descriptive analysis of Cochrane and Piazzesi (2005, 2008) and Duffee (2008), we develop methods for restricting expected excess returns

[4] To emphasize, our canonical form is key to *seeing* the result; due to observational equivalence, the result holds for *any* canonical form.

[5] Though one might conclude from reading the recent literature that enforcing no-arbitrage improves out-of-sample forecasts of bond yields, our theorems show that this is not the case. What underlies any documented forecast gains in these studies from using $GDTSM$s is the combined structure of no arbitrage *and* the auxiliary restrictions they impose on the $\mathbb{P}$ distribution of y.

to lie in a space of dimension $\mathcal{L}\,(< N)$, *without restricting a priori which of the N factors* $\mathcal{P}_t$ *represent priced risks*. If $\mathcal{L} < N$, then there are necessarily restrictions linking the historical and risk-neutral drifts of $\mathcal{P}_t$. In this case, the forecasts of future yields implied by a *GDTSM* are in principle different than those from an unrestricted VAR, and we investigate the empirical relevance of these constraints within three-factor ($N = 3$) *GDTSM*s.

Additionally, we show that our canonical form allows for computationally efficient estimation of *GDTSM*s. The conditional density of observed yields is fully characterized by $r_\infty^{\mathbb{Q}}$ and $\lambda^{\mathbb{Q}}$, as well as the parameters controlling any measurement errors in yields. Importantly, $(r_\infty^{\mathbb{Q}}, \lambda^{\mathbb{Q}})$ constitutes a low-dimensional, rotation-invariant (and thus economically meaningful) parameter space. Using standard search algorithms, we obtain near-instantaneous convergence to the global optimum of the likelihood function. Convergence is fast regardless of the number or risk factors or bond yields used in estimation, or whether the pricing factors $\mathcal{P}$ are measured with error.[6]

The rapid convergence to global optima using our canonical *GDTSM* makes it feasible to explore rolling out-of-sample forecasts. For a variety of *GDTSM*s – with and without measurement error in yield factors, and with and without constraints on the dimensionality $\mathcal{L}$ of risk premia – we compare the out-of-sample forecasting performance relative to a benchmark unconstrained VAR, and confirm our theoretical predictions in the data.

1 A Canonical *GDTSM* with *Observable* Risk Factors

In this section we develop our "JSZ" canonical representation of *GDTSM*s. Towards this end we start with a generic representation of a *GDTSM* in which the discrete-time evolution of the risk factors (state vector) $X_t \in \mathbb{R}^N$ is governed by the following equations:[7]

[6] To put this computational advantage into perspective, one needs to read no further than Duffee and Stanton (2007) and Duffee (2009) who highlight numerous computational challenges and multiple local optima associated with their likelihood functions. For example, Duffee reports that each optimization for his parametrization of a three-factor model takes about two days. In contrast, for the *GDTSM*(3) models examined in this paper, convergence to the global optimum of the likelihood function was typically achieved in about ten seconds, even though there are three times as many observations in our sample.

[7] All of our results apply equally to a continuous-time Gaussian model. Also, we assume that the risk factors, and hence the yield curve y_t, are first-order Markov. See the supplement to this paper (Joslin, Singleton, and Zhu (2010)) and Joslin, Le, and Singleton (2010) for relaxations of this assumption.

$$\Delta X_t = K_{0X}^{\mathbb{P}} + K_{1X}^{\mathbb{P}} X_{t-1} + \Sigma_X \epsilon_t^{\mathbb{P}}, \tag{1}$$

$$\Delta X_t = K_{0X}^{\mathbb{Q}} + K_{1X}^{\mathbb{Q}} X_{t-1} + \Sigma_X \epsilon_t^{\mathbb{Q}}, \tag{2}$$

$$r_t = \rho_{0X} + \rho_{1X} \cdot X_t, \tag{3}$$

where r_t is the one-period spot interest rate, $\Sigma_X \Sigma_X{}'$ is the conditional covariance matrix of X_t, and $\epsilon_t^{\mathbb{P}}, \epsilon_t^{\mathbb{Q}} \sim N(0, I_N)$. A canonical *GDTSM* is one that is maximally flexible in its parameterization of both the $\mathbb{Q}$ and $\mathbb{P}$ distributions of X_t, subject only to normalizations that ensure econometric identification. Before formally deriving our canonical *GDTSM* we briefly outline the basic idea. Variations of our canonical form, as well as some of its key implications for model specification and analysis, are discussed subsequently.

Suppose that N zero-coupon bond yields or N linear combinations of such yields, $\mathcal{P}_t$, are priced perfectly by the model (subsequently we relax this assumption). By a slight abuse of nomenclature we will refer to these linear combinations of yields as portfolios of yields. Applying invariant transformations[8] we show (i) the pricing factors X_t in (3) can be replaced by the observable $\mathcal{P}_t$, and (ii) the $\mathbb{Q}$ distribution of $\mathcal{P}_t$ can be fully characterized by the parameters $\Theta_{\mathcal{P}}^{\mathbb{Q}} \equiv (k_\infty^{\mathbb{Q}}, \lambda^{\mathbb{Q}}, \Sigma_{\mathcal{P}})$, where $\lambda^{\mathbb{Q}}$ is the vector of eigenvalues of $K_{1X}^{\mathbb{Q}}$ and $\Sigma_{\mathcal{P}} \Sigma_{\mathcal{P}}'$ is the covariance of innovations to the portfolios of yields.[9] When the model is stationary under $\mathbb{Q}$, $k_\infty^{\mathbb{Q}}$ is proportional to the risk-neutral long-run mean of the short rate $r_\infty^{\mathbb{Q}}$ and a *GDTSM* can be equivalently parameterized in terms of either parameter (see below).

The prices of all coupon bonds (as well as interest rate derivatives) are determined as functions of these observable pricing factors through no arbitrage. Importantly, though the pricing factors are now observable, the underlying parameter space of the $\mathbb{Q}$ distribution of $\mathcal{P}$ is still fully characterized by $\Theta_{\mathcal{P}}^{\mathbb{Q}}$. Moreover, the parameters of the $\mathbb{P}$ distribution of the (newly rotated and observable) state vector $\mathcal{P}_t$ are $(K_{0\mathcal{P}}^{\mathbb{P}}, K_{1\mathcal{P}}^{\mathbb{P}})$ along with $\Sigma_{\mathcal{P}}$. The remainder of this section fleshes out these ideas.

The model-implied yield on a zero-coupon bond of maturity m is an affine function of the

[8] Invariant transforms (Dai and Singleton (2000)) involve rotating, scaling and translating the state and parameter vectors to keep the short rate and bond prices unchanged (invariant), usually by a mapping $Y_t = AX_t + b$, where A is an invertible matrix. The transformed parameters are outlined in Appendix B.

[9] Duffie and Kan (1996) and Cochrane and Piazzesi (2005) also propose to use an identification scheme where the yields themselves are factors. Adrian and Moench (2008) explore a setting where the pricing factors are the portfolios themselves; however, they do not impose the internal consistency condition to make the factors equal to their no arbitrage equivalents and instead focus on the measurement errors. Our formulation offers an analytic parametrization and additionally makes transparent our subsequent results.

state X_t (Duffie and Kan (1996)):

$$y_{t,m} = A_m(\Theta_X^{\mathbb{Q}}) + B_m(\Theta_X^{\mathbb{Q}}) \cdot X_t, \tag{4}$$

where (A_m, B_m) satisfy well-known Riccati difference equations (see Appendix A for a summary), and $\Theta_X^{\mathbb{Q}} = (K_{0X}^{\mathbb{Q}}, K_{1X}^{\mathbb{Q}}, \Sigma_X, \rho_{0X}, \rho_{1X})$ is the vector of parameters from (2–3) relevant for pricing. We let $(m_1, m_2, \ldots, m_J)$ be the set of maturities (in years) of the bonds used in estimation of a *GDTSM*, $J > N$, and $y_t' = (y_{t,m_1}, \ldots, y_{t,m_J}) \in \mathbb{R}^J$ be the corresponding set of model-implied yields.

In general, (4) may be violated in the data due to market effects (e.g. bid-ask spreads or repo specials), violations of no arbitrage or measurement errors. We will collectively refer to all of these possibilities simply as measurement or pricing errors. To distinguish between model-implied and observed yields in the presence of pricing errors we let $y_{t,m}^o$ denote the yields that are observed *with measurement error*. To be consistent with the data, we must impose auxiliary structure on a *GDTSM*, beyond no arbitrage, in the form of a parametric distributional assumption for the measurement errors. We let $\{P^{\theta_m}\}_{\theta_m \in \Theta_m}$ denote the family of measures which describe the conditional distribution of $y_t - y_t^o$.

For any full-rank, portfolio matrix $W \in \mathbb{R}^{N \times J}$ we let $\mathcal{P}_t \equiv W y_t$ denote the associated N-dimensional set of portfolios of yields, where the i^{th} portfolio puts weight $W_{i,j}$ on the yield for maturity m_j. Applying (4) we obtain

$$\mathcal{P}_t = A_W(\Theta_X^{\mathbb{Q}}) + B_W(\Theta_X^{\mathbb{Q}})' X_t, \tag{5}$$

where $A_W = W[A_{m_1}, \ldots, A_{m_J}]'$ and $B_W = [B_{m_1}, \ldots, B_{m_J}]W\prime$. Note that $B_W(K_{1X}^{\mathbb{Q}}, \rho_1)$ depends only on the subset $(K_{1X}^{\mathbb{Q}}, \rho_1)$ of $\Theta_X^{\mathbb{Q}}$ (see (A3) in Appendix A).

Initially, we assume that there exist portfolios for which the no arbitrage pricing relations hold exactly:

Case P: There are N portfolios of bond yields $\mathcal{P}_t$, constructed with weights W, that are priced perfectly by the *GDTSM*: $\mathcal{P}_t^o = \mathcal{P}_t$.

We refer to the case where each portfolio consists of a single bond, so that N yields are priced perfectly, as Case **Y**. We defer until Section 6 the case where all bonds are measured with errors and estimation is accomplished by Kalman filtering.

We now state our main result for Case **P**:

Theorem 1. *Suppose that Case* **P** *holds for given fixed portfolio weights* W*. Then any canonical* GDTSM *is observationally equivalent to a unique* GDTSM *whose pricing factors*

$\mathcal{P}_t$ are the portfolios of yields $Wy_t = Wy_t^o$. Moreover, the $\mathbb{Q}$-distribution of $\mathcal{P}_t$ is uniquely determined by $(\lambda^{\mathbb{Q}}, k_\infty^{\mathbb{Q}}, \Sigma_{\mathcal{P}})$, where $\lambda^{\mathbb{Q}}$ is ordered.[10] *That is,*

$$\Delta\mathcal{P}_t = K_{0\mathcal{P}}^{\mathbb{P}} + K_{1\mathcal{P}}^{\mathbb{P}}\mathcal{P}_{t-1} + \Sigma_{\mathcal{P}}\epsilon_t^{\mathbb{P}} \tag{6}$$

$$\Delta\mathcal{P}_t = K_{0\mathcal{P}}^{\mathbb{Q}} + K_{1\mathcal{P}}^{\mathbb{Q}}\mathcal{P}_{t-1} + \Sigma_{\mathcal{P}}\epsilon_t^{\mathbb{Q}} \tag{7}$$

$$r_t = \rho_{0\mathcal{P}} + \rho_{1\mathcal{P}} \cdot \mathcal{P}_t, \tag{8}$$

is a canonical GDTSM*, where $K_{0\mathcal{P}}^{\mathbb{Q}}, K_{1\mathcal{P}}^{\mathbb{Q}}, \rho_{0\mathcal{P}}$ and $\rho_{1\mathcal{P}}$ are explicit functions of $(\lambda^{\mathbb{Q}}, k_\infty^{\mathbb{Q}}, \Sigma_{\mathcal{P}})$. Our canonical form is parametrized by $\Theta^{\mathcal{P}} = (\lambda^{\mathbb{Q}}, k_\infty^{\mathbb{Q}}, K_{0\mathcal{P}}^{\mathbb{P}}, K_{1\mathcal{P}}^{\mathbb{P}}, \Sigma_{\mathcal{P}})$.*

We refer to the *GDTSM* in Theorem 1 as the JSZ canonical form parametrized by $\Theta^{\mathcal{P}}$. Before formally proving Theorem 1, we outline the main steps. First, we want to show that any *GDTSM* is observationally equivalent to a model where the states are the observed bond portfolios $\mathcal{P}_t$ (with corresponding weights W.) Thus, for $\mathcal{G} = \{(K_0^{\mathbb{Q}}, K_1^{\mathbb{Q}}, \rho_0, \rho_1, K_0^{\mathbb{P}}, K_1^{\mathbb{P}}, \Sigma)\}$, the set of all possible *GDTSM*s,[11] we want to show that every $\Theta \in \mathcal{G}$ is observationally equivalent to some $\Theta_{\mathcal{P}} \in \mathcal{G}_{\mathcal{P}}^W$, where

$$\mathcal{G}_{\mathcal{P}}^W = \{(K_0^{\mathbb{Q}}, K_1^{\mathbb{Q}}, \rho_0, \rho_1, K_0^{\mathbb{P}}, K_1^{\mathbb{P}}, \Sigma) : \text{the factors are portfolios with weights } W\}.$$

This first step is easily established: for any *GDTSM* with latent state X_t, $\mathcal{P}_t$ satisfies (5). Following Dai and Singleton (2000) (DS), we can, by applying the change of variables outlined in Appendix B, compute the dynamics (under both $\mathbb{P}$ and $\mathbb{Q}$) of $\mathcal{P}_t$ and express r_t as an affine function of $\mathcal{P}_t$. The parameters after this change of variables give an observationally equivalent model where the states are the portfolios of yields.

Second, we establish uniqueness by showing that no two *GDTSM*s in $\mathcal{G}_{\mathcal{P}}^W$ are observationally equivalent. Clearly, if two *GDTSM*s are observationally equivalent and have the same observable factors, it must be that $(K_0^{\mathbb{P}}, K_1^{\mathbb{P}}, \Sigma)$ are the same. Intuitively, if the parameters $(K_0^{\mathbb{Q}}, K_1^{\mathbb{Q}}, \rho_0, \rho_1)$ are not the same, the price of some bonds would depend differently on the factors, a contradiction. In the second step, we formalize this intuition. Moreover, we show

[10] We fix an arbitrary ordering on the complex numbers such that 0 is the smallest number.

[11] More formally, we think of the set of *GDTSM*s as a set of stochastic processes for the yield curve rather than as a set of parameters governing the stochastic process of the yield curve. To see the correspondence, we define on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (with associated filtration $\{\mathcal{F}_t\}$) the processes $y : \Omega \times \mathbb{N} \to \mathbb{R}^{\mathbb{N}_+}$. Here $y_t^m(\omega)$ is the m-period yield at time t when the state is $\omega \in \Omega$. When our additional assumption that y is a Gaussian Markov process and no arbitrage is maintained (with risk premia at time t depending only on $\mathcal{F}_t$), these processes take the form of (1–3) and (4) for some parameters. In this way, we define a surjective map from the set of *GDTSM* parameters $(K_0^{\mathbb{Q}}, K_1^{\mathbb{Q}}, \rho_0, \rho_1, K_0^{\mathbb{P}}, K_1^{\mathbb{P}}, \Sigma)$ to the set of *GDTSM* stochastic processes. With this association, two *GDTSM*s are observationally equivalent when the corresponding stochastic processes have the same finite-dimensional distributions.

that for given $\lambda^{\mathbb{Q}}$ and $k_\infty^{\mathbb{Q}}$, there exists a unique $(K_0^{\mathbb{Q}}, K_1^{\mathbb{Q}}, \rho_0, \rho_1)$ consistent with no arbitrage and the states being the portfolios of yields $\mathcal{P}_t$. In the third and final step, we reparamatrize $\mathcal{G}_{\mathcal{P}}^W$ in terms of the free parameters $(k_\infty^{\mathbb{Q}}, r_\infty^{\mathbb{Q}}, \Sigma_{\mathcal{P}})$.

In the second step of our proof of Theorem 1, we will use the following analogue of the canonical form in Joslin (2007), proved in Appendix C.

Proposition 1. *Every canonical* GDTSM *is observationally equivalent to the canonical* GDTSM *with* $r_t = \iota \cdot X_t$,

$$\Delta X_t = K_{0X}^{\mathbb{Q}} + K_{1X}^{\mathbb{Q}} X_{t-1} + \Sigma_X \epsilon_t^{\mathbb{Q}}, \tag{9}$$

$$\Delta X_t = K_{0X}^{\mathbb{P}} + K_{1X}^{\mathbb{P}} X_{t-1} + \Sigma_X \epsilon_t^{\mathbb{P}}, \tag{10}$$

where ι *is a vector of ones,* Σ_X *is lower triangular (with positive diagonal),* $K_{1X}^{\mathbb{Q}}$ *is in ordered real Jordan form,* $K_{0X,1}^{\mathbb{Q}} = k_\infty^{\mathbb{Q}}$ *and* $K_{0X,i}^{\mathbb{Q}} = 0$ *for* $i \neq 1$*, and* $\epsilon_t^{\mathbb{Q}}, \epsilon_t^{\mathbb{P}} \sim N(0, I_N)$.

Here we specify the Jordan form with each eigenvalue associated with a single Jordan block (that is, each eigenvalue has geometric multiplicity of one.) Thus, when the eigenvalues are all real, $K_{1X}^{\mathbb{Q}}$ takes the form

$$K_{1X}^{\mathbb{Q}} = J(\lambda^{\mathbb{Q}}) \equiv \text{diag}(J_1^{\mathbb{Q}}, J_2^{\mathbb{Q}}, \ldots, J_m^{\mathbb{Q}}), \quad \text{where each } J_i^{\mathbb{Q}} = \begin{pmatrix} \lambda_i^{\mathbb{Q}} & 1 & \cdots & 0 \\ 0 & \lambda_i^{\mathbb{Q}} & \cdots & 0 \\ \vdots & \vdots & \ddots & 1 \\ 0 & \cdots & 0 & \lambda_i^{\mathbb{Q}} \end{pmatrix},$$

and where the blocks are in order of the eigenvalues. (See Appendix C for the real Jordan form when the eigenvalues are complex.) We refer to the set of Jordan canonical *GDTSM*s as $\mathcal{G}_J$ and it is parametrized by $\Theta^J = (\lambda^{\mathbb{Q}}, k_\infty^{\mathbb{Q}}, K_{0X}^{\mathbb{P}}, K_{1X}^{\mathbb{P}}, \Sigma_X)$. The eigenvalues of $\lambda^{\mathbb{Q}}$ may not be distinct and may be complex. We explore these possibilities empirically in Section 5.

Proof of Theorem 1: Having already established that we can rotate any model to one with $\mathcal{P}_t$ as the observed states, we proceed to prove the second step. Suppose that $\Theta_1, \Theta_2 \in \mathcal{G}_{\mathcal{P}}^W$ index two observationally equivalent canonical models. By the existence result in Proposition 1, each Θ_i is observationally equivalent to a *GDTSM*, Θ_i^J, which is in real ordered Jordan canonical form. Since

$$\mathcal{P}_t = A_W(\Theta_i^J) + B_W(\Theta_i)' X_{ti}^J, \tag{11}$$

where X_{ti}^J is the latent state for model Θ_i^J, it must be that

$$\Theta_i = A_W(\Theta_i^J) + B_W(\Theta_i^J)'\Theta_i^J. \tag{12}$$

Here, we use the notation that for a *GDTSM* with parameter vector Θ and state X_t, the observationally equivalent *GDTSM* with latent state $\hat{X}_t = C + DX_t$ has parameter vector $\hat{\Theta} = C + D\Theta$, as computed in Appendix B. Since observational equivalence is transitive, Θ_1^J is observationally equivalent to Θ_2^J; the uniqueness result in Proposition 1 implies that $\Theta_1^J = \Theta_2^J$. The equality in (12) then gives $\Theta_1 = \Theta_2$, which establishes our second step.

To establish the reparametrization in the third step, we focus on (11) and (12). The key is to show explicitly how given $(\lambda^{\mathbb{Q}}, k_\infty^{\mathbb{Q}})$ (from Θ_i^J) we can (i) choose the parameters $(K_{0J}^{\mathbb{P}}, K_{1J}^{\mathbb{P}}, \Sigma_J)$ to get any desired $(K_{0\mathcal{P}}^{\mathbb{P}}, K_{1\mathcal{P}}^{\mathbb{P}}, \Sigma_{\mathcal{P}})$ and (ii) construct the $(K_0^{\mathbb{Q}}, K_1^{\mathbb{Q}}, \rho_0, \rho_1)$ consistent with the factors being $\mathcal{P}_t$. Details are provided in Appendix D.

For reference, we summarize the transformations computed in the last step as:

Proposition 2. *Any canonical* GDTSM *with* $\mathbb{Q}$ *parameters* $(\lambda^{\mathbb{Q}}, k_\infty^{\mathbb{Q}}, \Sigma_{\mathcal{P}})$ *has the JSZ representation in Theorem 1 with*

$$K_{1\mathcal{P}}^{\mathbb{Q}} = BJ(\lambda^{\mathbb{Q}})B^{-1} \tag{13}$$

$$K_{0\mathcal{P}}^{\mathbb{Q}} = k_\infty^{\mathbb{Q}} Be_{m_1} - K_{1\mathcal{P}}^{\mathbb{Q}} A \tag{14}$$

$$\rho_{1\mathcal{P}} = (B^{-1})'\iota \tag{15}$$

$$\rho_{0\mathcal{P}} = -A \cdot \rho_{1\mathcal{P}} \tag{16}$$

where e_{m_1} *is a vector with all zeros except in the* m_1^{th} *entry, which is 1 (*m_1 *is the multiplicity of* $\lambda_1^{\mathbb{Q}}$*) and* $B = B_W(J(\lambda^{\mathbb{Q}}), \iota)'$*,* $A = A_W\left(0, J(\lambda^{\mathbb{Q}}), B^{-1}\Sigma_{\mathcal{P}}, k_\infty^{\mathbb{Q}}, \iota\right)$*, where* (A_W, B_W) *are defined in (5) and (A2–A3).*

Before proceeding, we discuss the interpretation of the parameter $k_\infty^{\mathbb{Q}}$. If X is stationary under $\mathbb{Q}$, then $k_\infty^{\mathbb{Q}}$ and $r_\infty^{\mathbb{Q}}$ (the long-run $\mathbb{Q}$-mean of the short rate) are related according to $r_\infty^{\mathbb{Q}} = k_\infty^{\mathbb{Q}} \sum_{i=1}^{m_1} (-\lambda_1^{\mathbb{Q}})^{-i}$, where m_1 is the dimension of the first Jordon block $J_1^{\mathbb{Q}}$ of $K_{1X}^{\mathbb{Q}}$. Thus, if $\lambda_1^{\mathbb{Q}}$ is not a repeated root ($m_1 = 1$), $r_\infty^{\mathbb{Q}}$ is simply $-k_\infty^{\mathbb{Q}}/\lambda_1^{\mathbb{Q}}$ in stationary models. This is the case in our subsequent empirical illustrations where we express our normalization in terms of the parameter $r_\infty^{\mathbb{Q}}$ owing to its natural economic interpretation.

That $k_\infty^{\mathbb{Q}}$ and $r_\infty^{\mathbb{Q}}$ are not always interchangeable in defining a proper canonical form for the set of all *GDTSM*s of form (1–3) can be seen as follows. In proceeding to the normalization of Proposition 1, a model with the factors normalized so that $r_t = \rho_0 + \iota \cdot X_t$ is further

normalized by a level translation ($X_t \mapsto X_t - \alpha$). Such level translations can always be used to enforce $\rho_0 = 0$, but they can only be used to enforce $K^{\mathbb{Q}}_{0X} = 0$ in the case that $K^{\mathbb{Q}}_{1X}$ is invertible (i.e., there are no zero eigenvalues).[12] When $m_1 = 1$ and there are no zero eigenvalues, the following two normalizations of $(K^{\mathbb{Q}}_{0\mathcal{P}}, \rho_0)$ are equivalent:

$$K^{\mathbb{Q}}_{0\mathcal{P}} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \text{ and } \rho_0 = \frac{-k^{\mathbb{Q}}_{\infty}}{\lambda^{\mathbb{Q}}_1} \quad \text{or} \quad K^{\mathbb{Q}}_0 = \begin{pmatrix} k^{\mathbb{Q}}_{\infty} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \text{ and } \rho_0 = 0. \tag{17}$$

Theorem 1 uses the form with $k^{\mathbb{Q}}_{\infty}$, and always applies regardless of the eigenvalues of $K^{\mathbb{Q}}_{1X}$.

2 $\mathbb{P}$-dynamics and Maximum Likelihood Estimation

Rather than defining latent states indirectly through a normalization on parameters governing the dynamics (under $\mathbb{P}$ or $\mathbb{Q}$) of latent states, the JSZ normalization has instead prescribed observable yield portfolios $\mathcal{P}$ and parametrized their $\mathbb{Q}$ distribution in a maximally flexible way consistent with no arbitrage. A distinctive feature of our normalization is that, in estimation, there is an inherent separation between the parameters of the $\mathbb{P}$ and $\mathbb{Q}$ distributions of $\mathcal{P}_t$. In contrast, when the risk factors are latent, estimates of the parameters governing the $\mathbb{P}$ distribution necessarily depends on those of the $\mathbb{Q}$ distribution of the state, since the pricing model is required to either invert the model for the fitted states (when N bonds are priced perfectly) or to filter for the unobserved states (when all bonds are measured with errors). This section formalizes this "separation property" of the JSZ normalization.

By Theorem 1, we can, without loss of generality, use N portfolios of the yields, $\mathcal{P}_t = \mathcal{P}^o_t \in \mathbb{R}^N$, as observed factors. Suppose that the individual bond yields, y_t, are to be used in estimation and that their associated measurement errors, $y^o_t - y_t$, have the conditional distribution P^{θ_m}, for some $\theta_m \in \Theta_m$. We require only that, for any P^{θ_m}, these errors are conditionally independent of lagged values of the measurement errors and satisfy the consistency condition $\mathbb{P}(Wy^0_t = \mathcal{P}_t | \mathcal{P}_t) = 1$.[13] Then the conditional likelihood function (under

[12] One implication of this observation is that setting both $k^{\mathbb{Q}}_{\infty}$ and $r^{\mathbb{Q}}_{\infty}$ to zero in the presence of a $\mathbb{Q}$-nonstationary risk factor, as was done by Christensen, Diebold, and Rudebusch (2007, 2009) in defining their arbitrage-free Nelson-Siegel model, amounts to imposing an over-identifying restriction on the drift of X_{1t}.

[13] Implicit in this formulation is the possibility that $Cov(y^o_t | \mathcal{P}_t; \lambda^{\mathbb{Q}}, k^{\mathbb{Q}}_{\infty}, \Sigma_{\mathcal{P}})$ is singular. This would be true in Case **Y**, where some yields are measured without errors, or when certain portfolios of y^o_t are priced perfectly, as with the use of principal components as observable factors or as in Chen and Scott (1995) who

$\mathbb{P}$) of the observed data (y_t^o) is

$$f(y_t^o|y_{t-1}^o;\Theta) = f(y_t^o|\mathcal{P}_t;\lambda^{\mathbb{Q}},k_\infty^{\mathbb{Q}},\Sigma_{\mathcal{P}},P^{\theta_m}) \times f(\mathcal{P}_t|\mathcal{P}_{t-1};K_{1\mathcal{P}}^{\mathbb{P}},K_{0\mathcal{P}}^{\mathbb{P}},\Sigma_{\mathcal{P}}). \tag{18}$$

Notice the convenient separation of parameters in the likelihood function. The conditional distribution of the yields measured with errors depends only on $(\lambda^{\mathbb{Q}},k_\infty^{\mathbb{Q}},\Sigma_{\mathcal{P}},P^{\theta_m})$ and not on $(K_{0\mathcal{P}}^{\mathbb{P}},K_{1\mathcal{P}}^{\mathbb{P}})$. In contrast, the conditional $\mathbb{P}$-density of the pricing factors $\mathcal{P}_t$ depends only on $(K_{1\mathcal{P}}^{\mathbb{P}},K_{0\mathcal{P}}^{\mathbb{P}},\Sigma_{\mathcal{P}})$, and not on $(\lambda^{\mathbb{Q}},k_\infty^{\mathbb{Q}})$. Using the assumption that $\mathcal{P}_t$ is conditionally Gaussian, the second term in (18) can be expressed as

$$f(\mathcal{P}_t|\mathcal{P}_{t-1};K_{1\mathcal{P}}^{\mathbb{P}},K_{0\mathcal{P}}^{\mathbb{P}},\Sigma_{\mathcal{P}}) = (2\pi)^{-N/2}|\Sigma_{\mathcal{P}}|^{-1}\exp\left(-\frac{1}{2}\|\Sigma_{\mathcal{P}}^{-1}\left(\mathcal{P}_t - E_{t-1}[\mathcal{P}_t]\right)\|^2\right), \tag{19}$$

where $E_{t-1}[\mathcal{P}_t] = K_{0\mathcal{P}}^{\mathbb{P}} + (I + K_{1\mathcal{P}}^{\mathbb{P}})\mathcal{P}_{t-1}$ and where for a vector x, $\|x\|^2$ denotes the euclidean norm squared: $\sum x_i^2$. The parameters $(K_{0\mathcal{P}}^{\mathbb{P}},K_{1\mathcal{P}}^{\mathbb{P}})$ that maximize the likelihood function f (conditional on $t=0$ information), namely

$$(K_{0\mathcal{P}}^{\mathbb{P}},K_{1\mathcal{P}}^{\mathbb{P}}) = \text{argmax}\sum_{t=1}^{T} f(y_t^o|y_{t-1}^o;K_{1\mathcal{P}}^{\mathbb{P}},K_{0\mathcal{P}}^{\mathbb{P}},\Sigma_{\mathcal{P}}) = \text{argmin}\sum_{t=1}^{T}\|\Sigma_{\mathcal{P}}^{-1}\left(\mathcal{P}_t^o - E_{t-1}[\mathcal{P}_t^o]\right)\|^2, \tag{20}$$

are the sample ordinary least squares (OLS) estimates, independent of $\Sigma_{\mathcal{P}}$ (Zellner (1962)). Summarizing these observations:

Proposition 3. *Under Case* **P** *the ML estimates of the* $\mathbb{P}$ *parameters* $(K_{0\mathcal{P}}^{\mathbb{P}},K_{1\mathcal{P}}^{\mathbb{P}})$ *are given by the OLS estimates of the conditional mean of* $\mathcal{P}_t$.

Absent constraints linking the $\mathbb{P}$ and $\mathbb{Q}$ dynamics, one can effectively separate the time-series properties ($\mathbb{P}$) of $\mathcal{P}_t$ from the cross-sectional constraints imposed by no arbitrage ($\mathbb{Q}$). The parameters governing $\mathbb{P}$-forecasts distribution thus can be estimated from time series alone, regardless of the cross-sectional restrictions. Furthermore, *independent* of $(\lambda^{\mathbb{Q}},k_\infty^{\mathbb{Q}},\Sigma_{\mathcal{P}})$, the OLS estimates of $(K_{0\mathcal{P}}^{\mathbb{P}},K_{1\mathcal{P}}^{\mathbb{P}})$ are by construction globally optimal. With $(K_{0\mathcal{P}}^{\mathbb{P}},K_{1\mathcal{P}}^{\mathbb{P}})$ at hand, we use the sample conditional variance of $\mathcal{P}_t$, $\hat{\Sigma}_{\mathcal{P}}\hat{\Sigma}_{\mathcal{P}}'$, computed from the OLS innovations as the starting value for the population variance $\Sigma_{\mathcal{P}}\Sigma_{\mathcal{P}}'$. Given $(\lambda^{\mathbb{Q}},k_\infty^{\mathbb{Q}})$, this starting value for $\Sigma_{\mathcal{P}}\Sigma_{\mathcal{P}}'$ is again by construction close to the global optimum. Therefore we

use different portfolios of yields as their factors. This setup also accommodates the case where both $\mathcal{P}$ and some of the individual components of y_t^o are priced perfectly by the *GDTSM*. Furthermore, the errors may be correlated, non-normal, or have time-varying conditional moments depending on $\mathcal{P}_t$. In practice, it has typically been assumed that the pricing errors are normally distributed.

have greatly reduced the number of parameters to be estimated. For instance, in a $GDTSM(3)$ model, the maximum number of parameters, excluding those governing P^{θ_m}, is 22 (3 for $\lambda^{\mathbb{Q}}$, 1 for $k_\infty^{\mathbb{Q}}$, 6 for $\Sigma_{\mathcal{P}}$, 3 for $K_{0\mathcal{P}}^{\mathbb{P}}$ and 9 for $K_{1\mathcal{P}}^{\mathbb{P}}$). With our normalization one can focus on only the 4 parameters $(\lambda^{\mathbb{Q}}, k_\infty^{\mathbb{Q}})$. This underlies the substantial improvement in estimation speed for the JSZ normalization over other canonical forms.

Key to our argument is the fact that we can parametrize of the conditional distribution of the yields measured with error independently of the parameters governing the $\mathbb{P}$-conditional mean of $\mathcal{P}$ in the sense of the factorization (18). For any $(K_{0\mathcal{P}}^{\mathbb{P}}, K_{1\mathcal{P}}^{\mathbb{P}}, \Sigma_{\mathcal{P}}, \lambda^{\mathbb{Q}}, k_\infty^{\mathbb{Q}})$, we have

$$\begin{aligned} &f(y_t^o|\mathcal{P}_t; \lambda^{\mathbb{Q}}, k_\infty^{\mathbb{Q}}, \Sigma_{\mathcal{P}}) \times f(\mathcal{P}_t|\mathcal{P}_{t-1}; K_{1\mathcal{P}}^{\mathbb{P}}, K_{0\mathcal{P}}^{\mathbb{P}}, \Sigma_{\mathcal{P}}) \\ &\quad \leq f(y_t^o|\mathcal{P}_t; \lambda^{\mathbb{Q}}, k_\infty^{\mathbb{Q}}, \Sigma_{\mathcal{P}}) \times f(\mathcal{P}_t|\mathcal{P}_{t-1}; K_{1\mathcal{P},OLS}^{\mathbb{P}}, K_{0\mathcal{P},OLS}^{\mathbb{P}}, \Sigma_{\mathcal{P}}), \end{aligned} \tag{21}$$

where we suppress the dependence on P^{θ_m}. This inequality follows from the observations that $(K_{0\mathcal{P}}^{\mathbb{P}}, K_{1\mathcal{P}}^{\mathbb{P}})$ has no effect on $f(y_t^o|\mathcal{P}_t)$ and that, for any $\Sigma_{\mathcal{P}}$, replacing $(K_{0\mathcal{P}}^{\mathbb{P}}, K_{1\mathcal{P}}^{\mathbb{P}})$ by its OLS estimate increases $f(\mathcal{P}_t|\mathcal{P}_{t-1})$.[14]

It is instructive to compare (18) with the likelihood function that arises in models with observable factors that parameterize the $\mathbb{P}$ distribution of $\mathcal{P}$ and the market prices of these risks. In this case the parameters are $(K_{0\mathcal{P}}^{\mathbb{P}}, K_{1\mathcal{P}}^{\mathbb{P}})$ and $(\rho_0, \rho_1, \Lambda_0, \Lambda_1, \Sigma_{\mathcal{P}})$, where $E_t^{\mathbb{P}}[\mathcal{P}_{t+1}] = E_t^{\mathbb{Q}}[\mathcal{P}_{t+1}] + \Sigma_{\mathcal{P}}(\Lambda_0 + \Lambda_1\mathcal{P}_t)$, for state-dependent market prices of risk $\Lambda_0 + \Lambda_1\mathcal{P}_t$. These parameters are subject to the internal consistency constraints $A_W = 0$ and $B_W = I_N$ that ensure that the model replicates the portfolios of yields $\mathcal{P}$. Moreover, analogous to (18), the factorization of the likelihood function takes the form

$$f(y_t^o|y_{t-1}^o; \Theta) = f(y_t^o|\mathcal{P}_t; K_{0\mathcal{P}}^{\mathbb{P}}, K_{1\mathcal{P}}^{\mathbb{P}}, \Sigma_{\mathcal{P}}, \rho_0, \rho_1, \Lambda_0, \Lambda_1) \times f(\mathcal{P}_t|\mathcal{P}_{t-1}; K_{1\mathcal{P}}^{\mathbb{P}}, K_{0\mathcal{P}}^{\mathbb{P}}, \Sigma_{\mathcal{P}}). \tag{22}$$

Replacing $(K_{0\mathcal{P}}^{\mathbb{P}}, K_{1\mathcal{P}}^{\mathbb{P}})$ with $(K_{0\mathcal{P},OLS}^{\mathbb{P}}, K_{1\mathcal{P},OLS}^{\mathbb{P}})$ again increases the second term, but now the first term is affected as well. Thus, within this parameterization, the fact that OLS recovers the ML estimates is completely obscured.[15]

[14] The last step requires observable factors – another important element of our argument. See Section 3 and (23).

[15] In fact, within a macro-$GDTSM$ with a similar parametrization of internally consistent market prices of risk and observable factors, Ang, Piazzesi, and Wei (2003) report that OLS estimates of $E^{\mathbb{P}}[\mathcal{P}_{t+1}|\mathcal{P}_t]$ are (slightly) different from their ML estimates. Our analysis generalizes to macro-$GDTSM$s (see Joslin, Le, and Singleton (2010)) and so, in fact, their OLS estimates *are* the ML estimates.

3 On the Relevance of No Arbitrage for Forecasting

The decomposition of the conditional likelihood function of the data in (18) leads immediately to several important insights about the potential roles of no-arbitrage restrictions for out-of-sample forecasting. First, Proposition 3 gives a general striking property of *GDTSM*s under Case **P**: the no-arbitrage feature of a *GDTSM* has no effect on the *ML* estimates of $K^{\mathbb{P}}_{0\mathcal{P}}$ and $K^{\mathbb{P}}_{1\mathcal{P}}$. This, in turn, implies that forecasts of future values of $\mathcal{P}$ are identical to those from an unconstrained $VAR(1)$ model for $\mathcal{P}_t$.[16] This result sharpens Duffee (2009)'s finding that the restrictions on a VAR implied by an arbitrage-free *GDTSM* cannot be rejected against the alternative of an unrestricted VAR.[17] When forecasting the N portfolios of yields $\mathcal{P}_t$, Proposition 3 shows *theoretically* that a similar result *must* hold insofar as Case **P** is (approximately) valid.

The JSZ normalization makes these observations particularly transparent. In contrast, in the (observationally equivalent) specification in (1–3), portfolio yield forecasts are

$$\begin{aligned} E_t[\mathcal{P}_{t+1}] - \mathcal{P}_t &= B_W(\Theta^{\mathbb{Q}}) \left(E_t[X_{t+1}] - X_t\right) = B_W(\Theta^{\mathbb{Q}}) \left(K^{\mathbb{P}}_{0X} + K^{\mathbb{P}}_{1X} X_t\right) \\ &= B_W(\Theta^{\mathbb{Q}}) \left(K^{\mathbb{P}}_{0X} + K^{\mathbb{P}}_{1X} (B_W P(\Theta^{\mathbb{Q}})^{-1} \left(\mathcal{P}_t - A_W(\Theta^{\mathbb{Q}})\right)\right). \end{aligned} \tag{23}$$

Thus, with latent states, the portfolio forecasts are expressed in terms of both the $\mathbb{P}$ and $\mathbb{Q}$ parameters of the model. From (23) it is not obvious that OLS recovers the ML estimates of $(K^{\mathbb{P}}_{0\mathcal{P}}, K^{\mathbb{P}}_{1\mathcal{P}})$. The JSZ normalization makes the implicit cancellations in (23) explicit.

Second, the structure of the likelihood function reveals that, in contrast to the pricing factors, no-arbitrage restrictions are potentially relevant for forecasting individual yields that are measured with error. The conditional density of y^o_t given $\mathcal{P}_t$ depends on the parameters of the risk-neutral distribution, and these are revealed through the cross-maturity restrictions implied by no arbitrage. In addition, diffusion invariance implies that $\Sigma_{\mathcal{P}}$ enters both terms of the likelihood function so efficient estimation of these parameters comes from imposing the structure of a *GDTSM*.

Finally, the structure of the density $f(y^o_t|\mathcal{P}_t)$ also reveals the natural alternative model for

[16] Note that, in principle, enforcing no-arbitrage restrictions may be relevant for the construction of forecast confidence intervals through the dependence on $\Sigma_{\mathcal{P}}$. However, empirically this effect is likely to be small.

[17] Duffee (2009) also shows theoretically that no arbitrage is *cross-sectionally* irrelevant in any affine model under the stochastically singular condition of no measurement errors. That is, if the model exactly fits the data without measurement errors, the cross-sectional loadings (A,B) of (4) are determined without reference to solving the Ricatti difference equations (A2–A3). Duffee does not theoretically explore the time series implications of the no measurement error assumption. In this case, not only would Proposition 3 apply (since Case **P** is a weaker assumption) so that the OLS estimates are the ML estimates of $(K^{\mathbb{P}}_{0\mathcal{P}}, K^{\mathbb{P}}_{1\mathcal{P}})$, but also $\Sigma_{\mathcal{P}}$ could be inferred from a sufficiently large cross-section of bond prices.

assessing gains in forecast precision from imposing no-arbitrage restrictions. The state-space representation of this unconstrained model reflects the presumption that bond yields have a low-dimensional factor structure, but it does not impose the restrictions implied by a no-arbitrage $DTSM$. Specifically, under Case **P** where $\mathcal{P}_t$ is priced perfectly by the $GDTSM$, the state equation is

$$\Delta X_{t+1} = K_{0X} + K_{1X} X_t + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma_X) \; i.i.d., \tag{24}$$

and the observation equation

$$\begin{pmatrix} \mathcal{P}_t \\ y_t^o \end{pmatrix} = C + DX_t + \begin{pmatrix} 0 \\ e_{mt} \end{pmatrix}, \quad e_{mt} \sim P^{\theta_m} \;\; i.i.d. \tag{25}$$

The parameter set is $\Theta_{SS} = \{(K_{0X}, K_{1X}, \Sigma_X, C, D, P^{\theta_m})\}$, where P^{θ_m} is an observation error distribution that is consistent with Case **P**.

No arbitrage requires that the observation equation parameters (C, D) must be of the form (4); that is, the dynamics are Gaussian under $\mathbb{Q}$. Additionally, no arbitrage enforces a link between the possible (C, D) and Σ_X (diffusion invariance). Since the parameters are not identified, one also imposes normalizations to achieve a just-identified model. Importantly, the choice of normalizations will in general affect the ML estimates of the parameters, Θ_{SS}, but will not affect the distribution of bond yields implied from the state space model (either in the cross section or time series.) For example, one could impose the identification scheme in Dai and Singleton (2000) either under the $\mathbb{P}$- or the $\mathbb{Q}$-measure. The estimates of (K_{0X}, K_{1X}) and (C, D) will be choice-specific, but these differences will be offset by changes in the latent states so that the fits to bond yields will be identical.

Notably, the unconstrained state-space representation (24)-(25) with parameter set Θ_{SS} is not the unconstrained J-dimensional VAR representation of y_t. The latter relaxes both the no-arbitrage (and any over-identifying restrictions) enforced in the $GDTSM$ *and* the assumed factor structure of bond yields (the dimension of X_t is less than the dimension of y_t^o). Consequently, gains in forecasting an individual yield using a $GDTSM$, relative to the forecasts from an unconstrained VAR model of y_t, may be due to the VAR being over-parametrized relative to the unconstrained factor model, the imposition of no-arbitrage restrictions within the $GDTSM$, or both. The role of no-arbitrage restrictions is an empirical issue that can be addressed by comparing the constrained and unconstrained versions of (24)-(25).

4 Irrelevance of Factor Structure for Forecasting

The *DTSM* literature considers a number of further constraints on the factor structure of a *GDTSM*, beyond those implied by the absence of arbitrage. In addition to making different identification assumptions, one can form a parsimonious model by restricting the distribution of certain variables (under either $\mathbb{P}$ or $\mathbb{Q}$) or by restricting the structure of risk premia. We first extend the results of Section 3 to characterize when this irrelevancy result does (and does not) hold in more general *GDTSM*s, and then we discuss the connection of our results to specific over-identified *GDTSM*s in the literature.

Within the state space model (24–25), the parameters (C, D) control the cross-sectional relationship among the yields, while P^{θ_m} controls the distribution of the measurement errors. The covariance matrix of the innovations of the latent states Σ_X is linked to $\Sigma_{\mathcal{P}}$ through the factor loadings (C, D). The restriction of no-arbitrage, for example, says both that only certain types of loadings (C, D) are feasible (those given by (4)) and that this feasible set depends on the particular value of Σ_X. Thus no arbitrage is a cross-parameter restriction on the feasible set of (C, D, Σ_X) in the general state space model. More generally, one might be interested in restrictions on a particular subset of the parameters $\eta \equiv (C, D, P_m^{\theta}, \Sigma_X)$, examples of which we discuss in subsequent subsections. The following theorem says that even if restrictions are imposed on η, as long as (K_{0X}, K_{1X}) are unrestricted, *OLS* will recover the *ML* estimates of $(K_{0\mathcal{P}}, K_{1\mathcal{P}})$. (K_{0X}, K_{1X}) will change in general with the restrictions imposed on η, but only through an affine transformation of the latent states.

Theorem 2. *Given the state space model (24– 25) and the portfolio matrix W determining the factors $\mathcal{P}_t$, let $\mathcal{H}$ be a subset of the admissible set of η where, for any $(C, D, \Sigma_X, P^{\theta_m}) \in \mathcal{H}$, the $N \times N$ upper left block of D is full rank. Consider the ML problem with η constrained to lie in the subspace $\mathcal{H}$:*

$$(K_{0X}^{\mathcal{H}}, K_{1X}^{\mathcal{H}}, \eta^{\mathcal{H}}) \in \underset{K_{0X}, K_{1X}; \eta \in \mathcal{H}}{\arg\max} f(\mathcal{P}_T, y_T, \ldots, \mathcal{P}_1, y_1 | \mathcal{P}_0, y_0).$$

Then $(K_{0X}^{\mathcal{H}}, K_{1X}^{\mathcal{H}}, \eta^{\mathcal{H}})$ are such that

$$K_{0\mathcal{P}} = D_{\mathcal{P}}^{\mathcal{H}} K_{0X}^{\mathcal{H}} - D_{\mathcal{P}}^{\mathcal{H}} K_{1X}^{\mathcal{H}} (D_{\mathcal{P}}^{\mathcal{H}})^{-1} C_{\mathcal{P}}^{\mathcal{H}}, \tag{26}$$

$$K_{1\mathcal{P}} = D_{\mathcal{P}}^{\mathcal{H}} K_{1X}^{\mathcal{H}} (D_{\mathcal{P}}^{\mathcal{H}})^{-1}, \tag{27}$$

where $C_{\mathcal{P}}^{\mathcal{H}}$ is the first N elements of $C^{\mathcal{H}}$, $D_{\mathcal{P}}^{\mathcal{H}}$ is the upper left $N \times N$ block of $D^{\mathcal{H}}$ and

$(K_{0\mathcal{P}}, K_{1\mathcal{P}})$ *are the OLS estimates of the regression*

$$\Delta\mathcal{P}_t = K_{0\mathcal{P}} + K_{1\mathcal{P}}\mathcal{P}_t + \epsilon_t^{\mathcal{P}}.$$

The proof is similar, though notationally more abstract, to the proof of Proposition 3 and is presented in Appendix E.

Using this result, we first illustrate the estimation of the general state space model of (24–25) when the possibility of arbitrage is not precluded. We next explore the implications of restrictions on the $\mathbb{Q}$ and $\mathbb{P}$ distributions, as well as on risk premia, for the conditional distribution of $\mathcal{P}_t$.

4.1 Factor Structure in Arbitrage Models

The factor model (24–25) is not necessarily consistent with the absence of arbitrage. This is because the loadings in (25) may not come from the solution of (4) for a given choice of $\Theta_X^{\mathbb{Q}}$. Nevertheless, this model is still of interest as it provides a baseline "factor structure" for the yield curve (cf. Duffee (2009)). Theorem 2 implies that, under Case **P**, the OLS estimates of the parameters governing (24) are identical to their counterparts from system ML estimation of (24–25) when the factors $\mathcal{P}_t$ are observed portfolios of bond yields.

Additionally when, in addition to Case **P**, the state space model has temporally i.i.d. normal pricing errors in (25), and these errors are orthogonal to the portfolio matrix W, the OLS regression of the observed yields onto the factors $\mathcal{P}$ give the ML estimates of the unconstrained ("with arbitrage") cross-sectional loadings (C, D) in (25). In this case the OLS regression estimates of $\Sigma_{\mathcal{P}}$ must also correspond (through the invariant transformation given in Theorem 2) to the ML estimates of Σ_X for the factor model. Taken together, these procedures provide a simple prescription for constructing alternative reference models (to arbitrage-free $GDTSM$s) that maintain the factor structure but do not impose no-arbitrage. In the empirical analysis in Section 5 we focus on comparisons of OLS forecasts of PCs with their forecasts from a variety of arbitrage-free models. These "with arbitrage" factor models provide a natural reference model when one is interested in forecasting yields.

4.2 Irrelevance of Constraints on the $\mathbb{Q}$-distribution of Yields

The JSZ normalization characterizes the state in terms of an observable portfolio of zero coupon yields. The conditional $\mathbb{Q}$ distribution of $\mathcal{P}_{t+\tau}$ (as a function of $\mathcal{P}_t$) is expressed in (7), which we have shown can be parametrized by $(\lambda^{\mathbb{Q}}, k_\infty^{\mathbb{Q}}, \Sigma_{\mathcal{P}})$. Within the model (that is,

without measurement errors), $\mathcal{P}$ is informative about the entire yield curve. Thus one type of restriction a researcher may be interested in imposing is on the conditional $\mathbb{Q}$-distribution of $\mathcal{P}_{t+\tau}$ (or $y_{t+\tau}$) as a function of $\mathcal{P}_t$ (or y_t).[18] Such constraints further restrict (beyond the no arbitrage restrictions) the cross-sectional loadings (C, D) in the general state space model as well as which innovation covariances are possible. Theorem 2 shows that restrictions on the $\mathbb{Q}$ distribution of $y_{t+\tau}$, as a function of y_t, are irrelevant for forecasting $\mathcal{P}_t$. Put differently, in the JSZ-normalized *GDTSM*, restrictions that only affect the parameters of the $\mathbb{Q}$-distribution of $\mathcal{P}_t$ ($\lambda^{\mathbb{Q}}, k_\infty^{\mathbb{Q}}$, as well as $\Sigma_{\mathcal{P}}$) are irrelevant for forecasting the portfolios of yields $\mathcal{P}_t$. Though latent-factor representations like (23) suggest that the $\mathbb{Q}$ parameters enter into $E_t^{\mathbb{P}}[\mathcal{P}_{t+1}]$, in fact absent restrictions *across* the $\mathbb{P}$ and $\mathbb{Q}$ parameters of the model, any $\mathbb{Q}$-restrictions must affect $(K_{0X}^{\mathbb{P}}, K_{1X}^{\mathbb{P}})$ in a manner that "cancels" their impact on $E_t^{\mathbb{P}}[\mathcal{P}_{t+1}]$.

One example of such a constraint in the literature is the arbitrage-free Nelson-Siegel (AFNS) model of Christensen, Diebold, and Rudebusch (2007). The AFNS model allows for a dynamically consistent *GDTSM* where, except for a convexity-induced intercept, the factor loadings correspond to those of Nelson and Siegel (1987). Since the AFNS model is the constrained special case of the JSZ normalization with $\lambda^{\mathbb{Q}} = (0, \lambda, \lambda)$ and $k_\infty^{\mathbb{Q}} = 0$,[19] an immediate implication of this observation is that *forecasts of $\mathcal{P}$ using an arbitrage-free Nelson-Siegel (AFNS) model are equivalent to forecasts based on an unconstrained VAR(1) representation of $\mathcal{P}$*. Proposition 3 implies that these restrictions do not affect the *ML* estimates of $K_{0\mathcal{P}}^{\mathbb{P}}$ and $K_{1\mathcal{P}}^{\mathbb{P}}$ and, hence, they *cannot* improve the forecasts of $\mathcal{P}$ relative to an unconstrained VAR(1). Thus, the forecast gains that Christensen, Diebold, and Rudebusch (2007) attribute to the structure of their AFNS pricing model are, instead, a consequence of the *joint* imposition of no-arbitrage and their constraints on the $\mathbb{P}$-distribution of bond yields.

4.3 Conditions for Irrelevance of Constraints on Latent Factors

A conclusion of Section 4.2 is that restrictions on the parameters governing $\mathbb{Q}$-distribution of yield factors are irrelevant for forecasts. In this section we address the question if, more generally, a parameter constraint on "$\mathbb{Q}$ parameters" within an identified *GDTSM* with latent factors affects forecasts. For example, a researcher may consider the following procedure. They begin with a *GDTSM* model with the normalizations of Dai and Singleton (2000) (DS) applied under $\mathbb{Q}$: $(K_{0X}^{\mathbb{P}}, K_{1X}^{\mathbb{P}})$ are free while $\Sigma_X = I$, $K_{0X}^{\mathbb{Q}} = 0$ and $K_{1X}^{\mathbb{Q}}$ is (ordered) lower

[18] More precisely, under $\mathbb{Q}$, $y_{t+\tau}|\mathcal{F}_t \sim N(\mu_t^\tau, \Sigma^\tau)$. If we express $\mu_t^\tau = \mu^\tau(y_t)$, restrictions on Σ^τ or the functional form μ^τ are irrelevant. More generally, since $E_t^{\mathbb{P}}[y_{t+s}] \in \mathcal{F}_t = \sigma(y_t)$, restrictions of the form $E_t^{\mathbb{Q}}[y_{t+\tau}] = g(E_t^{\mathbb{P}}[y_{t+\tau}])$ may affect forecasts.

[19] We show this formally in Joslin, Singleton, and Zhu (2010).

triangular (or real Schur to accommodate complex eigenvalues). After estimation, a more parsimonious model is obtained by taking any coefficients in $K_{1X}^{\mathbb{Q}}$ which are insignificantly different from zero and setting them to zero (or using an iterative AIC or BIC type procedure). A similar procedure is followed in, for example, Dai and Singleton (2002).

When $K_{0X}^{\mathbb{P}}$ and $K_{1X}^{\mathbb{P}}$ are unconstrained, constraints such as these on $\mathbb{Q}$-identified parameters are joint constraints on the cross-sectional properties of the yield curve and the covariance of innovations. To see this, one can invert the latent factors into the observable factors and observe that non-linear constraints within the JSZ normalization on $(\lambda^{\mathbb{Q}}, k_{\infty}^{\mathbb{Q}}, \Sigma_{\mathcal{P}})$ will hold. However, Theorem 2 directly shows that the resulting forecasts for $\mathcal{P}_t$ will be identical whether the constraints are imposed or not. The constraints in general *will change* the estimated $K_{0X}^{\mathbb{P}}$ and $K_{1X}^{\mathbb{P}}$, but they will also change the loadings and the latent states so that the forecasts of $\mathcal{P}_t$ will not change.

Alternatively, one could first apply a normalization under $\mathbb{P}$ and then restrict the parameters governing the $\mathbb{Q}$-conditional distribution of the implied latent states. For example, as above one could apply the DS normalization under $\mathbb{P}$ where $(K_{0X}^{\mathbb{P}}, K_{1X}^{\mathbb{P}})$ will be restricted while $(K_{0X}^{\mathbb{Q}}, K_{1X}^{\mathbb{Q}})$ are restricted. Duffee and Stanton (2007), for example, apply such a normalization. With this type of $\mathbb{P}$-identification, Theorem 2 no longer applies and it is easy to see that in general restrictions on the $\mathbb{Q}$-parameters (i.e. the $\mathbb{Q}$-conditional distribution of the latent factors as a function of the latent factors) will affect the forecasts of $\mathcal{P}_t$.

4.4 Relevance of Constraints on the Structure of Excess Returns

Central to the preceding irrelevance results is the absence of restrictions across the parameters of the $\mathbb{P}$ and $\mathbb{Q}$ distributions of $\mathcal{P}_t$. Such constraints would arise in practice if, for instance, the *GDTSM*-implied expected excess returns on bonds of different maturities lie in a space of dimension $\mathcal{L}$ less than $\dim(\mathcal{P}_t) = N$. Put another way, some risks in the economy may have either zero or constant risk premia. When $\mathcal{L} < N$, it also follows that time-variation in risk premia depends only on an $\mathcal{L}$-dimensional state variable. Cochrane and Piazzesi (2005, 2008) conclude that $\mathcal{L} = 1$ when conditioning risk premiums only on yield curve information. Joslin, Priebsch, and Singleton (2010) find that $\mathcal{L}$ is at least two when expected excess returns are conditioned on $\mathcal{P}_t$, inflation, and output growth. We explore the relevance for forecasting bond yields of imposing the constraint $\mathcal{L}$ within *GDTSM*s that condition risk premiums on the pricing factors $\mathcal{P}$. When this constraint is (approximately) valid, improved forecasts of y_t may arise from the associated reduction in the dimensionality of the parameter space.

To interpret this constraint, note from Cox and Huang (1989) and Joslin, Priebsch, and

Singleton (2010) that one-period, expected excess returns on portfolios of bonds with payoffs that track the pricing factors $\mathcal{P}_t$, say $xr\mathcal{P}_t$, are given by the components of

$$xr\mathcal{P}_t = K^{\mathbb{P}}_{0\mathcal{P}} - K^{\mathbb{Q}}_{0\mathcal{P}} + (K^{\mathbb{P}}_{1\mathcal{P}} - K^{\mathbb{Q}}_{1\mathcal{P}})\mathcal{P}_t. \tag{28}$$

That is, the i^{th} component of $(K^{\mathbb{P}}_{1\mathcal{P}} - K^{\mathbb{Q}}_{1\mathcal{P}})\mathcal{P}_t$ is the source of the risk premium for pure exposures to the i^{th} component of $\mathcal{P}_t$. Therefore, the constraint that the one-period expected excess returns on bond portfolios are driven by $\mathcal{L}$ linear combinations of the pricing factors $\mathcal{P}$ amounts to the constraint that the rank of $A_{RRP} = K^{\mathbb{P}}_{1\mathcal{P}} - K^{\mathbb{Q}}_{1\mathcal{P}}$ is $\mathcal{L}$.[20]

The reduced rank risk premium *GDTSM*s can be estimated through a concentration of the likelihood in the same spirit as (18). Given $(\lambda^{\mathbb{Q}}, k^{\mathbb{Q}}_{\infty}, \Sigma_{\mathcal{P}}, P^{\theta_m})$, the ML estimates of $(K^{\mathbb{P}}_{0\mathcal{P}}, K^{\mathbb{P}}_{1\mathcal{P}})$ can be computed as follows. First, compute (α, β) from the regression

$$\mathcal{P}_{t+1} - (K^{\mathbb{Q}}_{0\mathcal{P}} + K^{\mathbb{Q}}_{1\mathcal{P}}\mathcal{P}_t) = \alpha + \beta\mathcal{P}_t + \epsilon^{\mathcal{P}}_t, \tag{29}$$

where we fix the volatility matrix $\Sigma_{\mathcal{P}}$ of errors $\epsilon^{\mathcal{P}}_t$ and impose the constraint that β has rank $\mathcal{L}$. We show in Appendix F how one can compute the ML estimates of this constrained regression in closed form. For given $(\lambda^{\mathbb{Q}}, k^{\mathbb{Q}}_{\infty}, \Sigma_{\mathcal{P}}, P^{\theta_m})$, the ML estimates of the $\mathbb{P}$ parameters are then given by

$$K^{\mathbb{P}}_{0\mathcal{P}} = K^{\mathbb{Q}}_{0\mathcal{P}} + \hat{\alpha}, \quad K^{\mathbb{P}}_{1\mathcal{P}} = K^{\mathbb{Q}}_{1\mathcal{P}} + \hat{\beta}. \tag{30}$$

In comparison to the setting underlying Proposition 3 and Theorem 2, reduced rank risk premia enforce constraints across the parameters of the $\mathbb{P}$ and $\mathbb{Q}$ distributions. Consequently, the ML estimates of the $\mathbb{P}$ parameters are no longer given by their OLS counterparts. This, in turn, means that the implications of Proposition 3 discussed in Section 4.2 will, in general, no longer apply. Under the reduced-rank restrictions any further assumptions on the $\mathbb{Q}$-parameters (such as the constraints of the AFNS model) will directly affect the estimated $\mathbb{P}$ parameters as there is a link between the cross-section and time-series properties of yields. We explore the empirical implications of these observations in Section 5.

4.5 Relevance of Constraints on the $\mathbb{P}$-distribution of Yields

So far, we demonstrated that neither the imposition of no arbitrage nor restrictions on the $\mathbb{Q}$-dynamics have any effect on the ML estimates of $K^{\mathbb{P}}_{0\mathcal{P}}$ and $K^{\mathbb{P}}_{1\mathcal{P}}$. However, restrictions

[20] Alternatively, we could restrict the rank of $[K^{\mathbb{P}}_{0\mathcal{P}} - K^{\mathbb{Q}}_{0\mathcal{P}}, K^{\mathbb{P}}_{1\mathcal{P}} - K^{\mathbb{Q}}_{1\mathcal{P}}]$ to $\mathcal{L}$. This would enforce the stronger restriction that only $\mathcal{L}$ linear combination of the factors has non-zero expected excess return.

on risk premia, such as the reduced-rank assumption, link $\mathbb{P}$ and $\mathbb{Q}$ and interact with no arbitrage to affect estimates of $K^{\mathbb{P}}_{0\mathcal{P}}$ and $K^{\mathbb{P}}_{1\mathcal{P}}$. We now complete this discussion by examining whether no arbitrage affects the distribution of bond yields when one also imposes stand-alone restrictions on the $\mathbb{P}$-distribution of yields that do not impinge on the $\mathbb{Q}$ distribution, either directly or indirectly through risk premiums. Examples of such restrictions are that the yield portfolios are cointegrated or that the conditional mean of each portfolio yield does not depend on the other portfolio yields.[21] One can impose such restrictions without reference to a no-arbitrage model.

In these examples OLS no longer recovers the ML estimates of the parameters; rather, to obtain efficient estimates given $\Sigma_{\mathcal{P}}$, one must implement generalized least squares (GLS). Let $(K_0^{c*}(\Sigma_{\mathcal{P}}), K_1^{c*}(\Sigma_{\mathcal{P}}))$ denote the GLS estimates of $(K^{\mathbb{P}}_{0\mathcal{P}}, K^{\mathbb{P}}_{1\mathcal{P}})$ given $\Sigma_{\mathcal{P}}$:

$$(K_0^{c*}(\Sigma_{\mathcal{P}}), K_1^{c*}(\Sigma_{\mathcal{P}})) = \underset{K^{\mathbb{P}}_{0\mathcal{P}}, K^{\mathbb{P}}_{1\mathcal{P}}}{\arg\max} \sum_{t=1}^{T} f(\mathcal{P}_t^o | \mathcal{P}_{t-1}^o; K^{\mathbb{P}}_{1\mathcal{P}}, K^{\mathbb{P}}_{0\mathcal{P}}, \Sigma_{\mathcal{P}}), \tag{31}$$

where the arg max is taken over $(K^{\mathbb{P}}_{0\mathcal{P}}, K^{\mathbb{P}}_{1\mathcal{P}})$ satisfying the appropriate restriction on the $\mathbb{P}$-dynamics. In the presence of such restrictions, there is a non-degenerate dependence of (K_0^{c*}, K_1^{c*}) on $\Sigma_{\mathcal{P}}$. This dependence means that no arbitrage (which links $\Sigma_{\mathcal{P}}$ across $\mathbb{P}$ and $\mathbb{Q}$) affects the ML estimates of $(K^{\mathbb{P}}_{0\mathcal{P}}, K^{\mathbb{P}}_{1\mathcal{P}})$.

We explore the empirical implications of two types of restrictions on the $\mathbb{P}$-distribution of yields in Section 5: (1) a model with $K^{\mathbb{P}}_{1\mathcal{P}}$ constrained to be diagonal and (2) a model in which the $\mathcal{P}_t$ are cointegrated (with one unit root and no trend).

4.6 Comparing the JSZ Normalization to Other Canonical Models

The normalizations adopted by DS and Joslin (2007) preserve the latent factor structure in (9 – 10), in contrast to the rotation to observable pricing factors in the JSZ normalization. To our knowledge the only other normalization that has an "observable" state vector is the one explored by Collin-Dufresne, Goldstein, and Jones (2008) (CGJ). All three of these canonical models– DS, Joslin, and CGJ– are observationally equivalent.[22]

In the constant volatility subcase of the CGJ setup the state vector X_t is completely

[21] See Campbell and Shiller (1991) (among others) for empirical evidence on cointegration among bond yields. Diebold and Li (2006) adopt an assumption very similar to the second example.

[22] Different choices of normalizations, associated with different, unique matrix factorizations of the feedback matrix $K^{\mathbb{Q}}_{1X}$, give rise to observationally equivalent models, though models with different structure to their parameter sets. The JSZ normalization is based on the real Jordan factorization used in Proposition 1. CJG adopt the companion factorization. For any monic polynomial $p(x) = x^n - \mu_{n-1}x^{n-1} - \cdots - \mu_1 x - \mu_0$, the

defined by r_t and its first $N-1$ moments under $\mathbb{Q}$:

$$X_t = (r_t, \mu_{1t}, \mu_{2t}, \ldots, \mu_{N-1,t})', \tag{32}$$

where

$$\mu_{1t} = \frac{1}{dt} E^{\mathbb{Q}}(dr_t), \;\; \mu_{k+1,t} = \frac{1}{dt} E^{\mathbb{Q}}(d\mu_{kt}), \;\; k = 1, \ldots, N-2. \tag{33}$$

Under $\mathbb{Q}$, X_t follows

$$dX_t = (K^{\mathbb{Q}}_{0,CGJ} + K^{\mathbb{Q}}_{1,CGJ} X_t)dt + \Sigma_X dZ_t, \tag{34}$$

where Σ_X is lower triangular, $K^{\mathbb{Q}}_{0,CGJ} = (0, 0, \ldots, 0, \gamma)'$, and Z_t is the standard Brownian motion. By construction, the matrix $K^{\mathbb{Q}}_{1,CGJ}$ is the companion matrix factorization of the feedback matrix $K^{\mathbb{Q}}_{1X}$ in (9).

The sense in which X_t is observable in the CGJ normalization is quite different than in the JSZ normalization, and these differences may have practical relevance. First, it will not always be convenient to assume that the one-period short-rate r_t is observable. Duffee (1996) highlights various liquidity and "money-market" effects that might distort yields on short-term bond relative what is implied by a *GDTSM*. The *true* short rate– the one that implicitly underlies the pricing of long-term bonds– will not literally be observable absent an explicit model of these money-market effects. Second, actions by monetary authorities might necessitate the inclusion of additional risk factors or jumps in these factors when explicitly including short rates in the analysis of a *DTSM* (Piazzesi (2005)). Within the JSZ normalization one is free to define the portfolio matrix W so as to focus on segments of the yield curve away from the very short end, while preserving fully observable $\mathcal{P}$.

More subtly, the construction of the state vector in the CGJ normalization requires the parameters of the $\mathbb{Q}$ distribution. Therefore, any change in the implementation of a *GDTSM* that changes the implied $\mathbb{Q}$ parameters will necessarily change the observed pricing factors under the CGJ normalization. Fitting the same model to two overlapping sample periods could, for example, give rise to different values of the observed state variables during the overlapping period. In contrast, under the JSZ normalization we are led to identical values of

---

companion matrix is

$$C(p) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \mu_0 & \mu_1 & \mu_2 & \cdots & \mu_{n-1} \end{pmatrix}.$$

Given any matrix K, its monic characteristic polynomial is unique, and the matrix K is similar to its companion matrix $C(p(K))$.

$\mathcal{P}$ for all overlapping sample periods.

Full separation of the $\mathbb{P}$ and $\mathbb{Q}$ sides of the unrestricted model appears to be a unique feature of the JSZ normalization. It is this separation that clarifies the role of no-arbitrage restrictions in *GDTSM*s, and gives rise to the enormous computational advantages of our normalization relative to the DS, Joslin, and CGJ canonical models.

5 Empirical Results

We estimate the three-factor *GDTSM*s summarized in Table 1 by ML using the JSZ canonical form and the methods outlined in Section 3.[23] As all of our estimated models are stationary under $\mathbb{Q}$, we report our results in terms of $r_\infty^\mathbb{Q}$ instead of $k_\infty^\mathbb{Q}$. The data are end-of-month, Constant Maturity Treasury (CMT) yields from release Fed H.15 over the period from January 1990 to December 2007 (216 observations). The maturities considered are 6 months, and 1-, 2-, 3-, 5-, 7- and 10-years. From these coupon yields we bootstrap a zero-coupon curve assuming constant forward rates between maturities. Within Case **P**, we consider several subcases. With distinct real eigenvalues, we assume the first three principal components (*PC*s) are measured without error (RPC); or the 0.5-, 2-, and 10-year zero coupon yields are measured without error (RY). Additionally, we estimate models that price the first three *PC*s of the zero curve exactly under the constraints of repeated eigenvalues (JPC) and complex eigenvalues (CPC). Model JPC imposes the eigenvalue constraint of the AFNS model examined by Christensen, Diebold, and Rudebusch (2009). Finally, a subscript of "1" indicates the case of reduced-rank risk premiums ($\mathcal{L} = 1$) with the one-period expected excess returns being perfectly correlated across bonds. In all cases, except as noted, the component of measurement errors orthogonal to W are assumed to be normally distributed.[24] Although we derive portfolios from the principal components, one could also use portfolio loadings from

[23] $\bar{\lambda}_i^\mathbb{Q}$ denotes the complex conjugate of the i^{th} element of $\lambda^\mathbb{Q}$. Also, we defer discussion of case RKF, in which all yields are measured with error and Kalman filtering is applied, until Section 6.

[24] In Case **Y**, this assumption amounts to yield measurement errors being distributed i.i.d. $N(0, \sigma_p^2)$. When W comes from the principal components, the assumption is equivalent to the higher order *PC*s ($n > N$) being distributed $N(0, \sigma_p^2)$. In both of these cases, we can concentrate σ_p from the likelihood (conditional on $t = 1$ information) through $\hat{\sigma}_p^2 = \sum_{t=2,m}^{T} (y_{t,m}^o - y_{t,m})^2 / ((T-1) \times (J-N))$ where $y_{t,m}$ are the model yields which depend on all the other parameters. To be more precise about the error assumption: let $W_\perp \in \mathbb{R}^{(J-N)\times J}$ be a basis for the orthogonal complement of the row span of W. Then, since W has orthonormal rows, we can express y_t^o in terms of its projection onto W and the orthogonal complement to W as $y_t^o = W'Wy_t^o + (W_\perp)'W_\perp y_t^o = W'\mathcal{P}_t + (W_\perp)'W_\perp y_t^o$. We assume $y_t^o - y_t|\mathcal{P}_t$ has the degenerate distribution $N(W'\mathcal{P}_t, \sigma_p^2(W_\perp)'W_\perp)$ (which is rotation invariant in the sense that the likelihood is the same for alternative choices of base for the orthogonal complement to W.) Equivalently, the projection of y_t^o onto $W_\perp$ expressed in the coordinates $W_\perp$ is i.id. normal: $W_\perp y_t^o \sim N(0, \sigma_p^2 I_{J-N})$. This distribution satisfies $\mathbb{P}(Wy_t^o = \mathcal{P}_t|\mathcal{P}_t) = 1$.

various parametric splines for yields such as Nelson-Siegel loadings or polynomial loadings.

[Table 1 about here.]

An alternative measurement error structure arises when one supposes that coupon bonds are measured without error. In this case, portfolios of zero bond yields will necessarily incorporate measurement error. To that end, we consider

Case C: N coupon bonds are priced exactly and $J - N$ coupon bonds are measured with normally distributed errors in the *GDTSM*.

In implementing Case **C** with coupon-bond data, one can still select N portfolios of zero coupon yields and construct the rotation where these portfolios comprise the state vector. Even though such yields may not be observed, this rotation is still valuable because the portfolios of model-implied zero yields $\mathcal{P}_t$ can be approximated from the observed data. For example, one could bootstrap or spline an approximate zero coupon yield curve from the observed coupon bond prices and form an approximation of $\mathcal{P}_t$, call it $\mathcal{P}_t^a$. Importantly, the projection of $\mathcal{P}_t^a$ onto its own lag will recover reliable starting values for $K_{0\mathcal{P}}^{\mathbb{P}}$ and $K_{1\mathcal{P}}^{\mathbb{P}}$. However, because coupon bond yields are nonlinear functions of $\mathcal{P}$, the irrelevance propositions discussed in Section 3 do not apply to Case **C**. In our empirical implementation we consider the case of the 0.5-, 2-, and 10-year CMT yields measured without error, and the 1-, 3-, 5-, 7-year par coupon yields measured with errors (RCMT). Throughout, we report asymptotic standard errors for the maximum likelihood estimates that are computed using the outer product of the first derivative of the likelihood function to estimate the information matrix (see Berndt, Hall, Hall, and Hausman (1974))

In order to facilitate comparison of the estimates across models with different pricing factors, all of our results are presented in terms of the implied $\mathbb{P}$ distribution of the first three *PC*s of the zero yields.[25] Table 2 shows that these parameters are largely invariant to: (i) assumptions about the distribution of measurement errors, (ii) restrictions on the $\mathbb{Q}$-dynamics through restrictions on $\lambda^{\mathbb{Q}}$, and (iii) restrictions on the relation between the $\mathbb{Q}$- and $\mathbb{P}$-dynamics through the reduced rank assumption. The only mild exception is that model RCMT has a higher $r_\infty^{\mathbb{Q}}$, which is compensated for by slightly lower $\lambda_1^{\mathbb{Q}}$ and $\lambda_2^{\mathbb{Q}}$. The close alignment of results shows that the cross-section of bond yields provides a rich information set from which to extract the four relevant $\mathbb{Q}$-parameters, $r_\infty^{\mathbb{Q}}$ and $\lambda^{\mathbb{Q}}$.

[Table 2 about here]

[25] That is, under Case **Y** or when the CMT yields are priced perfectly by the *GDTSM*, after estimation, we impose the JSZ normalization based on the *PC*s of zero yields as the state variables.

Another notable feature of these estimates is that the results for model CPC are the same as those for model JPC. This is because, in the limit, as the complex part of the eigenvalues approach zero, the complex model approaches the Jordan model (see Appendix C). Thus we see that, for our data set, complex eigenvalues are not preferred over real eigenvalues.

Tables 3 and 4 present the parameters of the $\mathbb{P}$ distribution of $\mathcal{P}$. The final row presents parameters from a VAR (with no pricing involved) of the PCs. Table 4 reveals that initializing $\Sigma_{\mathcal{P}}$ using OLS residuals leads to very accurate starting values. By way of contrast, if we had instead used the Dai and Singleton (2000) (DS) canonical form, an accurate initialization of Σ_X would require a reliable initial value for $K_1^{\mathbb{Q}}$. The JSZ canonical form allows us to avoid this interplay between the values of Σ_X and $K_1^{\mathbb{Q}}$ by applying no-arbitrage constraints to determine $K_{1\mathcal{P}}^{\mathbb{Q}}$ independently of $\Sigma_{\mathcal{P}}$.

[Table 3 about here]

[Table 4 about here]

Across all specifications, the parameters are very comparable. Partly this is a consequence of Proposition 3: whether $\lambda^{\mathbb{Q}}$ is comprised of distinct real eigenvalues (RPC), complex eigenvalues (CPC), or repeated eigenvalues (JPC), the estimates of $K_{1\mathcal{P}}^{\mathbb{P}}$ and $K_{0\mathcal{P}}^{\mathbb{P}}$ are equal to each other and to the OLS estimates. However, stepping beyond this proposition, when we change whether it is PCs or individual yields (e.g., RPC versus RY) that are priced perfectly by the $GDTSM$ under Case $\mathbf{P}$, the parameters of the corresponding $\mathbb{P}$ distributions remain very similar. Imposing the reduced-rank risk premium constraint $\mathcal{L} = 1$ leads to generally similar, although in for some parameters there are measurable differences in estimates across corresponding models, particularly for some of the elements of $K_{1\mathcal{P}}^{\mathbb{P}}$.

Regarding the computational efficiency obtained using the the JSZ normalization, we stress that the only parameters that need to be estimated are $(r_\infty^{\mathbb{Q}}, \lambda^{\mathbb{Q}}, \Sigma_{\mathcal{P}})$ since, as discussed in Section 3, $(K_{0,\mathcal{P}}^{\mathbb{P}}, K_{1\mathcal{P}}^{\mathbb{P}})$ are determined by concentrating the likelihood and $(K_{0,\mathcal{P}}^{\mathbb{Q}}, K_{1,\mathcal{P}}^{\mathbb{Q}})$ are determined by no-arbitrage.[26] The models were estimated using sequential quadratic programming as implemented in Matlab's `fmincon`. Estimation under Case $\mathbf{P}$ using an informed guess of the $\mathbb{Q}$-eigenvalues took approximately 1.2 seconds.[27] Furthermore, 99%+ of the searches converged to the same likelihood value (to within the tolerance) with very

[26] The standard deviation of the pricing errors, σ_{pricing}, can be concentrated out as well, both when $\mathcal{L}$ equals 1 and 3.

[27] The computations were performed using a single-threaded application on a 2.4GHZ Intel Q6600 processor.

similar parameter estimates.[28] These computational advantages become even more important in the case where all yields are measured with error, which we consider in Section 6.

5.1 Statistical Inference Within the JSZ Canonical Form

There are two null hypotheses that are of particular interest given our observations in Section 3. The first test addresses the algebraic multiplicity of eigenvalues in the $GDTSM(3)$ model. As previously stated, the AFNS model of Christensen, Diebold, and Rudebusch (2007) is equivalent to the JSZ canonical form with three extra constraints, including a repeated eigenvalue of $K_1^{\mathbb{Q}}$. To assess the validity of the null hypothesis $\lambda_2^{\mathbb{Q}} = \lambda_3^{\mathbb{Q}}$, under the JSZ normalization, we perform a Likelihood Ratio (LR) test against the alternative that $\lambda^{\mathbb{Q}}$ is unconstrained. With this one linear constraint, the LR test statistic has an asymptotic χ^2 distribution with 1 degree of freedom, $\chi^2(1)$.

The second test of interest is the dimensionality of the one-period risk premium which, as discussed in Section 4.4, is captured by the rank of $A_{RRP} = K_{1\mathcal{P}}^{\mathbb{P}} - K_{1\mathcal{P}}^{\mathbb{Q}}$. To impose the constraint that $\mathcal{L} = 1$ we start with the singular value decomposition of A_{RRP}, UDV', where U and V are unitary matrices and D is diagonal with the diagonal sorted in decreasing order. The null hypothesis of interest– that A_{RRP} has rank 1– is therefore imposed by setting D_{22} and D_{33} to zero. To translate this representation into constraints on the parameter space, note that, for an N-factor $GDTSM$ with $\mathcal{L} = 1$,

$$DV'\mathcal{P}_t = D_{11}\sum_{j=1}^{N} V_{j1}\mathcal{P}_{jt}. \tag{35}$$

Therefore, the expected excess returns $xr\mathcal{P}_t$ (see Section 4.4) are given by

$$xr\mathcal{P}_t = \left(K_{0\mathcal{P}}^{\mathbb{P}} - K_{0\mathcal{P}}^{\mathbb{Q}}\right) + U_{\bullet 1}\cdot\left(D_{11}\sum_{j=1}^{N} V_{j1}\mathcal{P}_{jt}\right), \tag{36}$$

where $U_{\bullet 1}$ is the first column of U. The second term on the right-hand side of (36) expresses the time-varying components of $xr\mathcal{P}_t$ in terms of a common linear combination $V_{\bullet 1}'\mathcal{P}_t$ of the pricing factors. All of the parameters in (36) are econometrically identified by virtue

[28] An exception here is the Jordan form where typically there were two local extrema with either the smaller or the larger eigenvalue repeated. Another general consideration is that one must either optimizes over $k_\infty^{\mathbb{Q}}$ or alternatively imposes $\mathbb{Q}$-stationarity on the model if one desires to use $r_\infty^{\mathbb{Q}}$ in estimation. In fact, for estimation purposes, the issue of using $k_\infty^{\mathbb{Q}}$ versus $r_\infty^{\mathbb{Q}}$ is largely obviated by results in Joslin, Le, and Singleton (2010) who show how one can concentrate out $k_\infty^{\mathbb{Q}}$ under Case P.

of the facts that $V_{\bullet 1}' V_{\bullet 1} = 1$ (which identifies D_{11}) and $U_{\bullet 1}' U_{\bullet 1}$ (which identifies the weights on $D_{11} V_{\bullet 1}' \mathcal{P}_t$). Furthermore, given N, (36) implies $(N-1)^2$ cross-equation restrictions on the parameters of the conditional expectation $xr\mathcal{P}_t$. In our case, $N = 3$, so there are 4 cross-equation restrictions.

[Table 5 about here]

Tests for the equality of two eigenvalues are reported on top panel of Table 5, where a leading J means that model was estimated under the constraint that $\lambda_2^{\mathbb{Q}} = \lambda_3^{\mathbb{Q}}$ (consistent with the specifications of AFNS models). In the PC-based models this null hypothesis is not rejected, while for the yield-based models it is rejected at conventional significant levels. To interpret this difference across choices of risk factors, we note from Table 2 that the estimated $|\lambda_2^{\mathbb{Q}} - \lambda_3^{\mathbb{Q}}|$ is larger in model RY than in model RPC, with most of this difference being attributable to the larger value of $|\lambda_3^{\mathbb{Q}}|$ in model RY. The eigenvalue $\lambda_3^{\mathbb{Q}}$ governs the relatively high-frequency $\mathbb{Q}$-variation in yields and, thus, is particularly relevant for the behavior of the short end of the yield curve. Introducing the six-month yield directly as a pricing factor over weights the short end of the yield curve relative to having the PCs as pricing factors, as the latter are portfolios of yields along the entire maturity spectrum.

In the bottom panel we report tests of the reduced-rank, risk premium hypothesis that $\mathcal{L} = 1$. Under all model specifications this hypothesis cannot be rejected. This finding is consistent with the conclusions reached by Cochrane and Piazzesi (2005), though they effectively considered models with $N = 5$ as they examined $PC1$ through $PC5$.

5.2 Empirical Relevance of Constraints on $\mathbb{P}$-distribution of Yields

In Section 4.5, we demonstrated that imposing no arbitrage in addition to constraints on $\mathbb{P}$-distribution of yields affects the forecasts of yields. We now empirically explore the magnitude of the effect of the interaction of no arbitrage with (i) imposing $K_{1\mathcal{P}}^{\mathbb{P}}$ to be diagonal and (ii) imposing that $\mathcal{P}_t$ are cointegrated (with one unit root and no trend). In both cases, we assume risk premia have full rank and the $\mathbb{Q}$-distribution of yields is unconstrained.

[Table 6 about here]

[Table 7 about here]

Table 6 presents the estimation results when the constraint that $K_{1\mathcal{P}}^{\mathbb{P}}$ is diagonal in both the reference VAR as well as asymptotic standard errors. When the constraint of diagonal

$K^{\mathbb{P}}_{1\mathcal{P}}$ is imposed, no arbitrage has almost no effect on the parameters.[29] Additionally, the differences are small not only in magnitude, but also very small with respect to the standard errors.

Table 7 presents the estimation results for the VAR and no-arbitrage models when cointegration (without a trend) is imposed. Here we present standard errors computed by a parametric bootstrap due to the well-known non-standard asymptotics and small-sample bias associated with unit roots. The method which we used to bootstrap the standard errors is as follows: we randomly choose a data $t \in \{1, 2, \dots 216\}$ and initialize the state as the value of $\mathcal{P}$ on this date. Then, using the maximum likelihood estimate of the parameters, we simulate a path of the term structure for the sample size of 216 months and estimate the model based on this simulated data. These steps are repeated 1000 times. Although the no-arbitrage assumption has a somewhat larger effect than the diagonal case, the differences are again generally small. Taken together, these results suggest that although theoretically the no-arbitrage model may offer improved inference over the simple VAR model when stand-alone $\mathbb{P}$-constraints are imposed, such differences may, evidently, be small in practice.

5.3 Small-sample standard errors

Another feature of our normalization is that it facilitates the computation of smal-sample standard errors that can be compared to the asymptotic standard errors using the outer product of the first derivative of the likelihood function. We compare these results to bootstrapped standard errors computed with the procedure given in Section 5.2.

Table 8 presents the results for the model RPC. The asymptotic standard errors tend to overstate the precision with which we measure the effect of the level PC on the conditional means of the PCs $(K^{\mathbb{P}}_{1,11}, K^{\mathbb{P}}_{1,21}, K^{\mathbb{P}}_{1,31})$ by a factor of about two. These effects on standard errors for $K^{\mathbb{P}}_1$ and $\theta^{\mathbb{P}}$ are necessarily due to the small sample properties of OLS estimates in the VAR for $\mathcal{P}$ since, by Proposition 3, the full information ML estimates in the $GDTSM$ agree with the OLS estimates. Additionally, the precision with which we estimate the $\mathbb{Q}$ parameters is overstated by the asymptotic method by a factor of about 50%. Overall though, the asymptotic standard errors line up rather well with the bootsrapped standard errors.

[Table 8 about here]

[29] The average log-likelihood (across t) for the unconstrained no arbitrage model was 38.392, while for the diagonal-constrained model it was 38.291. The corresponding likelihood ratio test statistic is 44.0, far exceeding the 99% rejection region of 16.8 indicating a very strong rejection of this constraint.

5.4 Out-of-Sample Forecasting Results

An interesting question at this juncture is whether differences in parameter estimates translate into differences in the out-of-sample forecasting performance of these *GDTSM*s. We compute rolling re-estimation of each model using data from months $t = 1, \ldots, T$ $(T = 61, \ldots, 215)$ and use the model to predict, out of sample, the changes in the principal components over the next 1-, 3-, 6-, and 12-month periods. As a benchmark, we use the corresponding forecasts from an unconstrained VAR. As we noted in Section 3, *theoretically* the forecasts of $\mathcal{P}_t$ are the same across all models that assume these PCs are measured without error and that differ only in the constraints they impose on the $\mathbb{Q}$ distribution of $\mathcal{P}_t$. In particular, with $\mathcal{L} = 3$, whether we assume distinct real eigenvalues, complex eigenvalues or repeated eigenvalues (as in the AFNS model), the forecasts of $\mathcal{P}_t$ are all *exactly* the same as those from an unconstrained VAR. This explains the rows of zeros in Table 9.

[Table 9 about here]

Under the constraint $\mathcal{L} = 1$ (constrained risk premiums), there is an implicit constraint on $K^{\mathbb{P}}_{1\mathcal{P}}$ and, hence, enforcing the no-arbitrage constraints may improve forecasts. From Table 9 we see that there is a moderate improvement in forecasts for $PC1$ and $PC2$, particularly at longer horizons. Models RPC_1 and JPC_1 have different predictions (though only slightly). This is because the differences under $\mathbb{Q}$ implied by the repeated root assumption now propagate to the $\mathbb{P}$-dynamics through the restriction relating the $\mathbb{P}$- and $\mathbb{Q}$-drifts.

As further evidence on the empirical relevance of constraints on the $\mathbb{P}$ distribution of $\mathcal{P}$ for forecasting we pursue the examples of Section 5.2: constraining $K^{\mathbb{P}}_{1\mathcal{P}}$ to be diagonal (Table 6) or constraining $\mathcal{P}_t$ to have a common unit root (the cointegration example of Table 7).[30] The last four rows of Table 9 present the relative forecasting accuracy of VAR models with these constraints imposed, as well as their no-arbitrage counterparts with RPC being the unconstrained *GDTSM*. The constrained model $VAR + diag(K^{\mathbb{P}}_{1\mathcal{P}})$ shows notable improvements in out-of-sample forecast accuracy for the first and third PCs, particularly over longer horizons, but interestingly there is a deterioration in the forecast quality for $PC2$. This suggests that feedback from $(PC1, PC3)$ to $PC2$ is consequential for forecasting the slope of the yield curve. Imposing the cointegration constraint improves the forecasts of $PC1$ and, unlike in the prior example, also the forecasts of $PC2$.

Of most interest for our analysis is the finding that starting from either of the constrained VARs and then imposing the no-arbitrage restrictions has virtually no incremental effect

[30] For the cointegration example, we enforce the constraint that $[K^{\mathbb{P}}_{0\mathcal{P}}, K^{\mathbb{P}}_{1\mathcal{P}}]$ has a zero eigenvalue or, equivalently, there is a common unit root and no trend.

on forecast performance. Even though no-arbitrage restrictions can improve out-of-sample forecasts in these cases, in practice they have virtually no effect on the results in our data. The improvements in forecasting with either model RPC + $diag(K^{\mathbb{P}}_{1\mathcal{P}})$ or RPC + 1UR $[K^{\mathbb{P}}_{0\mathcal{P}}, K^{\mathbb{P}}_{1\mathcal{P}}]$ are entirely a consequence of imposing restrictions on the VAR model for $\mathcal{P}$.

It is instructive to place the findings of Christensen, Diebold, and Rudebusch (2007) for the AFNS model in the context of these results. They compare the forecast performance of an AFNS model with both $K^{\mathbb{P}}_{1X}$ and Σ_X in (1) constrained to be diagonal to Duffee (2002)'s canonical *GDTSM* based on the DS normalization (which is equivalent to our RPC model).[31] As with our examples, forcing $K^{\mathbb{P}}_{1X}$ to be diagonal is a direct constraint on the $\mathbb{P}$ distribution of $\mathcal{P}$ and, as such, may lead to more reliable forecasts than those from an unconstrained VAR model for $\mathcal{P}$. In fact, they report that their constrained AFNS model does outperform Duffee's model in forecasting bond yields, also with larger improvements over longer horizons. However, the results in Table 9 suggest that this improvement comes from the restrictions they imposed on the VAR model for $\mathcal{P}$ and not to the use of an AFNS pricing model.

6 Observable Factors with Measurement Errors

Up to this point we have assumed that N portfolios of yields are priced perfectly by the *GDTSM*. We turn next to the case where all of the zero-coupon yields used in estimation equal their *GDTSM*-implied values plus measurement errors. Under the assumption that the measurement errors are jointly normal, this is a Kalman filtering problem.

Case F: The yields on $J(> N)$ zero-coupon bonds equal their *GDTSM*-implied values plus mean zero, normally distributed errors, $y^o_t - y_t$.

A number of researchers (see, e.g., Duffee and Stanton (2007) and Duffee (2009)) have emphasized the computational challenges of estimation under Case **F**. Under the normalization of Dai and Singleton (2000) (DS), a researcher must estimate $(K^{\mathbb{Q}}_{1X}, K^{\mathbb{P}}_{0X}, K^{\mathbb{Q}}_{1X}, \rho_0, \rho_1)$, where $K^{\mathbb{Q}}_{1X}$ is lower triangular. In this parametrization, a researcher would likely have a diffuse prior on all of the parameters. Moreover, the states of the model depend on the parameters, so they too are unknown. We now show that our JSZ canonical representation extends to the setting of Case **F** and demonstrate its benefits both for interpretation and estimation of *GDTSM*s.

[31] Christensen, Diebold, and Rudebusch (2007) assume that all yields are measured with additive measurement errors, the case we turn to in Section 6. However, three-factor models price bonds quite accurately over the maturity range that they and we consider, so Theorem 2 should be informative about their findings.

Theorem 1 shows that any *GDTSM* is observationally equivalent to a model where the latent states are a given set of portfolios of yields, purged of measurement errors. In Case **P**, when the portfolios are assumed to be observed without measurement errors, this means the states are simply these portfolios of yields. In Case **F** we can maintain the interpretation that the latent states are portfolios of yields with known portfolio matrix W, though now constructed with the model-implied (measurement-error free) yields y_t. Equivalently, under Case **F**, one can view $\mathcal{P}_t = Wy_t$ as the "true" values of the pricing factors and view $\mathcal{P}_t^o = Wy_t^o$ as its observed counterpart.[32]

To set up the Kalman filtering problem for Case **F** we start with a given set of portfolio weights $W \in \mathbb{R}^{J\times N}$. From W and $(\lambda^{\mathbb{Q}}, r_\infty^{\mathbb{Q}}, \Sigma_{\mathcal{P}})$, we construct $(K_0^{\mathbb{Q}}, K_1^{\mathbb{Q}}, \rho_0, \rho_1)$ as prescribed in Proposition 2. From the no arbitrage relation (A2-A3) we then construct $A \in \mathbb{R}^J$ and $B \in \mathbb{R}^{J\times N}$ with $y_t = A + B\mathcal{P}_t$ and thus the relations

$$\Delta\mathcal{P}_t = K_{0\mathcal{P}}^{\mathbb{P}} + K_{1\mathcal{P}}^{\mathbb{P}}\mathcal{P}_t + \Sigma_{\mathcal{P}}\epsilon_t^{\mathbb{P}}\ , \tag{37}$$

$$y_t^o = A + B\mathcal{P}_t + \Sigma_Y\epsilon_t^{\mathrm{m}}, \tag{38}$$

where $\epsilon_t^{\mathbb{P}} \sim N(0, I_N)$ and $\epsilon_t^{\mathrm{m}} \sim N(0, I_M)$ are the measurement errors. Researchers have considered several parameterizations of the volatility matrix Σ_Y for ϵ_t^{m}. In our subsequent empirical examples we examine the cases of independent (diagonal Σ_Y) errors with distinct or common volatilities. These relations give the usual observation and state equations of the Kalman filter and they fully characterize the conditional distribution of the yield curve in terms of rotation-invariant parameters.

The computational benefits from using the JSZ normalization in this Case **F** arise, in part, from the observation that the least-squares projection of $\mathcal{P}_t^o$ onto $\mathcal{P}_{t-1}^o$ will nearly recover the *ML* estimates of $K_{0\mathcal{P}}^{\mathbb{P}}$ and $K_{1\mathcal{P}}^{\mathbb{P}}$ to the extent that $\mathcal{P}_t^o \approx \mathcal{P}_t$ (and we can choose portfolios, such as the principal components, to make these errors small).[33] Additionally, although not exact, we have nearly concentrated the likelihood in that the optimal $\mathbb{P}$ parameters will typically have weak dependence on the $\mathbb{Q}$ parameters owing to the fact that, as the $\mathbb{Q}$ parameters vary, the filtered states largely do not change.[34]

With the JSZ normalization, the parameter estimates are directly comparable across distributional assumptions on the measurement errors. That is, in analogy to Section 3,

[32] In fact, an equivalent characterization of the JSZ normalization is that, for a given portfolio matrix W, $A_W(\Theta^{\mathbb{Q}}) = 0$ and $B_W(\Theta^{\mathbb{Q}}) = I_N$.

[33] This approximation can be verified empirically by comparing $\mathcal{P}_t^o$ to $E_t^{\mathbb{P}}[\mathcal{P}_t]$ or $E_T^{\mathbb{P}}[\mathcal{P}_t]$.

[34] This is in contrast to, for example, the rotation of DS where, as the lower triangular $K_1^{\mathbb{Q}}$ is changed, the latent states vary as well. Thus, necessarily, so do the optimal $\mathbb{P}$ parameters given the specified $\mathbb{Q}$ parameters.

by fixing the yield portfolios, both measured with and without error, the $\mathbb{P}$ parameters are now directly comparable *regardless of the* $\mathbb{Q}$ *structure*. The parameters are also directly comparable across sample periods. When the $\mathbb{P}$-parameters are defined indirectly through a $\mathbb{Q}$-normalization, such comparisons will in general not be possible.

6.1 Empirical Implication

To illustrate Case **F** we estimate model RKF in which all J zero-coupon bonds used in estimation are measured with errors, and the eigenvalues of $K_1^{\mathbb{Q}}$ are all real. From Table 2 it is seen that the estimates of the $\mathbb{Q}$ parameters for model RKF are similar to those for models RPC and RY that are fit with N portfolios of yields priced exactly by the $GDTSM(3)$. Similarly, from Table 3 and Table 4 we see that the $\mathbb{P}$ parameters also generally match up across the models with and without filtering. An exception is the $\mathbb{P}$ distribution of $PC3$: when filtering, the volatility of $PC3$ is reduced by about 10%, and $PC3$ has a larger effect on the conditional mean of PC1 and PC2 (higher $K^{\mathbb{P}}_{1,13}, K^{\mathbb{P}}_{1,23}$). That is, $PC3$ both becomes a bit smoother and the model attributes a slightly greater affect of $PC3$ on forecasts of changes in the level and slope of the yield curve. For out-of-sample forecasts using model RKF, Table 9 shows that $PC1$ is better predicted by a simple VAR, while $PC2$ is predicted better than a VAR (though the differences are modest).

Also of interest in the presence of filtering are comparisons of the model-implied PCs with their corresponding sample estimates that, by assumption, are contaminated by measurement errors. Figure 1 plots the time series of the PCs computed from data against those from models RCMT, RY and RKF. For model RKF we plot the model-implied filtered $PCi^f_t = E_t[PCi_t]$. For all three models, the PCi^o are nearly identical to their model-implied counterparts. This is not surprising: if the model is accurately pricing the cross section of bonds, then it is almost a necessity that it will accurately match level, slope, and curvature. $PC3^f$ deviates slightly from $PC3^o$, and this is the source of the small differences seen in Figure 1.

[Figure 1 about here]

A quite different picture emerges when we increase the number of pricing factors to four or five using the JSZ normalization under Case **F**. For $i = 1, 2, 3$, PCi^f lines up well with PCi^o, as before. However from Figure 2 it is seen that $(PC4^f, PC5^f)$ appears to be a smoothed version of $(PC4^o, PC5^o)$, with the differences being substantial during some periods. To interpret these patterns we note that the likelihood function, through the Kalman filter, attempts to match both the cross-sectional pricing relationships and the time series variation

in excess returns. The higher order $PC4$ and $PC5$ only have small impacts on pricing since a three factor model already prices the cross-section of bonds well, but they do contain information about time-variation in expected returns.[35]

[Figure 2 about here]

Further insight into how ML addresses this dual objective is revealed by the estimated half-lives of the pricing factors under $\mathbb{Q}$ (computed from the estimated $\lambda^{\mathbb{Q}}$). In the five-factor $GDTSM$ the $\mathbb{Q}$ half-lives of $\mathcal{P}_t$ are (in years) $(15, 8.4, 2.4, 0.13, 0.08)$, whereas they are $(24, 1.2, 0.78)$ in the three-factor model. The presence of a factor with a very low half-life induces large movements in the short rate (the one-month rate in our discrete time formulation). Moreover, the sample average short rate is 23% which also results in large, wildly oscillating Sharpe ratios.

It is not the need to filter *per se* that gives rise to these fitting problems with a 5-factor model. When the first five PCs are priced perfectly by the $GDTSM$ (Model RPC), the properties of the short rate are now more plausible (see Table 10). However, the model-implied yields on bonds with maturities beyond those included in estimation are now wildly implausible. Furthermore, imposing the reduced rank restriction (Model RPC_1) does not materially improve the fit with five factors. For all of these error specifications with five factors, the Sharpe ratios for the higher-order PCs show substantial variation.[36] In contrast, the 3-factor specifications produce plausible values for these moments. We interpret this evidence as being symptomatic of over-fitting, of having too many pricing factors.

[Table 10 about here]

Does the accommodation of filtering substantially increase the computational complexity of estimation using the JSZ normalization? The parameters $(K^{\mathbb{P}}_{0,\mathcal{P}}, K^{\mathbb{P}}_{1\mathcal{P}})$ and σ_{pricing} are now included as part of the parameter search. As we argued for $\Sigma_{\mathcal{P}}$ in the Case **RP** case, we obtain very accurate starting points for $(K^{\mathbb{P}}_{0,\mathcal{P}}, K^{\mathbb{P}}_{1\mathcal{P}})$ *irrespective of any inaccuracies in* $(r^{\mathbb{Q}}_{\infty}, \lambda^{\mathbb{Q}})$. The additional cost of computing the Kalman filter as well as the lack of concentration of the likelihood function results in estimation times of approximately 10.4 seconds and, as without

[35] Cochrane and Piazzesi (2005, 2008) find that a portfolio of smoothed forward rates, that is correlated with $PC4$, predicts bond returns. Joslin, Priebsch, and Singleton (2010) find that smoothed growth in industrial production, which is also correlated with $PC4$, is an important determinant of excess returns for level and slope portfolios.

[36] See Duffee (2010) for a more extensive empirical evaluation of the properties of Sharpe ratios in $GDTSM$s. Joslin, Priebsch, and Singleton (2010) also investigate maximal Sharpe ratio variation within the context of macro-$GDTSM$s.

filtering, virtually all local optima are identical to within set tolerances. Using the results of the RP estimation as a starting point for the RF estimation decreased the estimation time to approximately 8.7 seconds. Thus, under the JSZ normalization, the estimation remains very fast even when all yields are measured with errors.

7 Conclusion

We derive a new canonical form for Gaussian dynamic term structure models. This canonical form allows for (essentially) arbitrary observable portfolios of zero-coupon yields to serve as the state variable. This allows us to characterize the properties of a *GDTSM* in terms of salient observables rather than latent states. Additionally, the risk-neutral distribution is parsimoniously characterized by the eigenvalues, $\lambda^{\mathbb{Q}}$, of the drift matrix and a constant that, under $\mathbb{Q}$-stationarity, is proportional to the long run mean of the short rate, $r_{\infty}^{\mathbb{Q}}$. Our canonical form reveals that simple OLS regression gives the maximum likelihood estimates of the parameters governing the physical distribution of bond yields. This results remains true even if additional restrictions of several types, such as restrictions on the risk-neutral condtional distribution of yields, are imposed. An immediate implication of this result is that constraints such as imposing the arbitrage-free Nelson Siegel model or imposing complex $\mathbb{Q}$-eigenvalues are irrelevant for forecasting bond yields. However, when one imposes structure on risk premia, such as the reduced-rank risk premium, a wedge from the unconstrained OLS estimates arises. Our canonical form allows us to easily overcome the challenge of empirical estimation of *GDTSM*s in the case of filtering. The empirical results suggest that either some caution should be exercised in interpreting higher-dimensional model or alternatively (perhaps preferably) care should be taken to avoid highly overparametrized models with implausible implications for either pricing or bond risk premia. Taken together, our results shed new light on estimation and interpretation of *GDTSM*s, and the effects of different specifications of the risk premiums and the risk-neutral distribution of bond yields on the observed dynamics of the yield curve.

Appendices

A Bond Pricing in *GDTSM*s

Under (1–3), the price of an m-year zero-coupon bond is given by

$$D_{t,m} = E_t^{\mathbb{Q}}[e^{-\sum_{i=0}^{m-1} r_{t+i}}] = e^{\mathcal{A}_m + \mathcal{B}_m \cdot X_t}, \tag{A1}$$

where $(\mathcal{A}_m, \mathcal{B}_m)$ solve the first-order difference equations

$$\mathcal{A}_{m+1} - \mathcal{A}_m = K_0^{\mathbb{Q}\prime}\mathcal{B}_m + \frac{1}{2}\mathcal{B}_m' H_0 \mathcal{B}_m - \rho_0 \tag{A2}$$

$$\mathcal{B}_{m+1} - \mathcal{B}_m = K_1^{\mathbb{Q}\prime}\mathcal{B}_m - \rho_1 \tag{A3}$$

subject to the initial conditions $\mathcal{A}_0 = 0, \mathcal{B}_0 = 0$. See, for example, Dai and Singleton (2003). The loadings for the corresponding bond yield are $A_m = -\mathcal{A}_m/m$ and $B_m = -\mathcal{B}_m/m$.

B Invariant Transformations of *GDTSM*s

As in DS, given the *GDTSM* with parameters as in (1–3) and latent state X_t, if we may apply the invariant transformation $\hat{X}_t = C + DX_t$ we then have an observationally equivalent *GDTSM* with latent state $\hat{X}_t$ and parameters given by

$$K_{0\hat{X}}^{\mathbb{Q}} = DK_{0X}^{\mathbb{Q}} - DK_{1X}^{\mathbb{Q}} D^{-1} C, \tag{A4}$$

$$K_{1\hat{X}}^{\mathbb{Q}} = DK_{1X}^{\mathbb{Q}} D^{-1}, \tag{A5}$$

$$\rho_{0\hat{X}} = \rho_{0X} - \rho_{1X}' D^{-1} C \tag{A6}$$

$$\rho_{1\hat{X}} = (D^{-1})'\rho_{1X}, \tag{A7}$$

$$K_{0\hat{X}}^{\mathbb{P}} = DK_{0X}^{\mathbb{P}} - DK_{1X}^{\mathbb{P}} D^{-1} C, \tag{A8}$$

$$K_{1\hat{X}}^{\mathbb{P}} = DK_{1X}^{\mathbb{P}} D^{-1}, \tag{A9}$$

$$H_{0\hat{X}} = DH_{0X}D' \tag{A10}$$

Given a parameter vector Θ, we denote the parameter vector of $\hat{X}_t$ as $C + D\Theta$.

C Proof of Proposition 1

We require a slight variation of the standard Jordan canonical form of a square matrix which maintains all real entries and bears a similar relation to the real Schur decomposition and the Schur decomposition.

Definition 1. *We refer to the* **real ordered Jordan form** *of a square matrix $A \in \mathbb{R}^{n\times n}$ with eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_m)$ with corresponding algebriac multiplicities $(m_1, m_2, \ldots, m_m)$ as*

$$A = J(\lambda) \equiv diag(J_1, J_2, \ldots, J_m),$$

where if λ_i is real, J_i is the $(m_i \times m_i)$ matrix

$$J_i = \begin{pmatrix} \lambda_i & 1 & \cdots & 0 \\ 0 & \lambda_i & \cdots & 0 \\ \vdots & \vdots & \ddots & 1 \\ 0 & \cdots & 0 & \lambda_i \end{pmatrix},$$

and if $|imag(\lambda_i)| > 0$, J_i is the $(2m_i \times 2m_i)$ matrix

$$J_i = \begin{pmatrix} R & I_2 & \cdots & 0 \\ 0 & R & \cdots & 0 \\ \vdots & \vdots & \ddots & I_2 \\ 0 & \cdots & 0 & R \end{pmatrix} \text{ with } R = \begin{pmatrix} real(\lambda_i) & -|imag(\lambda_i)| \\ |imagl(\lambda_i)| & real(\lambda_i) \end{pmatrix}$$

and otherwise the block is empty. Additionally, we apply an arbitrary ordering on $\mathbb{C}$ to order the blocks by their eigenvalues. In case there exist eigenvalues with geometric multiplicity greater than one, we also order the blocks by size.

Proof of Proposition 1: We first prove the existence by showing that a latent factor X_t with arbitrary $\mathbb{Q}$-dynamics

$$\Delta X_t = K_{0X}^{\mathbb{Q}} + K_{1X}^{\mathbb{Q}} X_{t-1} + \Sigma_X \epsilon_t^{\mathbb{Q}},$$

can be transformed to our desired from. By standard linear algebra, there exists matrix U so that $UK_{1X}^{\mathbb{Q}}U^{-1}$ is in the standard Jordan normal form. By Lemma 1 of the supplement to this paper (see Joslin, Singleton, and Zhu (2010)), we can further transform to have the real ordered form of Definition 1. Note that by Joslin (2007), each eigenvalue has geometric

multiplicity one and thus is associated with only one block due to the Markovian assumption. Now we separately consider the cases of real and imaginary Jordan blocks and show that we may transform the latent state to have $\rho_1 = \iota$.

1. A Jordan block J_i corresponds to real eigenvalues with algebraic multiplicity m_i (m_i could be 1). Then J_i is $m_i \times m_i$ matrix

$$J_i = \begin{pmatrix} \lambda_i & 1 & \cdots & 0 \\ 0 & \lambda_i & \cdots & 0 \\ \vdots & \vdots & \ddots & 1 \\ 0 & \cdots & 0 & \lambda_i \end{pmatrix}.$$

   Let $\rho_{1i} = (\rho_{1i}^{(1)}, \ldots, \rho_{1i}^{(k)})$ be the components of ρ_1 that correspond to the Jordan block J_i. We observe that $\rho_{1i}^{(1)} \neq 0$, for otherwise we can do without state variable $X_{ti}^{(1)}$, contradicting our assumption of an N-factor model. One can check that $B_i J_i B_i^{-1} = J_i$ if and only if B_i has the form

$$B_i = \begin{pmatrix} b_i^{(1)} & b_i^{(2)} & \cdots & b_i^{(m_i)} \\ 0 & b_i^{(1)} & \cdots & b_i^{(m_i-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & b_i^{(1)} \end{pmatrix}. \tag{A11}$$

   In particular, we can verify that the matrix

$$B_i = \begin{pmatrix} \rho_{1i}^{(1)} & \rho_{1i}^{(2)} - \rho_{1i}^{(1)} & \cdots & \rho_{1i}^{(m_i)} - \rho_{1i}^{(m_i-1)} \\ 0 & \rho_{1i}^{(1)} & \cdots & \rho_{1i}^{(m_i-1)} - \rho_{1i}^{(m_i-2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \rho_{1i}^{(1)} \end{pmatrix}$$

   satisfies $B_i J_i B_i^{-1} = J_i$ and $(B_i^{-1})' \rho_{1i} = \iota$.

2. A Jordan block J_i corresponds to complex eigenvalues with multiplicity m_i. Then J_i is

the $2m_i \times 2m_i$ matrix defined by

$$J_i = \begin{pmatrix} R & I_2 & \cdots & 0 \\ 0 & R & \cdots & 0 \\ \vdots & \vdots & \ddots & I_2 \\ 0 & \cdots & 0 & R \end{pmatrix} \text{ with } R = \begin{pmatrix} \text{real}(\lambda_i) & -|\text{imag}(\lambda_i)| \\ |\text{imagl}(\lambda_i)| & \text{real}(\lambda_i) \end{pmatrix}.$$

The proof is analogous to the real case, as the individual steps are the same but require lemmas to verify the intuitive steps hold with (2×2) block matrices replacing scalars. The details of the proof and subsequent steps for this case are available in Joslin, Singleton, and Zhu (2010).

We obtain the correct form of $K_{0X}^{\mathbb{Q}}$ as follows. We can demean the components of X corresponding to non-singular Jordan blocks by transforming $\hat{X}_t^b = X_t^b + \left(K_{1X}^{\mathbb{Q},b}\right)^{-1} K_{0X}^{\mathbb{Q},b}$. There can be at most one block corresponding to a zero eigenvector (which by our ordering would be the first), and the first $m_1 - 1$ entries of $K_{0X}^{\mathbb{Q}}$ can then be set to zero by translating to $\hat{X}_t^b = X_t^b - (K_{0X,2}^{\mathbb{Q},b}, K_{0X,3}^{\mathbb{Q},b}, , \ldots, K_{0X,m_1-1}^{\mathbb{Q},b}, , 0)'$. Finally, ρ_0 can then be set to zero by the translation $\hat{X}_{m_1,t} = X_{m_1,t} - \rho_0$.

The uniqueness of the canonical *GDTSM* stated in Proposition 1 follows from the uniqueness of an ordered Jordan decomposition and the fact that (i) the Jordan decomposition is maintained only by a block matrix where B has form (A11) and (ii) the only such B that satisifies $B'\iota = \iota$ is $B = I$. Furthermore, for $\theta \in \Theta_{JSZ}$ and any vector of parameters $a \neq 0$, either the translating by a violates the form of $K_{0X}^{\mathbb{Q}}$ (which happens if any state besides the last zero eigenvalue state (if one exists) is translated) or the translating violates $\rho_0 = 0$ (which happens if there is a zero eigenvalue and only the last such state is translated). This establishes the uniqueness and completes the proof of Proposition 1.

D Details of Step 3 in the Proof of Theorem 1

We have established that every *GDTSM* is observationally equivalent to a Jordan normalized model and the transformation relating the two models is found by computing the associated portfolio loadings:

$$\mathcal{G}_{\mathcal{P}}^{P} = \{A_W(\Theta^J) + B_W(\Theta^J)'\Theta^J : \Theta^J \in \mathcal{G}_J\}. \tag{A12}$$

Observe that since $\rho_1^J = \iota$, $B_W(\Theta^J)$ depends only on $\lambda^{\mathbb{Q}}$; let us denote $B_{\lambda^{\mathbb{Q}}} \equiv B_W(\Theta^J)'$. Similarly, let us denote $A_{\lambda^{\mathbb{Q}},\rho_0,\Sigma} \equiv A_W(\Theta^J)$. Since, for any $\lambda^{\mathbb{Q}}$, the map $s_{\lambda^{\mathbb{Q}}}(\Sigma) = B_{\lambda^{\mathbb{Q}}}^{-1}\Sigma$ is a bijection[37], we can reparametrize the conditional volatility by

$$\mathcal{G}_{\mathcal{P}}^P = \{A_{\Theta^J} + B_{\Theta^J}\Theta^J : \Theta^J = (k_\infty^{\mathbb{Q}} e_{m_1}, J(\lambda^{\mathbb{Q}}), 0, \iota, K_{0J}^{\mathbb{P}}, K_{1J}^{\mathbb{P}}, s_{\lambda^{\mathbb{Q}}}(\Sigma_{\mathcal{P}}))\}. \tag{A13}$$

Here we use $\Sigma_{\mathcal{P}}$ to denote the parameterization since, for $\Theta^J = (k_\infty^{\mathbb{Q}} e_{m_1}, J(\lambda^{\mathbb{Q}}), 0, \iota, K_{0J}^{\mathbb{P}}, K_{1J}^{\mathbb{P}}, B_{\lambda^{\mathbb{Q}}}^{-1}\Sigma_{\mathcal{P}})$, the transformed model $A_{\Theta^J} + B_{\Theta^J}\Theta^J$ (which has $\mathcal{P}_t$ as the factors since it is in $\mathcal{G}_{\mathcal{P}}$) has innovation volatility of $B_{\lambda^{\mathbb{Q}}}B_{\lambda^{\mathbb{Q}}}^{-1}\Sigma_{\mathcal{P}} = \Sigma_{\mathcal{P}}$.

Define the bijective map k on $\mathbb{R}^N \times \mathbb{R}^{N\times N}$ by:

$$k_{\lambda^{\mathbb{Q}},k_\infty^{\mathbb{Q}},\Sigma_{\mathcal{P}}}(K_0, K_1) = \left(B_{\lambda^{\mathbb{Q}}}K_0 - B_{\lambda^{\mathbb{Q}}}K_1 B_{\lambda^{\mathbb{Q}}}^{-1} A_{\lambda^{\mathbb{Q}},k_\infty^{\mathbb{Q}},\Sigma_{\mathcal{P}}},\ B_{\lambda^{\mathbb{Q}}}K_1 B_{\lambda^{\mathbb{Q}}}^{-1}\right). \tag{A14}$$

The function k maps (K_0, K_1) under the change of variables $X_t \mapsto A_{\lambda^{\mathbb{Q}},k_\infty^{\mathbb{Q}},\Sigma_{\mathcal{P}}} + B_{\lambda^{\mathbb{Q}}}X_t$. Using k, we further reparametrize $\mathcal{G}_{\mathcal{P}}^P$ by

$$\mathcal{G}_{\mathcal{P}}^P = \{A_{\Theta^J} + B_{\Theta^J}\Theta^J : \Theta^J = (k_\infty^{\mathbb{Q}} e_{m_1}, J(\lambda^{\mathbb{Q}}), 0, \iota, k^{-1}_{\lambda^{\mathbb{Q}},k_\infty^{\mathbb{Q}},\Sigma_{\mathcal{P}}}(K_{0\mathcal{P}}^{\mathbb{P}}, K_{1\mathcal{P}}^{\mathbb{P}}), s_{\lambda^{\mathbb{Q}}}(\Sigma_{\mathcal{P}}))\}. \tag{A15}$$

This gives our desired reparameterization of $\mathcal{G}_{\mathcal{P}}^P$ by $\Theta_{JSZ} = (\lambda^{\mathbb{Q}}, k_\infty^{\mathbb{Q}}, \Sigma_{\mathcal{P}}, K_{0\mathcal{P}}^{\mathbb{P}}, K_{1\mathcal{P}}^{\mathbb{P}})$. This is because, for $\Theta^J = \left(k_\infty^{\mathbb{Q}} e_{m_1}, J(\lambda^{\mathbb{Q}}), 0, \iota, k^{-1}_{\lambda^{\mathbb{Q}},k_\infty^{\mathbb{Q}},\Sigma_{\mathcal{P}}}(K_{0\mathcal{P}}^{\mathbb{P}}, K_{1\mathcal{P}}^{\mathbb{P}}), s_{\lambda^{\mathbb{Q}}}(\Sigma_{\mathcal{P}})\right)$,

$$\begin{aligned}\Theta^{\mathcal{P}} &= A_{\Theta^J} + B_{\Theta^J}\Theta^J \\ &= \left(k_{\lambda^{\mathbb{Q}},k_\infty^{\mathbb{Q}},\Sigma_{\mathcal{P}}}(0, J(\lambda^{\mathbb{Q}})),\ r_{\lambda^{\mathbb{Q}},k_\infty^{\mathbb{Q}},\Sigma_{\mathcal{P}}}(k_\infty^{\mathbb{Q}}, \iota),\ K_{0\mathcal{P}}^{\mathbb{P}}, K_{1\mathcal{P}}^{\mathbb{P}}, \Sigma_{\mathcal{P}}\right),\end{aligned} \tag{A16}$$

where $r_{\lambda^{\mathbb{Q}},k_\infty^{\mathbb{Q}},\Sigma_{\mathcal{P}}}$ maps (ρ_0, ρ_1) under the change of variables $X_t \mapsto A_{\lambda^{\mathbb{Q}},k_\infty^{\mathbb{Q}},\Sigma_{\mathcal{P}}} + B_{\lambda^{\mathbb{Q}}}X_t$:

$$r_{\lambda^{\mathbb{Q}},k_\infty^{\mathbb{Q}},\Sigma_{\mathcal{P}}}(\rho_0, \rho_1) = \left(\rho_0 - \rho_1' B_{\lambda^{\mathbb{Q}}}^{-1} A_{\lambda^{\mathbb{Q}},k_\infty^{\mathbb{Q}},\Sigma_{\mathcal{P}}},\ \left(B_{\lambda^{\mathbb{Q}}}^{-1}\right)'\rho_1\right). \tag{A17}$$

E Proof of Theorem 2

We first prove that (26–27) holds when $\mathcal{H}^0 = \{\eta_0 = (C^0, D^0, \Sigma_X^0, P^{\theta_m^0})\}$. Let

$$(K_{0X}^{\eta_0}, K_{1X}^{\eta_0}) = \underset{K_{0X},K_{1X}}{\arg\max}\, f(\mathcal{P}_T, y_T, \ldots, \mathcal{P}_1, y_1 | \mathcal{P}_0, y_0; \eta_0),$$

[37] For simplicity, we denote the Cholesky factorization, Σ, but we have in mind the covariance $\Sigma\Sigma'$.

which we subsequently show is uniquely maximized.

Let $(C^0_{\mathcal{P}}, D^0_{\mathcal{P}})$ denote the first N-element of C^0 and upper left $N \times N$ block of D^0, respectively. By our assumption of invertibility of $D^0_{\mathcal{P}}$, we have that $X_t = (D^0_{\mathcal{P}})^{-1}(\mathcal{P}_t - C^0_{\mathcal{P}})$. Thus, by our assumptions on the measurement errors,

$$f(\mathcal{P}_T, y_T, \ldots, \mathcal{P}_1, y_1 | \mathcal{P}_0, y_0; \eta_0, K_{0X}, K_{1X}) = f(\mathcal{P}_T, \ldots, \mathcal{P}_1 | \mathcal{P}_0; \eta_0, K_{0X}, K_{1X}) \times \prod_{t=1}^{T} f(e_{mt} | \mathcal{P}_t; \eta_0),$$

and so

$$(K^{\eta_0}_{0X}, K^{\eta_0}_{1X}) = \underset{K_{0X}, K_{1X}}{\arg\max}\, f(\mathcal{P}_T, \ldots, \mathcal{P}_1 | \mathcal{P}_0; \eta_0). \tag{A18}$$

Furthermore, substituting into (24) we have

$$\Delta \mathcal{P}_t = D_{0,\mathcal{P}} K_{1X} D^{-1}_{0,\mathcal{P}} \mathcal{P}_t + \left(D_{0,\mathcal{P}} K_{0X} - D^0_{\mathcal{P}} K_{1X} (D_{0,\mathcal{P}})^{-1} C_{0,\mathcal{P}} \right) + D\epsilon_t \,,\ \epsilon_t \sim \Sigma_X$$

It follows that the maximum value in (A18) is at most equal to the value of the likelihood corresponding to the OLS estimate. Note that although the value of the maximum likelihood depends on D, the argument that maximizes the value does not depend on D by the classic Zellner (1962) result. The OLS likelihood value is achieved by choosing (K_{0X}, K_{1X}) to satisfy (26–27) , which is feasible by the assumption that (K_{0X}, K_{1X}) is unconstrained and $D^0_{\mathcal{P}}$ is full rank.

This proves our result since $(K^{\mathcal{H}}_{0X}, K^{\mathcal{H}}_{1X}) = (K^{\eta_0}_{0X}, K^{\eta_0}_{1X})$ for some η_0 and we have shown that (26–27) hold for any η_0. Note that in the case that the parameters are under-identified, there will not be a unique maximum likelihood estimate in the sense that several η_0 may give the same likelihood, but (26–27) will hold for all possible choices. For some $\mathcal{H}$, there may not exist a maximizer, in which case the result holds vacuously. However, standard conditions and arguments, such as compactness, provide for the existence of a maximizer.

F ML Estimation of Reduced-Rank Regressions

Consider the regression as in (29) of the general form $Y_t = \alpha + \beta X_t + \epsilon_t$ subject to the constraint that β has rank r and where $\epsilon_t \sim N(0, \Sigma)$ i.i.d. with Σ known. That is, we wish to solve the program

$$(\alpha, \beta) = \underset{\text{rank}(\beta)=r}{\arg\min} \sum_t (Y_t - (\alpha + \beta X_t)' \Sigma^{-1} (Y_t - (\alpha + \beta X_t)).$$

It is easy to verify that by first de-meaning the variables we may assume without loss of generality that $\alpha \equiv 0$. Furthermore, by transforming the variables we may assume again without loss of generality that $\Sigma = I$ and $\sum_t X_t X_t' = I$. Under these assumptions, we wish to solve

$$\begin{aligned}
\beta &= \underset{\text{rank}(\beta)=r}{\arg\min} \operatorname{trace}\left((Y - X\beta')(Y - X\beta')'\right) \\
&= \underset{\text{rank}(\beta)=r}{\arg\min} \operatorname{trace}\left((Y - X\beta_{OLS}')(Y - X\beta_{OLS}')'\right) - 2\operatorname{trace}\left(X'(Y - X\beta_{OLS}')(\beta - \beta_{OLS})\right) \\
&\qquad + \operatorname{trace}(((X'X(\beta' - \beta_{OLS}'))(\beta - \beta_{OLS})) \\
&= \underset{\text{rank}(\beta)=r}{\arg\min} \|\beta - \beta_{OLS}\|_F
\end{aligned}$$

where Y and X are $(T \times N)$ and $(T \times M)$ matrices with the time series stacked vertically, $\beta_{OLS} = (X'X)^{-1}X'Y$, and F denotes the Frobenius norm: $\|A\|_F^2 = \sum_{i,j} |A_{i,j}|^2$. The above equalities repeatedly use the identity $\operatorname{trace}(AB) = \operatorname{trace}(BA)$. As in Keller (1962), this minimization problem has solution $\beta^* = UD_r^*V'$ where UDV' gives the singular value decomposition of β_{OLS} and D_r^* is the same as D except setting all of the singular values for $n > r$ to 0. This same proof applies again in the case where β is not square which would be the case where one assumes that only a single risk is priced (i.e. $[K_0^{\mathbb{P}}, K_1^{\mathbb{P}}] - [K_0^{\mathbb{Q}}, K_1^{\mathbb{Q}}]$ has reduced rank) rather than only a single risk has *time-varying* price of risk, as we do here.

References

Adrian, T., and E. Moench, 2008, "Pricing the Term Structure with Linear Regressions," working paper, Federal Reserve Bank of New York Staff Reports, 340.

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.

Ang, A., M. Piazzesi, and M. Wei, 2003, "What Does the Yield Curve Tell us About GDP Growth?," Working Paper, Columbia University.

Berndt, E., B. Hall, R. Hall, and J. Hausman, 1974, "Estimation Estimation and Inference in Nonlinear Structural Models," *Annals of Social Measurement*, 3(653–665).

Campbell, J., and R. Shiller, 1991, "Yield Spreads and Interest Rate Movements: A Bird's Eye View," *Review of Economic Studies*, 58, 495–514.

Chen, 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.

——— , 1995, "Interest Rate Options in Multifactor Cox-Ingersoll-Ross Models of the Term Structure," *Journal of Fixed Income*, Winter, 53–72.

Chernov, M., and P. Mueller, 2008, "The Term Structure of Inflation Expectations," *Working Paper, London Business School*.

Christensen, J. H., F. X. Diebold, and G. D. Rudebusch, 2007, "The Affine Arbitrage-Free Class of Nelson Siegel Term Structure Models," *Working Paper, Federal Reserve Bank of San Francisco*.

——— , 2009, "An Arbitrage-Free Generalized Nelson-Siegel Term Structure Model," *The Econometrics Journal*.

Cochrane, J., and M. Piazzesi, 2005, "Bond Risk Premia," *American Economic Review*, 95, 138–160.

——— , 2008, "Decomposing the Yield Curve," working paper, Stanford University.

Collin-Dufresne, P., R. Goldstein, and C. Jones, 2008, "Identification of Maximal Affine Term Structure Models," working paper, forthcoming, *Journal of Finance*.

Cox, J. C., and C. Huang, 1989, "Optimum Consumption and Portfolio Policies When Asset Prices Follow a Diffusion Process," *Journal of Economic Theory*, 49, 33–83.

Dai, Q., and K. Singleton, 2000, "Specification Analysis of Affine Term Structure Models," *Journal of Finance*, 55, 1943–1978.

——— , 2002, "Expectations Puzzles, Time-Varying Risk Premia, and Affine Models of the Term Structure," *Journal of Financial Economics*, 63, 415–441.

——— , 2003, "Term Structure Dynamics in Theory and Reality," *Review of Financial Studies*, 16, 631–678.

Diebold, F., and C. Li, 2006, "Forecasting the term structure of government bond yields," *Journal of Econometrics*, 130, 337–364.

Duffee, G., 1996, "Idiosyncratic Variation in Treasury Bill Yields," *Journal of Finance*, 51, 527–552.

——— , 2002, "Term Premia and Interest Rate Forecasts in Affine Models," *Journal of Finance*, 57, 405–443.

——— , 2008, "Information In (and Not In) The Term Structure," working paper, Working Paper, Johns Hopkins University.

——— , 2009, "Forecasting with the Term Structure: the Role of No-Arbitrage," *Working Paper, UC Berkeley*.

Duffee, G., 2010, "Sharpe ratios in term structure models," *Working Paper, John Hopkins University*.

Duffee, G., and R. Stanton, 2007, "Evidence on Simulation Inference for Near Unit-Root Processes with Implications for Term Structure Estimation," *Journal of Financial Econometrics*, 6(1), 108–142.

Duffie, D., and R. Kan, 1996, "A Yield-Factor Model of Interest Rates," *Mathematical Finance*, 6, 379–406.

Jardet, C., A. Monfort, and F. Pegoraro, 2009, "No-Arbitrage Near-Cointegrated VAR(p) Term Structure Models, Term Premiums and GDP Growth," working paper, Banque de France.

Joslin, S., 2007, "Pricing and Hedging Volatility in Fixed Income Markets," working paper, Working Paper, MIT.

Joslin, S., A. Le, and K. Singleton, 2010, "Maximum Likelihood Estimation of Gaussian Dynamic Term Structure Models," *Working Paper, Sloan School, MIT*.

Joslin, S., M. Priebsch, and K. Singleton, 2010, "Risk Premiums in Dynamic Term Structure Models with Unspanned Macro Risks," working paper, Stanford University.

Joslin, S., and K. Singleton, 2010, "What Gaussian Macro-$DTSM$s *Cannot* Tell us About the Macro Economy," working paper, Sloan School of Management.

Joslin, S., K. Singleton, and H. Zhu, 2010, "Supplement to 'A New Perspective on Gaussian $DTSM$s'," working paper, Sloan School, MIT.

Keller, J. B., 1962, "Factorization of Matrices by Least-Squares," *Biometrica*, 49, 239–242.

Nelson, C., and A. Siegel, 1987, "Parsimonious Modelling of Yield Curves," *Journal of Business*, 60, 473–489.

Pearson, N. D., and T. Sun, 1994, "Exploiting the Conditional Density in Estimating the Term Structure: An Application to the Cox, Ingersoll, and Ross Model," *Journal of Finance*, 49, 1279–1304.

Piazzesi, M., 2005, "Bond Yields and the Federal Reserve," *Journal of Political Economy*, 113, 311–344.

Zellner, A., 1962, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, 57, 348–368.

Tables

Table 1: Summary of Model Specifications

Model Name	Specification
RPC	Real $\lambda^{\mathbb{Q}\prime} = (\lambda_1^{\mathbb{Q}}, \lambda_2^{\mathbb{Q}}, \lambda_3^{\mathbb{Q}})$, $PC1$, $PC2$, $PC3$ priced exactly
RY	Real $\lambda^{\mathbb{Q}\prime} = (\lambda_1^{\mathbb{Q}}, \lambda_2^{\mathbb{Q}}, \lambda_3^{\mathbb{Q}})$, 0.5-, 2-, and 10-year zeros priced exactly
CPC	Complex $\lambda^{\mathbb{Q}\prime} = (\lambda_1^{\mathbb{Q}}, \lambda_2^{\mathbb{Q}}, \bar{\lambda}_2^{\mathbb{Q}})$, PC1, PC2, PC3 priced exactly
JPC	Real repeated $\lambda^{\mathbb{Q}\prime} = (\lambda_1^{\mathbb{Q}}, \lambda_2^{\mathbb{Q}}, \lambda_2^{\mathbb{Q}})$, PC1, PC2, PC3 priced exactly
RPC_1	RPC and rank 1 risk premia
RY_1	RY and rank 1 risk premia
RCMT_1	RCMT and rank 1 risk premia
JPC_1	JPC and rank 1 risk premia
RKF	Real distinct $\lambda^{\mathbb{Q}}$, and all yields are measured with error
RCMT	Real $\lambda^{\mathbb{Q}\prime} = (\lambda_1^{\mathbb{Q}}, \lambda_2^{\mathbb{Q}}, \lambda_3^{\mathbb{Q}})$, 0.5-, 2-, and 10-year CMTs priced exactly

Table 2: ML estimates of the risk-neutral parameters of the model-implied principal components.

Model	Parameter Estimate			
	$\lambda_1^{\mathbb{Q}}$	$\lambda_2^{\mathbb{Q}}$	$\lambda_3^{\mathbb{Q}}/\mathrm{im}(\lambda_2^{Q})$	$r_\infty^{\mathbb{Q}}$
RPC	-0.0024	-0.0481	-0.0713	8.61
	(0.000566)	(0.0083)	(0.0133)	(0.73)
RY	-0.00196	-0.0404	-0.0897	9.37
	(0.000378)	(0.00274)	(0.0073)	(0.789)
RKF	-0.00245	-0.0472	-0.0739	8.45
	(0.000567)	(0.00724)	(0.0125)	(0.678)
RCMT	-0.00178	-0.0372	-0.103	11.2
	(7e-005)	(0.000819)	(0.0029)	(0.346)
JPC	-0.00225	-0.0582	-0.0582	8.87
	(0.000409)	(0.00123)	(0.00123)	(0.536)
CPC	-0.00225	-0.0582	-0.0582	8.87
	(0.000409)	(0.00123)	(0.00123)	(0.536)
RPC_1	-0.00241	-0.0477	-0.0721	8.61
	(0.000559)	(0.00766)	(0.0126)	(0.715)
RY_1	-0.00197	-0.0403	-0.0902	9.37
	(0.000373)	(0.00269)	(0.00723)	(0.775)
RCMT_1	-0.00178	-0.0371	-0.103	11.2
	(6.92e-005)	(0.000828)	(0.003)	(0.345)
JPC_1	-0.00224	-0.0583	-0.0583	8.9
	(0.000405)	(0.00122)	(0.00122)	(0.54)

$r_\infty^{\mathbb{Q}}$ is normalized to percent per annum (by multiplying by 12×100). Asymptotic standard errors are given in parentheses.

Table 3: ML estimates of the physical parameters of the model-implied principal components.

Model	Parameter Estimate											
	$K^{\mathbb{P}}_{1,11}$	$K^{\mathbb{P}}_{1,12}$	$K^{\mathbb{P}}_{1,13}$	$K^{\mathbb{P}}_{1,21}$	$K^{\mathbb{P}}_{1,22}$	$K^{\mathbb{P}}_{1,23}$	$K^{\mathbb{P}}_{1,31}$	$K^{\mathbb{P}}_{1,32}$	$K^{\mathbb{P}}_{1,33}$	$\theta^{\mathbb{P}}_1$	$\theta^{\mathbb{P}}_2$	$\theta^{\mathbb{P}}_3$
RPC	-0.25	0.16	5.2	0.032	-0.32	4.2	-0.03	-0.028	-1.8	-0.11	0.025	0.0063
	(0.16)	(0.54)	(2.8)	(0.054)	(0.24)	(1.2)	(0.023)	(0.088)	(0.46)	(0.028)	(0.0075)	(0.00035)
RY	-0.25	0.11	5.5	0.037	-0.31	4.1	-0.03	-0.034	-1.8	-0.11	0.026	0.0061
	(0.15)	(0.55)	(2.7)	(0.054)	(0.22)	(1.2)	(0.023)	(0.091)	(0.47)	(0.027)	(0.0075)	(0.00035)
RKF	-0.12	0.33	6.7	0.0078	-0.35	4.7	-0.021	-0.007	-1.2	-0.14	0.026	0.0063
	(0.13)	(0.52)	(2.9)	(0.052)	(0.22)	(1.2)	(0.018)	(0.075)	(0.42)	(0.029)	(0.0055)	(0.00029)
RCMT	-0.25	0.11	5.7	0.037	-0.32	4.1	-0.031	-0.032	-1.8	-0.11	0.026	0.0062
	(0.15)	(0.55)	(2.6)	(0.056)	(0.23)	(1)	(0.02)	(0.071)	(0.43)	(0.044)	(0.0093)	(0.00052)
JPC	-0.25	0.16	5.2	0.032	-0.32	4.2	-0.03	-0.028	-1.8	-0.11	0.025	0.0063
	(0.15)	(0.54)	(2.7)	(0.054)	(0.24)	(1.2)	(0.023)	(0.087)	(0.46)	(0.027)	(0.0074)	(0.00035)
CPC	-0.25	0.16	5.2	0.032	-0.32	4.2	-0.03	-0.028	-1.8	-0.11	0.025	0.0063
	(0.15)	(0.55)	(2.7)	(0.052)	(0.24)	(1.2)	(0.023)	(0.092)	(0.46)	(0.099)	(0.088)	(0.014)
RPC_1	-0.24	-0.16	7.4	0.031	-0.14	3.3	-0.025	-0.061	-1.5	-0.11	0.025	0.0063
	(0.13)	(0.37)	(2)	(0.04)	(0.14)	(0.72)	(0.016)	(0.057)	(0.3)	(0.035)	(0.012)	(0.00039)
RY_1	-0.24	-0.14	7.3	0.038	-0.17	3.3	-0.026	-0.055	-1.6	-0.11	0.026	0.0061
	(0.13)	(0.38)	(1.8)	(0.04)	(0.14)	(0.64)	(0.018)	(0.062)	(0.29)	(0.03)	(0.011)	(0.00037)
$RCMT_1$	-0.25	-0.11	7.1	0.042	-0.18	3.3	-0.029	-0.045	-1.7	-0.11	0.025	0.0062
	(0.15)	(0.55)	(2.6)	(0.057)	(0.23)	(1.1)	(0.02)	(0.072)	(0.42)	(0.04)	(0.013)	(0.0005)
JPC_1	-0.23	-0.18	7.4	0.03	-0.14	3.3	-0.025	-0.064	-1.5	-0.11	0.025	0.0063
	(0.13)	(0.36)	(1.9)	(0.04)	(0.14)	(0.74)	(0.016)	(0.056)	(0.31)	(0.036)	(0.012)	(0.00039)

The long-run $\mathbb{P}$-mean of $\mathcal{P}$ is defined by $\theta^{\mathbb{P}} = -(K^{\mathbb{P}}_1)^{-1}K^{\mathbb{P}}_0$. $K^{\mathbb{P}}_1$ is annualized (by multiplying by 12). Asymptotic standard errors are given in parentheses.

Table 4: ML estimates of the conditional covariance of the model-implied principal components.

Model	Parameter Estimate					
	σ_1	σ_2	σ_3	ρ_{12}	ρ_{13}	ρ_{23}
RPC	2.2	0.884	0.373	-0.569	0.584	-0.422
	(0.126)	(0.0408)	(0.0164)	(0.0415)	(0.0485)	(0.0611)
RY	2.2	0.871	0.386	-0.566	0.57	-0.393
	(0.125)	(0.0426)	(0.0174)	(0.042)	(0.0502)	(0.0626)
RKF	2.21	0.837	0.313	-0.603	0.725	-0.631
	(0.127)	(0.0423)	(0.0205)	(0.044)	(0.0493)	(0.0668)
RCMT	2.23	0.73	0.316	-0.591	0.541	-0.362
	(0.423)	(0.0215)	(0.0278)	(0.0325)	(0.108)	(0.0504)
JPC	2.2	0.884	0.373	-0.569	0.584	-0.421
	(0.124)	(0.0408)	(0.0163)	(0.0413)	(0.0485)	(0.0605)
CPC	2.2	0.883	0.373	-0.569	0.581	-0.421
	(0.0407)	(0.0398)	(0.0152)	(0.0316)	(0.0401)	(0.0589)
RPC_1	2.21	0.888	0.374	-0.572	0.586	-0.424
	(0.123)	(0.0403)	(0.0155)	(0.0403)	(0.0479)	(0.0584)
RY_1	2.2	0.873	0.386	-0.568	0.571	-0.394
	(0.121)	(0.0419)	(0.0165)	(0.0411)	(0.0496)	(0.0608)
$RCMT_1$	2.23	0.731	0.316	-0.593	0.541	-0.362
	(0.424)	(0.0215)	(0.0278)	(0.0324)	(0.108)	(0.0507)
JPC_1	2.21	0.888	0.374	-0.572	0.586	-0.424
	(0.122)	(0.0402)	(0.0154)	(0.0402)	(0.0479)	(0.0579)

ρ_{ij} is the conditional correlations of the i^{th} and j^{th} components of the factors $\mathcal{P}_t$. Volatility estimates $\sigma_1, \sigma_2, \sigma_3$ are normalized to percent per annum (by multiplying by $100 \times \sqrt{12}$). Asymptotic standard errors are given in parentheses.

Table 5: Likelihood ratio tests.

$H_0 : \lambda_2^{\mathbb{Q}} = \lambda_3^{\mathbb{Q}}$

H_0	$\log L_0$	H_a	$\log L_a$	LR stats $\chi^2(1)$	p-value
JPC	38.3912	RPC	38.3921	0.375	0.540
JPC_1	38.3865	RPC_1	38.3876	0.463	0.496
JY	38.1679	RY	38.1863	7.906	0.005
JY_1	38.1638	RY_1	38.183	8.266	0.004
JRCMT	39.0123	RCMT	39.0414	12.513	0.000

$H_0 : \text{rank}\left(K_{1\mathcal{P}}^{\mathbb{P}} - K_{1\mathcal{P}}^{\mathbb{Q}}\right) = 1$

H_0	$\log L_0$	H_a	$\log L_a$	LR stats $\chi^2(4)$	p-value
RPC_1	38.3876	RPC	38.3921	1.9475	0.745
JPC_1	38.3865	JPC	38.3912	2.0358	0.729
RY	38.1863	RY_1	38.1830	1.4217	0.840
JY	38.1679	JY_1	38.1638	1.7819	0.776
RCMT_1	39.0387	RCMT	39.0414	1.161	0.884

The top panel reports tests equality of two eigenvalues, and bottom panel reports tests for rank-1 risk premium. The likelihood-ratio statistics are computed as $LR = -2(T-1)(\log L_0 - \log L_a)$, where $T = 216$ is sample size and $\log L_0$ and $\log L_a$ are the log-likelihoods under the null and alternative, respectively. All log-likelihoods are conditional on $t = 1$ and are time series averages across the $T - 1$ observations.

Table 6: The conditional mean parameters for the model with $K^{\mathbb{P}}_{1\mathcal{P}}$ constrained to be diagonal.

With No Arbitrage				Without No Arbitrage			
$K^{\mathbb{P}}_{0\mathcal{P}}$		$K^{\mathbb{P}}_{1\mathcal{P}}$		$K^{\mathbb{P}}_{0\mathcal{P}}$		$K^{\mathbb{P}}_{1\mathcal{P}}$	
-0.0129	-0.151			-0.0129	-0.151		
(0.0193)	(0.135)			(0.0188)	(0.131)		
0.00754		-0.286		0.00761		-0.289	
(0.00636)		(0.202)		(0.00635)		(0.201)	
0.013			-1.97	0.0129			-1.95
(0.00292)			(0.423)	(0.00292)			(0.421)

$K^{\mathbb{P}}_{1\mathcal{P}}$ is annualized by multiplying by 12. The left panel imposed no arbitrage and uses yield data for all maturities. The right panel does not use no arbitrage and simply computes the estimates of a VAR of $\mathcal{P}_t$ with $K^{\mathbb{P}}_{1\mathcal{P}}$ constrained to be diagonal through GLS.

Table 7: The conditional mean parameters for the model with cointegration with no trend and one unit root imposed.

With No Arbitrage				Without No Arbitrage			
$K^{\mathbb{P}}_{0\mathcal{P}}$		$K^{\mathbb{P}}_{1\mathcal{P}}$		$K^{\mathbb{P}}_{0\mathcal{P}}$		$K^{\mathbb{P}}_{1\mathcal{P}}$	
-0.0644	-0.258	0.113	5.22	-0.0668	-0.24	0.266	5.29
(0.0602)	(0.336)	(0.733)	(3.17)	(0.218)	(0.225)	(0.792)	(2.67)
-0.0189	0.0495	-0.112	4.32	-0.0172	0.0519	-0.168	4.32
(0.0236)	(0.124)	(0.288)	(1.28)	(0.0827)	(0.0824)	(0.31)	(1.03)
0.007	-0.0241	0.0482	-1.73	0.00713	-0.0184	0.0632	-1.71
(0.0105)	(0.0562)	(0.117)	(0.565)	(0.0326)	(0.0362)	(0.126)	(0.471)

The left panel imposed no arbitrage and uses yield data for all maturities. The right panel does not use no arbitrage and simply computes the estimates of a VAR of $\mathcal{P}_t$ with cointegration imposed so that $[K^{\mathbb{P}}_{0\mathcal{P}}, K^{\mathbb{P}}_{1\mathcal{P}}]$ has rank 2.

Table 8: The standard errors of the parameter estimates computed both by the asymptotic method and using a bootstrap method.

Parameter	Estimate	Asymptotic S.E.	Bootstrap S.E.
$K^{\mathbb{P}}_{1,11}$	-0.2543	(0.1551)	(0.2733)
$K^{\mathbb{P}}_{1,12}$	0.1595	(0.5428)	(0.8277)
$K^{\mathbb{P}}_{1,13}$	5.235	(2.761)	(3.1)
$K^{\mathbb{P}}_{1,21}$	0.03235	(0.05425)	(0.1057)
$K^{\mathbb{P}}_{1,22}$	-0.3153	(0.2359)	(0.3187)
$K^{\mathbb{P}}_{1,23}$	4.239	(1.212)	(1.233)
$K^{\mathbb{P}}_{1,31}$	-0.03047	(0.02263)	(0.04143)
$K^{\mathbb{P}}_{1,32}$	-0.02772	(0.08759)	(0.1314)
$K^{\mathbb{P}}_{1,33}$	-1.755	(0.4638)	(0.5337)
$\theta^{\mathbb{P}}_1$	-0.1109	(0.02762)	(0.02496)
$\theta^{\mathbb{P}}_2$	0.02539	(0.007469)	(0.00731)
$\theta^{\mathbb{P}}_3$	0.00631	(0.0003512)	(0.0003162)
$\lambda^{\mathbb{Q}}_1$	-0.002403	(0.0005662)	(0.0006167)
$\lambda^{\mathbb{Q}}_2$	-0.04813	(0.008296)	(0.007395)
$\lambda^{\mathbb{Q}}_3$	-0.07127	(0.0133)	(0.01162)
$r^{\mathbb{Q}}_\infty$	0.08606	(0.007302)	(0.01067)
σ_1	0.02205	(0.00126)	(0.001337)
σ_2	0.008838	(0.0004084)	(0.001508)
σ_3	0.003735	(0.0001643)	(0.0002803)
ρ_{21}	-0.5694	(0.04155)	(0.2268)
ρ_{31}	0.5842	(0.0485)	(0.1161)
ρ_{32}	-0.4218	(0.06114)	(0.156)

Here $\theta^{\mathbb{P}} = -(K^{\mathbb{P}}_1)^{-1}K^{\mathbb{P}}_0$ and ρ_{ij} is the conditional correlation between the i^{th} and j^{th} components of $\mathcal{P}_t$.

Table 9: The improvement in out-of-sample forecast accuracy relative to the forecasts from a $VAR(1)$.

Forecast Error Relative to	$PC1$				$PC2$				$PC3$			
Unconstrained $VAR(1)$ (%)	1m	3m	6m	12m	1m	3m	6m	12m	1m	3m	6m	12m
RPC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RY	-0.3	-0.5	-0.8	-0.7	0.2	0.4	0.4	0.0	0.1	0.8	1.3	0.8
RKF	0.9	3.0	5.9	12.9	-1.7	-4.7	-7.7	-10.0	1.2	3.3	7.7	10.6
JPC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0
CPC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0
RPC_1	-2.1	-4.3	-6.2	-7.1	-2.0	-3.8	-3.8	-1.6	-1.5	-2.7	-2.5	0.2
RY_1	-2.2	-4.8	-7.3	-8.8	-1.9	-3.9	-3.9	-1.8	-1.6	-2.7	-2.5	-1.0
JPC_1	-2.3	-4.7	-6.7	-8.2	-1.9	-3.7	-4.2	-2.7	-1.5	-2.6	-1.9	0.6
$VAR + diag(K^{\mathbb{P}}_{1\mathcal{P}})$	-5.3	-12.1	-18.6	-21.6	0.7	6.3	11.6	5.7	-2.4	-5.4	-9.1	-13.0
RPC + $diag(K^{\mathbb{P}}_{1\mathcal{P}})$	-5.3	-12.1	-18.6	-21.6	0.7	6.3	11.6	5.6	-2.4	-5.4	-9.1	-13.0
VAR + 1UR $[K^{\mathbb{P}}_{0\mathcal{P}}, K^{\mathbb{P}}_{1\mathcal{P}}]$	-5.3	-10.0	-12.9	-13.5	-2.3	-6.4	-8.9	-6.2	-1.0	-1.6	-1.7	-0.7
RPC + 1UR $[K^{\mathbb{P}}_{0\mathcal{P}}, K^{\mathbb{P}}_{1\mathcal{P}}]$	-5.3	-10.0	-13.0	-13.6	-2.3	-6.3	-8.8	-6.0	-1.0	-1.6	-1.8	-0.9

Forecast errors from a $VAR(1)$ is given by

$$\sqrt{\frac{1}{T-59}\sum_{t=60}^{T}\left(\Delta PCi_{t+1} - E_t[\Delta PCi_{t+1}]\right)^2},$$

where the expectation, E_t, is computed using the model estimated with data from time $\tau = 1, \ldots, t$. For example, a number of -5 implies that the model has 5% smaller out-of-sample RMSE than the unrestricted $VAR(1)$.

Table 10: This sample moments for three-factor and five-factor *GDTSM*s.

	3 Factor Models			5 Factor Models		
	RPC	RPC_1	RKF	RPC	RPC_1	RKF
mean 1-month rate	4.2%	4.2%	4.2%	4.3%	4.3%	23%
mean 30-year rate	5.8%	5.8%	5.9%	-31%	-39%	0.63%
$PC4$ Sharpe ratio mean	0.096	0.095	0.032	0.031	0.076	30
$PC4$ Sharpe ratio volatility	0.086	0.018	0.088	0.31	0.2	25
$PC5$ Sharpe ratio mean	0.096	0.095	0.032	0.031	0.076	30
$PC5$ Sharpe ratio volatility	0.086	0.018	0.088	0.31	0.2	25

Figures

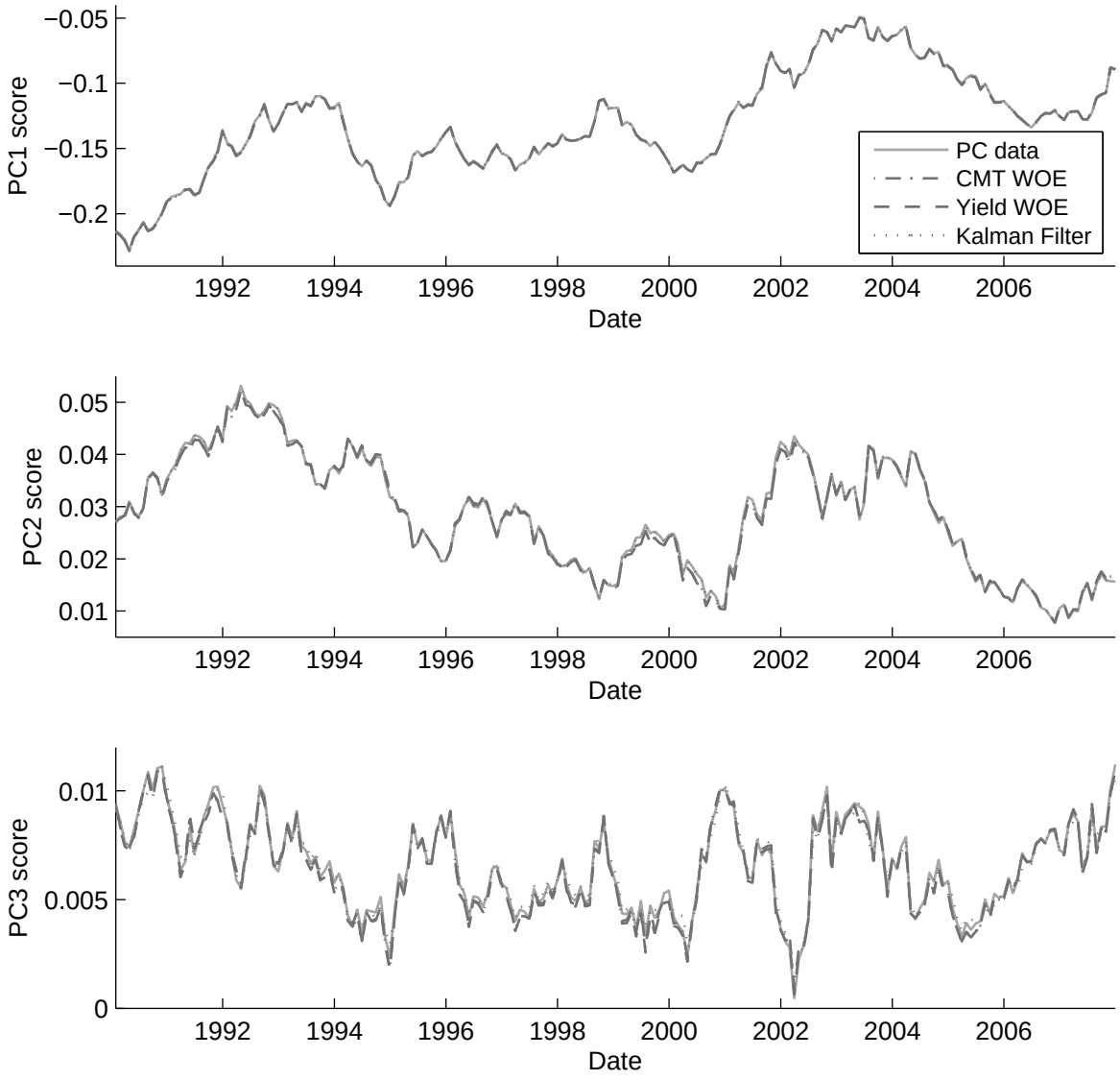

Figure 1: This figure plots the PCs implied by models RCMT, RY, and RKF against the estimated PCs from the data. All three models imply $PC1$ and $PC2$ that are almost indistinguishable from the data and from each other. The models imply slightly different $PC3$ but the difference is very small.

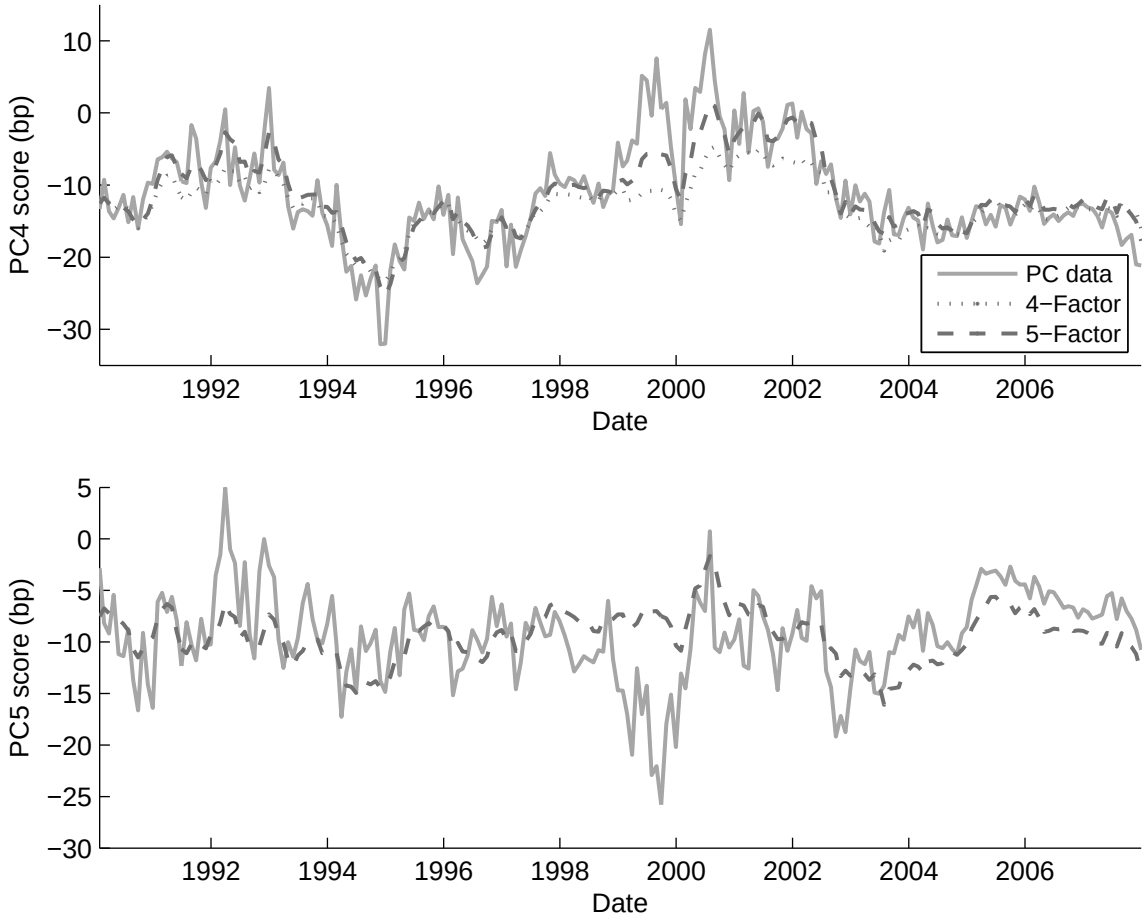

Figure 2: This figure plots the model implied and sample principal components for the fourth and fifth PCs when all PCs are assumed to be measured with normally distributed errors. High order PCs implied by the models are visibly different from the data.