Université de Montréal

Une méthode d'inférence bayésienne pour les modèles espace-état affines faiblement identifiés appliquée à une stratégie d'arbitrage statistique de la dynamique de la structure à terme des taux d'intérêt.

par
Sébastien Blais

Département de sciences économiques
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en sciences économiques

Avril, 2009

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée:

Une méthode d'inférence bayésienne pour les modèles espace-état affines faiblement identifiés appliquée à une stratégie d'arbitrage statistique de la dynamique de la structure à terme des taux d'intérêt.

présentée par:

Sébastien Blais

a été évaluée par un jury composé des personnes suivantes:

Marc Henry,	président-rapporteur
William J. McCausland,	directeur de recherche
Jean-Marie Dufour ,	membre du jury
Éric Jacquier,	examinateur externe
Onur Özgür,	représentant de l'examinateur externe
Martin Bilodeau ,	représentant du doyen de la FES

Thèse acceptée le 7 septembre 2009

RÉSUMÉ

Cette thèse porte sur l'inférence bayésienne pour les modèles affines de la structure à terme des taux d'intérêt. En particulier, elle met en évidence l'importance de la normalisation de l'espace des paramètres sur la qualité des prévisions générées par un espace-état linéraire gaussien. Puisque la vraisemblance de ces modèles est invariante par rapport à certaines transformations des paramètres, l'estimateur du maximum de vraisemblance des paramètres n'est pas unique. On normalise habituellement le modèle en considérant un sous-espace de l'espace des paramètres. Lorsque cette normalisation n'apporte pas l'identification globale des paramètres, elle est susceptible d'introduire des problèmes d'identification faible se traduisant par un estimateur ponctuel fortement biaisé. En comparaison, une densité prédictive bayésienne ne repose sur aucun estimateur ponctuel de paramètres, ce qui lui confère une certaine robustesse au problème d'identification faible. D'un point de vue méthodologique, je propose un nouvel échantillonneur de Monte Carlo par chaîne de Markov.

De plus, je démontre l'importance de la spécification des erreurs observationnelles sur l'inférence pour ces modèles. Je montre qu'une spécification courante où la matrice de covariance des erreurs n'est pas de plein rang peut produire des résidus fortement auto-corrélées. Au delà de ce cas particulier, je présente une analyse empirique de plusieurs autres restrictions imposées à la matrice de covariance des erreurs et je propose une nouvelle loi a priori pour cette matrice. Cette loi a priori permet de spécifier des restrictions souples sur un continuum entre des erreurs de matrice de covariance arbitraire et des erreurs indépendamment et identiquement distributées. J'évalue finalement l'utilité des modèles affines de la structure à terme dans un contexte de construction de stratégie d'arbitrage statistique. L'arbitrage statistique consiste à miser sur les déviations temporaires des valeurs de marchés par rapport à celles données par un modèle. Afin de neutraliser le risque par rapport aux facteurs communs, je construits des portefeuilles dont la valeur est approximativement non corrélée aux facteurs. Malgré un problème de spécification évident, la loi prédictive générée par le modèle permet de choisir des portefeuilles générant des gains économiquement significatifs.

Mots clés: modèles de la structure à terme, filtre de Kalman, prévisions bayésiennes, identification faible, sous-identification empirique, normalisation.

ABSTRACT

The suject of this is thesis is Bayesian inference for affine models of the term structure of interest rates. In particular, it highlights the critical role of normalization for the forecasting performance of Gaussian linear state-space models. Because the likelihood function of these models is invariant with respect to certain transformations of the parameter vector, the maximum likelihood parameter point estimator is not well defined. In general, one addresses transformation invariance by normalizing the parameter space. When this normalization does not provide global parameter identification, it can introduce weak identification problems, which can produce severely biased parameter point estimators. In contrast, Bayesian predictive densities do not rely on parameter point estimators. From a methodological point of view, I propose a novel MCMC sampler.

The thesis also demonstrates how observational error specification affects inference in these models. I show that one popular specification where the error covariance matrix does not have full rank can yield highly persistent residuals. Beyond that extreme particular case, I provide an empirical analysis of other strict restrictions on the covariance matrix and I propose a novel prior distribution for error covariance matrices. This prior allows the econometrician to specify soft restrictions on error cross-correlations and heteroscedasticity on a continuum between arbitrary and restricted covariance matrices.

Finally, I evaluate empirically the usefulness of affine term structure models for statistical arbitrage strategy construction. Statistical arbitrage exploits temporary deviations between market prices and fundamental values given by an economic model. In order to bet on temporary market price deviations from those implied by this model, I consider portfolios that are first-order hedged with respect to latent factors. In spite of obvious misspecification problems, I find that maximizing expected gains can be a profitable strategy for large institutional investors.

Keywords: dynamic term-structure models, Kalman filter, Bayesian forecasts, weak identification, empirical under-identification, normalization.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES

AR	Autoregressive
ARMA	Autoregressive-moving-average
ATSM	Affine term structure model
CIR	Cox-Ingersoll-Ross
CRSP	Center for Research in Security Prices
DTSM	Dynamic term structure model
EM	Expectation-maximization
EMM	Efficient method of moments
GMM	Generalized method of moments
IV	Instrumental variables
LSSM	Linear state-space model
MA	Moving-average
MAP	Maximum a posteriori
MCMC	Monte Carlo Markov chain
ML	Maximum likelihood
PC	Principal component
pdf	Probability density function
RMSE	Root mean square error
SDF	Stochastic discount factor
VAR	Vector autoregressive
VARMA	Vector autoregressive-moving-average

CHAPITRE 1

INTRODUCTION

Cette thèse porte sur l'inférence bayésienne pour les modèles affines de la structure à terme des taux d'intérêt (Dai et Singleton, 2002, offrent un survol de la littérature) et de leur utilisation dans un cadre décisionel d'investissement. En particulier, elle met en évidence l'importance de la normalisation de l'espace des paramètres et de la spécification des erreurs observationnelles sur la qualité des prévisions générées par un modèle affine.

Par leur parcimonie et leur solides assises théoriques, les modèles dynamiques de la structure à terme gagnent en popularité dans divers domaines. Outre la construction de portefeuille (Balı, Heidari, et Wu, 2006), on les utilise pour améliorer la prévision de variables macroéconomiques (Ang et Piazzesi, 2003), pour estimer des règles de politique monéraire (Ang, Dong, et Piazzesi, 2007), pour enrichir des modèles néo-keynésiens (Hördahl, Tristani, et Vestin, 2006; Bekaert, Cho, et Moreno, 2006; Dewachter et Lyrio, 2006), et pour estimer des paramètres structuraux, tels des paramètres de préférence (Garcia et Luger, 2007).

Le premier chapitre de cette thèse, *Forecasting with Weakly Identified Linear State-Space Models*, considère l'importance de la normalisation de l'espace des paramètres

sur la qualité des prévisions générées par un espace-état linéraire gaussien. Puisque la vraisemblance de ces modèles est invariante par rapport à certaines transformations des paramètres, l'estimateur du maximum de vraisemblance des paramètres n'est pas unique. On normalise habituellement le modèle en considérant un sous-espace de l'espace des paramètres. Lorsque cette normalisation n'apporte pas l'identification globale des paramètres, elle est susceptible d'introduire des problèmes d'identification faible se traduisant par un estimateur ponctuel fortement biaisé. Ces problèmes sont bien documentés dans de la littérature sur l'inférence pour les mélanges finis de distributions (Redner et Walker, 1984 ; Stephens, 2000 ; Frühwirth-Schnatter, 2001 ; Geweke, 2007 ; Hamilton, Waggoner et Zha, 2008), mais leur importance pour les modèles espace-état linéaires n'a pas été étudiée.

Une densité prédictive bayésienne ne repose sur aucun estimateur ponctuel, ce qui lui confère une certaine robustesse à ce problème d'identification faible. Par un exercice de simulation, je compare la performance des prévisions bayésiennes et celles obtenue par la méthode du maximum de vraisemblance, en termes d'erreur quadratique hors échantillon. Je montre que l'avantage des prévisions bayésiennes s'accentue lorsque l'identification des signes des facteurs devient plus faible.

D'un point de vue méthodologique, je propose un nouvel échantillonneur de Gibbs où les variables dont la loi a posteriori conditionelle n'est pas standard sont tirées

avec les variables latentes en un seul bloc. Je généralise aussi l'échantilloneur de permutations proposé par Frühwirth-Schnatter (2001) afin d'explorer efficacement les lobes symétriques de la loi a posteriori invariante des modèles espace-état linéaires.

Le second chapitre, *Bayesian Analysis of Affine Term Structure Models*, spécialise les résultats du premier aux modèles affines de la structure à terme. En particulier, j'utilise une normalisation novatrice du modèle apportant l'identification globale du modèle. De plus, je démontre l'importance de la spécification des erreurs observationnelles sur l'inférence pour ces modèles. Je montre qu'une spécification courante où la matrice de covariance des erreurs n'est pas de plein rang (Chen et Scott, 1993) peut produire des résidus plus fortement auto-corrélées qu'une spécification de plein rang (Chen et Scott, 1995). Puisque qu'un modèle affine décompose la structure à terme en un certain nombre de facteurs communs et idiosyncrasiques, la spécification de la matrice de covariance des erreurs affecte directement cette décomposition. Par exemple, la modélisation d'erreurs indépendamment et identiquement distribuées impose aux facteurs la lourde tâche de décrire à la fois la dynamique et la covariance contemporaine des taux d'intérêt, que cette dernière soit d'origine commune ou idiosyncrasique. En revanche, la modélisation d'erreurs de covariance arbitraire permet d'associer plus étroitement les facteurs aux composantes communes décrivant la dynamique de la structure à terme. Afin d'aller au-delà de ces cas particuliers, je propose nouvelle une loi a priori pour les matrices de covariance. Cette loi permet de spécifier des restrictions

souples sur un continuum entre des erreurs de matrice de covariance arbitraire et des erreurs indépendamment et identiquement distributées.

Il existe peu d'analyses empiriques bayésiennes des modéles affines de la structure à terme. Frühwirth-Schnatter et Geyer (1998), Lamoureux et Witte (2002), Müller et al. (2003), et Sanford et Martin (2005) considèrent des modèles de type CIR. Ang et al. (2007) utilisent échantionneur de Gibbs approximatif où les facteurs latents sont recentrés après chaque itération. Chib et Ergashev (2008) proposent un échantilloneur de Gibbs exact et numériquement efficace. Par contre, au meilleur de ma connaissance, il n'y existe aucune litérature qui tienne compte du problème d'identification faible. Je montre que ce problème empirique est présent dans un jeu de de données couramment utilisées en macro-économie (Ang et Bekaert, 2002 ; Dai et al., 2005 ; Ang et Piazzesi, 2003).

Le dernier chapitre, *Statistical Arbitrage with Affine Term Structure Models*, utilise les résultats des chapitres précédents pour évaluer l'utilité des modèles affines de la structure à terme dans un contexte de construction de stratégie d'arbitrage statistique. L'arbitrage statistique consiste à miser sur les déviations temporaires des valeurs de marchés par rapport à celles données par un modèle. Afin de neutraliser le risque par rapport aux facteurs communs, je construits des portefeuilles dont la valeur est approximativement non corrélée aux facteurs. J'obtiens ainsi un nombre fini de

stratégies, ce qui simplifie la solution numérique du problème d'optimisation. Malgré un problème de spécification évident, la loi prédictive générée par le modèle permet de choisir des portefeuilles générant des gains économiquement significatifs.

CHAPITRE 2

FORECASTING WITH WEAKLY IDENTIFIED LINEAR STATE-SPACE MODELS

Abstract

Normalizing models in empirical work is sometimes a more difficult task than commonly appreciated. Permutation invariance and local non-identification cause well-documented difficulties for maximum-likelihood and Bayesian inference in finite mixture distributions. Because these issues arise when some parameters are close to being unidentified, they are best described as weak identification (or empirical underidentification) problems. Although similar difficulties arise in linear state-space models, little is known about how they should be addressed. In this paper, I show that some popular normalizations do not provide global identification and yield parameter point estimators with undesirable finite-sample properties. At the computational level, I propose a novel posterior simulator for Gaussian linear state-space models, which I use to illustrate the relationship between forecasting performance and weak identification. In particular, Monte Carlo simulations show that taking into account parameter uncertainty reduces out-of-sample root mean square forecast errors when some parameters are weakly identified.
JEL classification: C11; C5; C52

2.1 Introduction

The likelihood function of many latent-variable models is invariant with respect to certain transformations of the parameters. For example, the likelihood function of certain finite mixture distributions is invariant with respect to permutation of the component distribution indices, leading to an inferential problem known as *label switching* in the literature (See Redner and Walker, 1984, for a survey). Consequently, some parameters in latent-variable (or unobserved-component) models are locally unidentified. For mixture distributions, component weights are unidentified in the parameter subspace where component distributions are identical.

Permutation invariance and local non-identification cause well-documented difficulties for likelihood-based inference in finite mixture distributions. Invariance with respect to a set of transformations is typically broken through normalization: one restricts attention to a particular parameter subspace. For a mixture of two normal distributions, one could consider the parameter subspace where the mean of the first distribution is larger than that of the second. It turns out that the choice of normalization has critical consequences for parameter point estimators in finite sample. Hamilton, Waggoner, and Zha (2007) [Summary] state that "poor normalizations can lead to multimodal distributions, disjoint confidence intervals, and very misleading characterizations of the true statistical uncertainty." Because these difficulties arise when some parameters are close to being unidentified, they can be described as *weak identification* problems in the econometrics literature, or *empirical underidentification* problems in the psychometrics literature.

Although weak permutation and reflection identification cause similar difficulties for inference in linear state-space models (LSSMs), it has received little attention. Jennrich (1978) shows that the likelihood function of linear factor models has symmetric lobes because it is invariant with respect to reflections across the axes of the coordinate system, which switch the state variables' sign. Frühwirth-Schnatter and Wagner (2008) stress some implications of reflection invariance for univariate linear state-space model selection. With respect to permutation invariance, Loken (2004) writes "The likelihood and posterior distributions for these models have some peculiar properties, and at the very least, researchers employing a Bayesian approach must recognize a multimodality problem in factor models analogous to the label-switching problem in mixture models." To the best of my knowledge, I provide the first empirical analysis of the finite-sample

implications of weak permutation and reflection identification for inference in LSSMs. Because these issues are well-known for mixture distributions and these models are simpler than LSSMs, I use the former as illustrative examples in this paper.

As the term suggests and is generally understood, a normalization is a restriction of the parameter space (*i.e.* a parameter subspace) that does not contain any information about the observables or the parameters. From that perspective, normalization is thus in sharp contrast with prior information specification. Therefore, although operationalizing normalization as a restriction of a prior distribution's support is common practice, I address normalization and prior specification separately. Doing so isolates the issues pertaining specifically to normalization, which affect both maximum likelihood (ML) and Bayesian inference. Being precise about a third modeling decision, namely parameterization, also proves useful in this paper. Reparameterization consists in defining a one-to-one mapping from one parameter space to another, and often takes the form of a change of coordinate system. Thus, like normalization, parameterization should not contain any information.

I propose a discussion of normalization which begins with the fundamental, if often side-stepped, question of whether (or when) normalization is strictly necessary. Because the likelihood function of LSSMs is invariant with respect to a certain set of parameter transformations, standard parameter point estimators are not defined uniquely. Thus, for instance, the maximum-likelihood problem defines a parameter set estimator rather than a point estimator. While normalizing the parameter space in order to obtain well-defined parameter point estimators is common practice, it should be emphasized that normalization is often not strictly necessary. In particular, parameter inference is possible as soon as the parameter set estimator is bounded (Manski, 2003),

which is the case if the likelihood function is invariant with respect to a finite set of parameter transformations.

Because point estimators are simpler from a computational as well as interpretational point of view, they are often preferred to set estimators in empirical applications. As there are many ways to normalize LSSMs, it is natural to ask how alternative normalizations should be compared. In general, normalizations do not merely ensure that parameter point estimators are well defined, they also have broader implications for inference. Building on the work of Hamilton, Waggoner, and Zha (2007), I argue that normalizations providing global identification are more likely to yield unimodal sampling distributions.

The difficulties associated with reflection local non-identification are closely related to the root-cancelation problem in autoregressive-moving-average (ARMA) models, which were discussed by Box and Jenkins (1976) and are the object ongoing research. Kleibergen and Hoek (2000) propose priors for a reparameterization of ARMA models in the context of order selection in order to penalize regions of the parameter space where roots are close to canceling out. Stoffer and Wall (1991) study the finite-sample properties of the ML estimator for a LSSM representation of ARMA processes when root-cancelation issues arise. They propose nonparametric Monte Carlo bootstrap standard errors and demonstrate their superiority over the usual asymptotic standard errors.

Point estimators are often less attractive quantities when their sampling distribution are multimodal. Parameter uncertainty thus plays an important role in such situations. In addition, a symmetric asymptotic approximations of a multimodal

sampling distribution can be unreliable. In contrast, Bayesian inference deals with parameter uncertainty in a consistent manner.

This paper is organized as follows.

In the first section, I describe normalization and weak identification in a general setting. This discussion introduces notation and addresses the questions of whether and when, loosely speaking, normalization is necessary, desirable and feasible. It then turns to comparing alternative normalizations and to practical implementation details.

The second section addresses the invariance of LSSMs with respect to transformations corresponding to linear transformations of the latent state variables. I show that a popular normalization described by Harvey (1989) does not provide global identification. Moreover, I show that it is observationally restrictive. I simplify the analysis of linear transformation invariance by considering elementary transformations: any linear transformation can be decomposed into scaling, rotation, permutation and reflection transformations. Ideal rotation and scale normalizations would preserve permutation and reflection invariance, allowing one to independently specify permutation and reflection normalizations. To the best of my knowledge, I provide the first observationally unrestrictive normalization of LSSMs that provides global identification.

In the third section, I propose permutation- and reflection-invariant prior distributions for the parameters of Gaussian LSSMs. Some of these priors rely on a reparameterization of the model that is easier to interpret. While normalization and parameterization do not change the informational content of the likelihood function, they might affect the interpretation of the parameters and, consequently, the specification

of prior information. I highlight situations in which one can inadvertently penalize reasonable regions of the parameter space.

In the fourth section, I describe a posterior simulator for Gaussian LSSMs and I explain how to implement reflection and permutation normalizations. I first present a Metropolis-within-Gibbs sampler for the parameters whose conditional posterior distributions are not standard. I draw these parameters and the latent state variables as a single block. Next, I extend the permutation sampler of Frühwirth-Schnatter (2001) in order to explore the symmetric lobes of reflection- and permutation-invariant posteriors. Although mixing over permutations and reflections is inferentially irrelevant, I argue that it helps monitoring the mixing properties of an MCMC simulator in other dimensions.

The fifth section considers the relationship between forecasting performance and weak identification. Because Bayesian predictive densities are reflection- and permutation-invariant, they are not affected by permutation and reflection normalization. Using simulations, I compare the performance of Bayesian and ML forecasts, on the basis of out-of-sample root mean square errors, and I find that the advantage of taking parameter uncertainty into account increases as reflection identification becomes weaker. I conclude with a research agenda for future research on these matters.

2.2 Weak identification

This section addresses normalization in latent state variable models from a general perspective. I consider likelihood-based inference methods, which rely on a parametric statistical model.

Definition 1 *A **parametric statistical model** is a triplet* $(\mathcal{Y}, \mathcal{F}, \Theta)$*, where* $\mathcal{Y}$ *is the sample*

space, $\mathcal{F} \equiv \{f(y|\,\theta) \,|\, y \in \mathcal{Y},\, \theta \in \Theta\}$ *is a set of parametric probability density functions on* $\mathcal{Y}$ *and* Θ *is the parameter set. The* ***likelihood function*** *of the model is the function* $l(\theta|\,y) = f(y|\,\theta)$.

The likelihood function of many latent-variable models is invariant with respect to sets of transformations.

Definition 2 *A function* $f : \Theta \to \mathbb{R}$ *is* ***invariant with respect a bijective transformation*** $T : \Theta \to \Theta$ *if* $f\,(T(\theta)) = f\,(\theta)$ *for all* $\theta \in \Theta$.

If $l(\theta|\,y)$ is invariant with respect to T on Θ for all $y \in \mathcal{Y}$ then we say that $T(\theta)$ and θ are **observationally equivalent**. We will also say that f is invariant with respect a set of bijective transformations $\mathcal{T}\,(\Theta)$ if it is invariant with respect to all of its elements. The notation $\mathcal{T}\,(\Theta)$ makes dependence on the set Θ explicit: $\mathcal{T}\,(\Theta)$ is a set of bijections on Θ. For example, for $\Theta' \subseteq \Theta$, $\mathcal{T}(\Theta') = \{T : \Theta' \to \Theta' | \, T \in \mathcal{T}(\Theta)\}$. I will omit this dependence and write $\mathcal{T}$ when this causes no confusion. The following examples illustrate this definition.

Example 1 (Normal mean) *Consider*

$$\mathbf{y} \;=\; b\mathbf{x} + \mathbf{e}, \quad \mathbf{x} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathcal{I}), \quad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathcal{I}),$$

for $(b, \sigma^2) \in \Psi = \mathbb{R} \times (0, \infty)$. *The likelihood function,*

$$l(b, \sigma^2 |\, \mathbf{y}) \;=\; \frac{1}{(2\pi b^2 \sigma^2)^{T/2}} \exp\left\{-\frac{1}{2b^2\sigma^2}\mathbf{y}'\mathbf{y}\right\},$$

satisfies $l(b, \sigma^2 |\, \mathbf{y}) = l\left(|Db|, \sigma^2/D^2 \big|\, \mathbf{y}\right)$ *for any* $D \neq 0$, *and it is therefore invariant with respect to*

$$\begin{aligned}
\mathcal{T}_D(\Theta) &= \left\{T_D : \Theta \to \Theta \middle| \, T_D(b,\sigma^2) = \left(Db, \sigma^2/D^2\right),\ D>0\right\} \\
\mathcal{T}_S(\Theta) &= \left\{T_S : \Theta \to \Theta \middle| \, T_S(b,\sigma^2) = \left(Sb, \sigma^2\right),\ |S|=1\right\} \\
\mathcal{T}_{SD}(\Theta) &= \left\{T_{SD} : \Theta \to \Theta \middle| \, T_{SD}(b,\sigma^2) = \left(SDb, \sigma^2/D^2\right),\ D>0, |S|=1\right\} \\
&= \left\{T_{SD} : \Theta \to \Theta \middle| \, T_{SD}(b,\sigma^2) = T_S\left(T_D(b,\sigma^2)\right)\right\}
\end{aligned}$$

The parameters b *and* σ^2 *enter the likelihood function as the product* $b^2\sigma^2$*. Transformations in* $\mathcal{T}_D$ *correspond to changing the scale of the unobserved factor* x *and reflect the fact that* $(Db)^2\frac{\sigma^2}{D^2} = b^2\sigma^2$ *for* $D \neq 0$*. Transformations in* $\mathcal{T}_S$ *correspond to reflections of* x *across the axis* $x = 0$*, which change its sign.*

Example 2 (Location mixture) *If the data is a sample from a finite mixture distribution whose* K *component distributions are from the same parametric family, then the likelihood has* $K!$ *symmetric lobes, each lobe corresponding to a permutation of the components' indices. Consider the following mixture of* $K = 2$ *normal distributions with common variance* σ^2 *and means* μ_1 *and* μ_2

$$f(\mathbf{y}|\mu_1,\mu_2,\pi,\sigma) = \pi\mathcal{N}(\mathbf{y}|\mu_1,\sigma) + (1-\pi)\mathcal{N}(\mathbf{y}|\mu_2,\sigma).$$

Label (or permutation) invariance refers to the likelihood function's invariance with respect to the re-labeling of the components. Here,

$$f(\mathbf{y}|\,\mu_1,\mu_2,\pi,\sigma) = f(\mathbf{y}|\,\mu_2,\mu_1,1-\pi,\sigma), \tag{2.1}$$

which establishes the invariance to the relabeling (or permuting) of component indices 1 *and* 2*. In matrix notation, a set of invariant transformations is*

$$\mathcal{T}_{\mathbf{P}}(\Theta) = \{T_{\mathbf{P}} : \Theta \to \Theta|\, T_{\mathbf{P}}(\mu,\Pi,\sigma) = (\mathbf{P}\mu, \mathbf{P}\Pi, \sigma)\}$$

with $\Theta = \mathbb{R}^2 \times \mathcal{S}^2 \times \mathbb{R}$*,* $\mathcal{S}^2$ *is the simplex of* $\mathbb{R}^2$*,* $\mu = [\mu_1\ \mu_2]'$*,* $\Pi = [\pi\ 1-\pi]'$ *and* $\mathbf{P}$ *is a permutation matrix,* i.e. *a matrix obtained by permuting the rows of an identity matrix.*

2.2.1 What is normalization?

In general, transformation invariance is addressed by normalizing the model.

Definition 3 *A normalization is a parameter subspace* $\Theta^N \subseteq \Theta$.

Normalizing a model thus defines a new model $\left(\mathcal{Y}, \mathcal{F}, \Theta^N\right)$.

Example 3 (Normal mean, continued) *Some normalizations are*

$$\begin{aligned}
\Theta^{\sigma_1^2} &= \left\{\theta \in \Theta \middle| \, \sigma^2 = 1\right\}, \\
\Theta^{b_{pos}} &= \left\{\theta \in \Theta \middle| \, b > 0\right\}, \\
\Theta^{b_{pos}\sigma_1^2} &= \Theta^{b_{pos}} \cap \Theta^{\sigma_1^2}.
\end{aligned}$$

The normalization $\Theta^{\sigma_1^2}$ *defines the following subsets of invariant transformations:*

$$\begin{aligned}
\mathcal{T}_S\left(\Theta^{\sigma_1^2}\right) &= \left\{T_S : \Theta^{\sigma_1^2} \to \Theta^{\sigma_1^2} \middle| \, T_S(b, \sigma^2) = \left(Sb, \sigma^2\right), |S| = 1\right\}, \\
\mathcal{T}_D\left(\Theta^{\sigma_1^2}\right) &= \left\{T_D : \Theta^{\sigma_1^2} \to \Theta^{\sigma_1^2} \middle| \, T_D(b, \sigma^2) = \left(Db, \sigma^2/D^2\right), D = 1\right\}, \\
\mathcal{T}_{SD}\left(\Theta^{\sigma_1^2}\right) &= \left\{T_{SD} : \Theta^{\sigma_1^2} \to \Theta^{\sigma_1^2} \middle| \, T_{SD}(b, \sigma^2) = \left(SDb, \sigma^2/D^2\right), D = 1, |S| = 1\right\}.
\end{aligned}$$

Note that the set $\mathcal{T}_D\left(\Theta^{\sigma_1^2}\right)$ *is a singleton, but that there are two transformations in the sets* $\mathcal{T}_S\left(\Theta^{\sigma_1^2}\right)$ *and* $\mathcal{T}_{SD}\left(\Theta^{\sigma_1^2}\right)$.

Definition 4 *Suppose that a function* $f : \Theta \to \mathbb{R}$ *is invariant with respect to a set of bijective transformations* $\mathcal{T}(\Theta)$. *A normalization* $\Theta^N \subseteq \Theta$ ***breaks the invariance*** *of* f *with respect to* $\mathcal{T}$, *which is denoted*

$$\mathcal{T}\left(\Theta^N\right) = \mathcal{T}_I,$$

if for all $T \in \mathcal{T}(\Theta)$,

$$\Theta^N \cap T(\Theta^N) \neq \emptyset \Rightarrow T \in \mathcal{T}_I,$$

where

$$\mathcal{T}_{\mathcal{I}} \;=\; \left\{T : \Theta \to \Theta \middle| \, T(\theta) = \theta\right\}$$

is a singleton: the identity transformation.

Note that $\Theta^N \subseteq \Theta \Rightarrow \mathcal{T}\left(\Theta^N\right) \subseteq \mathcal{T}\left(\Theta\right)$. Breaking invariance with respect to a set of bijective transformations $\mathcal{T}(\Theta)$ is thus considering a parameter subspace $\Theta^N \subseteq \Theta$ that is small enough to ensure that the only invariant bijection on that subspace is the identity transformation, $\mathcal{T}(\Theta^N) = \left\{T : \Theta^N \to \Theta^N \middle| \, T(\theta) = \theta\right\}$.

Example 4 (Normal mean, continued) *Consider the following scaling and reflection transformation sets:*

$$\begin{aligned}
\mathcal{T}_S\left(\Theta^b\right) &= \left\{T_S : \Theta^b \to \Theta^b \middle| \, T_S(b,\sigma^2) = \left(Sb, \sigma^2\right),\, S = 1\right\} = \mathcal{T}_{\mathcal{I}},\\
\mathcal{T}_S\left(\Theta^{\sigma_1^2}\right) &= \mathcal{T}_S\left(\Theta\right),\\
\mathcal{T}_D\left(\Theta^b\right) &= \mathcal{T}_D\left(\Theta\right),\\
\mathcal{T}_D\left(\Theta^{\sigma_1^2}\right) &= \left\{T_D : \Theta^{\sigma_1^2} \to \Theta^{\sigma_1^2} \middle| \, T_D(b,\sigma^2) = \left(Db, \sigma^2/D^2\right),\, D = 1\right\} = \mathcal{T}_{\mathcal{I}},\\
\mathcal{T}_{SD}\left(\Theta^b \cap \Theta^{\sigma_1^2}\right) &= \left\{T_{SD} : \Theta^b \cap \Theta^{\sigma_1^2} \to \Theta^b \cap \Theta^{\sigma_1^2} \middle| \right.\\
&\qquad \left. T_{SD}(b,\sigma^2) = \left(SDb, \sigma^2/D^2\right),\, S = 1, D = 1\right\} = \mathcal{T}_{\mathcal{I}}.
\end{aligned}$$

The normalization Θ^b breaks invariance with respect to reflection because T_S is not a bijection on Θ^b for $S \neq 1$. Similarly, $\Theta^{\sigma_1^2}$ breaks invariance with respect to scaling because T_D is not a bijection on $\Theta^{\sigma_1^2}$ for $D \neq 1$. Thus, $\Theta^b \cap \Theta^{\sigma_1^2}$ breaks invariance with respect to $\mathcal{T}_{SD}$.

Example 5 (Location mixture, continued) *One might contemplate one of the two following normalizations:*

$$\Theta^{\pi} = \{\theta \in \Theta | \, \pi > 0.5\}$$
$$\Theta^{\mu} = \{\theta \in \Theta | \, \mu_1 > \mu_2\}.$$

Each normalization would break permutation invariance as $\mathcal{T}(\Theta^{\pi}) = \mathcal{T}(\Theta^{\mu}) = \mathcal{T}_{\mathcal{I}}$.

One could consider normalizations of arbitrary form, but I restrict the following discussion to intersections of half spaces and hyper-planes,

$$\Theta^{N} = \bigcap_{i=1}^{I} \{\theta \in \Theta | \, \mathbf{g}_i'\theta > 0\} \cap \bigcap_{j=1}^{J} \{\theta \in \Theta | \, \mathbf{h}_j'\theta = 0\},$$

for some conformable real vectors $\mathbf{g}_1, \ldots, \mathbf{g}_I, \mathbf{h}_1, \ldots, \mathbf{h}_J$. For example, one would break invariance with respect to a set of $(I+1)!$ invariant transformations with a normalization consisting in the intersection of I half spaces. In contrast, the intersection of J hyper-planes would break invariance with respect to a set of invariant transformations that is equinumerous to $\mathbb{R}^J$ (*i.e* a set $\mathcal{T}$ that has the same cardinality as $\mathbb{R}^J$).

Example 6 (Normal mean, continued) *There are* $2!$ *transformations in* $\mathcal{T}_S(\Theta)$ *and the half space* $\{\theta \in \Theta | \, b > 0\}$ *breaks invariance with respect to reflections. The set* $\mathcal{T}_D(\Theta)$ *is equinumerous to* $\mathbb{R}$ (e.g. *the natural logarithm is a bijection from* $(0, \infty)$ *to* $\mathbb{R}$*) and the line* $\{\theta \in \Theta | \, \sigma^2 = 1\}$ *breaks scale invariance.*

Note that considering intersections of half spaces and hyper-planes is not as restrictive as it might seem. In particular, one can consider half spaces and hyper-planes in any space that is homeomorphic to Θ. In section 2.3, for example, I reparameterize some vectors in polar coordinates and I normalize in the space of angles and lengths.

2.2.2 Is normalization necessary?

When the likelihood function of a latent-variable model is invariant with respect to a certain set of parameter transformations, the maximum-likelihood problem defines a

parameter set estimator rather than a point estimator,

$$\left\{\theta \in \Theta \middle| \theta = \arg\max_{\theta' \in \Theta} l(\theta' | y)\right\}.$$

Example 7 (Location mixture, continued) *Permutation invariance implies that the likelihood function admits two equivalent global maxima, sitting at the summit* ***symmetric lobes****: if* $\left(\hat{\mu}, \hat{\Pi}, \hat{\sigma}\right)$ *is a global maximum, so is* $\left(\mathbf{P}\hat{\mu}, \mathbf{P}\hat{\Pi}, \hat{\sigma}\right)$*, for* $\mathbf{P} = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$.

While normalizing the parameter space in order to obtain well-defined ML parameter point estimators is common practice, it should be emphasized that normalization is often not strictly necessary. In particular, parameter inference is feasible as soon as the parameter set estimator is bounded (See Manski (2003) for a textbook treatment, and Chernozhukov et al. (2007) and Galichon and Henry (2009)). In terms of conditions on transformation sets, a sufficient condition is therefore that the set can be parameterized and that this parameter is bounded.

Definition 5 *A set of transformations*

$$\mathcal{T}_{\mathcal{J}}(\Theta) = \left\{T_j : \Theta \to \Theta \middle| j \in \mathcal{J} \subseteq \mathbb{R}\right\}$$

is ***bounded*** *if* $\mathcal{J}$ *is a bounded set.*

Example 8 (Normal mean, continued) *The ML parameter set estimator of* θ *on* $\Theta^{\sigma_1^2}$ *is bounded as* $\mathcal{T}_{SD}\left(\Theta^{\sigma_1^2}\right)$ *is bounded.*

From a Bayesian perspective, as long as priors are proper, posteriors are proper and the model is "identified" in that specific sense, but Bayesian inference is not immune to invariance issues. Although it is common practice to operationalize normalization through a truncation of the prior distribution, considering normalization and prior specification independently makes exposition clearer. I therefore consider priors such that $f(\theta) > 0$ for all $\theta \in \Theta$ in this paper.

I will argue in Section 2.4 that it is conceptually inconsistent to express prior beliefs over the relative plausibility of observationally equivalent parameter values. In other words, if the likelihood function is invariant with respect to a given transformation set, prior distributions should also be invariant with respect to that set.

Example 9 (Location mixture, continued) *The proper prior distribution* $f(\mu) = \mathcal{N}(\mu|\, \mathbf{0}, I)$ *is invariant with respect to* $\mathcal{T}_{\mathbf{P}}(\Theta)$.

This raises the question of whether there always exists a proper prior on Θ that is invariant with respect to any given set of transformations $\mathcal{T}(\Theta)$. This is obviously not the case. It is possible to specify a proper prior that is uninformative about (uniform over) observationally equivalent parameter values if and only if the transformation set $\mathcal{T}$ is bounded.

Example 10 (Normal mean, continued) $f(y|\, b, \sigma^2)$ *is invariant with respect to* $\mathcal{T}_S(\Theta)$ *and* $\mathcal{T}_D(\Theta)$. *Finding a proper joint prior distribution that is invariant with respect to* $\mathcal{T}_S(\Theta)$ *is straightforward. For example, a joint prior is invariant with respect to* $\mathcal{T}_S(\Theta)$ *as soon as its marginal prior* $f(b)$ *is symmetric and centered on zero. In contrast, there exists no proper joint prior* $f(b, \sigma^2)$ *such that* $f(b, \sigma^2) = f(ab, \sigma^2/a^2|\, a)$ *for all* $a \in \mathcal{A} = (0, \infty)$ *as this would require that the prior be constant over unbounded sets.*

Thus, Bayesian and ML inference for set estimators is possible when $\mathcal{T}$ is bounded. When $\mathcal{T}$ is not bounded, it is sometimes possible to write transformations in $\mathcal{T}$ as compositions of other transformations and identify a bounded subset of transformations.

Example 11 (Normal mean, continued) $\mathcal{T}_{SD}(\Theta)$ *is unbounded, but* $T_{SD}(b, \sigma^2) = T_S(T_D(b, \sigma^2))$ *and* $\mathcal{T}_S(\Theta)$ *is bounded. Therefore, one needs not break invariance with respect to reflections in order to make inference for* (b, σ^2). *For example, parameter set estimators are well-defined under* $\Theta^{\sigma_1^2}$ *as* $\mathcal{T}_{SD}(\Theta^{\sigma_1^2})$ *is bounded.*

This example illustrates that one can sometimes write an unbounded transformation set as the composition of smaller bounded and unbounded subsets. Breaking invariance with respect to the unbounded subset is sufficient for set estimators to be bounded.

2.2.3 What are the costs and benefits of normalization?

Because point estimators are simpler from a interpretational as well as computational point of view, they are often preferred to set estimators in empirical applications. Indeed, ML inference often calls for simulations methods (For example, Jacquier et al. (2007) show how a simple modification of the Bayesian MCMC algorithm produces the ML point estimate and its asymptotic variance covariance matrix.) and non connected confidence sets constitute a challenge to communication empirical results.

In the Bayesian framework, if prior distributions and the likelihood function are invariant with respect to a set of transformations $\mathcal{T}$, then so are posterior distributions. Invariant *proper* prior and posterior distributions are perfectly valid characterization of uncertainty. In cases where $\mathcal{T}$ is finite (but not a singleton), some posterior distributions are multimodal and thus cause interpretational difficulties. For example, if the bimodal posteriors of μ_1 and μ_2 in (3.5) are symmetric with respect to zero, the posterior means of these parameters are both equal to zero, $\mathbb{E}[\mu_1|\mathbf{y}] = \mathbb{E}[\mu_2|\mathbf{y}] = 0$ which is not particularly informative about the mixture components.

The model should therefore be normalized if the investigator uses mixtures or LSSMs as classification tools where the interpretation of the components or state variables is of interest[1]. When the parameters are not of direct interest however, such as when one uses latent-variable models as flexible parameterizations of the observables's

[1] Stephens (2000) discusses alternative, decision-theoretic approaches.

distribution, transformation invariance introduces no interpretational difficulty and normalization, beyond what is required in order to obtain well defined set estimators, is unnecessary.

In general, a multimodal posterior distribution also constitutes a computation challenge for a basic posterior simulator, but posteriors in latent-variable are no general multi-modal distributions: they are symmetric. Symmetry has two important implications. First, because any lobe contains all relevant information about the parameters, one can consider any single one of them. Second, because all lobes are equivalent, the mixing properties of a posterior simulator over permutations are irrelevant (Geweke, 2007).

2.2.4 Are the potential benefits of normalization always achievable?

In some cases, normalization can fail to provide its expected benefits and ML parameter point estimators can have multimodal sampling distributions, which causes concerns equivalent to those we have with set estimators. For example, multimodality can imply disjoint confidence intervals. Similarly, parameter posterior distributions can be multimodal. In such situations, interpretational benefits are lost. In addition, symmetric asymptotic approximation are unreliable and one must thus obtain sampling distributions by simulation methods.

These problems arise when some elements of θ are weakly identified. Except in the context of instrumental variables (IV) and the generalized method of moments (GMM), weak identification has not been defined precisely. Dufour and Hsiao (2008) write: "More generally, any situation where a parameter may be difficult to determine because we are close to a case where a parameter ceases to be identifiable may be called *weak identification*." Many common situations fit this description. For example,

multicollinearity issues arise in linear regression models when the sample covariance matrix of the regressors is "close" to being singular. If one restricts attention to ML inference, weak identification problems occur when the Fisher information matrix is close to being singular at the pseudo-true parameter values. Thus, weak identification is a joint property of both the model $(\mathcal{Y}, \mathcal{F}, \Theta^N)$ and the data y, as the term "empirical underidentification" used in psychometrics emphasizes. In this paper, I say that a parametric model is **weakly identified** if $\hat{\Theta}(\Theta^N)$ is close to Θ^l, where $\Theta^l \subseteq \Theta$ is the **singularity parameter subspace** where the Fisher information matrix is singular.

Example 12 (Location mixture, continued) *The information matrix is singular on* $\{\theta \in \Theta | \mu_1 = \mu_2\} \subset \Theta^\pi$, *where the probability* π *is unidentified. Thus we say that the model is weakly identified if the pseudo-true parameters* μ_1 *and* μ_2 *are too close to each other. In such situations, the lobes of the likelihood function are not well separated and they are not symmetric with respect to their respective mode. A symmetric normal approximation of the ML estimators's sampling distribution is thus unlikely to be accurate. Indeed, Dick and Bowden (1973) compare a Monte Carlo approximation of the parameter sampling variances to their asymptotic counterparts, which are approximated by a power series expansion of the information matrix developed by Hill (1963). They report that [Summary] "the sample variance of the estimates can be as much as three times greater than the estimated asymptotic variances".*

Weak identification has severe consequences for ML inference, which Dufour and Hsiao (2008) summarize thus:

> "...standard asymptotic distributional may remain valid, but they constitute very bad approximations to what happens in finite samples:
>
> 1. standard consistent estimators of structural parameters can be heavily biased and follow distributions whose form is far from the limiting

Gaussian distribution, such as bimodal distributions, even with fairly large samples (Nelson and Startz, 1990; Hiller, 1990; Buse, 1992);

2. standard tests and confidence sets, such as Wald-type procedures based on estimated standard errors, become highly unreliable or completely invalid (Dufour, 1997)"

How close is too close? As weak identification is a finite-sample concern, one might be tempted to believing it is only a small-sample concern. Even for fairly large sample sizes, however, the asymptotic approximation of the ML estimator's sampling distribution may be unreliable. Bound et al. (1995) present an IV situation in which weak identification difficulties persist even with 329000 observations. Intuitively, if the instruments were uncorrelated with the regressors in population, increasing the sample size would be futile. In practice however, the statistician never knows the pseudo-true parameter values and he should favor inferential methods that are robust to weak identification.

ML parameter point estimators are less attractive quantities when their sampling distribution are multimodal. For the same reasons that normalizations do not guarantee unimodal ML estimator sampling distributions, they do not guarantee unimodal posterior distributions (See Stephens (2000) for a discussion). Parameter uncertainty plays an important role is such situations. Bayesian inference deals with parameter uncertainty in a consistent manner. While standard ML forecasts rely on parameter point estimates, Bayesian forecasts average over the parameter posterior distribution. When weak identification issues arise and point estimators become unreliable, the simulations results presented in this paper reveal that the richer information content of posterior distributions yields better out of sample forecasts. Bayesian analysis has proved a useful framework for other weak identification problems. For example, Leamer (1973) provides an illuminating interpretation of multicollinearity. In this paper, I build on the

fact that global identification is unnecessary for forecasting purposes. Whether some parameters are subject to weak identification problems is therefore irrelevant.

2.2.5 How best to normalize?

Because there are many ways to normalize a model, it is natural to ask how alternatives should be compared. This paper proposes three criteria for choosing normalizations.

A first natural criterion is that a normalization should be observationally unrestrictive.

Definition 6 *Suppose $l(\theta|\, y)$ is the likelihood function of a parametric statistical model $(\mathcal{Y}, \mathcal{F}, \Theta)$. A normalization $\Theta^N \subseteq \Theta$ is* ***observationally unrestrictive*** *if there exists a transformation $g : \Theta \to \Theta^N$ such that $l(g(\theta)|\, y) = l(\theta|\, y)$ for all $y \in \mathcal{Y}$. A normalization is observationally restrictive otherwise.*

The two other criteria pertain to the shape of point estimator sampling or parameter posterior distributions. Obviously, multimodality issues will arise more often if the normalization is disconnected. A second criteria is thus that the normalization should be connected[2]. Note that intersections of half spaces and hyper-plans are connected spaces. Also, continuous bijections preserve connectedness (Royden, 1988).

Global identification does not only ensure ML estimator's uniqueness, it also affects its sampling distribution. If global identification is achieved through normalization, then normalization has implications for estimator sampling and parameter posterior distributions. Hiller (1990) shows how normalization in structural equations models affects the finite-sample distribution of ordinary least squares and two-stage least squares

[2]A space Θ^N is said to be connected if there do not exist two nonempty disjoint open sets O_1 and O_2 such that $\Theta^N = O_1 \cup O_2$.

estimators. Unfortunately, as Hamilton, Waggoner, and Zha (2007) note, "the fact that normalization can materially affect the conclusions one draws from likelihood-based methods is not widely recognized."

Hamilton, Waggoner, and Zha (2007) propose an *identification principle* as a general guideline for the choice of normalizations, advising that one should [p. 225] "make sure that the model is locally identified at all interior points". More generally, weak identification difficulties are amplified when the model is not globally identified. Global identification thus defines a third preorder on normalizations: Normalizations providing global identification are more likely to produce unimodal sampling distributions and thus alleviate weak identification issues.

In this paper, I use the following definition, which captures the three criteria described above:

Definition 7 *A normalization* $\Theta^N \subseteq \Theta$ *satisfies the* ***identification principle*** *if it*

$a)$ *is observationally unrestrictive;*

$b)$ *is connected;*

$c)$ *provides global identification.*

The following examples illustrate how disconnectedness and local non-identification can produce multimodal estimator sampling distributions.

Example 13 (Normal mean, continued) *The disconnected normalization* $\Theta^{b_{disc}} = \{\theta \in \Theta | \, b \in [-1, 0) \cup (1, \infty)\}$ *provides global identification and is observationally unrestrictive. However, it would produce a bimodal sampling distribution for* $\hat{b}$ *if the true parameter value of* b *were close to being equal to* 1.

Example 14 (Location mixture, continued) *The sampling distributions of the ML estimator of μ_1 and μ_2 can be multimodal under Θ^π. Intuitively, this normalization would perform poorly if the data came from a mixture distribution with $\pi = 0.5$ because component densities would be equiprobable. The identification principle rules out Θ^π because the Fisher information matrix is singular on $\{\theta \in \Theta | \, \mu_1 = \mu_2, \pi = 0.5\} \subset \Theta^\pi$. In contrast, the model is globally identified on Θ^μ.*

In the latter example, the identification principle yields a unique normalization, under which the ML estimator has a unimodal sampling distribution for any $\theta \in \Theta^\mu$. In slightly more general models, the identification principle is less straightforward to apply, as it may yield uncountably many normalizations. The practical guidance that the identification principle offers is thus incomplete. Moreover, there is no guarantee that any particular normalization ensures that the ML estimator has a unimodal sampling distribution.

Example 15 *Consider the location-and-scale mixture of normal distributions*

$$f(y_t|\mu_1, \mu_2, \pi, \sigma_1^2, \sigma_2^2) \;=\; \pi\phi(y_t|\mu_1, \sigma_1^2) + (1-\pi)\phi(y_t|\mu_2, \sigma_2^2).$$

The set where the information matrix is singular is not a line but a point,

$$\Theta^l \left\{\theta \in \Theta | \, \mu_1 = \mu_2\right\} \cap \left\{\theta \in \Theta | \, \sigma_1 = \sigma_2\right\}.$$

The identification principle still rules out restrictions based on π, but the singularity subspace no longer separates the parameter space into two symmetric half-spaces. Normalizations Θ^μ and

$$\Theta^\sigma = \{\theta \in \Theta | \, \sigma_1 > \sigma_2\}$$

both satisfy the identification principle, but neither ensures that all sampling distributions are unimodal. To illustrate, consider samples from a population with $\mu_1 = \mu_2$

and $\sigma_1 > \sigma_2$. Under Θ^{μ}, the ML estimator of σ_1 and σ_2 will have bimodal sampling distributions for sufficiently large samples (Geweke, 2007).

It cannot be overemphasized that the identification principle does not "solve" the weak identification problem. While it usefully defines a preorder on normalizations, it falls short of recommending a unique optimal normalization. Moreover, the identification principle does not ensure that standard asymptotics provide reliable approximations of ML estimator sampling distributions, and one should resort to simulation methods to accurately characterize the true statistical uncertainty of ML estimators.

Therefore, although unimodal sampling or posterior distribution may be desirable, they cannot be guaranteed. One can try a number of normalizations satisfying the identification principle and hope to find one that yields estimators with acceptable finite-sample properties. For ML inference, comparing competing normalizations can be impractical. For each, one should obtain sampling distributions by simulation methods. Because the ML estimator does not have a closed-form solution, this involves substantial computational costs.

Stephens (1997) shows that normalizations can equivalently be applied within a posterior sampler or as a post-simulation step on the output from an un-normalized sampler. Using the latter implementation, one can compare competing normalizations with negligible computational cost. Indeed, because normalizations truncate posteriors but do not change their informational content, they can be chosen a posteriori (Frühwirth-Schnatter, 2001). However, this exercise can be difficult in high-dimensional models.

2.2.6 Summary

Inference in latent variable models is possible as soon as the set of transformations with respect to which the likelihood function is invariant, is finite. One then has parameter set estimators. One can normalize further in order to obtain parameter point estimators. This might ease interpretation and computation, but can make inference sensitive to weak identification issues. In particular, parameter point estimates are unreliable quantities if the estimator's sampling distribution is multimodal, and parameter uncertainty should be taken into account. Also, exact sampling distributions are required in order to correctly describe statistical uncertainty. When parameter point estimates are of direct interest, connected normalizations that provide global parameter identification are more likely to produce unimodal sampling or posterior distributions. Comparing many such normalizations might prove useful. This is computational trivial in the Bayesian framework, but often impractical in the ML framework as estimators are not available in closed form.

2.3 Normalization of LSSMs

In the notation of Hamilton (1994), let $\mathbf{y}_t$ be a N-dimensional vector of observables at time t, ξ_t an latent (or unobserved) K-dimensional vector of latent state variables, and $\mathbf{x}_t$ a l-dimensional vector of observed exogenous variables. A Markovian Gaussian linear state-space model is defined by the system of equations

$$\xi_{t+1} = \mathbf{F}\xi_t + \mathbf{v}_t, \tag{2.2}$$

$$\mathbf{y}_t = \mathbf{B} + \mathbf{A}'\mathbf{x}_t + \mathbf{H}'\xi_t + \mathbf{w}_t, \tag{2.3}$$

where $\mathbf{v}_t$ and $\mathbf{w}_t$ are Gaussian white noises with covariance matrices $\mathbf{Q}$ and $\mathbf{R}$, respectively[3]. Equation (2.2) is referred to as the *state equation* and equation (2.3) as the *observation equation*. State variables are also known as *factors* and $\mathbf{H}$ as the matrix of *factor loadings*. For expositional clarity I consider only the case $\mathbf{A} = \mathbf{0}$ and $N \geq K$, but this is not a substantive restriction.

The likelihood function is invariant with respect to invertible linear transformations of the latent variables; for any invertible $\mathbf{M}$,

$$l(\mathbf{B}, \mathbf{M}'^{-1}\mathbf{H}, \mathbf{R}, \mathbf{MFM}^{-1}, \mathbf{MQM}', \mathbf{M}\xi_1|y_t) \equiv l(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{F}, \mathbf{Q}, \xi_1|y_t). \quad (2.4)$$

Thus, the system (2.2-2.3) can be written as

$$\tilde{\xi}_{t+1} = \tilde{\mathbf{F}}\tilde{\xi}_t + \tilde{\mathbf{v}}_t, \quad (2.5)$$

$$\mathbf{y}_t = \mathbf{B} + \mathbf{A}'\mathbf{x}_t + \tilde{\mathbf{H}}'\tilde{\xi}_t + \mathbf{w}_t, \quad (2.6)$$

where $\tilde{\mathbf{H}} = \mathbf{M}'^{-1}\mathbf{H}$, $\tilde{\mathbf{F}} = \mathbf{MFM}^{-1}$, $\tilde{\mathbf{Q}} = \mathbf{MQM}'$, $\tilde{\xi} = \mathbf{M}\xi$ and $\tilde{\mathbf{v}} = \mathbf{Mv}$.

2.3.1 Primitive transformations

In order to highlight the weak identification issues, it is useful to consider primitive transformations $\mathbf{M} = \{\mathbf{D}, \mathbf{O}, \mathbf{P}, \mathbf{S}\}$, where

- $\mathbf{D}$ is a diagonal, positive-definite *scaling* matrix;
- $\mathbf{O}$ is a *rotation* matrix;
- $\mathbf{P}$ is a *permutation* matrix;
- $\mathbf{S}$ is a diagonal *reflection* matrix with elements equal to 1 or -1.

[3]Appendix 2.8 shows how to generalize my results to LSSMs with correlated errors.

Let $\mathcal{T}_{\mathbf{D}}$, $\mathcal{T}_{\mathbf{O}}$, $\mathcal{T}_{\mathbf{P}}$ and $\mathcal{T}_{\mathbf{S}}$ denote these four sets of primitive transformations. Permutation and reflection matrices are orthogonal matrices, *i.e.* $\mathbf{P}'\mathbf{P} = \mathbf{S}'\mathbf{S} = \mathcal{I}$; rotation matrices are special orthogonal matrices, *i.e.* $\mathbf{O}'\mathbf{O} = \mathcal{I}$ and $|\mathbf{O}| = 1$. Any linear transformation can be decomposed into these primitive transformations in order to help clarifying invariance issues.

Note that the sets $\mathcal{T}_{\mathbf{S}}$, $\mathcal{T}_{\mathbf{P}}$ and $\mathcal{T}_{\mathbf{O}}$ are bounded. Indeed, $\mathcal{T}_{\mathbf{S}}$ contains 2^K transformations, $\mathcal{T}_{\mathbf{P}}$ contains $K!$ transformations, and $\mathcal{T}_{\mathbf{O}}$ can be parameterized by $K(K-1)/2$ angles. This means that breaking scale invariance is sufficient in order to make inference in LSSMs.

As permutation invariance in mixture distributions, permutation and reflection invariance makes the likelihood function of LSSMs multimodal and local non-identification introduces weak identification concerns. Intuitively, permutations are weakly identified when factors are too similar, and reflections are weakly identified when factor loadings are too small.

A much-cited reference on LSSM normalization is Harvey (1989). He writes (for the special case $\mathbf{F} = \mathcal{I}$) (p.451):

> "In order for the model to be identifiable, restrictions must be placed on [$\mathbf{Q}$] and [$\mathbf{H}$]. In classical factor analysis, the covariance matrix of the common factors is taken to be an identity matrix. However, this is not sufficient to make the model identifiable since if [M] is an orthogonal matrix, [(2.5-2.6)] still satisfies all the restrictions of the original model because [$\mathbf{Var}(\mathbf{M}\mathbf{v}_t) = \mathbf{M}\mathbf{M}' = \mathcal{I}$]. Some restrictions are needed on [M], and one way of imposing them is to require that the ij-th element of [M], [$\mathbf{M}_{ij}$], be zero for $j > i$,

$i = 1, \ldots, K-1$. Alternatively, $[\mathbf{Q}]$ can be set equal to a diagonal matrix while $[\mathbf{M}_{ij}] = 0$ for $j > i$ and $[\mathbf{M}_{ii}] = 1$ for $i = 1, \ldots, K$."

The proposed normalization are $\Theta^{\mathbf{Q}_{\mathcal{I}}} \cap \Theta^{\mathbf{H}_{lt}}$ and $\Theta^{\mathbf{Q}_{diag}} \cap \Theta^{\mathbf{H}_{lut}}$, where

$$\begin{aligned}
\Theta^{\mathbf{Q}_{diag}} &= \{\theta \in \Theta | \, \mathbf{Q} \text{ is diagonal}\}; \\
\Theta^{\mathbf{Q}_{\mathcal{I}}} &= \{\theta \in \Theta | \, \mathbf{Q} = \mathcal{I}\}; \\
\Theta^{\mathbf{H}_{lt}} &= \{\theta \in \Theta | \, \mathbf{H} \text{ has a lower triangular } K \times K \text{ block}\}; \\
\Theta^{\mathbf{H}_{lut}} &= \{\theta \in \Theta | \, \mathbf{H} \text{ has a lower unitriangular } K \times K \text{ block}\}.
\end{aligned}$$

A unitriangular matrix is triangular and has ones on the main diagonal. It is straightforward to show that these normalizations break invariance with respect to scaling, rotation, reflection and permutation. However, they do not provide global parameter identification. Moreover, these normalizations are observationally restrictive.

Both of Harvey's normalizations are observationally restrictive because they involve K^2 parameter restrictions while breaking scale invariance requires K restrictions and breaking rotation invariance requires $K(K-1)/2$ restrictions. Thus, $K(K-1)/2$ additional parameter restrictions reduce the model's flexibility. I will present several normalizations in this section, but consider a simple one here in order to illustrate how one can normalize scales and rotations with $K(K+1)/2$ restrictions. The central issue is that identity matrices are diagonal matrices with diagonal elements all set to 1, which implies that $\mathbf{M}\mathcal{I}\mathbf{M}' = \mathcal{I}$ if $\mathbf{M}$ is an orthogonal matrix. In contrast, consider diagonal matrices with diagonal elements set to distinct values, say $\mathbf{Q}_{kk} = k$ for $k = 1, \ldots, K$. Because $\mathbf{MQM}' \neq \mathbf{Q}$ when $\mathbf{M}$ is orthogonal, these $K(K+1)/2$ restrictions break rotation and scale invariance.

In order to see why Harvey's normalizations do not provide global parameter

identification, we first need to find a parameter subspace where some parameters are locally unidentified. Then, we need to show that the intersection of this subspace and the interior of the normalization is not empty. For example, consider the parameter subspace where the first column of $\mathbf{H}$ is a vector of ones and its other elements are all zeros. This subspace in strictly contained in both $\Theta^{\mathbf{H}_{lt}}$ and $\Theta^{\mathbf{H}_{lut}}$. Permutation invariance is broken by $\Theta^{\mathbf{H}_{lt}}$ or $\Theta^{\mathbf{H}_{lut}}$ because permuting the rows of a triangular matrix does not yield, *in general*, a triangular matrix. Thus, row permutation is not a bijective transformation on the space of triangular matrices. However, row permutation is a bijective transformation on the region described above: the first column of $\mathbf{PH}$ would be a vector of ones and its other elements would be all zeros. Thus $\Theta^{\mathbf{H}_{lt}}$ and $\Theta^{\mathbf{H}_{lut}}$ do not provide global identification.

I proceed to propose connected, observationally unrestrictive normalizations providing global identification. To the best of my knowledge, these are the first normalizations of LSSMs satisfying the identification principle. Although the concepts are easily extendable to other distributions, I present normalizations for Gaussian LSSMs for expositional clarity.

2.3.2 Breaking rotation invariance

The likelihood function of LSSMs is invariant to geometric rotations of state variables in Euclidean space: for given parameter values $\mathbf{Q}$, $\mathbf{H}$ and $\mathbf{F}$, any rotation matrix O defines observationally equivalent parameter values $\tilde{\mathbf{Q}} = \mathbf{O}\mathbf{Q}\mathbf{O}'$, $\tilde{\mathbf{H}} = \mathbf{O}'^{-1}\mathbf{H}$ and $\tilde{\mathbf{F}} = \mathbf{O}\mathbf{F}\mathbf{O}^{-1}$. Any K-dimensional rotation matrix can be parameterized by $\frac{K(K-1)}{2}$ angles.

Consider several normalization imposing $\frac{K(K-1)}{2}$ parameter restrictions:

$$
\begin{aligned}
\Theta^{\mathbf{Q}_{diag}} &= \{\theta \in \Theta | \, \mathbf{Q} \text{ is diagonal}\} \\
\Theta^{\mathbf{H}_{lt}} &= \{\theta \in \Theta | \, \mathbf{H} \text{ is lower triangular}\} \\
\Theta^{\mathbf{H}_{or}} &= \{\theta \in \Theta | \, \mathbf{HH}' \text{ is diagonal } (\mathbf{H} \text{ is row-orthogonal})\} \\
\Theta^{\mathbf{F}_{lt}} &= \{\theta \in \Theta | \, \mathbf{F} \text{ is lower triangular}\} \\
\Theta^{\mathbf{F}_{sym}} &= \{\theta \in \Theta | \, \mathbf{F} \text{ is symmetric}\}
\end{aligned}
$$

For the reasons given above, $\Theta^{\mathbf{H}_{lt}}$ does not provide global identification. Nor does $\Theta^{\mathbf{F}_{lt}}$, by similar arguments. This could lead one to consider $\Theta^{\mathbf{F}_{sym}}$ as simultaneous row-and-column permutation is a bijective transformation for symmetric matrices. However, this is observationally restrictive because the eigenvalue of real symmetric matrices are real. In particular, this would precludes latent variables with sinusoidal dynamics.

The off-diagonal elements of $\mathbf{Q}$ are not identified if $\mathbf{F} = \mathbf{0}$, which corresponds to the special case of static factor analysis. Thus $\Theta^{\mathbf{Q}_{diag}}$ does not provide global identification.

It therefore seems that $\Theta^{\mathbf{H}_{or}}$ is the only normalization considered above that provides global identification. However, one must parameterize row-orthogonal matrices with care in order to preserve permutation invariance. One way is in polar coordinates. In this parameterization, the K rows of $\mathbf{H}$ are points on N-dimensional spheres, each parameterized by $N-1$ angles. Let γ denote the $K \times N-1$ matrix of these angles and δ denote the K-dimensional vector of row lengths with elements

$$\gamma_{k,n} \equiv \arctan\left(\frac{H_{k,n+1}}{\sqrt{\sum_{i=1}^{n} H_{k,i}^2}}\right),$$

$$\delta_k \equiv \sqrt{\sum_{i=1}^{N} H_{k,i}^2},$$

for $k = 1, \ldots, K$ and $n = 1, \ldots, N-1$. Note that γ are not Euler angles.

In polar coordinates, I parameterize a row-orthogonal $K \times N$ factor loading matrix as

$$\mathbf{H}' = \mathbf{B}_1 \mathbf{B}_2 \ldots \mathbf{B}_K \mathbf{U},$$

where

$$\mathbf{B}_k = \rho_{k,k+1}\rho_{k,k+2}\cdots\rho_{k,N},$$

$$\rho_{i,j} = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & \cos\gamma_{i,j} & & & -\sin\gamma_{i,j} & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & \sin\gamma_{i,j} & & & \cos\gamma_{i,j} & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{bmatrix}_{N\times N},$$

$$\mathbf{U}_{N\times K} = \begin{bmatrix} \delta_1 & & \\ & \ddots & \\ & & \delta_K \\ \mathbf{0}_{N-K\times K} & & \end{bmatrix}$$

For future reference, let the following transformations denote the change of coordinate system defined above:

$$\mathbf{H} = f_{\mathbf{H}}(\gamma, \delta),$$

$$\gamma = f_{\gamma}(\mathbf{H}, \delta).$$

Note that $\mathcal{T}^{\mathrm{O}}$ is bounded and rotation normalization is therefore not necessary. However, in contrast to permutation and reflection, rotations are continuous functions and do not lead to multimodal sampling or posterior distributions. The cost and benefit analysis of not breaking rotation invariance is out of the scope of this paper.

2.3.3 Breaking scale invariance

There are two candidate parameters for breaking scale invariance, $\mathbf{H}$ and $\mathbf{Q}$, leading to what are respectively known as **centered** and **non-centered** scale parameterizations (Frühwirth-Schnatter, 2004):

$$\begin{aligned} \Theta^{\mathbf{Q}_{\mathcal{I}}} &= \{\theta \in \Theta | \, \mathbf{Q} = \mathcal{I}\} \\ \Theta^{\mathbf{H}_1} &= \{\theta \in \Theta | \; K \text{ columns of } \mathbf{H} \text{ have an element set to } 1\} \end{aligned}$$

Note that the centered scale parameterization can be generalized in two ways. First, one can break scale invariance by setting the diagonal elements to any value. For example,

$$\Theta^{\mathbf{Q}_k} = \{\theta \in \Theta | \, \mathbf{Q}_{kk} = k\}$$

would break rotation, scale and permutation invariance. From the discussion above, recall that this normalization does not provide global identification because the model is locally unidentified in the parameter subspace where $\mathbf{F} = \mathbf{0}$ and thus fails to break rotation invariance in that subspace.

Second, the off-diagonal elements of $\mathbf{Q}$ play no role in breaking scale invariance. For example,

$$\Theta^{\mathbf{Q}_{corr}} = \{\theta \in \Theta | \, \mathbf{Q} \text{ is a correlation matrix}\}$$

breaks scale invariance and provides global identification.

Breaking scale invariance through $\Theta^{\mathbf{H}_1}$ would also break rotation invariance, except on some parameter subspace as I discussed above. In polar coordinates, one can consider breaking scale invariance with

$$\Theta^{\mathbf{H}_\delta} = \{\theta \in \Theta | \, \delta_k = 1, \, k = 1, \ldots, K\} .$$

This normalization preserves rotation, permutation and reflection invariance. It also provides global identification.

2.3.4 Breaking permutation invariance

Weak permutation identification occurs in LSSMs when some factors are too similar to one another. Difficulties arise if the corresponding rows of $\mathbf{H}$, diagonal elements of $\mathbf{F}$ and diagonal elements of $\mathbf{Q}$ are too close pairwise. A set of permutation normalizations providing global identification in polar coordinates has the following form:

$$\begin{array}{lllll}
\alpha_1 f_1(\gamma_{1,1}-\gamma_{2,1}) + \ldots + & \alpha_{N-1} f_{N-1}(\gamma_{1,N-1}-\gamma_{2,N-1}) + & \alpha_N f_N(\mathbf{F}_{1,1}-\mathbf{F}_{2,2}) + & \alpha_{N+1} f_{N+1}(\mathbf{Q}_{1,1}-\mathbf{Q}_{2,2}) & > 0 \\
\alpha_1 f_1(\gamma_{2,1}-\gamma_{3,1}) + \ldots + & \alpha_{N-1} f_{N-1}(\gamma_{2,N-1}-\gamma_{3,N-1}) + & \alpha_N f_N(\mathbf{F}_{2,2}-\mathbf{F}_{3,3}) + & \alpha_{N+1} f_{N+1}(\mathbf{Q}_{2,2}-\mathbf{Q}_{3,3}) & > 0 \\
\vdots & & & & \\
\alpha_1 f_1(\gamma_{K-1,1}-\gamma_{K,1}) + \ldots + & \alpha_{N-1} f_{N-1}(\gamma_{K-1,N-1}-\gamma_{K,N-1}) + & \alpha_N f_N(\mathbf{F}_{K-1,K-1}-\mathbf{F}_{K,K}) + & \alpha_{N+1} f_{N+1}(\mathbf{Q}_{K-1,K-1}-\mathbf{Q}_{K,K}) & > 0
\end{array}$$

for set of odd bijections $\{f_1, \ldots, f_{N+1}\}$ on $\mathbb{R}$ and a vector $\alpha = (\alpha_1, \ldots, \alpha_{N+1})'$ in the simplex of $\mathbb{R}^{N+1}$.

Alternatively, in cartesian coordinates, a set of normalizations providing global identification has the form

$$\begin{array}{lllll}
\alpha_1 f_1(\mathbf{H}_{1,1} - \mathbf{H}_{2,1}) + \ldots + & \alpha_N f_N(\mathbf{H}_{1,N} - \mathbf{H}_{2,N}) + & \alpha_{N+1} f_{N+1}(\mathbf{F}_{1,1} - \mathbf{F}_{2,2}) & > 0 \\
\alpha_1 f_1(\mathbf{H}_{2,1} - \mathbf{H}_{3,1}) + \ldots + & \alpha_N f_N(\mathbf{H}_{2,N} - \mathbf{H}_{3,N}) + & \alpha_{N+1} f_{N+1}(\mathbf{F}_{2,2} - \mathbf{F}_{3,3}) & > 0 \\
\vdots & & & \\
\alpha_1 f_1(\mathbf{H}_{K-1,1} - \mathbf{H}_{K,1}) + \ldots + & \alpha_N f_N(\mathbf{H}_{K-1,N} - \mathbf{H}_{K,N}) + & \alpha_{N+1} f_{N+1}(\mathbf{F}_{K-1,K-1} - \mathbf{F}_{K,K}) & > 0
\end{array}$$

with $\{f_1, \ldots, f_{N+1}\}$ and α defined as above.

2.3.5 Breaking reflection invariance

Weak reflection identification concerns arise if any row of $\mathbf{H}$ is close to being a vector of zeros, which would make the information matrix close to being singular. Weak reflection identification issues also arise when any diagonal element of $\mathbf{Q}$ is close to zero and global identification is ensured if this subspace is excluded, *i.e.* if $\mathbf{Q}_{k,k} > 0$ for $k = 1, \ldots, K$. In cartesian coordinates, some reflection normalizations providing global identification are of the form

$$\begin{array}{ccccccccc}
\alpha_{1,1} f_{1,1}(\mathbf{H}_{1,1}) & + & \ldots & + & \alpha_{1,N} f_{1,N}(\mathbf{H}_{1,N}) & > & 0 \\
\alpha_{2,1} f_{2,1}(\mathbf{H}_{2,1}) & + & \ldots & + & \alpha_{2,N} f_{2,N}(\mathbf{H}_{2,N}) & > & 0 \\
 & & \vdots & & & & \\
\alpha_{K,1} f_{K,1}(\mathbf{H}_{K,1}) & + & \ldots & + & \alpha_{K,N} f_{K,N}(\mathbf{H}_{K,N}) & > & 0,
\end{array}$$

for any set of odd bijections $\{f_{1,1}, \ldots, f_{K,N}\}$ on $\mathbb{R}$ and vectors $\alpha_k = (\alpha_{k,1}, \ldots, \alpha_{k,N})'$ in the simplex of $\mathbb{R}^N$. In polar coordinates, one could break invariance with respect to reflections through

$$\begin{array}{ccccccccc}
\beta_1 & < & \alpha_{1,1}\gamma_{1,1} & + & \ldots & + & \alpha_{1,N-1}\gamma_{1,N-1} & < & \beta_1 + \pi \\
\beta_2 & < & \alpha_{2,1}\gamma_{2,1} & + & \ldots & + & \alpha_{2,N-1}\gamma_{2,N-1} & < & \beta_2 + \pi \\
\vdots & & & & \vdots & & & & \\
\beta_K & < & \alpha_{K,1}\gamma_{K,1} & + & \ldots & + & \alpha_{K,N-1}\gamma_{K,N-1} & < & \beta_K + \pi,
\end{array}$$

for any vectors $\alpha_k = (\alpha_{k,1}, \ldots, \alpha_{k,N-1})'$ in the simplex of $\mathbb{R}^{N-1}$. and $\beta' \in [0, \pi)^K$.

2.3.6 Root cancelation in the ARMA representation

Weak identification has an interesting interpretation in terms of root cancelation or redundant parameter issues. It is well-known that root cancelation can make parameter point estimators unreliable (Box and Jenkins, 1976). I first show that weak reflection identification in a one-factor LSSM implies root cancelation in its ARMA

representation. Because this particular LSSM is not a common representation of ARMA processes, I next show that root cancelation in an ARMA(1,1) process implies weak reflection identification for Aoki's canonical LSSM representation of the stochastic process.

The 1-factor LSSM for an univariate process

$$\begin{aligned} \xi_t &= F\xi_{t-1} + v_t, \qquad v_t \sim \mathcal{N}(0,1), \\ y_t &= B + H\xi_t + w_t, \qquad w_t \sim \mathcal{N}(0,R), \end{aligned}$$

has the ARMA(1,1) representation

$$y_t = \alpha + \rho y_{t-1} + \epsilon_t + \theta\epsilon_{t-1},$$

where

$$\begin{aligned} \theta &= -\frac{1 + F^2 + H^2/R^2 \pm \sqrt{(1 + F^2 + H^2/R^2)^2 - 4F^2}}{2F}, \\ \rho &= F. \end{aligned}$$

The factor reflection is weakly identified when the pseudo-true H is close to being equal to 0, which is also where the invertible moving-average root cancels out the autoregressive root as

$$\lim_{H \to 0} (\theta(H) + \rho) = 0.$$

This is not the most common LSSM representation of an ARMA(1,1). For example, Aoki (1987) proposes (See also Brockwell and Davis, 1991; Hamilton, 1994) the representation

$$\xi_t = \begin{bmatrix} \rho & 0 \\ 1 & 0 \end{bmatrix} \xi_{t-1} + \begin{bmatrix} v_{1,t} \\ 0 \end{bmatrix}, \tag{2.7}$$

$$y_t = \alpha + \begin{bmatrix} 1 & \theta \end{bmatrix} \xi_t. \tag{2.8}$$

Setting one factor loading to 1 in (2.8) breaks reflection invariance but do not provide global identification because the model is locally unidentified on the line $\rho = -\theta$. This is easily seen by substituting (2.7) into (2.8):

$$y_t = \alpha + (\rho + \theta)\xi_{1,t-1} + v_{1,t}.$$

Root cancelation occurs when the pseudo-true sum $H = \theta + \rho$ is close to being equal to 0, which is also where weak reflection identification issues arise. Aoki's canonical LSSM representation does not provide global identification. However, there exist other LSSM representations (Brockwell and Davis, 1991) of ARMA processes and some might have better finite sample properties than others. This investigation is out of the scope of this paper.

2.4 Prior Distributions

In this section, I propose permutation- and reflection-invariant prior distributions for the parameters of the LSSM (2.2-2.3). For finite mixture distributions, Geweke (2007, p. 3537) argues that "If the state labels have no substantive interpretation, then the prior density must also be permutation invariant." More generally, prior information should reflect the invariance property of the likelihood function and specifying prior beliefs on quantities that have no substantive interpretation is, at best, conceptually difficult to justify. Moreover, inference might be sensitive to prior specification if priors are informative with respect to reflection or permutation and some parameters are weakly identified.

I propose invariant conditionally conjugate priors when they are available. Any prior on $\mathbf{B}$ and $\mathbf{R}$ is permutation- and reflection-invariant. A normal prior on $\mathbf{B}$ is conditionally conjugate, as is an inverse Wishart on $\mathbf{R}$.

2.4.1 Permutation- and reflection-invariant priors

There are many ways to designing invariant priors, as all one needs to do is ensure that no information is provided with respect either reflections or permutations. The conceptually simplest approach is to specify arbitrary prior distributions and consider the equiprobable mixture of these priors over all possible permutation and reflection combinations.

Alternative approaches require some analysis in order to see how permutation or reflection affects each element of each parameter. Reparameterization sometimes helps in this analysis. Some parameters are naturally reflection invariant, *e.g.* $\mathbf{Q}_{kk}$ or $\mathbf{F}_{kk}$, and permutation invariance is obtained by any exchangeable prior distribution on the diagonal elements of $\mathbf{Q}$ or $\mathbf{F}$[4]. An exchangeable normal distribution has the form $\mathcal{N}(\mu\iota, \sigma^2((1-\rho)\mathcal{I} + \rho\iota\iota'))$. As another special case, i.i.d. univariate priors are permutation invariant. Priors that are symmetric with respect to 0 are reflection invariant. They are equivalently specified as priors on the absolute values of the parameters.

2.4.2 Normalization, parameterization, conditional conjugacy and prior information

Permutation- and reflection-invariant, proper priors provide no information with respect to permutation and reflection, but are informative in other dimensions. When

[4]This might sound tautological, as an exchangeable distribution defined as a permutation invariant distribution. However, permutations of the parameters need not correspond to permutations of the factors.

computational or other considerations leads one to specifying conditionally conjugate priors on the model's parameters, normalization and parameterization can have unexpected consequences for the resulting inference.

As an example, consider how scale normalization affects inference with conditionally conjugate priors for a simple LSSM with $N = K = 1$. Under the centered scale normalization $\Theta^{\mathbf{Q}_I}$, a zero-mean normal prior on factor loadings, $\mathbf{H} \sim \mathcal{N}(0, \sigma^2)$, is conditionally conjugate. This distributional assumption implies that $\mathbf{H}^2 \sim \mathcal{G}(0.5, 2\sigma)$. By scale invariance, this priors is equivalent to $\mathbf{Q} \sim \mathcal{G}(0.5, 2\sigma)$ under the non-centered scale normalization $\Theta^{\mathbf{H}_\delta}$, which is not conditionally conjugate. The standard conditionally conjugate prior for variances is an inverse Gamma distribution, which attributes much less weight to neighborhoods of 0 than a Gamma.

While the information matrix is singular at $\mathbf{Q} = 0$ (or equivalently $\mathbf{H} = 0$), it should be emphasized that the likelihood function and its first derivative with respect to $\mathbf{Q}$ are bounded. The prior's limiting behavior toward the singularity subspace can therefore have a strong influence on that of the posterior: if the prior and its first derivative go to zero, so do the posterior and its first derivative.

In general, and a fortiori in forecasting applications, no reasonable parameter value should be excluded. It might well be the case that a point in the singularity subspace provides a good description of the data. Conditionally conjugate priors under the centered scale normalization seem to be less informative about the singularity subspace than under the non-centered scale normalization. Frühwirth-Schnatter and Wagner (2008) investigate the role of scale parameterization for model selection.

2.4.3 Priors for F, ξ_1 and Q

Normal priors on F and ξ_1 are conditionally conjugate in this model. The diagonal elements of F are naturally reflection invariant, and exchangeable priors ensure permutation invariance. Off-diagonal elements require zero-mean exchangeable priors in order to ensure permutation and reflection invariance. With regard to ξ_1, zero-mean, exchangeable normal priors are permutation- and reflection-invariant.

Conditionally conjugate priors are available for Q. For example, an inverse Wishart prior distribution with scale parameter proportional to the identity matrix, $\mathbf{Q} \sim \mathcal{IW}(\nu, \alpha\mathcal{I})$, is permutation- and reflection-invariant.

2.4.4 Priors for γ

Normal priors on H are conditionally conjugate in this model. But my rotation normalization to the subspace of row-orthogonal factor loading matrices is parameterized through $K(K-1)/2$ rotation angles and my scale normalization sets the K row lengths to begin equal to one. Because permutation and reflection only change the direction and orientation of factor loadings, uniform priors on $[0, 2\pi)$ for each angle $\gamma_{k,n}$ ensure permutation and reflection invariance.

2.5 Posterior Simulation

This section describes posterior simulation for the LSSM (2.2-2.3). I propose a Metropolis-within-Gibbs sampler. In this sampler, parameters without standard conditional posteriors are drawn with the factors as a single block. Next, I propose an extension of Frühwirth-Schnatter's (2001) random permutation sampler to LSSMs and I and discusses the implementation of permutation and reflection normalizations.

2.5.1 Posterior simulator

Defining $\xi = \xi_{t=2:T}$, the Metropolis-Hastings update of the chain consists of the following cyle of parameter and state updates:

Given the state of the Markov chain at iteration $(m-1)$,

1. Generate $\mathbf{B}^{(m)} \sim p(\mathbf{B}|y, \gamma^{(m-1)}, \mathbf{R}^{(m-1)}, \mathbf{F}^{(m-1)}, \mathbf{Q}^{(m-1)}, \xi_1^{(m-1)}, \xi^{(m-1)})$
2. Generate $\mathbf{Q}^* \sim p(\mathbf{Q}|y, , \mathbf{B}^{(m)}, \gamma^{(m-1)}, \mathbf{R}^{(m-1)}, \mathbf{F}^{(m-1)}, \xi_1^{(m-1)}, \xi^{(m-1)})$
3. Generate $\mathbf{R}^{(m)} \sim p(\mathbf{R}|y, \mathbf{B}^{(m)}, \gamma^{(m-1)}, \mathbf{F}^{(m-1)}, \mathbf{Q}^*, \xi_1^{(m-1)}, \xi^{(m-1)})$
4. Generate $\mathbf{F}^* \sim p(\mathbf{F}|y, \mathbf{B}^{(m)}, \gamma^{(m-1)}, \mathbf{R}^{(m)}, \mathbf{Q}^*, \xi_1^{(m-1)}, \xi^{(m-1)})$
5. Generate ${\xi_1}^* \sim p(\xi_1|y, \mathbf{B}^{(m)}, \mathbf{R}^{(m)}, \mathbf{F}^*, \mathbf{Q}^*, \xi^{(m-1)})$
6. Generate $\gamma', \xi' \sim q(\gamma, \xi|y, \mathbf{B}^{(m)}, \mathbf{R}^{(m)}, \mathbf{F}^*, \mathbf{Q}^*, \xi_1^*)$
7. Take

$$(\gamma^*, \xi^*) = \begin{cases} (\gamma', \xi') & \text{with probability } \rho \\ \left(\gamma^{(m-1)}, \xi^{(m-1)}\right) & \text{with probability } 1-\rho \end{cases}$$

where

$$\rho = \min\left\{ \frac{p\left(\gamma', \xi' \middle| \mathbf{y}, \mathbf{B}^{(m)}, \mathbf{Q}^*, \mathbf{R}^{(m)}, \mathbf{F}^*, \xi_1^*\right)}{p\left(\gamma^{(m-1)}, \xi^{(m-1)} \middle| \mathbf{y}, \mathbf{B}^{(m)}, \mathbf{Q}^*, \mathbf{R}^{(m)}, \mathbf{F}^*, \xi_1^*\right)} \frac{p\left(\xi^{(m-1)} \middle| \mathbf{y}, \gamma^{(m-1)}, \mathbf{B}^{(m)}, \mathbf{Q}^*, \mathbf{R}^{(m)}, \mathbf{F}^*, \xi_1^*\right)}{p\left(\xi' \middle| \mathbf{y}, \gamma', \mathbf{B}^{(m)}, \mathbf{Q}^*, \mathbf{R}^{(m)}, \mathbf{F}^*, \xi_1^*\right)} \right\}$$

8. Generate $\mathbf{S}$ uniformly over the $K!$ reflection matrices
9. Generate $\mathbf{P}$ uniformly over the 2^K permutation matrices
10. Take

$$\begin{aligned}
\xi_1^{(m)} &= \mathbf{S}\mathbf{P}\xi_1^* \\
\xi^{(m)} &= \mathbf{S}\mathbf{P}\xi^* \\
\gamma^{(m)} &= f_\gamma\left(\mathbf{S}\mathbf{P} f_{\mathbf{H}}\left(\gamma^*\right)\right) \\
\mathbf{F}^{(m)} &= \mathbf{P}\mathbf{F}^*\mathbf{P}' \\
\mathbf{Q}^{(m)} &= \mathbf{S}\mathbf{P}\mathbf{Q}^*\mathbf{P}'\mathbf{S}'.
\end{aligned}$$

The full conditional posteriors of γ is not a standard distribution and this parameter is drawn jointly with the latent factors as a single block via the random-walk Metropolis-Hastings steps 6 and 7. Steps 8 to 10 constitute my mixture sampler. I detail both below.

2.5.2 Metropolis-Hastings-within-Gibbs

All parameters but γ admit conditionally conjugate priors and have standard conditional posteriors. I use a Gaussian random-walk Metropolis-Hastings step to draw this parameter jointly with the latent factors. Defining

$$\Phi \equiv \{\mathbf{B}, \mathbf{Q}, \mathbf{R}, \mathbf{F}, \xi_1\},$$

the proposal is

$$q\left(\gamma', \xi' \middle| \mathbf{y}, \gamma, \Phi, \Sigma_\gamma\right) = p\left(\xi' \middle| \mathbf{y}, \gamma', \Phi\right) \phi\left(\gamma' \middle| \gamma, \Sigma_\gamma\right), \tag{2.9}$$

where $p\left(\xi' \middle| \mathbf{y}, \gamma', \Phi\right)$ can be computed exactly using an algorithm developed independently by Carter and Kohn (1994) and Frühwirth-Schnatter (1994), and used by Kim and Nelson (1998), among others. The covariance matrix Σ_γ is to be specified by the investigator (See Robert and Casella, 2004, for a discussion).

Note that the joint proposal (2.9) does not depend on ξ, the current state of the factors. The Markov chain is less autocorrelated and therefore more efficient than if

it did. Because $p\left(\xi'\middle|\mathbf{y},\gamma',\Phi\right)$ is exact, the proposal can be close to its target for a relatively large Σ_γ.

In theory, one could simulate all parameters in a single block with the proposal

$$q\left(\gamma',\Phi',\xi'\middle|\mathbf{y},\gamma,\Phi,\Sigma_\gamma,\Sigma_\Phi\right) = p\left(\xi'\middle|\mathbf{y},\gamma',\Phi'\right)\phi\left(\gamma'\middle|\mathbf{Q},\Sigma_\gamma\right)\phi\left(\Phi'\middle|\Phi,\Sigma_\Phi\right).$$

However, the dimension of the parameter space, $K(N+K+2)+N(N+1)/2$, and the multimodality of the posterior would make the calibration of the random walk (the specification Σ_γ and Σ_Φ) challenging. In my experience, the efficiency costs associated with a inadequately calibrated random-walk proposal outweigh the benefits of single-move sampling (See Chib and Ergashev, 2008, for an alternative approach.)

2.5.3 Mixture sampler

From the discussion in the first section, whether normalizing the parameter space is desirable in the Bayesian framework depends on interpretational considerations. For instance, there is no need for normalization if one uses a LSSM as a flexible parametric model and latent variables are not of direct interest, as would be the case in a forecasting exercise. One would then consider the multimodal, permutation- and reflection-invariant posterior distributions.

Multimodal posteriors constitute a computational challenge for which tempering methods have proved useful (Robert and Casella, 2004, p. 540). However, the mixture sampler I describe in this paper takes advantage of the symmetry of the joint posterior in order to efficiently explore all of its $K!2^K$ lobes with high numerical efficiency. It generalizes Frühwirth-Schnatter's (2001) random permutation sampler in two ways.

First, (2.5-2.6) reveals that permutation invariance in LSSMs does not correspond to simple permutations of parameter indices. My mixture sampler deals with more general parameter transformations. Second, it addresses reflection invariance. The invariance property of the mixture sampler follows directly from that of the permutation sampler (See Frühwirth-Schnatter, 2001, Appendix, for a proof).

Implementation involves little programming effort. The K-dimensional diagonal reflection matrix $\mathbf{S}$ of step 9 has elements equal to 1 or -1 with probability 0.5. In step 10, one generates a random permutation K-dimensional vector $\mathbf{p}$ containing indices $\{1, \ldots, K\}$. The permutation matrix $\mathbf{P}$ is generated by placing the rows of an identify matrix in the order given by $\mathbf{p}$. If the joint posterior is invariant with respect to reflection and permutation invariance (e.i. of both the likelihood function and the priors are invariant with respect to reflection and permutation), the proposals in step 10 are accepted with probability one. Otherwise, one computes the acceptance probability.

Posterior symmetry as the other important implication that any single lobe contains all of the relevant information about the model. This implies that visiting all lobes is not a necessary condition for the posterior simulator to fully capture the informational content of the posterior distribution. Intuitively, because the proposals of the random mixture sampler are accepted with probability one, this device is redundant from a purely inferential point of view. This observation leads Geweke (2007) [Title] to state that "Simple MCMC works" unless [p. 3538] "there are mixing problems beyond those arising from permutation invariance of the posterior distribution." A basic MCMC simulator should reveal the posterior distribution just as efficiently. One could indeed skip steps 8 to 10 of the algorithm presented above and obtain equivalent forecasts.

If it is inferentially redundant, why then would anyone use the random permutation sampler? One answer is a practical one. Assessing the mixing properties of a MCMC sampler is no simple task. In LSSMs, standard methods of assessing convergence must take into account permutation and reflection invariance. For example, methods based on cumulative sums, like Brook's (1998), must consider permutation- and reflection-invariant quantities. Plotting the output of an MCMC sampler is another common way of doing a quick diagnosis of the generated chain. This exercise is complicated by reflection and permutation invariance. For example, for the state-space model (2.2-2.3) with three factors, there are six lobes any element of $\mathbf{B}$ can visit. A trend in the path of a parameter (which could indicate that the effect of initial conditions has not died out) can be difficult to see graphically when the chain keeps switching between lobes. Indeed, this is possibly what led Celeux et al. (2000) [p. 957] to assert that "we consider that almost the entirety of MCMC samplers implemented for mixture models has failed to converge!" The random mixture sampler ensures that all modes are visited. Geweke (2007) proposes to build permuted copies of the parameter vector as a post-simulation step. This approach is inferentially equivalent to the random permutation sampler.

If the interpretation of the components or factors is of direct interest and normalization is thus desirable, one can deterministically map the proposed parameter vector to the parameter sub-space satisfying the normalization. Paralleling Frühwirth-Schnatter's (2001) terminology, I refer to this implementation as the *constrained mixture sampler.* Note that this mapping can be carried out within the posterior simulator or applied as a post-simulation processing of the posterior sample (Stephens, 1997). Because many normalizations provide global identification but each can yield different parameter posterior distribution, one can try several alternatives until a normalization suiting one's

inferential objectives is found. In order to compare normalizations, the investigator can therefore efficiently use the output of an un-normalized posterior sampler.

2.6 Simulations Results

If the model is correctly specified, Bayesian out-of-sample forecasting RMSEs are smaller that those produced by the maximum likelihood method by construction: the Bayesian forecast constitutes the mathematical solution to the inferential problem of optimally updating the statistician's prior information with the data at hand in order to minimize an expected loss function, here the out-of-sample forecasting square error. Both frameworks are asymptotically equivalent, but the characteristics of the model and the nature of the data determine how much improvement the Bayesian approach yields in finite sample. I present Monte Carlo evidence showing that the forecast improvements for LSSMs is related to the weak identification problem described in this paper; the weaker the reflection identification, the larger the improvement.

I simulate artificial data sets from a one-factor representation of the dynamics of $N = 1$ variable observed for $T = 50$ periods,

$$\begin{aligned} \xi_t &= F\xi_{t-1} + v_t, \\ y_t &= B + H'\xi_t + w_t. \end{aligned}$$

I limit my empirical investigation to the impact of weak reflection identification, the nature of the weak permutation identification problem being similar. I set $B = 0$, $R = 1$, $Q = 1$, $F = 0.95$ and $\xi_1 = 0$ and look at the impact of varying H. A smaller factor loading H makes the reflection weak identification problem more severe because the Fisher information matrix is singular at $H = 0$. I consider $H = \{0.005, 0.01, 0.05, 0.1\}$. Note the sample size is irrelevant in itself as one would obtain similar results with a

larger T and smaller H's.

I compute one-period-ahead forecasts for $M = 1000$ artificial data sets. The Bayesian optimal forecast $\hat{y}_{Bayes,T+1}$ under mean root square error loss is the mean of the predictive density. I use a sample of 5000 iterations after a burn in phase of 500 iterations to construct the estimate. The observation error variance R is a priori $\mathcal{IG}(1,1)$. All other parameters have vagues priors that are centered over the singularity subspace: F, B and H have independent normal priors with mean 0 and variance 10^5.

I use the EM algorithm (Shumway and Stoffer, 1983; Watson and Engle, 1983) in order to find the maximum value of the likelihood. The exit condition is that the absolute difference in subsequent log-likelihood values is less than 0.0001% of its level. The ML forecast is $\hat{y}_{MLE,T+1} = \hat{B}_{MLE} + \hat{H}'_{MLE}\hat{\xi}_{T+1}$, where $\hat{\xi}_{T+1} = \mathbf{E}\left[\xi_{T+1}|y_T\right]$.

One usually reports RMSEs as measures of goodness-of-fit, and ratios of RMSEs as relative measures. I define RMSE_i as

$$\text{RMSE}_\text{i} = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}(y_{n,m,T+1} - \hat{y}_{\text{i},n,m,T+1})^2$$

for i={MLE,Bayes} and relative RMSE as $\text{RMSE}_{\text{Bayes}}/\text{RMSE}_{\text{MLE}}$. Ratios of MRSEs are more precisely estimated than individual RMSEs because the same data sets are used for Bayesian and MLE forecasts so that errors are correlated. I assume that errors are jointly Gaussian and compute parametric Monte Carlo standard errors for relative RMSEs.

Table 2.2 presents the relative RMSEs of out-of-sample forecasts ($\text{RMSE}_{\text{Bayes}}/\text{RMSE}_{\text{MLE}}$). For a significant proportion of samples, the EM algorithm converges to an AR(1) model. This would not be of particular concern in practice

Table 2.2: Out-of-sample relative performances and weak identification

H	All	ARMA(1,1)	AR(1)
0.005	0,795	0,817	0,655
	(0,020)	(0,022)	(0,048)
	M=1000	M=898	M=102
0.010	0,852	0,880	0,737
	(0,024)	(0,027)	(0,050)
	M=1000	M=884	M=116
0.050	0,883	0,919	0,748
	(0,024)	(0,028)	(0,051)
	M=1000	M=871	M=129
0.100	0,961	0,968	0,908
	(0,023)	(0,026)	(0,059)
	M=1000	M=876	M=124

The ratio of root mean square errors ($\mathrm{RMSE}_{\mathrm{Bayes}}/\mathrm{RMSE}_{\mathrm{MLE}}$) of out-of-sample one-period-ahead forecasts for various H, with parametric Monte Carlo standard errors in parentheses. M gives the number of data sets considered: all $M = 1000$ samples in the first column, samples for which the EM algorithm converged to an ARMA(1,1) process in the second column, and samples for which algorithm converged to an AR(1) process in the third column.

but I present these cases separately because a proper prior on R prevents this from happening in the Bayesian framework. Whether all data sets are considered together, or separated according to where the EM algorithm converged, the improvement increases as H approaches 0. Furthermore, the improvement stabilizes when H is large enough and the lobes are well separated.

2.7 Concluding Remarks

Inference for linear state-space models is complicated by a weak identification problem; if latent variables are too similar or if factor loadings are too small, the Fisher information matrix is close to being singular and the factors's reflection and permutation are weakly identified. I argue that a connected normalization providing global parameter identification is more likely to produce unimodal posterior distributions or a maximum-likelihood estimator with unimodal sampling distribution in finite sample, and I propose an observationally unrestrictive normalization of LSSM satisfying these conditions.

However, I stress that unimodal distribution cannot be ensured by an observationally unrestrictive normalization.

When some parameters are weakly identified, the Bayesian framework offers two advantages over the standard ML method. First, it yields better out-of-sample forecasts because it does not rely on biased parameter point estimators. The two approaches are only asymptotically equivalent, and this paper merely presents one setting in which taking into account parameter uncertainty proves useful. This suggests that taking into account parameter uncertainty in the ML framework could also yields benefits.

The second advantage, perhaps surprisingly, is computational. If factor interpretations are of direct interest, then one should compare competing normalizations in order to find one that yields parameter point estimators with good properties. Because the ML estimator's sampling distribution must be obtained by computationally expensive simulation methods, searching for a good normalization is impractical. In contrast, the Bayesian framework allows one to compare normalizations at negligible computational cost.

I leave many questions unanswered. First, whether there are benefits to preserving rotation invariance in LSSMs is an important empirical question. Because a rotation-invariant likelihood function would be smoother, it is possible that it provides significant computational benefits. Second, I argue that conditionally conjugate priors under the non-centered scale normalization might be too informative about the singularity set, but I don't provide any simulation experiment to quantify this problem. It would also be interesting to see how taking parameter uncertainty into account affects forecasts under various misspecification problems. Finally, one could verify

whether other popular LSSM representations of stationary ARMA processes satisfy the identification principle.

2.8 Appendix A - Invariance to linear transformations with correlated errors

This appendix explains how to generalize the results presented in this paper to LSSMs with correlated errors (Anderson and Moore, 1979). Next, it presents those results for the innovation representation of LSSMs.

LSSMs with correlated errors can be represented by the system of equations (2.2-2.3) with

$$\begin{bmatrix} \mathbf{v}_t \\ \mathbf{w}_t \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \mathbf{Q} & \mathbf{C} \\ \mathbf{C}' & \mathbf{R} \end{bmatrix}\right).$$

The likelihood function is invariant with respect to invertible linear transformations of the latent factors; for any invertible $\mathbf{M}$,

$$l(\mathbf{B}, \mathbf{M}'^{-1}\mathbf{H}, \mathbf{R}, \mathbf{M}\mathbf{F}\mathbf{M}^{-1}, \mathbf{M}\mathbf{Q}\mathbf{M}', \mathbf{M}\mathbf{C}, \mathbf{M}\xi_1|y_t) \equiv l(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{F}, \mathbf{Q}, \mathbf{C}, \xi_1|y_t).$$

In order to write the model in one of its popular representation, one parameterizes the covariance matrix as

$$\begin{bmatrix} \mathbf{Q} & \mathbf{C} \\ \mathbf{C}' & \mathbf{R} \end{bmatrix} = \begin{bmatrix} \mathbf{J} \\ \mathbf{G} \end{bmatrix} \begin{bmatrix} \mathbf{J} \\ \mathbf{G} \end{bmatrix}',$$

where $\mathbf{J}$ and $\mathbf{G}$ are respectively $K \times (K+N)$ and $N \times (K+N)$ matrices. In terms of $\mathbf{J}$ and $\mathbf{G}$, one can write the state-space system as

$$\begin{aligned} \xi_{t+1} &= \mathbf{F}\xi_t + \mathbf{J}\mathbf{u}_t, \\ \mathbf{y}_t &= \mathbf{B} + \mathbf{H}'\xi_t + \mathbf{G}\mathbf{u}_t, \end{aligned}$$

where $\mathbf{u}_t$ is a $(K+N)\times 1$ vector of standard normal random variables. For any invertible $\mathbf{M}$,

$$l(\mathbf{B}, \mathbf{M}'^{-1}\mathbf{H}, \mathbf{G}, \mathbf{MFM}^{-1}, \mathbf{MJ}, \mathbf{M}\xi_1|y_t) \equiv l(\mathbf{B}, \mathbf{H}, \mathbf{G}, \mathbf{F}, \mathbf{J}, \xi_1|y_t).$$

For stationary processes, the system has the following alternative innovation representation (Brockwell and Davis, 1991):

$$\begin{aligned} \xi_{t+1} &= \mathbf{F}\xi_t + \mathbf{Z}\mathbf{e}_t, \\ \mathbf{y}_t &= \mathbf{B} + \mathbf{H}'\xi_t + \mathbf{e}_t, \end{aligned}$$

with $\mathbf{e}_t \sim \mathcal{N}\left(\mathbf{0}, \bar{\mathbf{R}}\right)$ and $\mathbf{Z}$ is a $K \times N$ matrix. For any invertible $\mathbf{M}$,

$$l(\mathbf{B}, \mathbf{M}'^{-1}\mathbf{H}, \bar{\mathbf{R}}, \mathbf{MFM}^{-1}, \mathbf{MZ}, \mathbf{M}\xi_1|y_t) \equiv l(\mathbf{B}, \mathbf{H}, \bar{\mathbf{R}}, \mathbf{F}, \mathbf{Z}, \xi_1|y_t).$$

CHAPITRE 3

A BAYESIAN ANALYSIS OF AFFINE TERM STRUCTURE MODELS

Abstract

Dynamic term structure models are no-arbitrage structural economic factor models. In empirical applications, one specifies a statistical model for observational errors in order to accommodate the fact that the economic model imposes equality restrictions on observables that do not exactly hold in practice. Because term structure models involve unidentified structural parameters, they require normalization. This paper investigates the empirical importance of error modeling and normalization for inference for affine term structure models. At the methodological level, I propose and implement a new MCMC algorithm for Gaussian affine term structure models in which latent factors are drawn together with some parameters as a single block.

Comparing two popular approaches to modeling pricing errors, my analysis reveals that residuals from latent factor models have lower cross-correlations and autocorrelations than residuals from models where proxying factors are recovered by inverting the pricing equations. Because the latter models are special cases of the former in which some pricing errors are identically zero, this result implies that restrictions on error variances affect inference for factor dynamics. In order to investigate this issue, I compare latent factor models with homoscedastic and heteroscedastic errors: introducing heteroscedasticity further reduces residual cross-correlations and autocorrelations. While residuals from these independent-error models are correlated, modeling correlated errors increases residual autocorrelations. I use informative priors in order to obtain residuals that are compatible with the error correlation model but have low autocorrelations. I also propose an informative prior distribution for the dispersion of error standard deviations, which allows me to control the level of residual heteroscedasticity.

With respect to normalization, I provide evidence that factors are weakly identified from discount bond prices. This implies that a poor normalization can yield parameter point estimators with undesirable finite-sample properties. In particular, the maximum likelihood estimator can be severely biased and asymptotic confidence intervals unreliable. In contrast, Bayesian inference for pricing errors is valid. I demonstrate that Dai and Singleton's (2000, Journal of Finance) "canonical representation" makes inference particularly sensitive to these problems and I propose alternative normalizations.

3.1 Introduction

The topic of this paper is inference for *Dynamic Term Structure Models* (DTSM) (See Dai and Singleton, 2003, for a review). A DTSM is a factor model for the stochastic discount factor (SDF). Because the final nominal value of risk-free discount bonds is known with certainty, their prices are completely determined by the SDF. The joint specification of the factor physical dynamics and the functional forms of the short rate and the SDF defines a particular DTSM. The model I consider here is a discrete-time version of the Gaussian constant-diffusion essentially-affine DTSM of Duffee (2002), in which the short rate and the log-SDF are affine functions of factors that evolve as a Gaussian first-order vector autoregressive process under both the risk-neutral and physical measures. Under these assumptions, a N-dimensional vector y_t of discount rates at time t is affine in a K-dimensional vector X_t of factors,

$$y_t = \mathbf{A}(\psi) + \mathbf{B}(\psi)' X_t \tag{3.1}$$

$$X_t = X_{t-1} + \kappa^{\mathbb{P}}(\theta^{\mathbb{P}} - X_{t-1}) + \tilde{\epsilon}_t, \qquad \tilde{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \Sigma), \tag{3.2}$$

where $\mathbf{A}(\psi)$ and $\mathbf{B}(\psi)$ are functions of the model's structural parameter vector $\psi \in \Psi$, which includes $\kappa^{\mathbb{P}}$, $\theta^{\mathbb{P}}$ and Σ. From now on, I refer to this model as the *affine term structure model* (ATSM).

DTSMs are increasingly popular in economics and their applications are varied. They are parsimonious and theoretically consistent descriptions of the term structure of interest rates. They are used to improve macroeconomic forecasts (Ang and Piazzesi, 2003), to estimate monetary policy rules (Ang, Dong, and Piazzesi, 2007), to enrich new Keynesian models (Hördahl, Tristani, and Vestin, 2006; Bekaert, Cho, and Moreno, 2006; Dewachter and Lyrio, 2006)), and to estimate structural parameters, such as

preference parameters (Garcia and Luger, 2007).

One can see econometric inference for ATSM's as a sequence of five steps. First, economic theory gives a deterministic structural relationship, here given by equations (3.1-3.2), between the dynamics of the short rate and the term structure of interest rates. Second, to accommodate the fact that the model imposes equality restrictions on observables that do not exactly hold in practice, an observational error model is specified. This paper considers an additive Gaussian observational error vector $\tilde{e}_t$,

$$y_t = \mathbf{A}(\psi) + \mathbf{B}(\psi)' X_t + \tilde{e}_t, \qquad \tilde{e}_t \sim \mathcal{N}(\mathbf{0}, \Omega). \tag{3.3}$$

Then, the model requires normalization because it involves unidentified structural parameters. For example, affine models are invariant with respect to linear transformations of the factors: for any invertible matrix $\mathbf{M}$, there exists a function $g_{\mathbf{M}} : \Psi \to \Psi$ such that equations (3.3) and (3.2) can be equivalently written as

$$\begin{aligned} y_t &= \mathbf{A}(\psi) + \mathbf{B}(\psi)' \mathbf{M}^{-1} \mathbf{M} X_t + \tilde{e}_t, \\ &= \mathbf{A}\left(g_{\mathbf{M}}(\psi)\right) + \mathbf{B}\left(g_{\mathbf{M}}(\psi)\right) Z_t + \tilde{e}_t, \\ \mathbf{M}X_t &= \mathbf{M}X_{t-1} + \mathbf{M}\kappa^{\mathbb{P}}\mathbf{M}^{-1}(\mathbf{M}\theta^{\mathbb{P}} - \mathbf{M}X_{t-1}) + \mathbf{M}\tilde{\epsilon}_t, \\ &= Z_t = Z_{t-1} + \tilde{\kappa}^{\mathbb{P}}(\tilde{\theta}^{\mathbb{P}} - Z_{t-1}) + \tilde{\eta}_t, \qquad \tilde{\eta}_t \sim \mathcal{N}(\mathbf{0}, \tilde{\Sigma}), \end{aligned}$$

where $Z_t = \mathbf{M}X_t$, $\tilde{\kappa}^{\mathbb{P}} = \mathbf{M}\kappa^{\mathbb{P}}\mathbf{M}^{-1}$, $\tilde{\Sigma} = \mathbf{M}\Sigma\mathbf{M}'$ and $\tilde{\theta}^{\mathbb{P}} = \mathbf{M}\theta^{\mathbb{P}}$. Parameter vectors ϕ and $g_{\mathbf{M}}(\psi)$ are thus observationally equivalent and the unnormalized model is globally unidentified.

Fourth, an inferential method is selected to retrieve the information about the

economic model contained in the data. Finally, one chooses a loss function and makes decisions, in the form of estimates, tests or forecasts. This paper investigates the empirical importance of the error modeling and normalization steps for inference for ATSMs.

Error modeling

There are at least two distinct approaches to specifying errors in observations. The most popular approach in the macroeconomics and financial economics literatures[1] follows Chen and Scott (1993) and assumes that a number of discount rates equal to the number of factors is observed without error. This choice, where the error covariance matrix Ω has rank $N-K$, is justified on computational grounds, as the factors can then be recovered by inverting the pricing equation. I refer to this approach as the ***proxy*** modeling approach because it uses a deterministic function of observables to proxy latent factors. The second, ***latent-factor*** modeling approach follows Chen and Scott (1995) and assumes that all discount rates are observed with error. The errors covariance matrix Ω has rank N under this approach, which is popular in the empirical finance literature[2]. There are many theoretical reasons to prefer modeling errors on all yields: there is no need to make an arbitrary choice of which rates are observed without error; discount rates are often not observed but rather approximated from quoted coupon bond yields; and dynamics are more flexible.

To the best of my knowledge, these error modeling approaches have not been compared empirically. In this paper, I address the trade-off between model flexibility and computational ease. I find that the latent-factor approach yields residuals that are

[1]Some examples are: Dai and Singleton (2002); Ang and Piazzesi (2003); Duffee (2002); Cheridito et al. (2003); Duarte (2003).

[2]Examples are: Jegadeesh and Pennacchi (1996); Ball and Torous (1996); Babbs and Nowman (1999); Geyer and Pichler (1999); Lamoureux and Witte (2002); Ang et al. (2007).

significantly less correlated and autocorrelated than residuals from the proxy approach. This highlights the role of error modeling in the decomposition of observable dynamics into common and idiosyncratic components: restrictions on the latter affects inference for the former. In addition, a relatively higher pricing error on the short rate suggests that it might not be the best of proxying factors. Because DTSMs build on assumptions about short-rate dynamics, assuming that the short rate is observed without error is common practice; my results indicate that this particular modeling choice is not inferentially innocuous.

One often looks at residuals, as I do in this paper, for evidence of misspecification of the economic model. Finding such misspecification may lead the econometrician to more general error covariance specifications. On the other hand, an arbitrary covariance matrix may lead to over-parameterization. Because latent-factor models essentially decompose the dynamics of the observables into common and idiosyncratic components, error covariance modeling also allows the econometrician to specify which characteristics of the observables the common latent factors should capture. For example, if errors are modeled as i.i.d. random variables, factors are required to capture the heteroscedasticity, correlation and persistence of observables. In contrast, if errors are independent but heteroscedastic, factors might be better able to capture correlations and persistence. Factors might yet better capture persistence if errors are correlated.

The proxy approach is a special case of the latent-factor approach corresponding to a rather strong restriction on the covariance matrix Ω in which elements are equal to zero. Other restrictions are also likely to affect the factor-error decomposition. Imposing homoscedasticity and independence are examples of such restrictions. In order to consider these restrictions individually, I factorize the covariance matrix

into a correlation matrix and a diagonal matrix of precisions (the inverse of errors variances). In this paper, I propose priors that operationalize *soft* restrictions on the correlation and precision matrices. My empirical results show that using these priors for modeling mildly heteroscedastic and cross-correlated errors yields residuals with lower autocorrelations than strict homoscedastic or independent error models.

Normalization

I consider likelihood-based inference methods, which rely on a parametric statistical model.

Definition 8 *A **parametric statistical model** is a triplet $(\mathcal{Y}, \mathcal{F}, \Psi)$, where $\mathcal{Y}$ is the sample space, $\mathcal{F} \equiv \{f(y|\psi) | y \in \mathcal{Y}, \psi \in \Psi\}$ is a set of parametric probability density functions on $\mathcal{Y}$, and Ψ is the parameter set. The **likelihood function** of the model is the function $l(\psi|y) = f(y|\psi)$.*

The likelihood of ATSMs is invariant with respect to parameter transformations corresponding to affine transformations of the factors.

Definition 9 *A function $f : \Psi \to \mathbb{R}$ is **invariant with respect a bijective transformation** $T : \Psi \to \Psi$ if $f(T(\psi)) = f(\psi)$.*

If $l(\psi|y)$ is invariant with respect to T on Ψ for all $y \in \mathcal{Y}$ then we say that $T(\psi)$ and ψ are **observationally equivalent**. We will also say that f is invariant with respect a set of bijective transformations $\mathcal{T}(\Psi)$ if it is invariant with respect to all of its elements. The notation $\mathcal{T}(\Psi)$ makes dependence on the set Ψ explicit: $\mathcal{T}(\Psi)$ is a set of bijections on Ψ. For example, for $\Psi' \subseteq \Psi$, $\mathcal{T}(\Psi') = \{T : \Psi' \to \Psi' | T \in \mathcal{T}(\Psi)\}$. I will omit this dependence and write $\mathcal{T}$ when this causes no confusion. The following example illustrates this definition.

Example 16 *Consider*

$$y \;=\; bx + e, \quad x \sim \mathcal{N}(0, \sigma^2), \quad e \sim \mathcal{N}(0, 1),$$

for $(b, \sigma^2) \in \Psi = \mathbb{R} \times (0, \infty)$. *In that case the likelihood function satisfies* $l(b, \sigma^2 | \, y) = l\left(|Db|, \sigma^2/D^2 \middle| \, y\right)$ *for any* $D \neq 0$, *and it is therefore invariant with respect to*

$$\begin{aligned}
\mathcal{T}_D(\Psi) &= \left\{T : \Psi \to \Psi \middle| \, T_D(b, \sigma^2) = \left(Db, \sigma^2/D^2\right), \, D > 0\right\} \\
\mathcal{T}_S(\Psi) &= \left\{T : \Psi \to \Psi \middle| \, T_S(b, \sigma^2) = \left(Sb, \sigma^2\right), \, |S| = 1\right\} \\
\mathcal{T}_{SD}(\Psi) &= \left\{T : \Psi \to \Psi \middle| \, T_{SD}(b, \sigma^2) = \left(SDb, \sigma^2/D^2\right), \, D > 0, |S| = 1\right\} \\
&= \left\{T : \Psi \to \Psi \middle| \, T(\psi) = T_S\left(T_D(\psi)\right)\right\}
\end{aligned}$$

The parameters b *and* σ^2 *enter the likelihood function as the product* $b^2\sigma^2$. *Transformations in* $\mathcal{T}_D$ *correspond to changing the scale of the unobserved factor* x *and reflect the fact that* $(Db)^2 \frac{\sigma^2}{D^2} = b^2\sigma^2$ *for* $D > 0$. *Transformations in* $\mathcal{T}_S$ *correspond to reflections of* x *across* $x = 0$, *which change its sign.*

That the likelihood of ATSMs is invariant with respect to parameter transformations corresponding to affine transformations of the factors has the following meaning: If $\psi \in \Psi$ denotes the K-factor ATSM's parameter vector, y the observed panel of discount rates and $f(y | \, \psi, X)$ the probability density function of y conditional on a panel of factors X, then for any K-dimensional vector $\mathbf{t}$ and any invertible $K \times K$ matrix $\mathbf{M}$, there exists a bijection $T_{\mathbf{tM}}(\psi) : \Psi \to \Psi$ such that $f(y | \, T_{\mathbf{tM}}(\psi), \mathbf{M}(X - \mathbf{t})) = f(y | \, \psi, X)$.

In general, one addresses transformation invariance by normalizing the model.

Definition 10 *A normalization is a parameter subset* $\Psi^N \subseteq \Psi$.

A normalization $\Psi^N \subseteq \Psi$ breaks invariance with respect to a set of bijections $\mathcal{T}(\Psi)$ if $\mathcal{T}\left(\Psi^N\right) = \mathcal{T}_{\mathcal{I}}$, where

$$\mathcal{T}_{\mathcal{I}} = \left\{T : \Psi \to \Psi \middle| T(\psi) = \psi\right\} \tag{3.4}$$

is a singleton: the identity transformation. Note that $\Psi^N \subseteq \Psi \Rightarrow \mathcal{T}\left(\Psi^N\right) \subseteq \mathcal{T}(\Psi)$. Breaking invariance with respect to a set of bijective transformations $\mathcal{T}(\Psi)$ is thus considering a parameter subset $\Psi^N \subseteq \Psi$ that is small enough to ensure that the only invariant bijection on that subset is the identity transformation, $\mathcal{T}(\Psi^N) = \left\{T : \Psi^N \to \Psi^N \middle| T(\psi) = \psi\right\}$.

Example 17 (Example 1, continued) *Consider the normalizations*

$$\begin{aligned}\Psi^b &= \{\psi \in \Psi | \, b > 0\} \\ \Psi^{\sigma_1^2} &= \left\{\psi \in \Psi \middle| \, \sigma^2 = 1\right\}.\end{aligned}$$

Scaling and reflection transformations on these normalizations are:

$$\begin{aligned}\mathcal{T}_S\left(\Psi^b\right) &= \left\{T : \Psi^b \to \Psi^b \middle| \, T_S(b, \sigma^2) = \left(Sb, \sigma^2\right), \, S = 1\right\} = \mathcal{T}_{\mathcal{I}}, \\ \mathcal{T}_S\left(\Psi^{\sigma_1^2}\right) &= \mathcal{T}_S(\Psi), \\ \mathcal{T}_D\left(\Psi^b\right) &= \mathcal{T}_D(\Psi), \\ \mathcal{T}_D\left(\Psi^{\sigma_1^2}\right) &= \left\{T : \Psi^{\sigma_1^2} \to \Psi^{\sigma_1^2} \middle| \, T_D(b, \sigma^2) = \left(Db, \sigma^2/D^2\right), \, D = 1\right\} = \mathcal{T}_{\mathcal{I}}, \\ \mathcal{T}_{SD}\left(\Psi^b \cap \Psi^{\sigma_1^2}\right) &= \Big\{T : \Psi^b \cap \Psi^{\sigma_1^2} \to \Psi^b \cap \Psi^{\sigma_1^2} \Big| \\ &\qquad T_S(b, \sigma^2) = \left(SDb, \sigma^2/D^2\right), \, S = 1, D = 1\Big\} = \mathcal{T}_{\mathcal{I}}.\end{aligned}$$

Normalization Ψ^b breaks invariance with respect to reflections of factors because T_S is not a bijection on Ψ^b for $S \neq 1$. Similarly, $\Psi^{\sigma_1^2}$ breaks invariance with respect to $\mathcal{T}_D$ because T_D is not a bijection on $\Psi^{\sigma_1^2}$ for $D \neq 1$. Thus, $\Psi^b \cap \Psi^{\sigma_1^2}$ breaks invariance with respect to $\mathcal{T}_{SD}$.

One could consider normalization of arbitrary form, but it is natural to restrict attention to intersections of half hyper-spaces and hyper-planes,

$$\Psi^N = \bigcap_{i=1}^{I} \{\psi \in \Psi | \, \mathbf{g}_i'\psi > 0\} \cap \bigcap_{j=1}^{J} \{\psi \in \Psi | \, \mathbf{h}_j'\psi = 0\},$$

for some conformable real vectors $\mathbf{g}_1, \ldots, \mathbf{g}_I, \mathbf{h}_1, \ldots, \mathbf{h}_J$. For example, one would break invariance with respect to a set of $(I+1)!$ invariant transformations with a normalization consisting in the intersection of I half hyper-spaces. In contrast, the intersection of J half hyper-planes would break invariance with respect to a set of invariant transformations that is equinumerous to $\mathbb{R}^J$ (*i.e* $\mathcal{T}$ has the same cardinality as $\mathbb{R}^J$).

Example 18 (Example 1, continued) *There are* $2!$ *transformations in* $\mathcal{T}_S$ *and the half-space* $\{\psi \in \Psi | \, b > 0\}$ *breaks invariance with respect to reflections. The set* $\mathcal{T}_D$ *is equinumerous to* $\mathbb{R}$ (e.g. *the natural logarithm is a bijection from* $(0, \infty)$ *to* $\mathbb{R}$) *and the line* $\{\psi \in \Psi | \, \sigma^2 = 1\}$ *breaks scale invariance.*

Because there are many ways to normalize a model, it is natural to ask how alternatives should be compared. A first natural criterion for choosing a normalization is that it should be observationally unrestrictive.

Definition 11 *Suppose* $l(\psi | \, y)$ *is the likelihood function of a parametric statistical model* $(\mathcal{Y}, \mathcal{F}, \Theta)$. *A normalization* $\Psi^N \subseteq \Psi$ *is* ***observationally unrestrictive*** *if there exists a transformation* $g : \Psi \to \Psi^N$ *such that* $l(g(\psi) | \, y) = l(\psi | \, y)$ *for all* $y \in \mathcal{Y}$. *A normalization is observationally restrictive otherwise.*

Normalization does not merely ensure a parameter point estimator is well defined, it also affects its sampling distribution. For example, Hiller (1990) shows how normalization in structural equations models affects the finite-sample distribution of ordinary least squares and two-stage least squares estimators. Unfortunately, as Hamilton, Waggoner, and Zha (2007) note, "the fact that normalization can materially affect the conclusions

one draws from likelihood-based methods is not widely recognized."

In Blais (2008b), I show that it is in general not possible to ensure unimodal parameter point estimator sampling (or parameter posterior) distributions through an observationally unrestrictive normalization of the parameter space. Building on the work of Hamilton, Waggoner, and Zha (2007), I also argue that normalizations satisfying the following identification principle are more likely (over possible true parameter values) to produce unimodal distributions:

Definition 12 *A normalization $\Psi^N \subseteq \Psi$ satisfies the **identification principle** if it*

$a)$ *is observationally unrestrictive;*

$b)$ *is connected;*

$c)$ *provides global identification.*

Note that intersections of connected spaces are connected. Global identification can be difficult to verify and one often considers the weaker concept of local identification. Local identification can be equivalently defined in terms of the Fisher information matrix. Rothenberg (1971) shows that the parametric model $\left(\mathcal{Y}, \mathcal{F}, \Psi^N\right)$ is locally identified at $\psi_1 \in \Psi^N$ if the Fisher information matrix

$$\mathfrak{I}(\psi_1) \equiv \int_{y \in \mathcal{Y}} \frac{\partial \log l(y|\,\psi)}{\partial \psi} \frac{\partial \log l(y|\,\psi)}{\partial \psi'} f(y|\,\psi)\,\mathrm{d}y \Big|_{\psi_1}$$

is non-singular in a neighborhood of ψ_1. For future reference, let Ψ^l be defined as follows:

Definition 13 *The parameter subspace $\Psi^l \subseteq \Psi$ where the Fisher information matrix is singular or $\log l(y|\,\psi) = -\infty$ is the **singularity parameter subspace**.*

A third identification concept is that of weak identification (or empirical underidentification in the psychometrics literature). Except in the context of instrumental variables (IV) and the generalized method of moments (GMM), weak identification has not been defined precisely. Dufour and Hsiao (2008) write: "More generally, any situation where a parameter may be difficult to determine because we are close to a case where a parameter ceases to be identifiable may be called *weak identification*." Many common situations fit this description. For example, multicollinearity issues arise in linear regression models when the sample covariance matrix of the regressors is "close" to being singular. If one restricts attention to ML inference, weak identification problems occur when the Fisher information matrix is close to being singular at the pseudo-true parameter values. In this paper, I say that a parametric model $\left(\mathcal{Y}, \mathcal{F}, \Psi^N\right)$ is **weakly identified** if the pseudo-true parameter value $\psi^0 \in \Psi^N$ is "close" to Ψ^l.

Weak identification has severe consequences for ML inference, which Dufour and Hsiao (2008) summarize thus:

> "...standard asymptotic distributional may remain valid, but they constitute very bad approximations to what happens in finite samples:
>
> 1. standard consistent estimators of structural parameters can be heavily biased and follow distributions whose form is far from the limiting Gaussian distribution, such as bimodal distributions, even with fairly large samples (Nelson and Startz, 1990; Hiller, 1990; Buse, 1992);
> 2. standard tests and confidence sets, such as Wald-type procedures based on estimated standard errors, become highly unreliable or completely invalid (Dufour, 1997)"

How close is too close? As weak identification is a finite-sample concern, one might be tempted to believe it is only a small-sample concern. However, Bound et al. (1995) present an IV situation in which weak identification difficulties persist even with 329000 observations. Obviously, if the instruments were uncorrelated with the regressors in population, increasing the sample size would be futile. In practice however, the statistician never knows the pseudo-true parameter values and he should favor inferential methods that are robust to weak identification.

If a normalization provides global identification, then weak identification difficulties only arise when the pseudo-true parameter value is close to the normalization's boundary. In contrast, if a normalization does not provide global identification, then weak identification problems can occurs if the pseudo-true parameter value is close to the singularity subspace. Therefore, an econometrician should use an observationally unrestrictive normalization providing global identification when one exists.

Example 19 (Location mixture) *Consider the following mixture of* $K = 2$ *normal distributions with common variance* σ^2 *and means* μ_1 *and* μ_2

$$f(y|\mu_1, \mu_2, \pi, \sigma) = \pi\mathcal{N}(y|\mu_1, \sigma) + (1-\pi)\mathcal{N}(y|\mu_2, \sigma).$$

Label (or permutation) invariance refers to the likelihood function's invariance with respect to the re-labeling of the components. Here,

$$f(y|\,\mu_1, \mu_2, \pi, \sigma) = f(y|\,\mu_2, \mu_1, 1-\pi, \sigma),$$

which establishes the invariance to the re-labeling (or permuting) of component indices 1 *and* 2. *In matrix notation, this set of invariant transformations is*

$$\mathcal{T}_{\mathbf{P}}(\Psi) = \left\{ T : \Psi \to \Psi \middle| \, T_{\mathbf{P}}(\mu, \Pi, \sigma) = (\mathbf{P}\mu, \mathbf{P}\Pi, \sigma) \right\}$$

with $\mu = [\mu_1\ \mu_2]'$, $\Pi = [\pi\ 1-\pi]'$ *and* $\mathbf{P}$ *is a permutation matrix,* i.e. *a matrix obtained by permuting the rows of an identity matrix. Permutation invariance implies that the likelihood function admits two equivalent global maxima, sitting at the summit* ***symmetric lobes****: if* $\left(\hat{\mu}, \hat{\Pi}, \hat{\sigma}\right)$ *is a global maximum, so is* $\left(\mathbf{P}\hat{\mu}, \mathbf{P}\hat{\Pi}, \hat{\sigma}\right)$.

In order to break permutation invariance, one might contemplate one of the two following normalizations:

$$\begin{aligned} \Psi^{\pi} &= \{\psi \in \Psi | \, \pi > 0.5\} \\ \Psi^{\mu} &= \{\psi \in \Psi | \, \mu_1 > \mu_2\}. \end{aligned}$$

Either of these normalizations would break permutation invariance as

$$\begin{aligned} \mathcal{T}\left(\Psi^{\pi}\right) &= \left\{T : \Psi^{\pi} \to \Psi^{\pi} \middle| \, T_{\mathbf{P}}(\mu, \Pi, \sigma) = (\mathbf{P}\mu, \mathbf{P}\Pi, \sigma), \, \mathbf{P} = \mathcal{I}\right\} = \mathcal{T}_{\mathcal{I}} \\ \mathcal{T}\left(\Psi^{\mu}\right) &= \left\{T : \Psi^{\mu} \to \Psi^{\mu} \middle| \, T_{\mathbf{P}}(\mu, \Pi, \sigma) = (\mathbf{P}\mu, \mathbf{P}\Pi, \sigma), \, \mathbf{P} = \mathcal{I}\right\} = \mathcal{T}_{\mathcal{I}}. \end{aligned}$$

However, these normalizations yield different finite-sample inference. Assume one obtains ML estimates $\hat{\psi} = [\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}, \hat{\pi}]' = [1, 2, 1, 0.25]'$ *under* Ψ^{π}. $\mathcal{I}(\hat{\psi})$ *has full rank, which is can be verified numerically. However,* $\mathcal{I}(\psi)$ *is singular on* $\{\psi \in \Psi | \, \mu_1 = \mu_2\} \subset \Psi^{\pi}$. *Therefore, the sampling distributions of the ML estimator of* μ_1 *and* μ_2 *can be multimodal under* Ψ^{π}. *Intuitively, this normalization would perform poorly if the data came from a mixture distribution with* $\pi = 0.5$ *because component densities are equiprobable. The identification principle rules out* Ψ^{π} *because the information matrix is singular on* $\{\psi \in \Psi | \, \mu_1 = \mu_2\} \subset \Psi^{\pi}$. *In contrast, the model is globally identified on* Ψ^{μ}.

In the latter example, the identification principle yields a unique normalization, under which the ML estimator has a unimodal sampling distribution for any $\psi \in \Psi^{\mu}$. In slightly more general models, the identification principle is less straightforward to apply, as it

yields uncountably many normalizations. The practical guidance that the identification principle offers is thus incomplete. Moreover, there is no guarantee that any particular normalization ensures that the ML estimator has a unimodal sampling distribution.

Example 20 *Consider the location-and-scale mixture of normal distributions*

$$f(y_t|\mu_1,\mu_2,\pi,\sigma_1^2,\sigma_2^2) \;=\; \pi\phi(y_t|\mu_1,\sigma_1^2)+(1-\pi)\phi(y_t|\mu_2,\sigma_2^2).$$

The set where the information matrix is singular is not a line but a point,

$$\Psi^l=\{\psi\in\Psi|\,\mu_1=\mu_2\}\cap\{\psi\in\Psi|\,\sigma_1=\sigma_2\}\,.$$

The identification principle still rules out normalization Ψ^π, but the singularity set no longer separates the parameter space into two symmetric half-spaces. Normalizations Ψ^μ and

$$\Psi^\sigma=\{\psi\in\Psi|\,\sigma_1>\sigma_2\}$$

both satisfy the identification principle, but neither ensures that all sampling distributions are unimodal. To illustrate, consider samples from a population with $\mu_1=\mu_2$ and $\sigma_1>\sigma_2$. Under Ψ^μ, the MLE of σ_1 and σ_2 will both have bimodal sampling distributions for sufficiently large samples (Geweke, 2007).

In this paper, I show that Dai and Singleton's (2000) normalization of ATSMs violates the identification principle and therefore makes inference particularly sensitive to weak identification problems. Normalization of affine transformations is best understood by considering simple affine transformations: translation, scaling, rotation, permutation (labeling) and reflection (signing) of the factors. Permutation and reflection invariance make the likelihood function symmetric and introduce weak identification problems. The consequences of permutation invariance for inference for finite mixture distributions is well documented (Redner and Walker, 1984; Stephens, 2000; Celeux et al., 2000;

Frühwirth-Schnatter, 2001). I present evidence that these problems are empirically relevant for ATSMs and that normalization has important consequences for both ML and Bayesian inference. Furthermore, I propose uncountably many normalizations satisfying the identification principle.

Inference methodology

Because uncountably many normalizations of ATSMs satisfy the identification principle, the practical guidance that it offers is thus incomplete. Moreover, there is no guarantee that any particular normalization ensures that the ML estimator has a unimodal sampling distribution. For a given data set, some normalizations yield estimators with better finite-sample properties than others. Hamilton et al. (2007) suggest that one should " try several different normalizations" and "plot the small-sample distributions of parameters." Because one must resort to simulation methods (Stoffer and Wall, 1991), comparing normalizations can thus be computationally demanding or even intractable.

The Bayesian solution to this problem has computational and inferential advantages. Stephens (1997) shows that one can equivalently break permutation invariance within a posterior sampler or as a post-simulation step on the output from an un-normalized sampler. It turns out that his result also applies to reflection invariance. Observationally unrestrictive normalizations contain no information and Frühwirth-Schnatter (2001) thus proposes choosing a normalization by inspection of the posterior distribution. An econometrician can therefore generate a single sample form the permutation- and reflection-invariant posterior and then normalize the parameter space using the information contained in the data. Moreover, permutation- and reflection-invariant posterior distributions are perfectly valid characterization of uncertainty for permutation- and reflection-invariant quantities (Frühwirth-Schnatter, 2001; Geweke, 2007). In particular,

one can make valid and useful inference for observational errors using the permutation- and reflection-invariant posterior distributions. In contrast, inference for observational errors using ML parameter estimates can be completely invalid.

I thus proceed with a Bayesian analysis of the permutation- and reflection-invariant ATSM. I propose a Metropolis-within-Gibbs sampler in which latent factors are drawn together with some parameters as a single block. Because of the high correlations between latent factors and the parameters entering the pricing equations, the proposed sampler is numerically more efficient than one in which factors and parameters are drawn as separate blocks.

Few papers estimate DTSMs by Bayesian methods. Frühwirth-Schnatter and Geyer (1998),Lamoureux and Witte (2002), Müller et al. (2003), and Sanford and Martin (2005) consider CIR models. Ang et al. (2007) use an approximate Gibbs sampler where latent factors are de-meaned. Chib and Ergashev (2008) propose a numerically efficient, exact Gibbs sampler for ATSMs. To the best of my knowledge, none of the literature takes into account weak identification problems associated with permutation and reflection invariance. I use consider the reflection- and permutation-invariant posterior and show that some normalizations yield multimodal marginal posteriors.

This paper is organized as follows. In the first section, I briefly review some ATSM essentials, introduce notation and present the economic model. The second section pertains to error modeling and describes the proxy and latent-factor modeling approaches to error specification. I explain how normalization affects inference for ATSMs in the third section. Section 4 presents permutation- and reflection-invariant prior distributions. In the fifth section, I describe the posterior sampler and discuss its

implementation. In the final section, I present empirical results.

3.2 Economic modeling

The dynamics of the SDF constrain the term structure. This section first presents sufficient conditions for obtaining affine discount rates. Conditions for analytical pricing are easier to express in terms of the risk-neutral measure . They impose a tight relationship between the SDF and the physical factor dynamics, while allowing for considerable flexibility in other dimensions (see Dai, Le, and Singleton (2006) for an analysis of ATSMs in continuous time, and Bertholon, Monfort, and Pegoraro (2007) for more general pricing models). Given the risk-free dynamics, choosing physical factor dynamics fixes the risk premium, and *vice versa*. In this paper, I opt for a discrete-time version of Duffee's (2002) conditionally-Gaussian factor model, in which the SDF is exponential-affine in these factors.

3.2.1 Pricing discount bonds

In discrete time, given the nominal SDF at $t+1$, M_{t+1}, the price at time t of a discount bond maturing n periods from t, $P_{n,t}$, satisfies the difference equation

$$P_{n,t} = \mathrm{E}_t^{\mathbb{P}}\left[P_{n-1,t+1}M_{t+1}\right], \tag{3.5}$$

with boundary conditions

$$P_{0,t} = 1, \forall t, \tag{3.6}$$

and where the operator $\mathrm{E}_t^{\mathbb{P}}[.]$ refers to the conditional expectation at t under the physical measure $\mathbb{P}$. For future reference, I define the the log price of the n-period discount

bond, $p_{n,t} \equiv \ln P_{n,t}$, and the continuously-compounded yield to maturity of the n-period discount bond as $y_{n,t} \equiv -\frac{p_{n,t}}{n}$. Equation (3.5) may or may not admit an analytical solution, depending on the joint physical dynamics of the prices and the SDF.

In discrete time, markets with are incomplete and the functional form of the SDF must be specified. In this model, the state of the economy at time t is completely specified by a K-dimensional vector of factors X_t. Following Gouriéroux, Monfort, and Polimenis (2002), the log SDF, m_{t+1}, is written in the simplest affine manner as

$$-m_{t+1} = \Lambda_t X_{t+1} + \gamma_t, \tag{3.7}$$

where Λ_t is the time-dependent price of risk. They show how to determine the time-dependent intercept γ_t from (3.5) for compound autoregressive processes, of which the Gaussian process I specify below is a special case.

Under the physical measure $\mathbb{P}$, the latent factors are given by

$$X_{t+1} = \mu_t^{\mathbb{P}} + \Sigma_t^{1/2} \epsilon_{t+1}. \tag{3.8}$$

where $\mu_t^{\mathbb{P}}$ and Σ_t are the conditional mean and covariance of the factor vector, $\Sigma_t^{1/2}$ is the upper Cholesky factor of Σ_t and ϵ_{t+1} is a vector of independent standard normal random variables.

Writing (3.5) for the one-period bond,

$$\mathrm{e}^{-y_{1,t}} = \mathbf{E}^{\mathbb{P}}_{X_{t+1}|X_t}\left[\mathrm{e}^{m_{t+1}}\right],$$

and solving for γ_t yields

$$\gamma_t = y_{1,t} - \Lambda_t \mu_t^{\mathbb{P}} + \frac{1}{2}\Lambda_t \Sigma_t \Lambda_t'. \tag{3.9}$$

Substituting γ_t in the expression for the log-SDF (3.7) gives

$$-m_{t+1} = y_{1,t} + \Lambda_t \left(X_{t+1} - \mu_t^{\mathbb{P}}\right) + \frac{1}{2}\Lambda_t \Sigma_t \Lambda_t'.$$

One still has to specify $\mu_t^{\mathbb{P}}$, Σ_t, $y_{1,t}$ and Λ_t in such a way that (3.5) has a simple solution. It turns out to be much easier to work under the risk-neutral measure $\mathbb{Q}$, for which (3.5) is written as

$$P_{n,t} = \mathrm{e}^{-y_{1,t}} \mathrm{E}_t^{\mathbb{Q}} \left[P_{n-1,t+1}\right], \tag{3.10}$$

and specify the short rate and the risk-neutral factor dynamics in a way that facilitates pricing. DTSMs are therefore often appropriately referred to as *short rate models*. Two such assumptions are the following:

Assumption 1 *A1. The short rate is affine in the factors. That is,*

$$Y_{1,t} \equiv Y_1(X_t) = \tilde{A}_1 + \tilde{\mathbf{B}}_1' X_t, \tag{3.11}$$

where $\tilde{A}_1$ and $\tilde{\mathbf{B}}_1$ are constants.

Assumption 2 *A2. Under the risk-neutral measure, the factor dynamics are given by an Gaussian VAR(1) process*

$$X_{t+1} = X_t + \kappa^{\mathbb{Q}}(\theta^{\mathbb{Q}} - X_t) + \Sigma_t^{1/2} \epsilon_{t+1} \tag{3.12}$$

Under equivalent assumptions, Ang and Piazzesi (2003) solve[3] (3.10) and show that prices of discount bonds maturing in $n > 0$ periods satisfy:

$$\begin{aligned} P_{n,t} &= \exp\{-\tilde{A}_n - \tilde{\mathbf{B}}_n' X_t\} \\ \text{with} \quad \tilde{A}_{n+1} &= \tilde{A}_1 + \tilde{A}_n + \theta^{\mathbb{Q}'}\kappa^{\mathbb{Q}'}\tilde{\mathbf{B}}_n - \frac{1}{2}\tilde{\mathbf{B}}_n'\Sigma_t\tilde{\mathbf{B}}_n \\ \text{and} \quad \tilde{\mathbf{B}}_{n+1} &= \tilde{\mathbf{B}}_1 + (\mathcal{I} - \kappa^{\mathbb{Q}'})\tilde{\mathbf{B}}_n \end{aligned} \tag{3.13}$$

with the boundary conditions (3.6) $\tilde{A}_0 = 0$ and $\tilde{B}_0 = \mathbf{0}$.

Dai, Singleton, and Yang (2005) show[4] how to link physical and risk-neutral measures. Since the price $P_{n,t}$ of a *any* cash flow c_{t+1} can be calculated under (3.5) or (3.10) :

$$\mathbf{E}^{\mathbb{P}}_{X_{t+1}|X_t}\left[\mathrm{e}^{m_{t+1}}c_{t+1}\right] = \mathrm{e}^{-y_{1,t}}\mathbf{E}^{\mathbb{Q}}_{X_{t+1}|X_t}\left[c_{t+1}\right],$$

one can identify the risk-neutral measure

$$\mathrm{d}\mathbb{Q} = \mathrm{e}^{y_{1,t}+m_{t+1}}\mathrm{d}\mathbb{P},$$

and compute the risk-neutral unconditional mean,

$$\mu_t^{\mathbb{Q}} \equiv \mathbf{E}^{\mathbb{Q}}_{X_{t+1}|X_t}\left[X_{t+1}\right] = \mu_t^{\mathbb{P}} - \Sigma_t\Lambda_t'. \tag{3.14}$$

Since the change of measure concerns only the conditional mean, Λ_t completely specifies the passage between physical and risk-neutral measure. Alternatively, it is completely

[3]Proof is provided in Appendix 3.10 for completeness.

[4]See Appendix 3.11.

specified by $\mu_t^{\mathbb{P}}$ and $\mu_t^{\mathbb{Q}}$ as

$$\Lambda_t = \Sigma_t^{-1}\left(\mu_t^{\mathbb{P}} - \mu_t^{\mathbb{Q}}\right). \tag{3.15}$$

3.2.2 Physical dynamics and risk premia

At this point, the conditionally Gaussian physical dynamics of factors (3.8) is still somewhat general and one needs to specify $\mu_t^{\mathbb{P}}$ and Σ_t to complete the model. Any function of X_t will do. This choice will determine the functional form of the risk premium. For example, regime switching processes can be considered in this framework, as in Dai, Singleton, and Yang (2005) or Monfort and Pegoraro (2007).

In this study, I consider a simple but popular VAR(1) process,

$$X_{t+1} = X_t + \kappa^{\mathbb{P}}(\theta^{\mathbb{P}} - X_t) + \Sigma^{1/2}\epsilon_{t+1}. \tag{3.16}$$

Using (3.15), the risk premium is

$$\begin{aligned}\Lambda_t &= \Sigma^{-1}\left(\left(\kappa^{\mathbb{P}} - \kappa^{\mathbb{Q}}\right)X_t + \kappa^{\mathbb{P}}\theta^{\mathbb{P}} - \kappa^{\mathbb{Q}}\theta^{\mathbb{Q}}\right) \\ &= \Sigma^{-1}\left(\lambda_0 + \lambda_1 X_t\right),\end{aligned}$$

where

$$\lambda_0 \equiv \kappa^{\mathbb{P}}\theta^{\mathbb{P}} - \kappa^{\mathbb{Q}}\theta^{\mathbb{Q}} \tag{3.17}$$

$$\lambda_1 \equiv \kappa^{\mathbb{P}} - \kappa^{\mathbb{Q}}. \tag{3.18}$$

These relations imply that there are only two K-dimensional vectors (from $\left\{\lambda_0, \theta^{\mathbb{P}}, \theta^{\mathbb{Q}}\right\}$) and two $K \times K$ matrices (from $\left\{\lambda_1, \kappa^{\mathbb{P}}, \kappa^{\mathbb{Q}}\right\}$) to specify.

3.3 Error modeling

The economic model presented in the previous section gives a deterministic relationship between the state variables and observed discount rates. The state variables consisting of K factors, the covariance matrix of $N > K$ discount rates has rank K. The econometrician must thus model observational errors in order to obtain a non-singular likelihood.

There is no standard terminology for the modeling of pricing errors in the literature. In this paper, I use *proxy* and *latent-factor* for the error modeling approaches. The *proxy* modeling approach, most often used in the macroeconomics or financial economics literature[5], follows Chen and Scott (1993) and assumes that only $N - K$ yields are observed with error. This is computationally convenient. But it is obviously awkward and theoretically unjustified to maintain, for example, that the model prices 5-year bonds exactly and 4-year bonds with error. Modeling errors on all yields is proposed by Chen and Scott (1995) and is consistent with the fact that the model is a mere simplification of reality and describes it imperfectly. This *latent-factor* modeling approach, is popular in the empirical finance literature[6]. There are two more reasons to preferring the latent-factor approach.

Even if the model were indeed *true*, the construction of discount rates from observable coupon bond yields introduces errors in the former. Coupon bond yields are non-linear functions of discount rates. Using quoted yields directly in statistical inference thus presents a computational challenge and the standard approach is thus

[5]Examples are: Dai and Singleton (2000); Duffee (2002); Ang and Piazzesi (2003); Evans (2003) and Garcia and Luger (2007).

[6]See Jegadeesh and Pennacchi (1996); Geyer and Pichler (1999) and Babbs and Nowman (1999) for frequentist examples; and Frühwirth-Schnatter and Geyer (1998), Lamoureux and Witte (2002) and Ang et al. (2007)) for Bayesian studies.

building discount rates in an *ad hoc* manner, before statistical inference and without reference to the model[7]. Adding pricing errors to the model is a way to account for such data pre-processing. As one alternative to pre-processing bond yields, one can use strip bonds (Lamoureux and Witte, 2002) which give discount rates directly. However, these are arguably less liquid bonds which leads to other problems.

From a statistical point a view, the vector autoregressive moving-average representation of the rate dynamics is quite different under the two error specifications. Simple algebra reveals[8] that the proxy approach implies a VAR(1) representation of the rate dynamics, while latent-factor approach implies a more general VARMA(1,1) representation.

Because one does inference for the economic and the statistical models jointly, a restrictive error model could lead one to wrongly infer that an ATSM provides an inappropriate description of the term structure. In this paper, I compare the proxy and latent-factor approaches by looking at the statistical properties of the residuals under these two specifications.

For notational convenience, let $A_n \equiv \tilde{A}_n/n$ and $\mathbf{B}_n \equiv \tilde{\mathbf{B}}_n/n$ denote the standardized pricing coefficients, which I stack in matrices $\mathbf{A}$ and $\mathbf{B}$. One observes N rates at time t, stacked in a vector y_t. Under the latent-factor modeling approach, pricing and measurement errors add up to a multivariate normal error of covariance Ω and one

[7]Bliss (1997) explains and compare several such methods. The problem is potentially more important for methods that impose some smoothness to the curve, as the cubic spline method of McCulloch (1975). Imposing a structure actually adds information not contained in the data but it also removes some information as bonds are not priced exactly.

[8]See Appendix 3.12.

writes the system as

$$\begin{aligned} y_t &= \mathbf{A} + \mathbf{B}' X_t + \Omega^{1/2} u_t \\ X_t &= X_{t-1} + \kappa^{\mathbb{P}}(\theta^{\mathbb{P}} - X_{t-1}) + \Sigma^{1/2}\epsilon_t. \end{aligned} \tag{3.19}$$

Under the proxy modeling approach, the N yields are partitioned into sets of K perfectly observed yields, y_t^p, and $N - K$ imperfectly observed yields, y_t^i, resulting in the system

$$y_t^p = \mathbf{A}^p + \mathbf{B}^{p\prime} X_t$$

$$y_t^i = \mathbf{A}^i + \mathbf{B}^{i\prime} X_t + \Omega^{1/2} u_t \tag{3.20}$$

$$X_t = X_{t-1} + \kappa^{\mathbb{P}}(\theta^{\mathbb{P}} - X_{t-1}) + \Sigma^{1/2}\epsilon_t. \tag{3.21}$$

The likelihood is then computed by substituting $X_t = \mathbf{B}^{i\prime -1}(y_t^p - \mathbf{A}^p)$ into (3.20-3.21), which highlights that proxying factors are affine transformations of the perfectly observed yields.

Because latent-factor models essentially decompose the dynamics of the observables into common and idiosyncratic components, error covariance modeling also allows the econometrician to specify which characteristics of the observables the common latent factors should capture. The proxy approach is a special case of the latent-factor approach corresponding to a rather strong restriction on the covariance matrix Ω in which elements are equal to zero. Other restrictions are also likely to affect the factor-error decomposition. Imposing homoscedasticity and independence are examples of such restrictions. In order to consider these restrictions individually, I factorize the covariance matrix into a correlation matrix $\mathbf{R}$ and a diagonal matrix ξ of

precisions,

$$\Omega = \mathbf{DRD}' \tag{3.22}$$

$$\xi^{-1} = \mathbf{DD}', \tag{3.23}$$

where $\mathbf{D}$ is the diagonal matrix of standard deviations. In this paper, I propose priors on $\mathbf{R}$ and ξ that operationalize *soft* restrictions on the correlation and precision matrices.

3.4 Normalization

Let $\psi \in \Psi$ denote the K-factor ATSM's parameter vector,

$$\psi = \{\mathbf{A}_1, \mathbf{B}_1, \theta^{\mathbb{P}}, \theta^{\mathbb{Q}}, \kappa^{\mathbb{P}}, \kappa^{\mathbb{Q}}, \Sigma, \Omega\}.$$

For any K-dimensional vector $\mathbf{t}$ and any invertible $K \times K$ matrix $\mathbf{M}$,

$$\begin{aligned} f\left(y \middle| T_{\mathbf{tM}}(\psi), \mathbf{M}(X - \mathbf{t})\right) &= f\left(y \middle| \psi, X\right) \\ f\left(y \middle| T_{\mathbf{tM}}(\psi)\right) &= f\left(y \middle| \psi\right), \end{aligned}$$

where

$$T_{\mathbf{tM}}\left(\mathbf{A}_1, \mathbf{B}_1, \theta^{\mathbb{P}}, \theta^{\mathbb{Q}}, \kappa^{\mathbb{P}}, \kappa^{\mathbb{Q}}, \Sigma, \Omega\right) =$$

$$\left(\mathbf{A}_1 - \mathbf{B}_1'\mathbf{t}, \mathbf{M}'^{-1}\mathbf{B}_1, \mathbf{M}(\theta^{\mathbb{P}} - \mathbf{t}), \mathbf{M}(\theta^{\mathbb{Q}} - \mathbf{t}), \mathbf{M}\kappa^{\mathbb{P}}\mathbf{M}^{-1}, \mathbf{M}\kappa^{\mathbb{Q}}\mathbf{M}^{-1}, \mathbf{M}\Sigma\mathbf{M}', \Omega\right).$$

We thus say that the density of discount rates is invariant with respect to $\mathcal{T}_{\mathbf{tM}}(\Psi)$ and the parameter vectors ψ and $T_{\mathbf{tM}}(\psi)$ are observationally equivalent.

Decomposing affine transformations into simpler ones clarifies the identification

problem. If an function is invariant with respect to some set of transformations $\mathcal{T}_f$ and $T_f(\psi) = T_g(T_h(\psi))$, then a normalization breaking invariance with respect to $\mathcal{T}_h$ and $\mathcal{T}_g$ breaks invariance with respect to $\mathcal{T}_f$. Several decompositions are possible, but a finer decomposition yields more insight than a coarser one. Here, I decompose affine transformations into translations, scaling, rotations, permutations and reflections:

$$
\begin{aligned}
\mathcal{T}_{\mathbf{t}} &= \left\{T_{\mathbf{tM}} \in \mathcal{T}_{\mathbf{tM}} \middle| \, \mathbf{M} = \mathcal{I}\right\} \\
\mathcal{T}_{\mathbf{D}} &= \left\{T_{\mathbf{tM}} \in \mathcal{T}_{\mathbf{tM}} \middle| \, \mathbf{t} = \mathbf{0}, \mathbf{M} = \mathbf{D}, \mathbf{D}_{ii} > 0, \mathbf{D}_{ij} = 0, j \neq j\right\} \\
\mathcal{T}_{\mathbf{O}} &= \left\{T_{\mathbf{tM}} \in \mathcal{T}_{\mathbf{tM}} \middle| \, \mathbf{t} = \mathbf{0}, \mathbf{M} = \mathbf{O}, \mathbf{O}\mathbf{O}' = \mathcal{I}, |\mathbf{O}| = 1\right\} \\
\mathcal{T}_{\mathbf{P}} &= \left\{T_{\mathbf{tM}} \in \mathcal{T}_{\mathbf{tM}} \middle| \, \mathbf{t} = \mathbf{0}, \mathbf{M} = \mathbf{P}, \mathbf{P}_{ij} \in \{0,1\}, \iota'\mathbf{P} = \iota', \mathbf{P}\iota = \iota\right\} \\
\mathcal{T}_{\mathbf{S}} &= \left\{T_{\mathbf{tM}} \in \mathcal{T}_{\mathbf{tM}} \middle| \, \mathbf{t} = \mathbf{0}, \mathbf{M} = \mathbf{S}, |\mathbf{S}_{ii}| = 1, \mathbf{S}_{ij} = 0, j \neq j\right\}.
\end{aligned}
$$

These transformations have the following geometrical interpretations:

- $X + \mathbf{t}$ translates columns of X by $\mathbf{t}$;
- $\mathbf{D}$ is a diagonal scaling matrix with positive elements, $\mathbf{D}X$ changes the scale of columns of X;
- $\mathbf{O}$ is a rotation matrix, $\mathbf{O}X$ rotates the columns of X in Euclidean space;
- $\mathbf{P}$ is a permutation matrix, $\mathbf{P}X$ swaps the rows of X;
- $\mathbf{S}$ is diagonal reflection (or signing) matrix elements 1 or -1, $\mathbf{S}X$ changes the signs of columns of X;

3.4.1 Breaking invariance

A normalization Ψ^N breaks invariance with respect $\mathcal{T}_{\mathbf{tM}}$ if

$$\mathcal{T}_{\mathbf{t}}(\Psi^N) = \mathcal{T}_{\mathbf{P}}(\Psi^N) = \mathcal{T}_{\mathbf{D}}(\Psi^N) = \mathcal{T}_{\mathbf{S}}(\Psi^N) = \mathcal{T}_{\mathbf{O}}(\Psi^N) = \mathcal{T}_{\mathcal{I}}$$

where $\mathcal{T}_{\mathcal{I}}$ is defined by equation (3.4).

Dai and Singleton (2000) propose the following normalization of affine models:

$$\Psi^{DS} = \Psi^{\theta^{\mathbb{P}}} \cap \Psi^{\kappa^{\mathbb{P}}_{tri}} \cap \Psi^{\Sigma_{\mathcal{I}}} \cap \Psi^{B_1},$$

where

$$\begin{aligned} \Psi^{\theta^{\mathbb{P}}} &= \left\{\psi \in \Psi \middle| \, \theta^{\mathbb{P}} = \mathbf{0}\right\} \\ \Psi^{\kappa^{\mathbb{P}}_{tri}} &= \left\{\psi \in \Psi \middle| \, \kappa^{\mathbb{P}} \text{ is lower triangular}\right\} \\ \Psi^{\Sigma_{\mathcal{I}}} &= \left\{\psi \in \Psi \middle| \, \Sigma = \mathcal{I}\right\} \\ \Psi^{B_1} &= \left\{\psi \in \Psi \middle| \, B_1 > \mathbf{0}\right\}. \end{aligned} \tag{3.24}$$

It is straightforward to show that

$$\mathcal{T}_{\mathbf{P}}(\Psi^{\kappa^{\mathbb{P}}_{tri}}) = \mathcal{T}_{\mathbf{O}}(\Psi^{\kappa^{\mathbb{P}}_{tri}}) = \mathcal{T}_{\mathbf{D}}(\Psi^{\Sigma_{\mathcal{I}}}) = \mathcal{T}_{\mathbf{S}}(\Psi^{B_1}) = \mathcal{T}_{\mathbf{t}}(\Psi^{\theta^{\mathbb{P}}}) = \mathcal{T}_{\mathcal{I}},$$

which confirms that $\mathcal{T}_{\mathbf{tM}}(\Psi^{DS}) = T_{\mathcal{I}}$. For example, showing that $T_{\mathbf{P}}$ is not bijective on $\Psi^{\kappa^{\mathbb{P}}_{tri}}$ only requires a counterexample. Here, one looks for a lower triangular $\kappa^{\mathbb{P}}$ such

that $\mathbf{P}\kappa^{\mathbb{P}}\mathbf{P}'$ is not lower triangular: the only permutation matrix $\mathbf{P}$ such that

$$\mathbf{P}\begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ 1 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & 0 \\ 1 & \dots & \dots & 1 & 1 \end{bmatrix}\mathbf{P}'$$

is lower triangular is the identity matrix, *i.e.* $\mathbf{P} = \mathcal{I}$, which confirms that $\mathcal{T}_{\mathbf{P}}(\Psi^{\kappa^{\mathbb{P}}_{tri}}) = \mathcal{T}_{\mathcal{I}}$.

3.4.2 Weak identification

Dai and Singleton's normalization break invariance with respect to affine transformations but do not satisfy the identification principle, which makes inference sensitive to weak identification issues. In particular, difficulties arise in the region about $\Psi^{\kappa^{\mathbb{P}}_{diag}} = \left\{\psi \in \Psi \middle| \kappa^{\mathbb{P}} \text{ is diagonal}\right\} \subset \Psi^{\kappa^{\mathbb{P}}_{tri}}$, where the permutation normalization $\Psi^{\kappa^{\mathbb{P}}_{tri}}$ becomes ineffective. Indeed, the likelihood is invariant with respect to permutations on $\Psi^{\kappa^{\mathbb{P}}_{diag}}$,

$$\mathcal{T}_{\mathbf{P}}(\Psi^{\kappa^{\mathbb{P}}_{diag}}) \;=\; \mathcal{T}_{\mathbf{P}}(\Psi) \neq \mathcal{T}_{\mathcal{I}}.$$

Reflection invariance introduce weak identification inferential difficulties too. In both cases, difficulties arise because the Fisher information matrix is singular on the following parameter subspace, where some parameters are locally unidentified:

$$\tilde{\Psi} \;=\; \left(\bigcup_{(i,j)\in\{1,..,K\},j\neq j} \Psi^{\kappa^{\mathbb{P}}_{ij}} \cap \Psi^{\kappa^{\mathbb{Q}}_{ij}} \cap \Psi^{B_{1,ij}} \cap \Psi^{\Sigma_{ij}}\right) \cup \left(\bigcup_{k=1}^{K} \Psi^{B_{k,0}}\right), \quad (3.25)$$

where

$$
\begin{aligned}
\Psi^{\kappa^{\mathbb{Q}}_{ij}} &= \left\{\psi \in \Psi \middle|\, \kappa^{\mathbb{Q}}_{ii} = \kappa^{\mathbb{Q}}_{jj}\right\} \\
\Psi^{\kappa^{\mathbb{P}}_{ij}} &= \left\{\psi \in \Psi \middle|\, \kappa^{\mathbb{P}}_{ii} = \kappa^{\mathbb{P}}_{jj}\right\} \\
\Psi^{\mathbf{B}_{1,ij}} &= \left\{\psi \in \Psi \middle|\, \mathbf{B}_{1,i} = \mathbf{B}_{1,j}\right\} \\
\Psi^{\Sigma_{ij}} &= \left\{\psi \in \Psi \middle|\, \Sigma_{ii} = \Sigma_{jj}\right\} \\
\Psi^{\mathbf{B}_{k,0}} &= \left\{\psi \in \Psi \middle|\, \text{the } k^{th} \text{ row of } \mathbf{B} \text{ is a vector of zeros}\right\}. \qquad (3.26)
\end{aligned}
$$

Intuitively, if any factor k contains too little information about the discount rates or if any two factors (i, j) are too similar then identification problems arise. The latter situation is known as the *label switching* problem in the finite mixture literature (Redner and Walker, 1984; Celeux et al., 2000; Frühwirth-Schnatter, 2001) because it is then difficult to break permutation invariance. The former could then be referred to as *sign switching*. In this case, it is difficult to break reflection invariance, which corresponds to changes in factor signs. In that sense, permutation and reflection invariance introduce weak identification issues.

With respect to permutation invariance, the subset (3.25) suggests that any of the following normalizations satisfy the identification principle and would thus yield estimators with better finite-sample properties than $\Psi^{\kappa^{\mathbb{P}}_{tri}}$:

$$
\begin{aligned}
\Psi^{B_{1,ord}} &= \left\{\psi \in \Psi \middle|\, (\mathbf{B}_{1,1} - \mathbf{B}_{1,2}) > 0, \ldots, (\mathbf{B}_{1,K-1} - \mathbf{B}_{1,K}) > 0\right\} \\
\Psi^{\kappa^{\mathbb{P}}_{ord}} &= \left\{\psi \in \Psi \middle|\, (\kappa^{\mathbb{P}}_{1,1} - \kappa^{\mathbb{P}}_{2,2}) > 0, \ldots, (\kappa^{\mathbb{P}}_{K-1,K-1} - \kappa^{\mathbb{P}}_{K,K}) > 0\right\} \\
\Psi^{\kappa^{\mathbb{Q}}_{ord}} &= \left\{\psi \in \Psi \middle|\, (\kappa^{\mathbb{Q}}_{1,1} - \kappa^{\mathbb{Q}}_{2,2}) > 0, \ldots, (\kappa^{\mathbb{Q}}_{K-1,K-1} - \kappa^{\mathbb{Q}}_{K,K}) > 0\right\}.
\end{aligned}
$$

For example, $\Psi^{\kappa^{\mathbb{P}}_{ord}} \cap \Psi^{\kappa^{\mathbb{P}}_{diag}}$ satisfies the identification principle and breaks invariance with respect to rotation and permutation as

$$\begin{aligned} \mathcal{T}_{\mathbf{O}}(\Psi^{\kappa^{\mathbb{P}}_{diag}}) &= \mathcal{T}_{\mathcal{I}} \\ \mathcal{T}_{\mathbf{P}}(\Psi^{\kappa^{\mathbb{P}}_{ord}}) &= \mathcal{T}_{\mathcal{I}}. \end{aligned} \tag{3.27}$$

From the the factor loading equation (3.13), reflection normalizations satisfying the identification principle $\Psi^{\mathbf{B}_{k,0}}$ involve elements of $\mathbf{B}_1$ and $\kappa^{\mathbb{Q}}$. For example, Ψ^{B_1} and $\{\psi \in \Psi | \, \mathbf{B}_{1k} > 0, \, \mathbf{B}_{2j} > 0, \, j \neq k\}$ satisfy the identification principle. In terms of the structural parameters $\mathbf{B}_1$ and $\kappa^{\mathbb{Q}}$, this normalization is

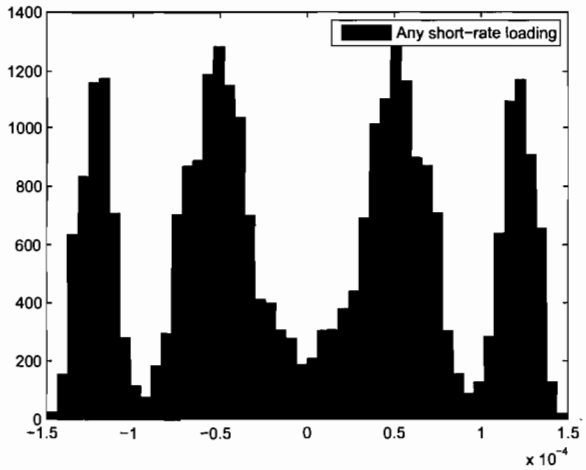
$$\{\psi \in \Psi | \, \mathbf{B}_{1i} > 0, \, \mathbf{B}_{2j} > 0, \, j \neq i\} = \left\{\psi \in \Psi | \, \mathbf{B}_{1i} > 0, \, \kappa^{\mathbb{Q}} | \mathbf{B}_{1,i} > 0, \, j \neq i\right\}.$$

For the data set I consider in this paper, Figure 3.1 shows the histogram of a sample from the permutation- and reflection-invariant posterior of $\mathbf{B}_1$. While one factor can perhaps be identified, the other factors cannot be identified from the posterior of $\mathbf{B}_1$ so normalization $\Psi^{B_{1,ord}}$ would not break permutation invariance effectively. Moreover, there is some significant posterior probability in the region about zero and $\Psi^{\mathbf{B}_1}$ would thus not break reflection invariance effectively.

Normalization $\Psi^{\mathbf{B}_1} \cap \Psi^{B_{1,ord}}$ would of course yield unimodal marginal posteriors for $\mathbf{B}_1$ (Figure 3.2). However, it does not yield unimodal marginal posteriors for every parameters. For example, Figure 3.3 shows that normalization $\Psi^{\mathbf{B}_1} \cap \Psi^{B_{1,ord}}$ yields bimodal marginal posteriors for two elements of the diagonal of $\kappa^{\mathbb{Q}}$.

There are situations where none of the few normalizations described above yields

Figure 3.1: Sample form the permutation- and reflection-invariant posterior distribution of B_1.



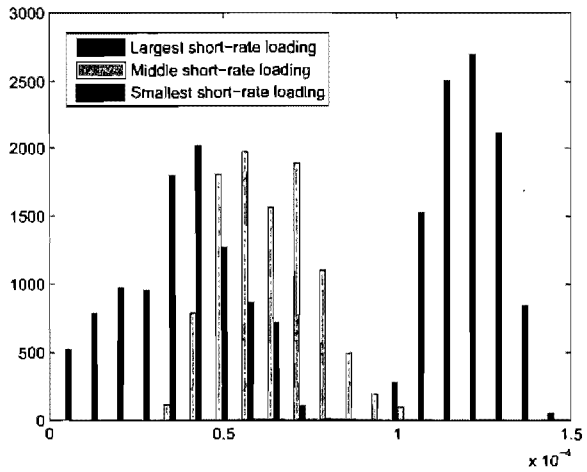
unimodal posterior distributions (Geweke, 2007). Fortunately, uncountably many normalizations satisfy the identification principle. While there is no guarantee that there exists a normalization ensuring that posteriors are unimodal, an uncountably large set is more likely to contain one than a small finite set.

In order to obtain one family of normalizations satisfying the identification, consider, for example, the following hyperplanes:

$$\left\{\psi \in \Psi \middle| \, \kappa_{1,1}^{\mathbb{P}} - \kappa_{2,2}^{\mathbb{P}} = 0\right\}$$

$$\left\{\psi \in \Psi \middle| \, \mathbf{B}_{1,1} - \mathbf{B}_{1,2} = 0\right\}.$$

Figure 3.2: Sample form the normalized posterior distribution of B_1.



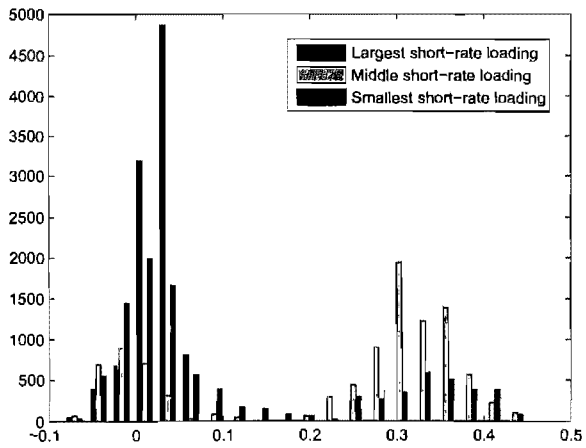
Each of these hyperplanes defines two half-spaces, and normalizations satisfying the identification principle consist in one of these half-spaces, *e.g.*

$$\left\{\psi \in \Psi \middle|\, \kappa_{1,1}^{\mathbb{P}} > \kappa_{2,2}^{\mathbb{P}}\right\}$$
$$\left\{\psi \in \Psi \middle|\, \mathbf{B}_{1,1} > \mathbf{B}_{1,2}\right\}.$$

These normalizations satisfy the identification because their frontier includes the singularity set and their interiors do not intersect with the singularity set (Definition 13). Note that, for any odd bijections $g_i : \mathbb{R} \to \mathbb{R}$, $i = 1, 2$, the half-spaces defined by the following hyperplanes also satisfy the identification principle:

$$\left\{\psi \in \Psi \middle|\, g_1(\kappa_{1,1}^{\mathbb{P}} - \kappa_{2,2}^{\mathbb{P}}) = 0\right\}$$
$$\left\{\psi \in \Psi \middle|\, g_2(\mathbf{B}_{1,1} - \mathbf{B}_{1,2}) = 0\right\}.$$

Figure 3.3: Sample form the normalized posterior distribution of $\kappa^{\mathbb{Q}}_{kk}$, $k = 1, \ldots, 3$.



Examples of useful bijections are those defining changes of coordinate system. Furthermore, the half-spaces defined by any convex combination of the above hyperplanes,

$$\left\{\psi \in \Psi \middle| \alpha g_1(\kappa^{\mathbb{P}}_{1,1} - \kappa^{\mathbb{P}}_{2,2}) + (1-\alpha) g_2(\mathbf{B}_{1,1} - \mathbf{B}_{1,2}) = 0, \alpha \in [0,1]\right\}$$

also satisfy the identification principle.

3.4.3 Observational restrictions

One might remark that a triangular matrix has more non-zero elements than a diagonal matrix of the same dimension and conclude that $\Psi^{\kappa^{\mathbb{P}}_{diag}}$ imposes observational restrictions. This is indeed the case. Note, however, that Dai and Singleton's (2000) normalization is already observationally restrictive as breaking scale invariance does not require restrictions on the off-diagonal elements of Σ,

$$\mathcal{T}_{\mathbf{D}}(\Psi^{\Sigma_{\mathcal{I}}}) \;=\; \mathcal{T}_{\mathbf{D}}(\Psi^{\Sigma_{corr}}) = \mathcal{T}_{\mathcal{I}},$$

where

$$\Psi^{\Sigma_{corr}} = \{\psi \in \Psi | \, \Sigma \text{ is a correlation matrix}\}. \tag{3.28}$$

The empirical investigation I present in this paper does not address whether these restrictions reduce model flexibility in a significant manner. I estimate a model that has the same number of parameters as Dai and Singleton's (2000) canonical representation, and I do not break permutation- and reflection-invariance. My normalization is thus

$$\Psi^{B} = \Psi^{\theta^{\mathbb{P}}} \cap \Psi^{\kappa^{\mathbb{P}}_{diag}} \cap \Psi^{\Sigma_{corr}},$$

for which

$$\begin{aligned} \mathcal{T}_{\mathbf{S}}(\Psi^{B}) &= \mathcal{T}_{\mathbf{S}}(\Psi) \\ \mathcal{T}_{\mathbf{P}}(\Psi^{B}) &= \mathcal{T}_{\mathbf{P}}(\Psi). \end{aligned}$$

Other normalizations impose observational restrictions. For example, Ang, Dong, and Piazzesi (2007) use

$$\Psi^{ADP} = \Psi^{\theta^{\mathbb{P}}} \cap \Psi^{\kappa^{\mathbb{P}}_{tri}} \cap \Psi^{\Sigma_{diag}} \cap \Psi^{B_{1,\iota}}$$

where

$$\Psi^{B_{1,\iota}} = \{\psi \in \Psi | \, B_1 = \iota\}$$

The normalization $\Psi^{B_{1,\iota}}$ is observationally restrictive because the short rate is then affine in all K factors, *i.e.* models where an element of $\mathbf{B}_1$ is zero and the corresponding factor drives only risk premium are ruled out.

Since the work of Litterman and Scheinkman (1991), many econometricians look for factor interpretations in terms of level, slope and curvature (LSC) of the term structure. For a simpler term structure model, Gouriéroux et al. (2002) argue that rotation invariance implies that such factors must be looked for in an uncountable set, which is impractical. Indeed, by continuously rotating factors, one might find ones with interpretations close to level, slope and curvature of the term structure. Christensen, Diebold, and Rudebusch (2007) show that Nelson-Siegel term structure models are observationally restrictive affine models with LSC factors where

$$\kappa^{\mathbb{Q}} = \begin{bmatrix} 0 & \kappa & -\kappa \\ 0 & 0 & \kappa \\ 0 & 0 & 0 \end{bmatrix}.$$

From equation (3.13) however, LSC factors can be obtained more generally through the following parameter restriction:

$$\Psi^{\kappa^{\mathbb{Q}}_{LSC}} = \left\{ \psi \in \Psi \,\middle|\, \kappa^{\mathbb{Q}} = \begin{bmatrix} 0 & \kappa^{\mathbb{Q}}_{1,2} & \kappa^{\mathbb{Q}}_{1,3} \\ 0 & 0 & \kappa^{\mathbb{Q}}_{2,3} \\ 0 & 0 & \kappa^{\mathbb{Q}}_{3,3} \end{bmatrix} \right\}. \tag{3.29}$$

Because $\Psi^{\kappa^{\mathbb{Q}}_{LSC}}$ breaks invariance with respect to rotations and permutations,

$$\mathcal{T}_{\mathbf{O}}(\Psi^{\kappa^{\mathbb{Q}}_{LSC}}) = \mathcal{T}_{\mathbf{P}}(\Psi^{\kappa^{\mathbb{Q}}_{LSC}}) = \mathcal{T}_{\mathcal{I}},$$

the following normalization breaks invariance with respect to affine transformations:

$$\Psi^{LSC} = \Psi^{\theta^{\mathbb{P}}} \cap \Psi^{\kappa^{\mathbb{Q}}_{LSC}} \cap \Psi^{\Sigma_{corr}} \cap \Psi^{B_1}.$$

3.5 Parameterization and prior specification

3.5.1 Parameterization

Reparameterization consists in defining a one-to-one mapping from a parameter space to another one, which often takes the form of a change of coordinate system. Some parameterizations yield parameters that are easier to interpret than others, which facilitates prior specification: I would prefer correlations to covariances on that basis. Other parameterizations may affect the numerical efficiency or stability of some algorithms. For example, one often considers the logarithm of a standard deviation in maximization routines because it maps the real line onto the positive half-line. Parameterization also affects the performance of posterior simulator (See Frühwirth-Schnatter, 2004, for an application to state space models.) Note that while I could use one parameterization for prior specification and another one for numerical efficiency reasons, this does not seem to be necessary here.

3.5.1.1 Long term discount rate factor loadings

My reparameterization of $\kappa^{\mathbb{Q}}$ is based on a novel analytic solution of the factor loadings recurrence equation (3.13). Assuming that $\kappa^{\mathbb{Q}}$ is eigendecomposable,

$$\begin{aligned}
\mathbf{B}_n &= \sum_{i=0}^{n-1}(\mathcal{I}-\kappa^{\mathbb{Q}'})^i\mathbf{B}_1 \\
&= [\mathcal{I}-(\mathcal{I}-\kappa^{\mathbb{Q}'})^n][\mathcal{I}-(\mathcal{I}-\kappa^{\mathbb{Q}'})]^{-1}\mathbf{B}_1 \\
&= [\mathcal{I}-(\delta\gamma\delta^{-1})^n][\mathcal{I}-(\delta\gamma\delta^{-1})]^{-1}\mathbf{B}_1 \\
&= [\delta\delta^{-1}-(\delta\gamma\delta^{-1})^n][\delta\delta^{-1}-(\delta\gamma\delta^{-1})]^{-1}\mathbf{B}_1 \\
&= \delta[\mathcal{I}-\gamma^n][\mathcal{I}-\gamma]^{-1}\delta^{-1}\mathbf{B}_1, \qquad (3.30)
\end{aligned}$$

where the third line uses the eigendecomposition of $\mathcal{I} - \kappa^{\mathbb{Q}'}$, i.e.

$$\mathcal{I} - \kappa^{\mathbb{Q}'} = \delta\gamma\delta^{-1}. \tag{3.31}$$

These eigenvalues, γ, thus play a central role in long-term factor loadings via γ^n, and one should expect the data to be informative about these quantities. Note that this reparameterization is not a bijection: it assumes that $\kappa^{\mathbb{Q}}$ is eigendecomposable, *i.e.* that it has K distinct eigenvalues. Recall that invertibility does not ensure eigendecomposability. Furthermore, I restrict the parameter space to matrices with real-valued eigenvalues. Complex eigenvalues would generate a sinusoidal pattern in factor loadings, in which, for example, odd-month maturities could be more sensitive to some factor than even-month maturities.

Eigenvector are defined up to a scalar multiplication so I consider normalized unit-length eigenvectors with positive first-element, which I parameterize in polar coordinates, omitting the radial coordinate. Define the matrix of angles, $\phi \equiv [\phi_1 \dots \phi_K]$, where the vector $\phi_j \in (-\frac{\pi}{2}, \frac{\pi}{2}]^{K-1}$, $j = 1, \dots, K$, contains the angles associated with the eigenvector δ_j:

$$\phi_{k,j} \equiv \arctan\left(\frac{\delta_{k+1,j}}{\sqrt{\sum_{i=1}^{k}\delta_{i,j}^2}}\right) \quad \text{for} \quad k = 1, \dots, K-1. \tag{3.32}$$

For the benchmark model I estimate in this paper, the average posterior correlations of the elements of $[\gamma, \phi]$ and those of $\kappa^{\mathbb{Q}}$ are respectively 0.26 and 0.38. Although I do not investigate the effect of this parameterization on the numerical efficiency of the posterior sampler, lower posterior correlation may result in better mixing.

3.5.1.2 Short rate factor loadings

For $K > 1$, I use a parameterization (ζ, σ) of the short rate factor loadings $\mathbf{B_1}$ in polar coordinates. I define ζ to be the $K-1$-vector of angles $[\zeta_1, \ldots, \zeta_{K-1}] \in (0, 2\pi] \times (-\frac{\pi}{2}, \frac{\pi}{2}]^{K-2}$, where

$$\zeta_k \equiv \arctan\left(\frac{\mathbf{B}_{1,k+1}}{\sqrt{\sum_{i=1}^{k} \mathbf{B}_{1,i}^2}}\right) \quad \text{for} \quad k = 1, \ldots, K-1, \tag{3.33}$$

and σ to be the logarithm of the Euclidean norm of $\mathbf{B}_1$,

$$\sigma \equiv \log\left(\sqrt{\mathbf{B}_1'\mathbf{B}_1}\right). \tag{3.34}$$

This parameterization in natural logarithm results from a computational consideration: the posterior distribution of the norm of $\sqrt{\mathbf{B}_1'\mathbf{B}_1}$ has thick tails which the posterior sampler has difficulties exploring. Considering σ improves mixing considerably.

From equation (3.30), note that e^{σ} can be interpreted as the common standard deviation of factor innovations: rates can be written as

$$\begin{aligned} Y_{n,t} &= A_n + \mathbf{B}_1^{*\prime}\delta^{-1'}[\mathcal{I} - \gamma]^{-1}[\mathcal{I} - \gamma^n]\delta' X_t^* \\ X_t^* &= (\mathrm{I} - \kappa^{\mathbb{P}})X_{t-1}^* + e_t^*, \end{aligned} \tag{3.35}$$

with $\mathbf{B}_1^* = \mathrm{e}^{-\sigma}\mathbf{B}_1$ a unit-length factor-loading vector, and $e_t^* \sim \mathcal{N}(0, \mathrm{e}^{\sigma}\Sigma)$, where Σ is a correlation matrix.

3.5.1.3 Error covariance matrix

I parameterize the error covariance matrix as a diagonal matrix of precisions ξ and a correlation matrix $\mathbf{R}$ as in (3.22), which I repeat here:

$$\begin{aligned} \Omega &= \mathbf{DRD}' \\ \xi^{-1} &= \mathbf{DD}', \end{aligned}$$

where $\mathbf{D}$ is the diagonal matrix of standard deviations.

3.5.1.4 Stationarity

Over short horizons, the dynamics of interest rates might not be well described by stationary processes. In order increase flexibility, I do not impose factor stationarity and consider the level of factors at $t = 1$ as an extra parameter, X_1. I therefore allow for co-integration, as yields could share a common unit-root factor.

3.5.1.5 Parameterization summary

To summarize the parameterizations and normalizations I use in this paper, descriptions of the parameters are given in Table 3.1. Because I make inference for permutation- and reflection-invariant ATMSs, these normalizations do not break permutation or reflection invariance.

3.5.1.6 Mapping parameters between parameter subspaces

Because permutation- and reflection-invariance implies that the likelihood function has $K!2^K$ symmetric modes, an observationally unrestrictive normalization consists in an element of a partition of the parameter space into $K!2^K$ observationally equivalent subspaces. The next section describes an extension of Frühwirth-Schnatter's (2001)

Table 3.1: Summary of parameterization and restrictions.

	Estimated parameters
A_1	Mean short-rate; positive.
(ζ, σ)	Spherical parameterization (log-radius) of $\mathbf{B}_1$, (3.33-3.34).
$\kappa^{\mathbb{P}}$	Physical mean-reversion diagonal matrix (3.27).
Σ	Factor correlation matrix (3.28).
(γ, ϕ)	Eigendecomposition of $\kappa^{\mathbb{Q}}$; spherical parameterization (unit radius) of eigenvectors (3.31-3.32).
λ_0	Mean risk premium.
ξ	Pricing error precisions (3.22).
$\mathbf{R}$	Pricing error correlation matrix (3.22).
X_1	Factor vector at $t = 1$.
	Derived parameters
$\mathbf{B}_1$	Short-rate factor loading (3.33-3.34).
$\kappa^{\mathbb{Q}}$	Risk-neutral mean-reversion matrix (3.31-3.32).
λ_1	Factor coefficient in risk premium (3.18).
$\theta^{\mathbb{Q}}$	Risk-neutral factor mean (3.17).
Ω	Pricing error covariance (3.22).
	Fixed parameters
$\theta^{\mathbb{P}}$	Physical factor mean (3.24).

Estimated parameters are used directly in the inference and have prior distributions associated to them. *Derived parameters* are functions of the *estimated parameters*. *Fixed parameters* are constrained by normalizations.

permutation sampler that maps a parameter vector $\psi \in \Psi$ to $\psi^N \in \Psi^N$, where Ψ^N has one of the two following interpretations. It can be a normalization, in which case the algorithm is used in order to normalize a sample from an un-normalized posterior sampler. It could also be a randomly chosen element of the partition associated with a normalization, in which case the algorithm is used in order to efficiently explore all $K!2^K$ observationally equivalent subspaces. Because permutation and reflection matrices are orthogonal matrices, mapping one parameter subspace to another is achieved by the following transformation

$$T_{\mathbf{SP}}\left(\left\{\mathbf{B}_1, \Lambda_0, \Sigma, \kappa^{\mathbb{P}}, \kappa^{\mathbb{Q}}, X\right\}\right)$$

$$= \left\{\mathbf{SPB}_1, \mathbf{SP}\Lambda_0, \mathbf{SP}\kappa^{\mathbb{P}}\mathbf{P}'\mathbf{S}', \mathbf{SP}\kappa^{\mathbb{Q}}\mathbf{P}'\mathbf{S}', \mathbf{SP}\Sigma\mathbf{P}'\mathbf{S}', \mathbf{SP}X\right\} \tag{3.36}$$

where $\mathbf{S}$ is a reflection matrix and $\mathbf{P}$ is a permutation matrix.

A few properties of this mapping should be noted, as I use them in order to propose permutation- and reflection-invariant prior distributions. First, pre-multiplying a vector by a reflection matrix $\mathbf{S}$ changes its direction and pre-multiplying it by $\mathbf{P}$ changes its orientation. In both cases, the Euclidean norm is preserved. In particular,

$$\sigma = \log\left(\sqrt{\mathbf{B}_1'\mathbf{B}_1}\right) = \log\left(\sqrt{(\mathbf{SPB}_1)'(\mathbf{SPB}_1)}\right). \tag{3.37}$$

Second, if the eigendecomposition of a matrix $\mathbf{A}$ is $\mathbf{A} = \delta\gamma\delta^{-1}$, then

$$\mathbf{SPAP}'\mathbf{S}' = (\mathbf{SP}\delta)\gamma(\mathbf{SP}\delta)^{-1}. \tag{3.38}$$

So the mapping changes the direction and orientation of the eigenvectors of $\mathbf{A}$ and leave its eigenvalues unchanged.

3.5.2 Prior distributions

I propose permutation- and reflection-invariant priors. For finite mixture distributions, Geweke (2007, p. 3537) argues that "If the state labels have no substantive interpretation, then the prior density must also be permutation invariant." His argument applies to reflection invariance as well. Prior distribution hyper-parameters are given in Appendix 3.9.

There are many ways to designing permutation- and reflection-invariant priors, as all one needs to do is ensure that no information is provided with respect either permutations or reflections. The conceptually simplest approach is to specify arbitrary prior distributions and consider the equiprobable mixture of these priors over all $K!2^K$ permutation and reflection combinations. Alternative approaches require some analysis to see how each element of each parameter is affected by permutation and reflection. Reparameterization sometimes helps in this analysis. Some parameters are naturally reflection-invariant, *e.g.* γ or the diagonal of $\kappa^{\mathbb{P}}$. Exchangeable prior distributions are permutation-invariant for some parameters, *e.g.* the diagonal elements of $\kappa^{\mathbb{P}}$[9]. As a special case, i.i.d. univariate priors are permutation-invariant. Priors that are symmetric with respect to 0 are reflection-invariant. They are equivalently specified as priors on the absolute values of the parameters.

In this section, I propose conditionally conjugate priors when they are available. An exchangeable normal distribution has the form $\mathcal{N}(\mu\iota, \sigma^2((1-\rho)\mathcal{I} + \rho\iota\iota'))$.

3.5.2.1 Prior distribution of ξ

I use a hierarchical prior for $\xi \equiv \text{diag}(\Omega^{-1})$ that is a Inverse-Gamma-scale-mixture of an N-dimensional vector of conditionally independent Gamma distributions. Specifically,

$$p(\xi|\gamma_\Omega^0, \nu_\Omega^0, \beta_\Omega^0) = \int_0^\infty \prod_{n=1}^N \mathcal{G}\left(\xi_n|\gamma_\Omega^0, \frac{\eta}{\gamma_\Omega^0}\right) \mathcal{IG}\left(\eta|\nu_\Omega^0, \beta_\Omega^0\right) \mathrm{d}\eta.$$

[9]This might perhaps sound tautological, as an exchangeable distribution defined as a permutation invariant distribution. However, permutations of the parameters need not correspond to permutations of the factors.

One can integrate η out and write this mixture in closed form as (See appendix 3.13.)

$$p(\xi|\gamma_\Omega^0,\nu_\Omega^0,\beta_\Omega^0) = \frac{{\gamma_\Omega^0}^{N\gamma_\Omega^0}{\beta_\Omega^0}^{\nu_\Omega^0}\Gamma(N\gamma_\Omega^0+\nu_\Omega^0)}{\Gamma(\nu_\Omega^0)\Gamma(\gamma_\Omega^0)^N}\frac{\prod_{n=1}^{N}\xi_n^{\gamma_\Omega^0-1}}{\left(\beta_\Omega^0+\gamma_\Omega^0\sum_{n=1}^{N}\xi_n\right)^{N\gamma_\Omega^0+\nu_\Omega^0}}. \quad (3.39)$$

This prior allows one to express separately prior knowledge about the global scale of the precisions and their dispersion. A large value of γ_Ω^0 corresponds to strong belief that errors are nearly identically distributed. A small value of ν_Ω^0 expresses little knowledge about the scale of precisions. β_Ω^0 is a level parameter that centers precisions around $\beta_\Omega^0/(\nu_\Omega^0-1)$.

Note that this prior is conditionally conjugate in any Gaussian model.

3.5.2.2 Prior distribution of R and Σ

R and Σ are correlation matrices and I use a prior distribution proposed by Barnard, McCulloch, and Meng (2000). They obtain this distribution by integrating the standard deviations out of an inverse-Wishart-distributed covariance matrix with identity matrix scale parameter. Defining the one-to-one mapping $g(\mathbf{R},\mathbf{D}) = \mathbf{DRD}' = \Omega$, which decomposes a covariance matrix Ω into a diagonal matrix of standard deviations D and a correlation matrix R, the distribution is

$$\begin{aligned} p(\mathbf{R}|\tau) &= \int \mathcal{IW}(\mathbf{DRD}'|\mathcal{I},\tau)\ |\mathcal{J}(\mathbf{D},\mathbf{R})|\ \mathrm{d}\mathbf{D} \\ &= |\mathbf{R}|^{\frac{1}{2}(\tau-1)(N-1)-1}\left(\prod_{i=1}^{N}|\mathbf{R}_{(ii)}|\right)^{-\frac{\tau}{2}}, \end{aligned}$$

where $\mathcal{IW}(\mathbf{DRD}'|\mathbf{W},\tau)$ is the Inverse Wishart distribution with shape τ and scale $\mathbf{W}$, $\mathcal{J}(\mathbf{D},\mathbf{R})$ is the Jacobian of the mapping $g(\cdot)$ and $\mathbf{R}_{(ii)}$ is the ith principal sub-matrix of $\mathbf{R}$. It has the property that individual correlations have Beta marginal distributions $\mathcal{B}eta(\frac{\tau-N+1}{2},\frac{\tau-N+1}{2})$ extended to $[-1,1]$ (*i.e.* $(\mathbf{R}_{ij}+1)/2$ has a Beta marginal distribution), which is uniform over $[-1,1]$ for $\tau = N+1$. My priors are thus $p(\mathbf{\Sigma}|\tau_{\Sigma}^{0})$ and $p(\mathbf{R}|\tau_{\mathbf{R}}^{0})$. Note that $p(\mathbf{\Sigma}|\tau_{\Sigma}^{0})$ is permutation- and reflection-invariant because all correlations have identical marginal priors that are symmetric with respect to 0.

3.5.2.3 Joint prior distribution of A_1 and λ_0

The system of difference equations (3.13) that the pricing coefficients satisfy introduces non-linearities that are generally viewed as preventing an analytical expression of the risk premium's conditional posterior distribution. I propose the following novel solution of (3.13) in order to obtain the conditional posterior of $\mathbf{a} \equiv [A_1 \quad \lambda_0']'$.

Write the pricing equations as

$$\begin{aligned}\tilde{A}_n &= \mathbf{a}' \begin{bmatrix} n \\ \sum_{i=1}^{n-1} \tilde{\mathbf{B}}_i \end{bmatrix} - \frac{1}{2}\left(\sum_{i=1}^{n-1} \tilde{\mathbf{B}}_i' \mathbf{\Sigma} \tilde{\mathbf{B}}_i\right) \\ \tilde{\mathbf{B}}_n &= \mathbf{B}_1 + (\mathrm{I} - \kappa^{\mathbb{Q}'})\tilde{\mathbf{B}}_{n-1},\end{aligned}$$

and define

$$\begin{aligned}\Delta_{1(K+1\times N)} &= \begin{bmatrix} 1 & \cdots & 1 \\ \frac{\sum_{i=1}^{n_1-1}\tilde{\mathbf{B}}_i}{n_1} & \cdots & \frac{\sum_{i=1}^{n_N-1}\tilde{\mathbf{B}}_i}{n_N} \end{bmatrix} \\ \Delta_{2(N\times 1)} &= \frac{1}{2}\begin{bmatrix} \frac{\sum_{i=1}^{n_1-1}\tilde{\mathbf{B}}_i'\mathbf{\Sigma}\tilde{\mathbf{B}}_i}{n_1} & \cdots & \frac{\sum_{i=1}^{n_N-1}\tilde{\mathbf{B}}_i'\mathbf{\Sigma}\tilde{\mathbf{B}}_i}{n_N} \end{bmatrix}'.\end{aligned}$$

The conditional posterior is then

$$p(\mathbf{a}|y, X, \Omega, \Sigma, \kappa^{\mathbb{Q}}, \mu_{\lambda_0}, \Sigma_{\lambda_0}) = N\left(\mathbf{a}|\hat{\mu}_{\mathbf{a}}, \hat{\Sigma}_{\mathbf{a}}\right) p(\mathbf{a}|\mu_{\mathbf{a}}, \Sigma_{\mathbf{a}}),$$

where

$$\hat{\Sigma}_{\mathbf{a}}^{-1} = T\Delta_1\Omega^{-1}\Delta_1' \tag{3.40}$$

$$\hat{\mu}_{\mathbf{a}} = \hat{\Sigma}_{\mathbf{a}}\Delta_1\Omega^{-1}\left[T\Delta_2 - T\iota_N + \sum_{t=1}^{T} y_t - \mathbf{B}'X_t\right], \tag{3.41}$$

and $\{n_1, n_2, \ldots, n_N\}$ is the set of N maturities, which shows that the conditional posterior admits conjugate Gaussian priors.

The rank of $\hat{\Sigma}_{\mathbf{a}}$ is at most equal to that of Δ_1, which is $K + 1$ unless loadings are constant over maturities for one factor ($\mathbf{B}_{n,k} = b_k$ for $n = 1, \ldots, N$), or the entire term structure is identically sensitive to two factors ($\mathbf{B}_{n,k} = \mathbf{B}_{n,j}$, for $n = 1, \ldots, N$ and $K \neq j$). The latter case corresponds to a parameter subspace $\Psi^{\mathbf{B}_{k,0}}$ defined by equation (3.26).

The former case corresponds to a situation where one factor has a pure level interpretation: a change in that factor shifts the entire term structure. Asymptotically, one would expect the sample mean of this factor to be equal to its population mean, which is zero by my normalization $\Psi^{\theta^{\mathbb{P}}}$ (see equation 3.24), and expect the factor to describe time variations in the short rate mean around A_1. In finite sample however, A_1 is imprecisely estimated: because factors are highly correlated, the factor's sample mean $\bar{X}$ can be significantly different from zero, its population value by normalization. This provides an interesting explanation of the poor performance of Gibbs samplers for ATSMs. The

sample mean of discount rates is

$$\bar{y} = \mathbf{a}'\Delta_1 - \frac{1}{2}\Delta_1 + \mathbf{B}'\bar{X}.$$

Because $\mathbf{a}$ and $\bar{X}$ play similar roles in the description of average discount rates, simulations not reported in this paper reveal that Gibbs sampling schemes where $\mathbf{a}$ and X are drawn as separate blocks result in poor mixing. One can overcome this inferential difficulty by fixing the value of A_1 to some reasonable value (which is observationally restrictive), or by constraining the factors' sample mean to being zero (Ang et al., 2007) (so that factors are not longer drawn from their full conditional posterior). The posterior sampler I describe in the next section, which draws $\mathbf{a}$ and X as a single block, is exact and mixes much better than the Gibbs schemes described above.

My priors are

$$\begin{aligned} p(A_1) &= \mathcal{N}(A_1|\mu^0_{A_1}, \Sigma^0_{A_1})\mathbf{1}_{A_1>0} \\ p(\lambda_{0,k}) &= \mathcal{N}\left(\lambda_0|\mu^0_{\lambda_0}, \Sigma^0_{\lambda_0}\right), \end{aligned}$$

for $k = 1, \ldots, K$, where the truncation $\mathbf{1}_{A_1>0}$ reflects my personal belief that the mean nominal short rate considered in this paper is positive.

3.5.2.4 Prior distribution of σ

The logarithm of the Euclidean norm of $\mathbf{B}_1$ (the global scale of factor innovations, see equations 3.34 and 3.35) is normally distributed

$$p(\sigma|\mu^0_\sigma, \Sigma^0_\sigma) \equiv \mathcal{N}(\sigma|\mu^0_\sigma, \Sigma^0_\sigma).$$

From equation (3.37), any prior on σ is permutation- and reflection-invariant.

3.5.2.5 Prior distribution of ζ

The short-rate vector of factor loadings is a priori uniformly distributed on a K-dimensional hyper-sphere with radius e^{σ}, which implies the following prior distribution on the angles:

$$p(\zeta) \equiv \frac{1}{2\pi}\prod_{k=2}^{K}\frac{1}{4\pi}\cos(\zeta_k).$$

Since permutations and reflections only change the direction and orientation of the factor loadings, this prior is permutation- and reflection-invariant.

3.5.2.6 Prior distribution of $\kappa^{\mathbb{P}}$

I do not impose stationarity and use a i.i.d. normal distribution

$$p(\kappa^{\mathbb{P}}_{k,k}) = \mathcal{N}(\kappa^{\mathbb{P}}_{k,k}|\mu^0_{\kappa^{\mathbb{P}}}, \Sigma^0_{\kappa^{\mathbb{P}}}),$$

for $k = 1, \ldots, K$.

3.5.2.7 Prior distribution of γ

The eigenvalues of $\mathcal{I} - \kappa^{\mathbb{Q}'}$ are a priori i.i.d. normally distributed

$$p(\gamma_k|\mu^0_\gamma, \Sigma^0_\gamma) \equiv \mathcal{N}(\gamma_k|\mu^0_\gamma, \Sigma^0_\gamma).$$

From (3.38), any prior is permutation- and refection-invariant.

3.5.2.8 Prior distribution of ϕ

Eigenvectors are defined up to a scalar multiplication so I consider the normalized unit-length eigenvectors with positive first-element. The K eigenvectors of $\mathcal{I} - \kappa^{\mathbb{Q}'}$ are a priori uniformly distributed on the unit half-sphere, which implies the following prior distribution on the angles:

$$p(\phi_k) \equiv \frac{1}{\pi} \prod_{j=2}^{K} \frac{1}{4\pi} \cos(\phi_{k,j}),$$

for $k = 1, \ldots, K$. Again, permutations and reflections only affects the eigenvector directions and orientations, and this prior is thus permutation- and reflection-invariant.

3.6 Posterior simulator

This section describes a Metropolis-within-Gibbs sampler combined with a extension of Früwirth-Schnatter's (2001) permutation sampler.

3.6.1 MCMC algorithm

Defining the parameter vector

$$\vartheta \equiv \{A_1, \lambda_0, \Omega, \sigma, \zeta, \gamma, \phi, \Sigma\},$$

my Metropolis-Hastings update of the chain consists of the following cycle of parameter and state updates:

Given the state of the Markov chain at iteration $(m-1)$,

1. Generate $\kappa^{\mathbb{P}*} \sim p\left(\kappa^{\mathbb{P}} \middle| y, X_1^{(m-1)}, \vartheta^{(m-1)}, X_{t=2:T}^{(m-1)}\right)$.

2. Generate $X_1^* \sim p\left(X_1 \middle| y, \kappa^{\mathbb{P}*}, \vartheta^{(m-1)}, X_{t=2:T}^{(m-1)}\right)$.

3. Generate $(\vartheta', X'_{t=2:T}) \sim q\left(\vartheta, X_{t=2:T} \middle| y, \kappa^{\mathbb{P}*}, X_1^*\right)$.

4. Take

$$(\vartheta^*, X_{t=2:T}^*) = \begin{cases} (\vartheta', X'_{t=2:T}) & \text{with probability } \rho \\ \left(\vartheta^{(m-1)}, X_{t=2:T}^{(m-1)}\right) & \text{with probability } 1-\rho \end{cases},$$

where

$$\rho = \min\left\{ \frac{p\left(\vartheta', X'_{t=2:T} \middle| y, \kappa^{\mathbb{P}*} X_1^*\right)}{p\left(\vartheta^{(m-1)}, X_{t=2:T}^{(m-1)} \middle| y, \kappa^{\mathbb{P}*}, X_1^*\right)} \frac{q\left(\vartheta^{(m-1)}, X_{t=2:T}^{(m-1)} \middle| y, \kappa^{\mathbb{P}*}, X_1^*\right)}{q\left(\vartheta', X'_{t=2:T} \middle| y, \kappa^{\mathbb{P}*}, X_1^*\right)} \right\}.$$

5. Generate $\mathbf{S}$ uniformly over the $K!$ signing matrices.

6. Generate $\mathbf{P}$ uniformly over the 2^K permutation matrices.

7. Take (see equation 3.36)

$$\left\{\mathbf{B}_1^{(m)}, \Lambda_0^{(m)}, \Sigma^{(m)}, \kappa^{\mathbb{P}(m)}, \kappa^{\mathbb{Q}(m)}, X^{(m)}\right\} = T_{\mathbf{SP}}\left(\left\{\mathbf{B}_1^*, \Lambda_0^*, \Sigma^*, \kappa^{\mathbb{P}*}, \kappa^{\mathbb{Q}*}, X^*\right\}\right).$$

The proposal in the Metropolis-Hastings defined by steps (3-4) is

$$q\left(\vartheta', X'_{t=2:T} \middle| y, \vartheta, \kappa^{\mathbb{P}}, X_1\right) = p\left(X'_{t=2,\ldots,T} \middle| y, \vartheta', \kappa^{\mathbb{P}}, X_1\right) \mathcal{N}\left(\vartheta' \middle| \vartheta, \Sigma_\vartheta\right), \quad (3.42)$$

where the density $p\left(X'_{t=2,\ldots,T} \middle| y, \vartheta', \kappa^{\mathbb{P}}, X_1\right)$ can be computed exactly using an algorithm independently suggested by Carter and Kohn (1994) and Frühwirth-Schnatter (1994), and used extensively, among others, by Kim and Nelson (1998). The parameter Σ_ϑ is chosen by the econometrician (See Robert and Casella (2004) for a discussion).

3.6.2 Mixture sampler

Steps (5-7) define a mixture sampler that generalizes Frühwirth-Schnatter's (2001) permutation sampler to permutation- and reflection-invariant linear state space models. Like the permutation sampler, the mixture sampler comes in two flavors. As used above, it allows to efficiently explore all symmetric modes of the posterior distribution. Alternatively, it can operationalize a normalization in the following manner.

As already mentioned, an observationally unrestrictive normalization consists in an element of a partition of the parameter space into $K!2^K$ observationally equivalent subspaces,

$$\Psi = \bigcup_{i=1}^{K!2^K} \Psi_i^N.$$

Assume one considers normalization $\Psi^N = \Psi_1^N$. In order to map Ψ onto Ψ_1^N, for $i = 1, \ldots, K!2^K - 1$, take $\mathbf{P}_i$ and $\mathbf{S}_i$ such that $T_{\mathbf{P}_i\mathbf{S}_i}(\psi) \in \Psi_1^N$ for $\psi \in \Psi_i^N$. Steps (5-6) are thus replaced by

(5a) Take $\mathbf{S} = \mathbf{S}_i$ such that $T_{\mathbf{S}_i}(\psi^*) \in \Psi^N$.

(6a) Take $\mathbf{P} = \mathbf{P}_i$ such that $T_{\mathbf{P}_i}(\psi^*) \in \Psi^N$.

In order to facilitate the interpretation of the parameters, one would like to find a normalization which yields unimodal parameter posterior distributions. Hamilton et al. (2007) show that normalizations satisfying the identification principle are more likely to yield such posteriors. The search can thus be restricted to normalizations satisfying the identification principle. However, there are uncountably many such normalizations. Because permutation and reflection normalizations can be implemented as a post-simulation step (Stephens, 1997; Geweke, 2007), a Bayesian analysis makes comparing a large number of normalizations computationally feasible. In contrast, one must obtain the sampling

distribution of the ML estimator by simulations methods (See Stoffer and Wall (1991) for an application to linear state space models) in order to see whether a particular normalization produces unimodal sampling distributions.

3.7 Empirical results

In this section, I investigate the empirical role of error modeling. I use a panel of monthly sampled continuously-compounded discount rates from the Fama CRSP data files. Maturities are 1, 3, 12, 36 and 60 months, and the 204 observations of the curve run from January 1988 to December 2004. The 1- and 3-month rates are from the CRSP Risk Free Rates File and the longer maturities are from the Fama-Bliss Discount Bonds File. Discount bond rates were originally built from bootstrapping a filtered set of observed coupon Treasuries and are used by Ang and Bekaert (2002), Dai et al. (2005) and Ang and Piazzesi (2003), among many others.

My benchmark model is the 3-latent-factor affine model with homoscedastic errors, which I label $A^L_{\Omega=\omega\mathcal{I}}$. I compare it to four alternatives (which I summarize in Table 3.2 for clarity): the 3-latent-factor affine model with heteroscedastic errors, $A^L_{\Omega=\text{diag}(\xi^{-1})}$; the 3-latent-factor affine model with heteroscedastic and correlated errors, $A^L_{\Omega=\mathbf{DRD}'}$; the 3-proxying-factor affine model with homoscedastic errors on the 3-month and 3-year rates ($A^P_{\Omega=\omega\mathcal{I}}$); and the 3-principal-component model (PC)[10]. I compare these models through the posterior distribution of several statistics of interest.

Because residuals are functions of the parameter vector, they are random vectors too. It is therefore possible to consider the posterior distribution of residual statistics, which I approximate using a sample from my posterior simulator. For each model, the

[10]This model is presented in Appendix 3.14 for completeness.

posterior sampler runs for 500 000 iterations, of which I keep every 100th iteration to lighten some computations. For example, I obtain a posterior sample for the mean short-rate residual by computing

$$\bar{e}_1^{(m)} = \frac{1}{T}\sum_{t=1}^{T} y_t - \mathbf{A}(\psi^{(m)}) - \mathbf{B}(\psi^{(m)})' X_t^{(m)}$$

for $m = 1, \ldots, 50\,000$, while I compute the posterior median of $\bar{e}_1$ as

$$\text{median}(\bar{e}_1) = \arg\max_e \left\{ \frac{1}{50\,000} \sum_{m=1}^{50\,000} \mathbf{1}(\bar{e}_1^{(m)} < e) \leq \frac{1}{2} \right\}.$$

Tables in this section report the posterior median and 95%-inter-quantile credibility intervals for such statistics. For expositional briefness, I will say that a parameter is significant if its 95%-inter-quantile credibility interval does not include zero.

Table 3.2: Model notation

Model	Description
$A^L_{\Omega=\omega\mathcal{I}}$	3 latent factors; homoscedastic errors (Benchmark).
$A^L_{\Omega=\text{diag}(\xi^{-1})}$	3 latent factors; heteroscedastic errors.
$A^L_{\Omega=\mathbf{DRD}'}$	3 latent factors; heteroscedastic and correlated errors.
$A^P_{\Omega=\omega\mathcal{I}}$	3 proxying factors; homoscedastic errors.
PC	3 principal components.

3.7.1 Observational errors

Table 3.3 reports the 95-%-inter-quantile credibility intervals for pricing error and absolute error statistics for affine models, and sample statistics for the principal components model. In order to compare the benchmark affine model with homoscedastic errors $A^L_{\Omega=\omega\mathcal{I}}$ (Panel a) to model $A^L_{\Omega=\text{diag}(\xi^{-1})}$ (Panel b), I use a relatively uninformative

prior on the dispersion of precisions ($\gamma_\Omega^0 = 1.01$). Allowing for high heteroscedasticity reveals that the short rate is relatively mispriced by the economic model, with errors in the order of 10 basis points (bp) on average that exhibit a standard deviation of almost 40 bp, while the errors on other maturities are not significantly different from zero. Absolute errors confirm this pattern.

Because DTSMs are derived from an hypothesis on the short rate (see equation 3.11), this is of central concern. This hypothesis justifies the general use of the short rate as a proxying factor. It therefore seems that the short rate is badly "measured" in some way, compared to other maturities. Note that one obtains even larger pricing errors on the short rate from the Fama Treasury Bill Term Structure Files derived from 6-month Treasury Bills. One possible explanation is that the bootstrapping method used to extract the 1-month rate is bound to result in higher measurement errors than for other maturities. In the 6-month Fama Treasury Bill Term Structure Files, the maximum maturity mismatch is 4 days, which is more significant on the 1-month rate than on the 3-month rate. Short-rate residuals would also be larger than other maturities if there were an omitted short-rate-specific factor. For example, if short term instruments are held for liquidity reasons, then there would be some priced liquidity factor: investors would accept a return lower than the pure time-value return for the liquidity services provided by these instruments.

Comparing the proxy (Panel d) and latent-factor modeling approaches (Panel a), errors on $N-K$ yields are now distributed on N rates, and residuals are accordingly smaller in absolute terms (approximately 2 bp *versus* 3 bp) and are less variable (approximately 14 bp *versus* 17 bp). It is perhaps surprising that three rates from the $A^L_{\Omega=\text{diag}(\xi^{-1})}$ model (Panel b) have relatively small residuals. This might lead one to

Table 3.3: Pricing errors (in basis points) statistics - covariance modeling.

	Maturity	1	3	12	36	60
Errors	Median	0.97 (−1.38, 3.31)	−2.37* (−4.67, −0.13)	2.22* (0.05, 4.38)	−1.19 (−3.44, 1.03)	0.46 (−1.82, 2.74)
	Mean	1.25 (−0.62, 3.10)	−2.60* (−4.34, −0.86)	2.19* (0.43, 3.93)	−1.25 (−3.07, 0.57)	0.43 (−1.45, 2.31)
	Std dev	14.64* (13.14, 16.28)	14.66* (13.29, 16.08)	12.78* (11.46, 14.17)	12.80* (11.49, 14.17)	12.79* (11.47, 14.17)
Abs errors	Median	9.62* (8.10, 11.29)	9.75* (8.26, 11.37)	8.75* (7.37, 10.24)	8.68* (7.32, 10.16)	8.64* (7.28, 10.15)
	Mean	11.57* (10.34, 12.93)	11.73* (10.58, 12.94)	10.36* (9.24, 11.52)	10.27* (9.17, 11.44)	10.22* (9.11, 11.39)
	Max	47.78* (36.49, 67.12)	48.33* (37.46, 65.56)	38.13* (30.69, 50.25)	37.85* (30.35, 49.99)	37.59* (30.20, 49.46)
	Std dev	9.05* (7.94, 10.29)	9.16* (8.14, 10.25)	7.81* (6.89, 8.82)	7.75* (6.84, 8.74)	7.71* (6.79, 8.70)
Panel a: $A^L_{\Omega=\omega\mathcal{I}}$						
Errors	Median	11.24* (5.81, 16.75)	−0.00 (−0.12, 0.10)	1.24 (−1.07, 3.68)	−0.07 (−1.37, 1.25)	0.00 (−0.02, 0.02)
	Mean	13.65* (9.57, 17.75)	−0.00 (−0.10, 0.08)	1.20 (−0.71, 3.20)	−0.15 (−1.21, 0.90)	−0.00 (−0.01, 0.02)
	Std dev	38.58* (34.94, 42.45)	0.57* (0.28, 1.08)	13.70* (11.99, 15.52)	7.71* (6.63, 8.70)	0.10* (0.08, 0.17)
Abs errors	Median	24.65* (20.82, 28.81)	0.38* (0.18, 0.74)	9.12* (7.56, 10.87)	5.08* (4.18, 6.06)	0.07* (0.05, 0.12)
	Mean	30.96* (27.89, 34.22)	0.45* (0.22, 0.86)	10.90* (9.48, 12.45)	6.11* (5.24, 6.93)	0.08* (0.06, 0.14)
	Max	162.64* (126.48, 208.30)	1.66* (0.78, 3.34)	42.64* (33.38, 57.17)	23.58* (18.43, 31.55)	0.29* (0.22, 0.53)
	Std dev	26.73* (23.78, 29.82)	0.34* (0.17, 0.65)	8.40* (7.24, 9.65)	4.71* (3.99, 5.41)	0.06* (0.05, 0.10)
Panel b: $A^L_{\Omega=\text{diag}(\xi^{-1})}, \gamma^0_\Omega = 1.01$						
Errors	Median	−6.16* (−10.19, −1.77)	2.10 (−0.38, 4.51)	−1.52 (−4.18, 1.03)	−1.44 (−5.32, 2.11)	−0.87 (−4.61, 2.47)
	Mean	−9.10* (−12.69, −4.57)	2.19* (0.06, 4.41)	−1.45 (−3.64, 0.66)	−1.39 (−4.79, 1.40)	−1.52 (−4.79, 1.29)
	Std dev	25.94* (24.01, 28.25)	12.20* (10.30, 14.79)	13.49* (11.06, 16.79)	20.49* (15.98, 25.39)	20.09* (15.28, 24.84)
Abs errors	Median	14.30* (12.00, 17.03)	8.32* (6.72, 10.42)	9.07* (7.11, 11.47)	13.85* (10.33, 18.06)	13.71* (9.93, 17.85)
	Mean	19.22* (17.43, 21.43)	9.89* (8.33, 12.08)	10.77* (8.86, 13.37)	16.36* (12.72, 20.54)	16.12* (12.13, 20.15)
	Max	134.34* (113.95, 156.42)	37.51* (28.71, 51.48)	42.91* (30.95, 61.69)	63.01* (45.10, 86.83)	60.52* (42.58, 82.37)
	Std dev	19.64* (17.47, 21.85)	7.54* (6.26, 9.30)	8.29* (6.70, 10.54)	12.41* (9.60, 15.46)	12.13* (9.15, 15.04)
Panel c: $A^L_{\Omega=\mathbf{DRD}'}, \tau^0_\Omega = 50, \gamma^0_\Omega = 5$						

Posterior medians of the median, mean, maximum and standard deviation of errors for the 1-, 3-, 12-, 36- and 60-month posterior discount rates. 95% credibility intervals are presented in parentheses and a * indicates that credibility interval does not include 0.

Table 3.3: Pricing errors (in basis points) statistics - covariance modeling (Continued).

	Maturity	1	3	12	36	60
Errors	Median		−4.33* (−6.99, −1.61)		−1.97 (−4.83, 0.90)	
	Mean		−4.44* (−6.47, −2.45)		−2.13 (−4.50, 0.36)	
	Std dev		19.58* (18.02, 21.22)		15.20* (13.73, 16.74)	
Abs errors	Median		13.12* (11.31, 15.18)		10.29* (8.72, 12.03)	
	Mean		15.75* (14.40, 17.21)		12.23* (10.99, 13.56)	
	Max		71.58* (54.73, 95.55)		46.20* (36.98, 61.44)	
	Std dev		12.43* (11.16, 13.75)		9.29* (8.25, 10.47)	
			Panel d: $A^P_{\Omega=\omega\mathcal{I}}$			
Errors	Median	−0.1	0.2	−0.0	−0.5	−0.3
	Std dev	4.9	9.2	4.9	5.6	4.5
Abs errors	Median	2.9	5.3	3.0	3.9	3.4
	Mean	3.7	7.0	3.7	4.6	3.7
	Max	21.4	41.1	22.7	18.0	11.9
	Std dev	3.1	5.9	3.2	3.3	2.6
			Panel e: PC			

Posterior medians of the median, mean, maximum and standard deviation of errors for the 1-, 3-, 12-, 36- and 60-month posterior discount rates. 95% credibility intervals are presented in parentheses and a * indicates that credibility interval does not include 0.

Table 3.4: Pricing errors (in basis points) statistics - precision modeling.

	Maturity	1	3	12	36	60
Errors	Median	−4.46* (−7.80, −1.11)	2.66* (0.86, 4.52)	−1.47 (−3.26, 0.32)	0.57 (−1.18, 2.38)	−0.28 (−2.13, 1.58)
	Mean	−7.26* (−10.30, −4.13)	2.75* (1.21, 4.33)	−1.36 (−2.87, 0.16)	0.65 (−0.85, 2.22)	−0.23 (−1.84, 1.36)
	Std dev	24.12* (22.62, 25.64)	10.25* (9.13, 11.43)	9.32* (8.30, 10.42)	9.31* (8.15, 10.74)	9.31* (8.12, 10.81)
Abs errors	Median	12.84* (10.93, 14.93)	7.09* (5.96, 8.34)	6.27* (5.23, 7.40)	6.19* (5.11, 7.46)	6.22* (5.09, 7.54)
	Mean	17.45* (16.09, 18.87)	8.45* (7.48, 9.49)	7.47* (6.60, 8.44)	7.40* (6.44, 8.59)	7.41* (6.42, 8.64)
	Max	124.52* (108.29, 142.26)	31.27* (25.09, 40.46)	29.71* (23.10, 40.07)	28.89* (22.50, 38.72)	28.40* (22.10, 38.58)
	Std dev	18.15* (16.57, 19.91)	6.43* (5.62, 7.35)	5.75* (5.02, 6.53)	5.71* (4.91, 6.69)	5.68* (4.88, 6.69)
Panel b: $A^L_{\Omega=\text{diag}(\xi^{-1})}, \gamma^0_\Omega = 2$						
Errors	Median	−4.20* (−7.70, −0.59)	2.73* (0.93, 4.57)	−1.57 (−3.44, 0.20)	0.73 (−1.12, 2.67)	−0.32 (−2.32, 1.63)
	Mean	−6.87* (−10.21, −3.25)	2.84* (1.34, 4.35)	−1.45 (−3.06, 0.06)	0.75 (−0.81, 2.44)	−0.27 (−1.99, 1.43)
	Std dev	23.82* (22.22, 25.40)	10.34* (9.12, 11.80)	9.46* (8.30, 10.73)	9.79* (8.34, 11.33)	9.87* (8.20, 11.50)
Abs errors	Median	12.76* (10.89, 14.88)	7.17* (5.97, 8.56)	6.37* (5.29, 7.63)	6.49* (5.27, 7.90)	6.57* (5.23, 8.02)
	Mean	17.26* (15.89, 18.69)	8.53* (7.47, 9.79)	7.60* (6.61, 8.70)	7.78* (6.60, 9.08)	7.85* (6.49, 9.20)
	Max	121.75* (103.44, 140.29)	31.61* (25.23, 41.15)	30.13* (23.24, 41.01)	30.08* (23.26, 40.29)	29.92* (22.59, 40.97)
	Std dev	17.78* (15.90, 19.76)	6.49* (5.62, 7.54)	5.83* (5.03, 6.73)	6.00* (5.04, 7.04)	6.01* (4.92, 7.12)
Panel b: $A^L_{\Omega=\text{diag}(\xi^{-1})}, \gamma^0_\Omega = 5$						
Errors	Median	−2.41 (−5.63, 0.74)	3.07* (1.18, 4.91)	−1.98* (−4.02, −0.01)	1.08 (−1.12, 3.39)	−0.35 (−2.71, 1.97)
	Mean	−4.61* (−7.27, −1.77)	3.26* (1.61, 4.80)	−1.86* (−3.59, −0.16)	1.08 (−0.76, 2.92)	−0.35 (−2.41, 1.64)
	Std dev	22.43* (20.95, 24.11)	10.95* (9.78, 12.20)	10.22* (9.17, 11.38)	10.97* (9.79, 12.32)	11.10* (9.76, 12.65)
Abs errors	Median	12.23* (10.43, 14.45)	7.62* (6.39, 8.87)	6.98* (5.84, 8.27)	7.38* (6.14, 8.76)	7.43* (6.20, 8.78)
	Mean	16.22* (14.90, 17.66)	9.08* (8.09, 10.12)	8.27* (7.38, 9.30)	8.77* (7.76, 9.93)	8.87* (7.77, 10.11)
	Max	110.72* (94.81, 129.22)	34.06* (27.29, 43.49)	32.99* (25.94, 45.27)	33.24* (26.76, 43.16)	33.20* (26.57, 44.08)
	Std dev	16.14* (14.84, 17.55)	6.93* (6.13, 7.92)	6.32* (5.58, 7.18)	6.69* (5.87, 7.63)	6.75* (5.86, 7.74)
Panel c: $A^L_{\Omega=\text{diag}(\xi^{-1})}, \gamma^0_\Omega = 50$						

Posterior medians of the median, mean, maximum and standard deviation of errors for the 1-, 3-, 12-, 36- and 60-month posterior discount rates. 95% credibility intervals are presented in parentheses and a * indicates that credibility interval does not include 0.

conclude that the proxy approach might not be too restrictive if one happens to pick the right rates to proxy latent factors. However, residual size is not necessarily the best metric to evaluate a model if the objective is extracting *common* factors from a panel of interest rates.

Introducing correlation in addition to a limited dispersion of precisions (Panel c) does not change the overall picture: the short rate still has larger and more variable residuals than longer rates. Note that a slightly more informative prior on precision dispersion, from $\gamma_\Omega^0 = 1.01$ (Panel b) to $\gamma_\Omega^0 = 5$ (Panel c), is sufficient to keep precisions within some common range. For example, the standard deviation of the 60-month residuals is 0.1 bp for $\gamma_\Omega^0 = 1.01$ while all five standard deviations are between 12 and 26 bp for $\gamma_\Omega^0 = 5$. Table 3.4 investigates this issue in more detail.

Principal components, introduced in the term structure literature by Litterman and Scheinkman (1991), are often presented as the standard benchmark and are indeed hard to beat in terms of various measures of error size, but the latent-factor modeling approach is definitely a serious competitor. Once one sets the short rate apart as a special mispriced rate, the models fare equally well with respect to most metrics. For example, the principal components model and the heteroscedastic model $A^L_{\Omega=\text{diag}(\xi^{-1})}$ (Panel b) yield residuals and absolute residuals with medians and standard deviations of the same order.

Table 3.4 presents a sensibility analysis with respect to the prior precision dispersion parameter, γ_Ω^0, when errors are uncorrelated. As γ_Ω^0 goes from 1.01 (Table 3.3, Panel b) to 2 (Panel a), the prior allows the standard deviation of the short rate to be singled out, as it gets more than twice as high as any other maturity. Also, the mean

residual on the 3-month rate is now significantly different from zero. Increasing γ_Ω^0 to 5 (Panel b) yield similar results: significant residual means for the short and 3-month rates, and a high short-rate residual standard deviation. But a further increase, to 50, somewhat changes the pattern: while the short-rate residual standard deviation is still higher than that of the other maturities, residual means are similar to those of the heteroscedastic model $A^L_{\Omega=\omega\mathcal{I}}$ (Panel a).

3.7.2 Cross-section properties

I next examine the correlations of pricing residuals. I consider posterior sample covariance matrices for models in which the covariance matrix is at most diagonal. Table 3.5 shows low but significant cross-correlations between some adjacent maturities for the benchmark homoscedastic model (Panel a). None of the correlations from the heteroscedastic model are significant when precision dispersion is a priori high (Panel b), but lower dispersion yields significant correlations between adjacent maturities (Panel c). One could argue that *all* the correlations from the proxying-factor model (Panel d) are significantly different from zero, but as there is only one such correlation, that would arguably be abusive. A larger number of rates would be necessary to reach any meaningful conclusion. Note that cross-correlations for all affine models are especially small compared with those from the principal components model (Panel e), which confirms that looking exclusively at measures of residual size does not tell the whole story.

That correlations decrease as one allows heteroscedasticity highlights the role of error modeling in the statistical decomposition of observables into common and idiosyncratic components. The observables' heteroscedasticity is thus partially captured by the idiosyncratic component, which allows factors to better capture other dimen-

sions. Here, factors better capture correlations. In terms of prior specification, the main message from Table 3.5 is the following. While a very relatively uninformative prior precision dispersion ($\gamma_{\Omega}^{0} = 1.01$) is compatible with uncorrelated errors, a slightly more informative prior ($\gamma_{\Omega}^{0} = 5$) is not. Therefore, if an informative prior is used in order to keep precisions within some common range, then the error model should allow for some degree of correlation.

3.7.3 Time-series properties

Table 3.6 shows the posterior error autocorrelations and partial autocorrelations for all maturities (rows) and the first 3 orders (columns) for models presented in Table 3.3. Principal components (Panel d) do not model dynamics and PC consequently presents the worst performance. In spite of the fact that model $A^{P}_{\Omega=\text{diag}(\xi^{-1})}$ is a "dynamic term structure model", it produces residuals that are as autocorrelated (Panel c). In both cases, patterns in the coefficients suggest high-order ARMA structures. Latent-factor modeling of pricing errors (Panels a and b) seems more in line with the error model's assumption of time serial independence, although there still exists some residual dynamics. These results are consistent with the implied richer VARMA(1,1) representation of latent-factor models. Comparing the i.i.d error model (Panel a) to the more general model with heteroscedastic and correlated errors (Panel b), prior precision dispersion and correlation modeling is likely to affect the residual dynamics. I investigate these issues in Tables 3.7 and 3.8.

Table 3.5: Sample covariance of pricing errors.

14.66* (13.14 , 16.28)				
−0.15* (-0.27 , -0.02)	14.66* (13.29 , 16.08)			
0.04 (-0.10 , 0.17)	−0.01 (-0.14 , 0.13)	12.79* (11.46 , 14.17)		
0.08 (-0.06 , 0.21)	−0.14* (-0.27 , -0.01)	0.13 (-0.01 , 0.26)	12.82* (11.49 , 14.17)	
−0.04 (-0.18 , 0.10)	0.07 (-0.07 , 0.20)	−0.02 (-0.15 , 0.12)	0.11 (-0.03 , 0.24)	12.80* (11.47 , 14.17)

Panel a: $A^L_{\Omega=\omega\mathcal{I}}$

38.60* (34.94 , 42.45)				
0.00 (-0.13 , 0.14)	0.58* (0.28 , 1.08)			
−0.11 (-0.23 , 0.02)	0.00 (-0.14 , 0.14)	13.72* (11.99 , 15.52)		
0.07 (-0.05 , 0.20)	−0.00 (-0.14 , 0.14)	−0.02 (-0.14 , 0.11)	7.69* (6.63 , 8.70)	
−0.00 (-0.14 , 0.14)	−0.00 (-0.14 , 0.14)	0.00 (-0.14 , 0.14)	−0.00 (-0.14 , 0.14)	0.11* (0.08 , 0.17)

Panel b: $A^L_{\Omega=\mathrm{diag}(\xi^{-1})}, \gamma^0_\Omega = 1.01$

23.82* (22.22 , 25.40)				
0.18* (0.07 , 0.29)	10.37* (9.12 , 11.80)			
−0.12* (-0.23 , -0.00)	0.26* (0.13 , 0.39)	9.47* (8.30 , 10.73)		
0.08 (-0.03 , 0.19)	−0.13 (-0.26 , 0.01)	0.22* (0.07 , 0.37)	9.79* (8.34 , 11.33)	
−0.02 (-0.14 , 0.10)	−0.08 (-0.21 , 0.06)	0.10 (-0.04 , 0.24)	0.60* (0.48 , 0.71)	9.84* (8.20 , 11.50)

Panel c: $A^L_{\Omega=\mathrm{diag}(\xi^{-1})}, \gamma^0_\Omega = 5$

19.59* (18.02 , 21.22)	
−0.15* (-0.27 , -0.03)	15.21* (13.73 , 16.74)

Panel d: $A^P_{\Omega=\omega\mathcal{I}}$

4.85				
-1.00	9.19			
0.90	-0.91	4.89		
0.57	-0.53	0.14	5.63	
-0.75	0.73	-0.38	-0.97	4.52

Panel e: PC

Posterior medians and 95%-inter-quantile credibility intervals for the standard-deviation (diagonal, in basis points) and correlations of pricing errors. A * indicates that credibility interval does not include 0

Table 3.6: Sample autocorrelations and partial autocorrelations of pricing errors.

	Autocorrelation			Partial autocorrelation		
Order	1	2	3	1	2	3
e_1	0.05 (−0.07, 0.18)	0.07 (−0.06, 0.19)	−0.01 (−0.14, 0.11)	0.05 (−0.07, 0.18)	0.06 (−0.06, 0.19)	−0.02 (−0.15, 0.11)
e_3	0.39* (0.30, 0.47)	0.24* (0.15, 0.34)	0.20* (0.10, 0.30)	0.39* (0.30, 0.47)	0.11* (0.02, 0.20)	0.09 (−0.00, 0.19)
e_{12}	0.23* (0.10, 0.36)	0.08 (−0.05, 0.22)	0.09 (−0.06, 0.21)	0.24* (0.11, 0.37)	0.02 (−0.10, 0.16)	0.06 (−0.07, 0.19)
e_{36}	0.38* (0.26, 0.48)	0.27* (0.15, 0.38)	0.22* (0.10, 0.34)	0.38* (0.26, 0.48)	0.14* (0.02, 0.26)	0.10 (−0.03, 0.22)
e_{60}	0.24* (0.10, 0.37)	0.15 (−0.00, 0.29)	0.16* (0.01, 0.31)	0.24* (0.10, 0.37)	0.09 (−0.05, 0.22)	0.11 (−0.03, 0.24)

Panel a: $A^L_{\Omega=\omega\mathcal{I}}$

	Autocorrelation			Partial autocorrelation		
Order	1	2	3	1	2	3
e_1	0.27* (0.18, 0.35)	0.11* (0.02, 0.20)	−0.06 (−0.15, 0.03)	0.27* (0.19, 0.35)	0.04 (−0.04, 0.12)	−0.11* (−0.18, −0.04)
e_3	0.27* (0.13, 0.42)	0.10 (−0.04, 0.24)	0.07 (−0.06, 0.21)	0.28* (0.13, 0.42)	0.02 (−0.10, 0.14)	0.04 (−0.09, 0.17)
e_{12}	0.23* (0.09, 0.36)	0.00 (−0.13, 0.13)	0.01 (−0.13, 0.15)	0.23* (0.09, 0.37)	−0.06 (−0.18, 0.07)	0.02 (−0.11, 0.15)
e_{36}	0.42* (0.28, 0.54)	0.13 (−0.01, 0.26)	0.05 (−0.06, 0.18)	0.42* (0.28, 0.54)	−0.07 (−0.16, 0.04)	0.03 (−0.07, 0.13)
e_{60}	0.36* (0.17, 0.50)	0.05 (−0.11, 0.20)	0.02 (−0.09, 0.15)	0.36* (0.18, 0.50)	−0.10 (−0.19, 0.00)	0.04 (−0.05, 0.15)

Panel b: $A^L_{\Omega=\mathbf{DRD}'}, \tau^0_\Omega = 50, \gamma^0_\Omega = 5$

	Autocorrelation			Partial autocorrelation		
Order	1	2	3	1	2	3
e_3	0.46* (0.42, 0.51)	0.33* (0.28, 0.39)	0.25* (0.19, 0.31)	0.46* (0.42, 0.51)	0.15* (0.13, 0.17)	0.07* (0.05, 0.10)
e_{36}	0.72* (0.71, 0.75)	0.55* (0.52, 0.60)	0.44* (0.40, 0.51)	0.72* (0.72, 0.75)	0.05* (0.02, 0.09)	0.06* (0.03, 0.09)

Panel c: $A^O_{\Omega=\omega\mathcal{I}}$

	Autocorrelation			Partial autocorrelation		
Order	1	2	3	1	2	3
e_1	0.46	0.33	0.24	0.46	0.15	0.07
e_3	0.44	0.32	0.23	0.44	0.15	0.07
e_{12}	0.34	0.23	0.17	0.34	0.13	0.07
e_{36}	0.74	0.57	0.48	0.74	0.07	0.07
e_{60}	0.71	0.55	0.44	0.71	0.09	0.06

Panel d: PC

Posterior medians and 95%-inter-quantile credibility intervals of pricing errors sample autocorrelations (right panels) and partial autocorrelations (left panels) for the first three orders (columns) in the 3-factor models, for each maturity (rows). A * indicates that credibility interval does not include 0.

Table 3.7 presents a sensibility analysis with respect to correlation prior specification. Comparing all models, it seems that correlations do not affect residual dynamics much when errors are heteroscedastic: there is little residual dynamics when errors are uncorrelated (Panel a). However, first-order partial autocorrelations do increase as prior correlations are less informative.

Table 3.8 considers the impact of precision dispersion prior specification on residual dynamics. An uninformative prior (Panel a) singles out two rates: the residuals on the 3-month and 60-month rates look serially independent. This suggests that two factors are closely associated with these maturities. This is confirmed by small mean residuals (Table 3.3, Panel b). In contrast, more informative priors (Panels b and c) yield residuals that have more similar dynamics across maturities. In the extreme case of homoscedastic errors (Table 3.6, Panel a), the short rate is singled out as a factor, which leaves no residual dynamics. However, other maturities have significant high-order residual dynamics, which further suggests that this factor contains information that is specific to the short rate.

Table 3.7: Sample autocorrelations and partial autocorrelations of pricing errors - Correlation modeling.

	Autocorrelation			Partial autocorrelation		
Order	1	2	3	1	2	3
e_1	0.24* (0.15, 0.32)	0.09* (0.01, 0.17)	−0.09 (−0.16, 0.00)	0.24* (0.15, 0.32)	0.04 (−0.03, 0.11)	−0.12* (−0.19, −0.05)
e_3	0.25* (0.11, 0.38)	0.10 (−0.04, 0.23)	0.10 (−0.03, 0.23)	0.25* (0.11, 0.38)	0.04 (−0.08, 0.15)	0.07 (−0.05, 0.19)
e_{12}	0.27* (0.14, 0.40)	0.10 (−0.04, 0.23)	0.07 (−0.07, 0.20)	0.28* (0.14, 0.40)	0.02 (−0.10, 0.14)	0.04 (−0.08, 0.16)
e_{36}	0.25* (0.13, 0.36)	0.11 (−0.02, 0.23)	0.10 (−0.02, 0.23)	0.25* (0.13, 0.36)	0.04 (−0.08, 0.16)	0.07 (−0.05, 0.19)
e_{60}	0.15* (0.01, 0.28)	0.01 (−0.13, 0.16)	0.09 (−0.04, 0.23)	0.15* (0.01, 0.28)	−0.02 (−0.15, 0.12)	0.10 (−0.03, 0.22)

Panel a: $A^L_{\Omega=\mathrm{diag}(\xi^{-1})}, \gamma^0_\Omega = 5$

	Autocorrelation			Partial autocorrelation		
Order	1	2	3	1	2	3
e_1	0.25* (0.17, 0.34)	0.10* (0.01, 0.19)	−0.08 (−0.16, 0.01)	0.25* (0.17, 0.34)	0.04 (−0.04, 0.11)	−0.12* (−0.19, −0.05)
e_3	0.25* (0.11, 0.38)	0.09 (−0.05, 0.23)	0.08 (−0.05, 0.21)	0.25* (0.11, 0.38)	0.03 (−0.09, 0.15)	0.06 (−0.07, 0.18)
e_{12}	0.20* (0.07, 0.33)	0.02 (−0.11, 0.16)	0.03 (−0.10, 0.17)	0.21* (0.07, 0.33)	−0.02 (−0.14, 0.10)	0.03 (−0.09, 0.16)
e_{36}	0.24* (0.12, 0.36)	0.03 (−0.09, 0.15)	0.03 (−0.09, 0.16)	0.25* (0.12, 0.36)	−0.04 (−0.15, 0.08)	0.04 (−0.08, 0.15)
e_{60}	0.14* (0.01, 0.26)	−0.07 (−0.19, 0.05)	0.02 (−0.10, 0.14)	0.14* (0.01, 0.26)	−0.10 (−0.21, 0.02)	0.04 (−0.08, 0.16)

Panel b: $A^L_{\Omega=\mathbf{DRD}'}, \tau^0_\Omega = 250, \gamma^0_\Omega = 5$

	Autocorrelation			Partial autocorrelation		
Order	1	2	3	1	2	3
e_1	0.25* (0.16, 0.33)	0.08 (−0.01, 0.17)	−0.10* (−0.19, −0.01)	0.25* (0.16, 0.33)	0.01 (−0.07, 0.09)	−0.13* (−0.21, −0.06)
e_3	0.34* (0.20, 0.46)	0.14 (−0.01, 0.28)	0.09 (−0.06, 0.23)	0.34* (0.20, 0.46)	0.03 (−0.10, 0.15)	0.04 (−0.09, 0.17)
e_{12}	0.31* (0.19, 0.42)	0.02 (−0.10, 0.15)	−0.03 (−0.15, 0.10)	0.32* (0.19, 0.42)	−0.09 (−0.20, 0.02)	−0.01 (−0.13, 0.11)
e_{36}	0.56* (0.47, 0.63)	0.26* (0.14, 0.37)	0.13* (0.02, 0.25)	0.56* (0.47, 0.64)	−0.08 (−0.16, 0.01)	0.03 (−0.06, 0.12)
e_{60}	0.53* (0.43, 0.62)	0.22* (0.09, 0.35)	0.12 (−0.00, 0.25)	0.53* (0.43, 0.62)	−0.08 (−0.17, 0.00)	0.06 (−0.03, 0.14)

Panel c: $A^L_{\Omega=\mathbf{DRD}'}, \tau^0_\Omega = 6, \gamma^0_\Omega = 5$

Posterior medians and 95%-inter-quantile credibility intervals of pricing errors sample autocorrelations (right panels) and partial autocorrelations (left panels) for the first three orders (columns) in the 3-factor models, for each maturity (rows). A * indicates that credibility interval does not include 0.

Table 3.8: Sample autocorrelations and partial autocorrelations of pricing errors - Precision modeling.

	Autocorrelation			Partial autocorrelation		
Order	1	2	3			
e_1	0.17* (0.04, 0.29)	0.10 (−0.02, 0.22)	0.03 (−0.09, 0.16)	0.17* (0.04, 0.29)	0.07 (−0.06, 0.19)	0.01 (−0.12, 0.13)
e_3	−0.01 (−0.14, 0.13)	−0.00 (−0.14, 0.13)	−0.00 (−0.14, 0.13)	−0.01 (−0.14, 0.13)	−0.01 (−0.15, 0.13)	−0.00 (−0.14, 0.13)
e_{12}	0.20* (0.07, 0.33)	0.11 (−0.03, 0.24)	0.10 (−0.03, 0.24)	0.20* (0.07, 0.33)	0.07 (−0.06, 0.20)	0.08 (−0.06, 0.21)
e_{36}	0.19* (0.05, 0.32)	0.10 (−0.03, 0.24)	0.12 (−0.01, 0.26)	0.19* (0.05, 0.33)	0.06 (−0.06, 0.19)	0.10 (−0.04, 0.23)
e_{60}	−0.00 (−0.14, 0.13)	−0.01 (−0.14, 0.13)	−0.00 (−0.14, 0.13)	−0.00 (−0.14, 0.13)	−0.01 (−0.15, 0.13)	−0.00 (−0.14, 0.13)

Panel a: $A^L_{\Omega=\text{diag}(\xi^{-1})}, \gamma^0_\Omega = 1.01$

	Autocorrelation			Partial autocorrelation		
Order	1	2	3	1	2	3
e_1	0.24* (0.15, 0.32)	0.09* (0.01, 0.17)	−0.09 (−0.16, 0.00)	0.24* (0.15, 0.32)	0.04 (−0.03, 0.11)	−0.12* (−0.19, −0.05)
e_3	0.25* (0.11, 0.38)	0.10 (−0.04, 0.23)	0.10 (−0.03, 0.23)	0.25* (0.11, 0.38)	0.04 (−0.08, 0.15)	0.07 (−0.05, 0.19)
e_{12}	0.27* (0.14, 0.40)	0.10 (−0.04, 0.23)	0.07 (−0.07, 0.20)	0.28* (0.14, 0.40)	0.02 (−0.10, 0.14)	0.04 (−0.08, 0.16)
e_{36}	0.25* (0.13, 0.36)	0.11 (−0.02, 0.23)	0.10 (−0.02, 0.23)	0.25* (0.13, 0.36)	0.04 (−0.08, 0.16)	0.07 (−0.05, 0.19)
e_{60}	0.15* (0.01, 0.28)	0.01 (−0.13, 0.16)	0.09 (−0.04, 0.23)	0.15* (0.01, 0.28)	−0.02 (−0.15, 0.12)	0.10 (−0.03, 0.22)

Panel b: $A^L_{\Omega=\text{diag}(\xi^{-1})}, \gamma^0_\Omega = 5$

	Autocorrelation			Partial autocorrelation		
Order	1	2	3	1	2	3
e_1	0.18* (0.11, 0.27)	0.06 (−0.02, 0.13)	−0.11* (−0.18, −0.03)	0.19* (0.11, 0.27)	0.02 (−0.05, 0.10)	−0.13* (−0.20, −0.06)
e_3	0.30* (0.19, 0.43)	0.15* (0.01, 0.27)	0.13 (−0.00, 0.26)	0.30* (0.19, 0.43)	0.06 (−0.06, 0.16)	0.08 (−0.04, 0.20)
e_{12}	0.28* (0.16, 0.40)	0.10 (−0.03, 0.22)	0.07 (−0.06, 0.20)	0.29* (0.16, 0.41)	0.02 (−0.10, 0.14)	0.04 (−0.08, 0.16)
e_{36}	0.24* (0.13, 0.35)	0.08 (−0.04, 0.21)	0.09 (−0.03, 0.21)	0.24* (0.13, 0.35)	0.02 (−0.10, 0.13)	0.07 (−0.05, 0.19)
e_{60}	0.17* (0.03, 0.29)	0.01 (−0.14, 0.15)	0.09 (−0.04, 0.24)	0.17* (0.03, 0.29)	−0.02 (−0.16, 0.10)	0.10 (−0.03, 0.23)

Panel c: $A^L_{\Omega=\text{diag}(\xi^{-1})}, \gamma^0_\Omega = 50$

Posterior medians and 95%-inter-quantile credibility intervals of pricing errors sample autocorrelations (right panels) and partial autocorrelations (left panels) for the first three orders (columns) in the 3-factor models, for each maturity (rows). A * indicates that credibility interval does not include 0.

While this clearly merits further investigation, with perhaps a larger number of rates, it seems that a general error model with relatively informative priors can help the econometrician affect the decomposition of observables into common components and idiosyncratic errors. In particular, for the data set considered here, hyperparameter values $\tau_\Omega^0 = 50$ and $\gamma_\Omega^0 = 5$ yield residuals that are roughly consistent with the error model, although they still leave some residual dynamics. More generally, the following facts emerge:

1. Low prior correlations are inconsistent with residual correlations when prior heteroscedasticity is low (Table 3.5, Panel c);

2. Less informative correlation priors increase first-order residual partial autocorrelations (Table 3.7, Panel c);

3. Low prior heteroscedasticity increases residual dynamics for all maturities except the short rate (Table 3.8, Panels a and c).

4. Less informative precision dispersion priors yield rate-specific factors (Table 3.3, Panel b);

3.8 Concluding remarks

Modeling observational errors on all discount rates is desirable on both theoretical and empirical grounds, and is computationally feasible. Assuming that some rates are observed without error is observationally restrictive and yields residuals with high variances, cross-correlations and autocorrelations. Because factor models decompose observable dynamics into common and idiosyncratic components, error modeling is not inferentially innocuous. Extreme error modeling choices illustrate the relevant issues: the likelihood function is singular if all rates are observed without error, while

the model is globally unidentified if the error dynamics are as rich as those of the factors.

Between the extremes, the econometrician has considerable room for modeling common and idiosyncratic components. For example, modeling heteroscedastic errors highlights that some rates are better proxying factor candidates than others. Because these errors capture part of observable heteroscedasticity, factors can better capture cross-correlations and autocorrelation. However, modeling heteroscedastic errors shares some drawbacks with the proxying-factor approach: factors better describe some discount rates at the expense of others. I show how an informative heteroscedasticity prior specification mitigates this problem and yield factors describing features that are common to the entire panel of discount rate rather than a small subset thereof. In addition, modeling low cross-correlations through an informative prior helps further reduce residuals autocorrelations.

Inference for affine models is complicated by weak identification problems, which make the Bayesian methodology particularly appealing for at least two reasons. Because one may have to evaluate a large number of normalizations before one that yields estimators with good finite-sample properties is found, the fact that ML estimator sampling distributions must be obtained by simulations methods makes this approach computationally prohibitive. In contrast, normalizations can be compared at little computational cost using a sample from the un-normalized posterior distribution. Moreover, Bayesian inference for observational errors does not rely on biased parameter point estimators and therefore provides valid diagnostics of model adequacy. These computational and inferential considerations make the proposed methodology an ideal candidate for empirical macroeconomic work, especially with relatively small data sets, of the order of a few hundred months or quarters.

I leave a number of important empirical questions unanswered. How binding are parameter restrictions that yield factors with level, slope and curvature interpretations? Does the role of error modeling changes as one observes a larger number of maturities? As for the modeling of factors, the scale normalization and short rate factor loadings parameterization proposed in this paper yields a parameter, σ, which can be interpreted as the factors's common variance. This parameterization lends itself to the specification of a simple stochastic volatility model with a single common volatility factor.

3.9 Appendix A - Prior distribution hyperparameters

Table 3.9: Prior distribution parameters for the 3-factor models.

Parameter	Value
$\mu^0_{A_1}$	4e-003
$\Sigma^0_{A_1}$	1e-005
$\mu^0_{\sigma_0}$	-5
$\Sigma^0_{\sigma_0}$	1e+003
$\mu^0_{\lambda_0}$	0
$\Sigma^0_{\lambda_0}$	100
μ^0_{γ}	0
Σ^0_{γ}	5
$\mu^0_{\kappa^{\mathbb{P}}}$	0
$\Sigma^0_{\kappa^{\mathbb{P}}}$	5
τ	10
ν^0_{Ω}	1.01
γ^0_{Ω}	1.01
β^0_{Ω}	1e-013

3.10 Appendix B - Solution to the pricing difference equation

The solution to the pricing difference equation (3.10) is due to Ang and Piazzesi (2003) and is provided here for completeness. Assume the solution is of the form $P_{n,t} = \exp\{\tilde{A}_n + \tilde{B}'_n X_t\}$,

$$
\begin{aligned}
P_{n,t} &= \exp\{\tilde{A}_1 + \tilde{\mathbf{B}}'_1 X_t\}\mathbf{E}^{\mathbb{Q}}_t[P_{n-1,t+1}] \\
\exp\{\tilde{A}_n + \tilde{\mathbf{B}}'_n X_t\} &= \exp\{\tilde{A}_1 + \tilde{\mathbf{B}}'_1 X_t\}\mathbf{E}^{\mathbb{Q}}_t[\exp\{\tilde{A}_{n-1} + \tilde{\mathbf{B}}'_{n-1} X_{t+1}\}] \\
&= \exp\{\tilde{A}_1 + \tilde{\mathbf{B}}'_1 X_t\} \\
&\qquad \times \exp\{\tilde{A}_{n-1} + \tilde{\mathbf{B}}'_{n-1}\left(X_t + \kappa^{\mathbb{Q}}(\theta^{\mathbb{Q}} - X_t)\right) + \frac{1}{2}\tilde{\mathbf{B}}'_{n-1}\Sigma\tilde{\mathbf{B}}_{n-1},
\end{aligned}
$$

and match the coefficients to get the recursions (3.13).

3.11 Appendix C - From physical drift to risk-neutral drift in a conditionally Gaussian model with log-linear SDF

This proof is based on Dai, Singleton, and Yang (2005). Since the price of a *any* cash flow c_{t+1} can be calculated under both measure, i.e.

$$\mathbf{E}^{\mathbb{P}}_{X_{t+1}|X_t}\left[\mathrm{e}^{m_{t+1}}c_{t+1}\right] = \mathrm{e}^{-y_{1,t}}\mathbf{E}^{\mathbb{Q}}_{X_{t+1}|X_t}\left[c_{t+1}\right],$$

we can identify the risk-neutral measure,

$$\mathrm{d}\mathbb{Q} = \mathrm{e}^{y_{1,t}+m_{t+1}}\mathrm{d}\mathbb{P}.$$

We can then compute the risk-neutral trend,

$$\begin{aligned}
\mu_t^{\mathbb{Q}} &\equiv \mathbf{E}^{\mathbb{Q}}_{X_{t+1}|X_t}\left[X_{t+1}\right] \\
&= \int_{\mathbb{R}^K} X_{t+1} \exp\left\{-\Lambda_t\left(X_{t+1}-\mu_t^{\mathbb{P}}\right) - \frac{1}{2}\Lambda_t\Sigma_t\Lambda_t'\right\} \mathrm{d}\mathbb{P} \\
&= \exp\left\{\Lambda_t\mu_t^{\mathbb{P}} - \frac{1}{2}\Lambda_t\Sigma_t\Lambda_t'\right\} \\
&\quad \times \int_{\mathbb{R}^K} X_{t+1}\mathrm{e}^{-\Lambda_t X_{t+1}} \mathrm{d}\mathbb{P} \\
&= \exp\left\{\Lambda_t\mu_t^{\mathbb{P}} - \frac{1}{2}\Lambda_t\Sigma_t\Lambda_t'\right\} \\
&\quad \times \frac{\partial}{\partial\Lambda_t} \int_{\mathbb{R}^K} \mathrm{e}^{-\Lambda_t X_{t+1}} \mathrm{d}\mathbb{P} \\
&= -\exp\left\{\Lambda_t\mu_t^{\mathbb{P}} - \frac{1}{2}\Lambda_t\Sigma_t\Lambda_t'\right\} \\
&\quad \times \frac{\partial}{\partial\Lambda_t} \exp\left\{-\Lambda_t\mu_t^{\mathbb{P}} + \frac{1}{2}\Lambda_t\Sigma_t\Lambda_t'\right\} \\
&= \mu_t^{\mathbb{P}} - \Sigma_t\Lambda_t'. \qquad (3.43)
\end{aligned}$$

3.12 Appendix D - VARMA-representation of yields

To simplify exposition, consider an N-factor model, where there are just as many latent factors as there are observed yields. The K-factor model, with $K < N$ can then viewed as a constrained N-factor model. I first rewrite (3.16) with $\nu_t \equiv \Sigma^{1/2}\epsilon_t$

$$X_{t+1} = X_t + \kappa^{\mathbb{P}}(\theta^{\mathbb{P}} - X_t) + \nu_{t+1}$$

and (3.19) with $\zeta_t \equiv \Omega^{1/2}u_t$

$$y_t = \mathbf{A} + \mathbf{B}'X_t + \zeta_t.$$

When yields are observed without any measurement error and the model is assumed to be perfect, one can inverse the pricing equations to solve for the yields and obtain

$$y_{t+1} = (I - \mathbf{B}'(I - \kappa^{\mathbb{P}})\mathbf{B}'^{-1})\mathbf{A} + \mathbf{B}'(I - \kappa^{\mathbb{P}})\mathbf{B}'^{-1}y_t + \mathbf{B}'\nu_{t+1}$$

When all yields are subject to measurement with errors or when the model describes reality imperfectly, one obtains the same VAR(1) process but with measurement errors in the variables

$$(y_{t+1} - \zeta_{t+1}) = (I - \mathbf{B}'(I - \kappa^{\mathbb{P}})\mathbf{B}'^{-1})\mathbf{A} + \mathbf{B}'(I - \kappa^{\mathbb{P}})\mathbf{B}'^{-1}(y_t - \zeta_t) + \mathbf{B}'\nu_{t+1}$$

Such a process is equivalent to a VARMA(1,1) (Box, Jenkins, and Reinsel, 1994).

3.13 Appendix E - Inverse-Gamma-mixture of Gammas

Our hierarchal prior for $\xi \equiv \text{diag}\,(\Omega^{-1})$ is a Inverse-Gamma-mixture of Gamma densities. Specifically,

$$p(\xi|\gamma,\nu,\beta) = \int_0^\infty \prod_{n=1}^{N} G\left(\xi_n|\gamma,\frac{\eta}{\gamma}\right) IG\,(\eta|\nu,\beta)\ \mathrm{d}\eta$$

with

$$G\left(\xi_n|\gamma,\frac{\eta}{\gamma}\right) = \frac{(\frac{\gamma}{\eta})^{\gamma}}{\Gamma(\gamma)}\Xi_n^{\gamma-1}\exp\left(-\frac{\gamma}{\eta}\Xi_n\right)$$

and

$$IG\,(\eta|\nu,\beta) = \frac{\beta^{\nu}}{\Gamma(\nu)}\frac{\exp\,(-\beta/\eta)}{\eta^{\nu+1}}.$$

One can write this mixture in closed form as

$$p(\xi|\gamma,\nu,\beta) = \frac{p(\Xi,\eta|\gamma,\nu,\beta)}{p(\eta|\Xi,\nu,\beta)}$$

since

$$\begin{aligned} p(\eta|\Xi,\gamma,\nu,\beta) &\propto p(\Xi|\gamma,\eta)p(\eta|\nu,\beta) \\ &\propto \left(\frac{\gamma}{\eta}\right)^{N\gamma} \exp\left(-\frac{\gamma}{\eta}\sum_{n=1}^{N}\Xi_n\right)\frac{\exp\left(-\beta/\eta\right)}{\eta^{\nu+1}} \\ &\propto IG\Big(\eta\Big|N\gamma+\nu,\beta+\gamma\sum_{n=1}^{N}\Xi_n\Big). \end{aligned}$$

Explicitly,

$$p(\xi|\gamma,\nu,\beta) = \frac{\gamma^{N\gamma}\beta^{\nu}\Gamma(N\gamma+\nu)}{\Gamma(\nu)\Gamma(\gamma)^N}\frac{\prod_{n=1}^{N}\xi_n^{\gamma-1}}{\left(\beta+\gamma\sum_{n=1}^{N}\xi_n\right)^{N\gamma+\nu}}$$

The mean of ξ_n is

$$\begin{aligned} \mathbf{E}[\xi_n] &= \int_0^\infty \xi_n \left[\int_0^\infty G\left(\xi_n|\gamma,\frac{\eta}{\gamma}\right) IG\left(\eta|\nu,\beta\right)\,\mathrm{d}\eta\right]\,\mathrm{d}\xi_n \\ &= \int_0^\infty \left[\int_0^\infty \xi_n G\left(\xi_n|\gamma,\frac{\eta}{\gamma}\right)\,\mathrm{d}\xi_n\right] IG\left(\eta|\nu,\beta\right)\,\mathrm{d}\eta \\ &= \int_0^\infty \eta IG\left(\eta|\nu,\beta\right)\,\mathrm{d}\eta \\ &= \frac{\beta}{\nu-1} \end{aligned}$$

Note that this prior is conditionally conjugate in a Gaussian model.

3.14 Appendix G - Principal components

To build orthogonal factors, one takes the singular-value decomposition of the yield sample covariance matrix

$$\hat{\Sigma}_y^2 = \delta\gamma\delta'.$$

For $K < N$ principal components, consider the K first columns of δ, call it δ_K and take

$$\hat{y}_{PC} = \bar{y} + (y - \bar{y})\delta_K\delta_K'$$

where $\bar{y}$ is the sample mean.

CHAPITRE 4

A STATISTICAL ARBITRAGE STRATEGY ON THE TERM STRUCTURE OF INTEREST RATES.

Abstract

This article develops and implements a Bayesian decision framework for constructing fixed income statistical arbitrage strategies. Statistical arbitrage exploits temporary deviations between market prices and fundamental values given by an economic model. The Bayesian decision framework is ideally suited for statistical arbitrage strategy construction and risk evaluation as it allows an investor to combine economic theory and prior beliefs about reasonable parameter values with data in order to produce predictive densities. In contrast to a normal approximation of risk that limits possible strategies to functions of mean and covariance, predictive densities are rich objects that give the flexibility to formulate complex risk and return objectives in the form of a utility function. In addition, Bayesian predictive densities are robust to weak permutation and reflection identification problems that affect inference in affine term structure models and can be more accurate than forecasts based on maximum-likelihood parameter point estimates (Blais, 2008b,a).

I illustrate this framework using a simple affine term structure model of Government of Canada bond prices. In order to bet on temporary market price deviations from those implied by this model, I consider portfolios that are first-order hedged with respect to latent factors. The optimal portfolio maximizes expected gains, subject to the constraint that the initial price deviation is sufficiently large to cover reasonable execution costs. I find that this simple strategy can be profitable for large institutional investors.

JEL classification: C11; C5; C32; G11; G12

4.1 Introduction

This article develops and implements a Bayesian decision framework for constructing fixed income statistical arbitrage strategies. *Statistical arbitrage* differs from *risk-free arbitrage*, which is simply referred to as *arbitrage* in the academic literature. An arbitrage portfolio is one that generates non-negative gains with certainty and requires an initial investment of at most zero. In contrast, statistical arbitrage exploits temporary deviations between market prices and fundamental values given

by an economic model. Positive gains from investing in such mispriced portfolios are expected on average, but are not certain.

In practice, the cumulative gain over a long investment horizon is not the only relevant characteristic of an arbitrage strategy. Although risk is diversified away over time, short horizons do matter to an investor. For example, his bonus could be based on his annual performance.

If an investor adheres to a number of axioms[1], his preferences have an expected utility representation and strategy construction consists of the following maximization problem:

$$\pi^* = \arg\max_{\pi \in \mathcal{I}} \int u(z, \pi)\ f(z)\ \mathrm{d}z, \tag{4.1}$$

where $u(\cdot)$ is the Bernoulli utility function representing the investor's preferences over monetary gains, z is a vector of random state variables with predictive density $f(z)$, and $\mathcal{I}$ is a set of strategies available to the investor.

The specification of these three components determines whether this maximization problem is feasible: the set of strategies over which maximization is done, the predictive density used in order to compute the expectation, and the objective (utility) function. In order to obtain a practical solution, one could specify a simple utility function over gains or a simple econometric model for z. The approach I take in this paper is considering a relatively small, finite set of strategies $\mathcal{I}$. This allows me to use sophisticated utility functions and econometric models while finding optimal strategies at a small computational cost.

[1]See Mas-Colell et al. (1995) for an formal exposition. The investor's preferences over portfolios would have to be complete, transitive, continuous and satisfy the independence axiom.

A simple utility function is unappealing if it does not represent the investor's preferences with enough truthfulness, as quadratic functions that treat downside and upside risks symmetrically. A simple econometric model will yield unreliable predictive densities and unreliable risk measures if it does not capture important features of the data. In contrast, the worst consequence of considering a simple set of strategies is that the optimal strategy might have little interest: there might be no profitable strategy in a relatively small set. However, I favor this latter approach, admitting that finding a good set can be challenging.

The Bayesian decision framework is ideally suited for statistical arbitrage strategy construction and risk evaluation. The first reason is theoretical coherence: information from economic theory and the investor's prior beliefs about reasonable parameter values are combined with the data to obtain predictive densities in a manner that rests on solid decision-theoretic foundations. In contrast to a normal approximation of risk that limits possible strategies to functions of mean and covariance, predictive densities are rich objects that allow an investor to formulate complex risk and return objectives in the form of a utility function. The predictive density is, by definition, the investor's best guess about the distribution of future gains. While academics are sometimes reluctant to rely on prior knowledge, no fixed income portfolio manager would consider parameter values that he considers highly implausible. Indeed, some investors rely primarily on prior beliefs. The Bayesian framework is simply the probabilistic way to use such beliefs in an optimal manner. A second reason is practical. In Blais (2008b,a), I demonstrate that Bayesian predictive densities are robust to weak permutation and reflection identification problems that affect inference in affine term structure models. In contrast, standard ML parameter point estimators can be severely biased and asymptotic

sampling distributions unreliable.

My objective is illustrative of how one can cast an investment decision problem into a Bayesian expected-utility framework, and I am not committed to a particular econometric model or utility function. I apply this framework to Government of Canada bond prices. I use a simple affine term structure model to estimate the predictive density of the term structure of discount rates. As bond prices are functions of discount rates, the predictive density of bond portfolio values is easily obtained by Monte Carlo Markov Chain simulations.

I consider a finite set of strategies which builds on the factor-risk neutral strategy proposed by Balı, Heidari, and Wu (2006) (BHW). Their statistical arbitrage strategy involves three steps: building a portfolio that is first-order factor-risk neutral (delta-hedged) to the strongly persistent dynamic factors but fully exposed to pricing errors; buying this portfolio in proportion to its risk-adjusted price deviation from model-implied values; and holding the position for four weeks. They provide some evidence that this strategy can generate significant cumulative gains out of sample over several months of consecutive weekly trading. In order to take into account execution costs that may vary with market conditions, I express profitability in terms of gains per basis point of portfolio bid-ask spreads. I consider the delta-hedged portfolio which maximizes expected gains, among those that are a posteriori likely to be sufficiently mispriced at inception to cover execution costs.

This paper is structured as follows. In the second section, I present the econometric model that I use to obtain the predictive density of bond portfolios. The third section describes a flexible parametric family of utility functions. In the forth section,

I explain how to obtain first-order factor-risk neutral portfolios, which I refer to delta-hedged portfolios, and how I take into account execution costs. I present the data set and empirical results in the final section of this paper.

4.2 Inference for fixed income portfolios

In this section, I present the econometric model that I use to estimate the predictive density of bond portfolio values. The economic model is a discrete-time affine term structure model, which is a Gaussian linear state-space model where the coefficients of the state equation are subject to non-linear restrictions imposed by an economic model. These constraints substantially reduce the number of free parameters to be estimated and can produce better forecasts than unconstrained linear state-space models (Ang and Piazzesi, 2003).

4.2.1 An affine term structure model

The no-arbitrage factor model I use in this paper rests on three assumptions (See Blais, 2008a, for a detailed presentation). First, the short rate $Z_{1,t}$ (the one-period risk-free rate) is an affine function of a K-dimensional vector of latent factors X_t,

$$Z_{1,t} = A_1 + B_1' X_t. \tag{4.2}$$

Second, I assume there are no arbitrage opportunities. This implies that there exists (See Cochrane, 2005) a risk-neutral probability measure $\mathbb{Q}$ under which the price $D_{n,t}$ of a discount bond maturing n periods from t is equal to its conditional expected future value,

$$D_{n,t} = \mathbf{E}_t^{\mathbb{Q}} \left[D_{n-1,t+1} \right]. \tag{4.3}$$

The risk-neutral measure is defined through the factors dynamics, which are first-order Gaussian vector autoregressive,

$$X_t = X_{t-1} + \kappa^{\mathbb{Q}}(\theta^{\mathbb{Q}} - X_{t-1}) + \epsilon_t, \tag{4.4}$$

where ϵ_{t+1} is K-dimensional vector of serially independent Gaussian random variables with covariance Σ. Given (4.2) and (4.4), the solution to (4.3) is $D_{n,t} = \exp\{-Z_{n,t}n\}$, where

$$\begin{aligned} Z_{n,t} &= \frac{\tilde{A}_n}{n} + \frac{\tilde{B}_n}{n}' X_t, \\ &= A_n + B_n' X_t, \\ \text{with} \quad \tilde{A}_{n+1} &= \tilde{A}_1 + \tilde{A}_n + (\kappa^{\mathbb{Q}}\theta^{\mathbb{Q}})'\tilde{B}_n - \frac{1}{2}\tilde{B}_n'\Sigma\tilde{B}_n, \\ \text{and} \quad \tilde{B}_{n+1} &= \tilde{B}_1 + (\mathrm{I} - \kappa^{\mathbb{Q}'})\tilde{B}_n. \end{aligned}$$

The economic model is fully specified by (4.2-4.4). I now describe the statistical model, which consists of the specification of the rate dynamics under the physical measure, the normalization of the parameter space for the purposes of identification, and the specification of prior distributions.

Under the physical measure, factors have first-order Gaussian vector autoregressive dynamics,

$$X_t = X_t + \kappa^{\mathbb{P}}(\theta^{\mathbb{P}} - X_{t-1}) + \epsilon_t. \tag{4.5}$$

I observe a N-dimensional vector of discount rates z_t measured with errors,

$$z_t = Z_t + e_t = A + B'X_t + e_t,$$

where u_t is a vector of serially independent Gaussian random variables with covariance matrix Ω.

For future reference, let θ denote the parameter

$$\theta \equiv \left\{A_1, B_1, \theta^{\mathbb{Q}}, \kappa^{\mathbb{Q}}, \kappa^{\mathbb{P}}, \Sigma, \Omega, X_1\right\} \in \Theta;$$

and $l(\theta|z)$, the likelihood function.

Affine state space models are identified up to affine transformations of the factors: for any invertible $K \times K$ matrix M, there exists a bijection $g_M(\theta) : \Theta \rightarrow \Theta$ such that $l(g_M(\theta)|z) = l(\theta|z)$ (Dai and Singleton, 2000). Examples of affine transformation are rotations, scalings, translations, permutations and reflections across the coordinate axes. The model is said to be *invariant* with respect to these transformations, and one normalizes the parameter space for the purpose of identification. In particular, permutation and reflection invariance implies that the likelihood function has $K!2^K$ symmetric and perfectly equivalent modes (Blais, 2008a). In this paper, $\theta^{\mathbb{P}} = 0$, $\kappa^{\mathbb{P}}$ is diagonal, $\kappa^{\mathbb{Q}}$ is arbitrary, Σ is a correlation matrix. This normalization breaks invariance with respect to translation, rotation and scaling, but preserves permutation and reflection invariance.

Because the affine term structure model decomposes discount rates dynamics into common (X_t) and idiosyncratic (e_t) components, I argue in (Blais, 2008a) that error modeling, *i.e.* the specification of Ω, has critical implications for this decomposition. In particular, I argue that modeling moderately heteroscedastic errors can produce better forecasts than modeling homoscedastic or highly heteroscedastic errors. Also, modeling correlated errors can reduce the latent factors' explanatory poser. Accordingly, here, errors are uncorrelated but moderately heteroscedastic. I achieve this modeling with the following hierarchical prior for $\xi \equiv \text{diag}\,(\Omega^{-1})$, which is an Inverse-Gamma-scale-

mixture of vectors of conditionally independent Gamma distributions. Specifically,

$$f(\xi|\gamma_\Omega^0,\nu_\Omega^0,\beta_\Omega^0) \;=\; \int_0^\infty \prod_{n=1}^N \mathcal{G}\left(\xi_n|\gamma_\Omega^0,\frac{\eta}{\gamma_\Omega^0}\right)\mathcal{IG}\left(\eta|\nu_\Omega^0,\beta_\Omega^0\right)\,\mathrm{d}\eta.$$

One can integrate η out and write this mixture in closed form as (Blais, 2008a)

$$f(\xi|\gamma_\Omega^0,\nu_\Omega^0,\beta_\Omega^0) \;=\; \frac{{\gamma_\Omega^0}^{N\gamma_\Omega^0}{\beta_\Omega^0}^{\nu_\Omega^0}\Gamma(N\gamma_\Omega^0+\nu_\Omega^0)}{\Gamma(\nu_\Omega^0)\Gamma(\gamma_\Omega^0)^N}\frac{\prod_{n=1}^N \xi_n^{\gamma_\Omega^0-1}}{\left(\beta_\Omega^0+\gamma_\Omega^0\sum_{n=1}^N\xi_n\right)^{N\gamma_\Omega^0+\nu_\Omega^0}}. \quad (4.6)$$

This prior gives the flexibility to express prior knowledge about the common scale of the variances and their dispersion, independently. A large value of γ_Ω^0 corresponds to a strong belief that errors are close to being homoscedastic. Here, 250 is a relatively large value which implies that the prior 95%-inter-quantile range of $\frac{\Omega_{ii}-\Omega_{jj}}{\Omega_{ii}}$, $i \neq j$, is approximately $[-0.18, 0.18]$. In other words, error variances are within 18% of each other with prior probability 0.95. A small value of ν_Ω^0 expresses little prior knowledge about the level of variances, while β_Ω^0 is a scale parameter. Priors for the other parameters and hyper-parameter values are given in Appendix 4.6.

4.2.2 A word on notation

For notational convenience, I use subscripts in place of function arguments when this causes no confusion. For example, I use $\dot{Z}_{n,t} \equiv Z_n(X_t) \equiv Z(X_t, A_n, B_n)$. Also, I drop subscripts of certain variables in order to define vectors and matrices, *e.g.* Z_t denotes the vector $[Z_{1,t} : Z_{N,t}]'$, while Z denotes the matrix $[Z_1 : Z_t]$.

Throughout, $f(\cdot)$ stands for a probability density function. For future reference, Table 4.1 lists the main indices used in this paper.

Table 4.1: Summary of indices.

Indices	
$i = 1, \ldots, I = 16$	Bonds used to form portfolios.
$n = 1, \ldots, N = 30$	Discount rate maturities; the observables.
$t = 1, \ldots, T = 362$	Time periods; weeks.
$l = 1, \ldots, L = \frac{I}{K+1} = 1820$	subsets of $K+1$-bond chosen from $\{1, \ldots, I\}$; portfolios.
$m = 1, \ldots, M = 500\,000$	Iterations of the posterior simulator.
$k = 1, \ldots, K = 3$	Latent factors.
$h = 1, \ldots, H = 10$	Holding periods.

4.2.3 Misspecification problems

This state-space model does not capture important features of rate dynamics. In particular, estimated pricing errors are autocorrelated. This is not surprising as one does not expect a small number of factors (here $K = 3$) with relatively simple dynamics (4.5) to *fully* capture the dynamics of a large cross-section ($N = 30$) of discount rates. One could consider a larger number of factors, with perhaps richer dynamics, but a certain degree of market segmentation (Modigliani and Sutch, 1966, 1967) is likely to yield persistent maturity-specific errors. For example, some bonds provide liquidity services for which investors willingly pay a premium over bonds with otherwise identical characteristics.

Modeling persistent errors gives rise to identification issues. However, from a Bayesian perspective, proper priors ensure proper posteriors and parameters are *identified* in that specific sense. For example, one could model errors as a first-order vector autoregressive process,

$$e_t = \Phi e_{t-1} + v_t, \tag{4.7}$$

with N-dimensional serially independent $v_t \sim \mathcal{N}(0, \Xi)$, and specify strongly informative priors on Φ and Ξ. Because the inference objective is extracting common factors that capture most of discount rate dynamics, priors should be informative about the unconditional error covariance matrix Υ defined through $\Upsilon = \Phi\Upsilon\Phi' + \Xi$. One could therefore specify priors on (Φ, Υ), and use a prior on Υ that is concentrated around $\varphi^2 I$, with φ in the order of a few basis points. This could be achieved in a flexible manner through the decomposition of Υ into a diagonal matrix of variances and a correlation matrix. The variance matrix could have a prior of the form (4.6), while the prior proposed by Barnard, McCulloch, and Meng (2000) (see Appendix 4.6) could be used for the correlation matrix.

This approach would require a comprehensive sensitivity analysis and I use an approximation in this paper. For each iteration m of the posterior simulator, I compute the ordinary least squares estimates $\Phi^{(m)}$ and $\Xi^{(m)}$ of (4.7) for $e_{1:t}^{(m)}$. I denote the distribution of this sample by $\tilde{f}(\Phi, \Xi | e_{1:t})$, where the notation $\tilde{f}$ indicates that this distribution is approximate.

I thus approximate the predictive density in the following manner. Let $\tilde{\theta}$ denote the following parameters:

$$\tilde{\theta} \equiv \left\{ A_1, B_1, \theta^{\mathbb{Q}}, \kappa^{\mathbb{Q}}, \kappa^{\mathbb{P}}, \Sigma, X_1 \right\} \in \tilde{\Theta}.$$

I define

$$\tilde{f}(z_{t+h}, X_{t+h}, e_{t+h-1}, \theta, \Phi, \Xi | z_{1:t}) \equiv$$
$$\int f(z_{t+h} | \tilde{\theta}, \Phi, \Xi, X_{t+h}, e_{t+h-1}) f(X_{t+h} | z_{1:t}, \theta)$$
$$\times f(e_{t+h-1} | \Phi, \Xi, e_{1:t}) \tilde{f}(\Phi, \Xi | e_{1:t}) f(e_{1:t} | z_{1:t}, \theta) f(\theta | z_{1:t}) \, \mathrm{d}e_{1:t}.$$

Using this approximate posterior, I compute

$$\tilde{f}(z_{t+h}, \theta, \Phi, \Xi | z_{1:t}) \approx \int \tilde{f}(z_{t+h}, X_{t+h}, e_{t+h-1}, \theta, \Phi, \Xi | z_{1:t}) \, \mathrm{d}X_{t+h} \, \mathrm{d}e_{t+h-1}.$$

In what follows, I do not make the distinction between $f(z_{t+h}, \theta | z_{1:t})$ and $\tilde{f}(z_{t+h}, \tilde{\theta}, \Phi, \Xi | z_{1:t})$, and I use the former for exposition clarity.

4.3 Bayesian statistical arbitrage

In order to make an investment decision of the form (4.1), an investor would formulate a utility function over possible gains and losses. This can be challenging in practice. One approach is selecting a parametric family of utility functions and determining parameter values by introspection. For example, an investor could find a particular value γ such that the power utility function $u(g_{t,h}) = g_{t,h}^{1-\gamma}$ best approximates his preferences, where a larger value corresponds to a higher level of risk aversion.

Other parametric families might be more easily interpretable. For example, the function

$$u(g_{t,h}) = \gamma_1 g_{t,h} + \gamma_2 \left(g_{t,h} - \mu\right)^2 + \gamma_3 \mathbf{1}\left(g_{t,h} < \alpha\right)$$

yields the optimal portfolio

$$\pi_{l,t}^{*} = \arg\max_{\pi \in \mathcal{I}} \gamma_1 \mu + \gamma_2 \sigma^2 + \gamma_3 \mathrm{Prob}\left(g_{t,h} < \alpha | z_{1:t}\right), \tag{4.8}$$

where $\mu = \mathbf{E}[g_{t,h}|z_{1:t}]$, $\sigma = \mathbf{E}[(g_{t,h} - \mu)^2 | z_{1:t}]^{1/2}$ and $\mathbf{1}(A) = 1$ if A is true and 0 otherwise. The investor would determine values of $\gamma_{1:3}$ and α which best approximate his relative preferences over portfolios in terms of expectation, variance and the probability that gains are below α. Arguably, such values of $\gamma_{1:3}$ and α can be difficult to obtain. In this paper, I set $\gamma_1 = 1$ and $\gamma_2 = \gamma_3 = 0$.

4.4 Strategy set

Investing is a dynamic problem: it involves both investment and disinvestment decisions. Often, the timing of transactions is not know in advance and will most likely depend on future market developments. For instance, if the position is large relative to daily trading volumes, an investor may opt to open and close the position in a sequence of small transactions over several days. From a game theoretic point of view, a strategy is a complete state-contingent plan (see Mas-Colell et al., 1995, for an introduction). In simple words, a strategy consists of all the trades one would execute to open and close the positions, specified for *all* possible scenarios at *all* times in the future.

Although this definition is a reasonable description of how portfolio managers think about investing, I make a number of simplifications in order to obtain a mathematically tractable decision problem. In this section, I restrict the set of strategies that are considered, $\mathcal{I}$.

First, I consider a market with I of bonds. Thus, portfolios are I-dimensional vectors $\pi_t \in \mathbb{R}^I$. Next, I assume that trading takes place once per period and a maximum investment horizon of H periods. For simplicity, I also assume that all relevant information is contained in discount rates z_t, but this is not a substantive restriction. In this setting, a strategy is a $I\times(H+1)$-dimensional matrix-valued function

$$\pi_{t,0:H} \equiv [\pi_{t,0}, \ldots, \pi_{t,H}], \tag{4.9}$$

whose columns are I-dimensional vector-valued functions of the N discount rates observed at time t. Thus, the information that becomes available after the initial investment is not used in constructing the strategy. Note the following notation abuse: I use π for both portfolios and strategies. This is justified and should cause

no confusion as a portfolio is the value a strategy function takes for a particular argument.

Despite these simplifying assumptions, the set of strategies described above is still uncountably large and I propose a finite subset of strategies in this section in order to make the numerical maximization problem practical. In this section, I first describe BHM's strategy, which defines a finite set of delta-hedged portfolios. Second, because my objective is exploiting temporary valuation gaps between market and model-implied values, I restrict attention to portfolios that are a posteriori likely to be mispriced. Finally, I describe how I take execution costs into account.

Other restrictions could be imposed on the set of strategies. For example, institutional portfolio managers are typically subject to a number of restrictions in terms of risk exposures: value-at-risk, issuer, industry sector, country, credit or deviation-from-benchmark limits. These could be operationalized as restrictions on the set of strategies available to the investor in a straightforward manner.

4.4.1 Delta-hedged portfolios

My dynamic term structure model decomposes interest rate dynamics into highly persistent common factors and serially independent idiosyncratic errors. In fact, factor dynamics are relatively close to random walks and therefore difficult to predict with any accuracy. Balı, Heidari, and Wu (2006) make the point that statistical arbitrage might be profitable. In order to bet on temporary deviations from model-impled values, they consider portfolios that are delta-hedged (first-order factor-risk neutral) with respect to the persistent factors. As they need only $K+1$ bonds[2] to hedge K factors, they note that they can build $L = \binom{I}{K+1}$ delta-hedged portfolios, which defines a set of portfolios

[2]BHW used swaps instead of bonds, but swaps are mathematically equivalent to bonds trading at par.

that they can use to evaluate the empirical performance of their idea. They find that the average (over L) risk-neutral portfolio yields significant gains on average (*i.e.* over time).

Their set of delta-hedged portfolios is finite, which make it a good candidate for numerical optimization. In particular, maximization over a finite set involves no numerical derivative computation. A delta-hedged portfolio is a $K+1$-dimensional vector $\pi_{l,t}$ which satisfies

$$\frac{\partial}{\partial X'} \pi'_{l,t} p_l(A + B'X + e_t) \Big|_{X=X_t} = \mathbf{0}$$

where p_l denotes a $K+1$-dimensional vector of bond price functions

$$p_{l,t} = p_l(A + B'X_t + e_t) = p_l(Z_t + e_t), \tag{4.10}$$

and l keeps track of the bonds maturities and coupons[3].

In order to write this portfolio concisely, let $H_{l,t}$ denote the first-order derivative of (4.10) with respect to factor X_t,

$$H_{l,t}(\theta, X_t) = \frac{\partial}{\partial X'} p_l (A + B'X + e_t) \Big|_{X=X_t}.$$

The first-order factor-risk neutral portfolios is then

$$\pi_{l,t} \equiv \Delta_l(\theta, X_t)^{-1}\mathbf{e} \tag{4.11}$$

with

$$\Delta_l(\theta, X_t) \equiv \begin{bmatrix} H_{l,t}(\theta, X_t) & \mathbf{e} \end{bmatrix}',$$

[3]I use Mathworks' Financial Toolbox to compute prices, accrued interests and coupon payments, using standard Canadian market conventions and calendar.

$\mathbf{e}$ is the K+1-dimensional vector $\mathbf{e} \equiv [0 \ldots 0\,1]'$, and where the last element of π is 1 by normalization.

Delta-hedged portfolios are functions of the model parameter vector θ and factor levels X_t, both of which are unknown. BHW plug in their ML parameter estimate $\hat{\theta}$. They report that $H_{l,t}$ does not vary much with X_t, and thus plug in a constant factor value $\bar{X} \equiv \frac{1}{T}\sum_{t=1:T} \hat{X}_t$, which is the sample mean of the smoothed factors. Their portfolios,

$$\pi_l^{BHW} \equiv \Delta_l(\hat{\theta}, \bar{X})^{-1}\mathbf{e}, \tag{4.12}$$

are therefore constant over time.

BHW's portfolios are delta-hedged if $\theta = \hat{\theta}_{MLE}$ and $X_t = \bar{X}$. In fact, because θ and X_t are unknown, so is $\pi_{l,t}$ in (4.11). Looking for delta-hedged portfolios can be formalized as a decision problem where the objective function is the sum of squared sensitivities:

$$\begin{aligned}
\pi_{l,t}^{Bayes} &= \arg\min_{\pi\in\mathbb{R}^I} \int (\Delta(\theta, X_t)\pi - \mathbf{e})' (\Delta(\theta, X_t)\pi - \mathbf{e})\, f\left(\theta, X_t \middle| z_{1:t}\right) \mathrm{d}\theta\, \mathrm{d}X_t, \\
&= \int \pi_{l,t}(\theta, X_t)\, f\left(\theta, X_t \middle| z_{1:t}\right) \mathrm{d}\theta\, \mathrm{d}X_t,
\end{aligned}$$

where $f\left(\theta, X_t \middle| z_{1:t}\right)$ is the joint posterior of θ and X_t, and the second line follows from the fact that the posterior expectation is the Bayesian estimator associated with quadratic objective functions (Robert, 2001). These $L \times T$ portfolios are thus a posteriori delta-hedged at any given time. For future reference, let Π_t denote the set of portfolios thus defined.

4.4.2 Mispriced portfolios with execution costs

BHM propose to invest in proportion of the difference between the market value and the estimated model-impled value, which they scale by the variance of the estimate. At each trading period, they invest in proportion to

$$\omega_{l,t} \equiv \frac{\pi_l^{BHW'}\left(p_l(z_t) - p_l(\hat{Z}_t)\right)}{\hat{\mathbf{Var}}\left[\pi^{BHW'}\left(p_l(z_t) - p_l(\hat{Z}_t)\right)\right]}, \tag{4.13}$$

where $\hat{\mathbf{Var}}[X]$ is the sample variance of X[4]. Intuitively, they invest in proportion to a risk-adjusted measure of market mispricing.

Price differentials do not ensure gains on average: price differentials must be significantly larger than execution costs. In bond markets, these are typically expressed in basis point of yield. For example, one could buy a bond i at price $p_i(z_t - \delta_{i,t})$ and sell it at $p_i(z_t + \delta_{i,t})$, reflecting a bid-ask spread of $2\delta_{i,t}$. Bid-ask spreads are bond-specific and change over time. Rather than assuming a particular bid-ask spread value, I express relevant quantities in basis points of a common bid-ask spread (bpbas), for expositional clarity. For a portfolio $\pi_{l,t}$, the cost in bpbas, $c_{l,t}$ is

$$c_{l,t} \equiv c(\pi_{l,t}^{Bayes}, z_t) = \text{abs}\left(\pi_{l,t}^{Bayes}\right)' \left(p_l(z_t - \delta) - p_l(z_t + \delta)\right),$$

where δ is one half of a basis point. Let $d_{l,t}$ denote the valuation differential for the delta-hedged portfolio $\pi_{l,t}$,

$$d_{l,t}(\theta, X_t) \equiv \frac{\pi_{l,t}^{Bayes'}\left(p_l(Z_t) - p_l(z_t)\right)}{c_{l,t}},$$

where explicit dependence on θ and X_t stresses that fact that valuation differentials depend on these unknown quantities. An investor would like $d_{l,t}$ to be above the

[4]The authors do not justify the use of variance instead of standard deviation, which would yield a unit-free measure.

execution costs he faces in the actual market conditions.

Similarly, I define net gains in bpbas as

$$g_{t,h}(z_{t+h}, \pi_{l,t}^{Bayes}) \equiv \frac{\pi_{l,t}^{Bayes'}\left(p_l(z_{t+h}) - p_l(z_t) - m_{l,t}\right)}{c_{l,t}}, \tag{4.14}$$

where the term $m_{l,t}$ captures the financing of cash balances related to initial investment and coupon re-investment. I assume a constant rate r applies to both long and short cash positions.

In this paper, I restrict the strategy set to portfolios whose posterior probability that the valuation differential is larger than the costs an investor would incur to get in and out of the position at 0.25 bpbas per way is at least 50%. I denote the set of portfolios satisfying this criteria by

$$\mathcal{I}_t \equiv \left\{\pi_{l,t} \in \Pi_t \,\middle|\, l = 1:L,\ \text{Prob}\left(d_{l,t} > 0.5 \middle| z_{1:t}\right) > 0.5\right\}. \tag{4.15}$$

I will refer to portfolios in $\mathcal{I}_t$ as *mispriced portfolios*. A particular portfolios is excluded, for example, if the posterior density of its value at time t is too vague. It turns out that the predictive density of this portfolio value if often equally vague, which makes this investment relatively risky. The opposite relation also holds: portfolios that are mispriced with certainty are typically risk-less portfolios with limited upside possibilities.

4.4.3 Investment horizon

Based on the estimated persistence of estimated errors, $\hat{u}_t = z_t - \hat{Z}_t$, BHW propose an investment horizon of four weeks, which is somewhat longer than the half-life of errors based on their average first-order autocorrelation.

In terms of the general strategies defined in (4.9), BHW's is relatively simple and has the following form:

$$\pi_{t,0:4}^{BHW} = \left[\omega_{l,t}\pi_l^{BHW} \quad \mathbf{0} \quad \mathbf{0} \quad \mathbf{0} \quad -\omega_{l,t}\pi_l^{BHW}\right],$$

for *any* combination l of $K+1$ bonds. In particular, the decision to open a position only depends on a risk-adjusted measure of price differential (4.13), without consideration of expected gains. The decision to close the position depends on the half-life of pricing errors, without consideration of whether any residual pricing differential is expected after three, four, or more weeks.

I maximize, over mispriced portfolios, expected gains per bpbas after a holding horizon of h periods,

$$\pi_{t,h}^{*} = \max_{\pi \in \mathcal{I}_t} \int g_{t,h}(z_{t+h}, \pi)\, f\left(z_{t+h} \middle| z_{1:t}\right) \mathrm{d}z_{t+h}, \tag{4.16}$$

The strategy I use in the empirical part of this paper can be thus summarized:

1. Form $L = \binom{I}{K+1}$ delta-hedged portfolio (4.13);
2. Find those that are likely to be sufficient mispriced to cover execution costs (4.15);
3. Select an investment horizon h;
4. Invest in portfolio (4.16);
5. Close the position after h weeks.

In terms of the general strategies defined in (4.9), this strategy has the following form:

$$\pi_{t,0:h}^{*} = \left[\pi_{t,h}^{*} \quad \underbrace{\mathbf{0} \quad \cdots \quad \mathbf{0}}_{h-1 \text{ times}} \quad -\pi_{t,h}^{*}\right].$$

Instead of selecting a fixed investment horizon in Step 3, one can treat h as an additional variable in the optimization problem. I did not adopt this approach but I present detailed

empirical results for all horizons up to 10 weeks.

4.5 Empirical results

I build discount rates by bootstrapping (see Campbell, Lo, and MacKinlay, 1997) Government of Canada bond closing prices from PC-Bond's DEX indices, for 362 Wednesdays from 16 January 2001 to 18 December 2007. When multiple quotes are available, I consider the bond whose coupon is closest to its yield. This is a standard procedure which takes into account the fact that bonds trading with large premia or discounts are relatively old, illiquid issues. I obtain 30 discount rates for each date, with maturities from 1 to 30 years.

I use discount rates as observables in order to simplify computations, which constitutes a common approximation (see Dai and Singleton, 2003, for a survey). For computation considerations, I do not update parameter posteriors as new information becomes available after the first investment period. I use the first 311 weeks to compute posterior distributions and keep the remaining 51 for out-of-sample evaluation.For $t > 311$, errors are introduced because $f(\theta|z_{1:t}) \neq f(\theta|z_{1:311})$. Accuracy losses as t becomes larger than 311 could thus suggest that parameters are unstable. I simulate posterior distribution using the method I describe in Blais (2008a).

4.5.1 Out of sample performance

I report statistics on the out-of-sample performance of sets of mispriced portfolios ($\mathcal{I}_t$) and optimal ($\pi^*_{t,h}$) portfolios. Computations apply a rate of 4.375% to long and short cash positions $m_{l,t}$ in (4.14), with daily compounding. This is the average internal financing rate that applied at CDP Capital for the period considered, which is assumed to be known in advance.

To present my results, I use the following quantities:

$$\begin{aligned}
\bar{g}_{t,h} &= \frac{1}{\#(\mathcal{I}_t)} \sum_{\pi \in \mathcal{I}_t} g_{t,h}(z_{t+h}, \pi), \\
\hat{\rho}_{t,h} &= \frac{1}{\#(\mathcal{I}_t)} \sum_{\pi \in \mathcal{I}_t} \mathbf{1}(g_{t,h}(z_{t+h}, \pi) > 0), \\
\hat{\alpha}_{t,h} &= \min_{\alpha} \left\{ \frac{1}{\#(\mathcal{I}_t)} \sum_{\pi \in \mathcal{I}_t} \mathbf{1}(g_{t,h}(z_{t+h}, \pi) < \alpha) \geq 0.05 \right\}, \\
g^*_{t,h} &= g_{t,h}(z_{t+h}, \pi^*_{t,h}), \\
p^*_{t,h} &= \text{Prob}(g^*_{t,h} < g_{t,h}(z_{t+h}, \pi^*_{t,h}) | z_{1:t}), \\
\alpha^*_{t,h} &= \inf_{\alpha} \left\{ \text{Prob}(g_{t,h}(z_{t+h}, \pi^*_{t,h}) < \alpha | z_{1:t}) \geq 0.05 \right\},
\end{aligned}$$

where $\#(\mathcal{A})$ is the number of elements in set $\mathcal{A}$.

The first three statistics pertain to mispriced portfolio sets $\mathcal{I}_t$, and are respectively the sample mean ($\bar{g}_{t,h}$), sample proportion ($\hat{\rho}_{t,h}$) of portfolios with positive gains and the 5^{th} percentile ($\hat{\alpha}_{t,h}$) of the empirical distribution. The fourth statistic is the actual gains from the optimal portfolios ($g^*_{t,h}$). The fifth quantity is the estimated predictive cumulative probability ($p^*_{t,h}$) of realized gains (the posterior counterpart of the p-value), which gives an indication of risk measurement accuracy. The final quantity is the posterior 5%-value-at-risk of optimal portfolio gains.

Table 4.2: Mispriced portfolio statistics

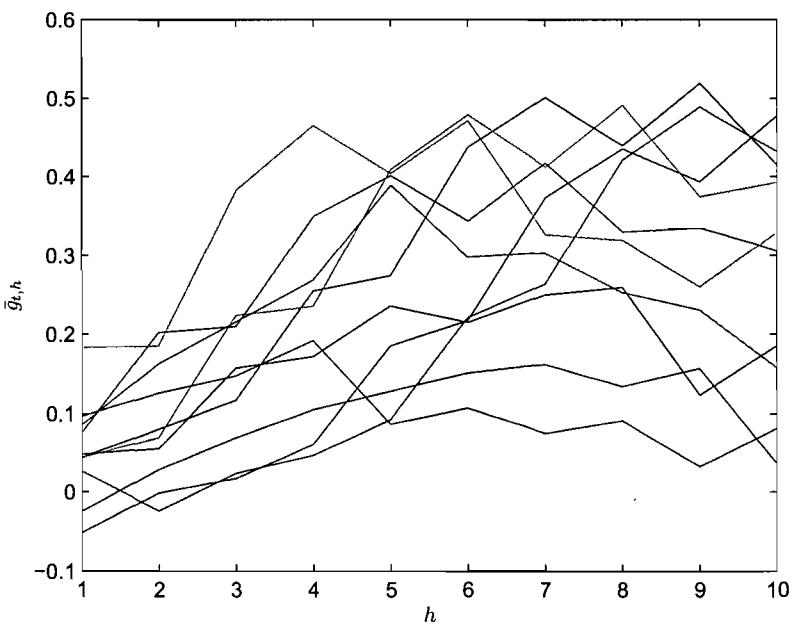
	h=	1	2	3	4	5	6	7	8	9	10
$t = 311$	$\bar{g}_{t,h}$	0.027	-0.024	0.024	0.048	0.092	0.221	0.264	0.421	0.489	0.432
	$\hat{\rho}_{t,h}$	0.713	0.420	0.603	0.593	0.604	0.897	0.840	0.896	0.907	0.912
	$\hat{\alpha}_{t,h}$	-0.078	-0.221	-0.152	-0.142	-0.134	-0.041	-0.116	-0.082	-0.106	-0.032
$t = 312$	$\bar{g}_{t,h}$	-0.052	-0.001	0.018	0.061	0.185	0.217	0.373	0.435	0.393	0.478
	$\hat{\rho}_{t,h}$	0.261	0.491	0.434	0.518	0.920	0.810	0.906	0.927	0.917	0.934
	$\hat{\alpha}_{t,h}$	-0.176	-0.176	-0.101	-0.115	-0.033	-0.057	-0.032	-0.081	-0.064	-0.022
$t = 313$	$\bar{g}_{t,h}$	0.045	0.081	0.117	0.255	0.274	0.438	0.501	0.440	0.519	0.415
	$\hat{\rho}_{t,h}$	0.687	0.726	0.788	0.973	0.965	0.978	0.938	0.889	0.924	0.780
	$\hat{\alpha}_{t,h}$	-0.065	-0.075	-0.070	0.022	0.012	0.032	-0.017	-0.121	-0.164	-0.102
$t = 314$	$\bar{g}_{t,h}$	0.044	0.070	0.224	0.236	0.409	0.479	0.412	0.491	0.374	0.392
	$\hat{\rho}_{t,h}$	0.602	0.693	0.946	0.887	0.955	0.931	0.888	0.914	0.712	0.814
	$\hat{\alpha}_{t,h}$	-0.103	-0.087	-0.003	-0.025	0.008	-0.048	-0.132	-0.174	-0.148	-0.093
$t = 315$	$\bar{g}_{t,h}$	0.076	0.203	0.210	0.349	0.401	0.343	0.418	0.330	0.335	0.306
	$\hat{\rho}_{t,h}$	0.707	0.943	0.943	0.946	0.898	0.873	0.896	0.778	0.898	0.769
	$\hat{\alpha}_{t,h}$	-0.168	-0.011	-0.012	-0.006	-0.130	-0.214	-0.273	-0.156	-0.029	-0.097
$t = 316$	$\bar{g}_{t,h}$	0.183	0.185	0.383	0.466	0.404	0.471	0.326	0.319	0.260	0.329
	$\hat{\rho}_{t,h}$	0.939	0.935	0.969	0.969	0.931	0.964	0.760	0.843	0.708	0.695
	$\hat{\alpha}_{t,h}$	-0.017	-0.014	0.037	0.041	-0.042	0.034	-0.108	-0.068	-0.113	-0.241
$t = 317$	$\bar{g}_{t,h}$	0.086	0.163	0.216	0.269	0.389	0.298	0.303	0.253	0.231	0.157
	$\hat{\rho}_{t,h}$	0.617	0.793	0.822	0.866	0.921	0.680	0.686	0.663	0.590	0.582
	$\hat{\alpha}_{t,h}$	-0.110	-0.096	-0.128	-0.091	-0.072	-0.211	-0.160	-0.223	-0.354	-0.383
$t = 318$	$\bar{g}_{t,h}$	0.097	0.126	0.148	0.192	0.087	0.107	0.075	0.091	0.033	0.082
	$\hat{\rho}_{t,h}$	0.652	0.786	0.791	0.882	0.626	0.668	0.584	0.522	0.529	0.536
	$\hat{\alpha}_{t,h}$	-0.144	-0.277	-0.115	-0.129	-0.138	-0.105	-0.135	-0.299	-0.304	-0.277
$t = 319$	$\bar{g}_{t,h}$	-0.025	0.029	0.070	0.105	0.128	0.151	0.162	0.134	0.157	0.037
	$\hat{\rho}_{t,h}$	0.530	0.689	0.752	0.595	0.638	0.631	0.676	0.687	0.682	0.523
	$\hat{\alpha}_{t,h}$	-0.197	-0.203	-0.253	-0.218	-0.153	-0.262	-0.382	-0.425	-0.399	-0.582
$t = 320$	$\bar{g}_{t,h}$	0.049	0.056	0.157	0.172	0.236	0.215	0.250	0.260	0.123	0.186
	$\hat{\rho}_{t,h}$	0.653	0.660	0.655	0.774	0.763	0.769	0.777	0.779	0.709	0.714
	$\hat{\alpha}_{t,h}$	-0.253	-0.194	-0.222	-0.167	-0.259	-0.379	-0.424	-0.396	-0.577	-0.556
$t = 321$	$\bar{g}_{t,h}$	0.005	0.071	0.106	0.149	0.226	0.235	0.237	0.176	0.236	0.326
	$\hat{\rho}_{t,h}$	0.586	0.727	0.762	0.749	0.774	0.784	0.776	0.729	0.741	0.772
	$\hat{\alpha}_{t,h}$	-0.150	-0.387	-0.303	-0.416	-0.558	-0.602	-0.585	-0.770	-0.734	-0.641
$t = 322$	$\bar{g}_{t,h}$	0.100	0.103	0.161	0.247	0.247	0.240	0.184	0.200	0.278	0.308
	$\hat{\rho}_{t,h}$	0.734	0.750	0.738	0.770	0.760	0.778	0.699	0.672	0.715	0.691
	$\hat{\alpha}_{t,h}$	-0.318	-0.236	-0.355	-0.483	-0.522	-0.483	-0.682	-0.646	-0.558	-0.915
$t = 323$	$\bar{g}_{t,h}$	0.066	0.054	0.134	0.128	0.134	0.102	0.110	0.147	0.072	0.137
	$\hat{\rho}_{t,h}$	0.864	0.629	0.747	0.729	0.722	0.669	0.662	0.722	0.576	0.640
	$\hat{\alpha}_{t,h}$	-0.059	-0.081	-0.190	-0.237	-0.202	-0.409	-0.373	-0.321	-0.695	-1.035
$t = 324$	$\bar{g}_{t,h}$	0.010	0.081	0.076	0.074	-0.007	0.022	0.091	-0.032	-0.010	-0.212
	$\hat{\rho}_{t,h}$	0.424	0.584	0.565	0.576	0.463	0.510	0.598	0.446	0.452	0.418
	$\hat{\alpha}_{t,h}$	-0.140	-0.276	-0.335	-0.304	-0.502	-0.487	-0.440	-0.974	-1.242	-1.744
$t = 325$	$\bar{g}_{t,h}$	0.050	0.038	0.050	-0.029	-0.008	0.033	-0.134	-0.130	-0.328	-0.292
	$\hat{\rho}_{t,h}$	0.631	0.646	0.683	0.428	0.468	0.597	0.360	0.431	0.360	0.357
	$\hat{\alpha}_{t,h}$	-0.258	-0.255	-0.191	-0.438	-0.392	-0.526	-1.117	-1.368	-1.693	-2.056
$t = 326$	$\bar{g}_{t,h}$	0.048	0.088	-0.066	-0.005	0.011	-0.160	-0.398	-0.616	-0.472	-0.374
	$\hat{\rho}_{t,h}$	0.613	0.695	0.197	0.436	0.567	0.344	0.277	0.285	0.338	0.308
	$\hat{\alpha}_{t,h}$	-0.114	-0.094	-0.255	-0.249	-0.332	-0.907	-1.299	-1.710	-1.800	-1.392
$t = 327$	$\bar{g}_{t,h}$	0.036	-0.067	-0.064	-0.065	-0.323	-0.396	-0.665	-0.614	-0.468	-0.368
	$\hat{\rho}_{t,h}$	0.733	0.185	0.199	0.447	0.180	0.253	0.244	0.247	0.211	0.228
	$\hat{\alpha}_{t,h}$	-0.029	-0.251	-0.220	-0.373	-0.910	-1.316	-1.613	-1.789	-1.367	-0.991
$t = 328$	$\bar{g}_{t,h}$	-0.129	-0.121	-0.109	-0.432	-0.599	-0.914	-0.817	-0.639	-0.520	-0.585
	$\hat{\rho}_{t,h}$	0.098	0.044	0.380	0.098	0.108	0.085	0.163	0.125	0.115	0.146
	$\hat{\alpha}_{t,h}$	-0.303	-0.248	-0.478	-0.940	-1.457	-1.655	-1.853	-1.413	-1.055	-1.271
$t = 329$	$\bar{g}_{t,h}$	0.056	0.083	-0.097	-0.265	-0.445	-0.273	-0.182	-0.118	-0.154	-0.294
	$\hat{\rho}_{t,h}$	0.818	0.729	0.455	0.306	0.348	0.442	0.400	0.435	0.425	0.359
	$\hat{\alpha}_{t,h}$	-0.111	-0.186	-0.747	-1.144	-1.557	-1.628	-1.221	-0.925	-1.156	-1.081
$t = 330$	$\bar{g}_{t,h}$	0.027	-0.184	-0.349	-0.477	-0.245	-0.164	-0.052	-0.018	-0.195	-0.334
	$\hat{\rho}_{t,h}$	0.645	0.324	0.222	0.241	0.453	0.410	0.455	0.507	0.391	0.341
	$\hat{\alpha}_{t,h}$	-0.250	-0.681	-1.144	-1.471	-1.596	-1.167	-0.873	-1.092	-1.104	-1.263

Table 4.2: Mispriced portfolio statistics (continued)

	h=	1	2	3	4	5	6	7	8	9	10
$t = 331$	$\bar{g}_{t,h}$	-0.163	-0.264	-0.378	-0.198	-0.127	0.031	0.043	-0.108	-0.191	-0.191
	$\hat{\rho}_{t,h}$	0.285	0.290	0.380	0.448	0.411	0.510	0.507	0.483	0.378	0.391
	$\hat{\alpha}_{t,h}$	-0.584	-1.019	-1.435	-1.493	-1.079	-0.843	-1.025	-1.067	-1.140	-1.140
$t = 332$	$\bar{g}_{t,h}$	-0.080	-0.171	-0.147	-0.003	0.148	0.166	-0.016	-0.079	-0.030	0.174
	$\hat{\rho}_{t,h}$	0.260	0.334	0.411	0.411	0.546	0.514	0.425	0.357	0.352	0.603
	$\hat{\alpha}_{t,h}$	-0.611	-0.942	-0.962	-0.546	-0.385	-0.520	-0.611	-0.748	-0.711	-0.400
$t = 333$	$\bar{g}_{t,h}$	0.000	0.257	0.360	0.509	0.565	0.393	0.205	0.205	0.414	0.367
	$\hat{\rho}_{t,h}$	0.589	0.596	0.735	0.830	0.769	0.696	0.617	0.631	0.895	0.864
	$\hat{\alpha}_{t,h}$	-0.573	-0.702	-0.337	-0.184	-0.319	-0.353	-0.483	-0.455	-0.095	-0.235
$t = 334$	$\bar{g}_{t,h}$	0.248	0.389	0.507	0.519	0.368	0.192	0.203	0.468	0.403	0.935
	$\hat{\rho}_{t,h}$	0.684	0.875	0.956	0.919	0.860	0.754	0.772	0.823	0.790	0.968
	$\hat{\alpha}_{t,h}$	-0.631	-0.236	0.022	-0.105	-0.171	-0.491	-0.365	-0.309	-0.473	0.129
$t = 335$	$\bar{g}_{t,h}$	0.235	0.410	0.374	0.281	0.262	0.280	0.523	0.423	0.673	0.456
	$\hat{\rho}_{t,h}$	0.863	0.910	0.860	0.799	0.743	0.781	0.849	0.823	0.882	0.913
	$\hat{\alpha}_{t,h}$	-0.128	-0.112	-0.130	-0.221	-0.304	-0.223	-0.311	-0.469	-0.259	-0.074
$t = 336$	$\bar{g}_{t,h}$	0.199	0.167	0.071	0.105	0.107	0.328	0.246	0.435	0.280	0.103
	$\hat{\rho}_{t,h}$	0.749	0.634	0.582	0.511	0.465	0.847	0.793	0.831	0.883	0.722
	$\hat{\alpha}_{t,h}$	-0.160	-0.275	-0.313	-0.445	-0.416	-0.367	-0.520	-0.210	-0.233	-0.645
$t = 337$	$\bar{g}_{t,h}$	-0.094	-0.165	-0.190	-0.118	0.173	0.043	0.331	0.132	0.099	-0.134
	$\hat{\rho}_{t,h}$	0.187	0.076	0.203	0.328	0.775	0.676	0.827	0.759	0.750	0.430
	$\hat{\alpha}_{t,h}$	-0.297	-0.345	-0.755	-0.710	-0.285	-0.543	-0.261	-0.403	-0.534	-0.830
$t = 338$	$\bar{g}_{t,h}$	0.001	0.010	0.091	0.337	0.202	0.360	0.199	0.198	0.070	0.232
	$\hat{\rho}_{t,h}$	0.406	0.460	0.524	0.855	0.752	0.795	0.775	0.764	0.619	0.797
	$\hat{\alpha}_{t,h}$	-0.261	-0.604	-0.505	-0.213	-0.386	-0.490	-0.299	-0.301	-0.716	-0.375
$t = 339$	$\bar{g}_{t,h}$	-0.064	0.049	0.300	0.176	0.515	0.336	0.261	0.032	0.239	0.215
	$\hat{\rho}_{t,h}$	0.476	0.599	0.823	0.746	0.886	0.860	0.819	0.563	0.753	0.713
	$\hat{\alpha}_{t,h}$	-0.508	-0.387	-0.252	-0.444	-0.183	-0.173	-0.548	-0.652	-0.573	-0.470
$t = 340$	$\bar{g}_{t,h}$	0.090	0.355	0.239	0.751	0.494	0.526	0.191	0.430	0.406	0.407
	$\hat{\rho}_{t,h}$	0.649	0.820	0.803	0.930	0.912	0.909	0.720	0.886	0.799	0.806
	$\hat{\alpha}_{t,h}$	-0.127	-0.291	-0.362	-0.069	-0.066	-0.079	-0.455	-0.253	-0.301	-0.212
$t = 341$	$\bar{g}_{t,h}$	0.280	0.155	0.729	0.438	0.494	0.080	0.332	0.287	0.297	0.361
	$\hat{\rho}_{t,h}$	0.764	0.691	0.929	0.926	0.916	0.643	0.871	0.787	0.796	0.885
	$\hat{\alpha}_{t,h}$	-0.258	-0.360	-0.070	-0.035	-0.072	-0.536	-0.253	-0.256	-0.205	-0.076
$t = 342$	$\bar{g}_{t,h}$	-0.106	0.469	0.333	0.265	-0.152	0.143	0.137	0.153	0.221	-0.089
	$\hat{\rho}_{t,h}$	0.182	0.824	0.770	0.728	0.363	0.642	0.510	0.556	0.546	0.344
	$\hat{\alpha}_{t,h}$	-0.322	-0.198	-0.468	-0.517	-0.772	-0.599	-0.595	-0.676	-0.563	-0.839
$t = 343$	$\bar{g}_{t,h}$	0.546	0.386	0.360	-0.028	0.263	0.233	0.262	0.325	0.023	0.172
	$\hat{\rho}_{t,h}$	0.836	0.799	0.828	0.472	0.739	0.593	0.629	0.649	0.561	0.580
	$\hat{\alpha}_{t,h}$	-0.275	-0.333	-0.307	-0.556	-0.411	-0.426	-0.495	-0.344	-0.566	-0.545
$t = 344$	$\bar{g}_{t,h}$	0.082	-0.026	0.152	0.143	0.084	0.036	0.072	0.169	0.196	0.275
	$\hat{\rho}_{t,h}$	0.676	0.533	0.568	0.561	0.565	0.501	0.526	0.616	0.645	0.607
	$\hat{\alpha}_{t,h}$	-0.628	-0.773	-1.017	-0.856	-0.866	-0.961	-0.813	-1.040	-0.999	-0.617
$t = 345$	$\bar{g}_{t,h}$	0.070	0.061	0.137	0.023	0.022	0.076	0.082	0.119	0.216	0.221
	$\hat{\rho}_{t,h}$	0.676	0.453	0.609	0.496	0.469	0.580	0.441	0.485	0.630	0.613
	$\hat{\alpha}_{t,h}$	-0.414	-0.663	-0.426	-0.485	-0.475	-0.338	-0.662	-0.585	-0.319	-0.408
$t = 346$	$\bar{g}_{t,h}$	0.041	0.068	0.052	0.024	0.123	0.109	0.201	0.235	0.271	0.530
	$\hat{\rho}_{t,h}$	0.455	0.538	0.374	0.456	0.577	0.464	0.560	0.600	0.567	0.681
	$\hat{\alpha}_{t,h}$	-0.693	-0.384	-0.437	-0.487	-0.389	-0.724	-0.685	-0.365	-0.532	-0.347
$t = 347$	$\bar{g}_{t,h}$	0.218	0.233	0.237	0.344	0.074	0.211	0.235	0.225	0.403	0.408
	$\hat{\rho}_{t,h}$	0.856	0.749	0.667	0.766	0.622	0.665	0.855	0.742	0.865	0.781
	$\hat{\alpha}_{t,h}$	-0.230	-0.175	-0.184	-0.197	-0.263	-0.279	-0.094	-0.225	-0.170	-0.278
$t = 348$	$\bar{g}_{t,h}$	0.035	-0.007	0.118	0.019	0.106	0.174	0.197	0.419	0.498	0.925
	$\hat{\rho}_{t,h}$	0.412	0.440	0.622	0.405	0.504	0.670	0.574	0.738	0.713	0.837
	$\hat{\alpha}_{t,h}$	-0.248	-0.289	-0.280	-0.526	-0.510	-0.226	-0.418	-0.262	-0.440	-0.336
$t = 349$	$\bar{g}_{t,h}$	0.017	0.175	0.027	0.107	0.186	0.132	0.354	0.372	0.651	0.593
	$\hat{\rho}_{t,h}$	0.511	0.705	0.542	0.528	0.792	0.656	0.796	0.806	0.840	0.823
	$\hat{\alpha}_{t,h}$	-0.196	-0.123	-0.413	-0.432	-0.147	-0.356	-0.102	-0.440	-0.352	-0.470
$t = 350$	$\bar{g}_{t,h}$	0.122	0.099	0.154	0.269	0.298	0.506	0.591	0.938	0.830	0.944
	$\hat{\rho}_{t,h}$	0.771	0.699	0.613	0.814	0.738	0.856	0.851	0.874	0.876	0.894
	$\hat{\alpha}_{t,h}$	-0.101	-0.399	-0.343	-0.186	-0.421	-0.162	-0.211	-0.221	-0.293	-0.140
$t = 351$	$\bar{g}_{t,h}$	-0.002	0.034	0.190	0.270	0.450	0.587	1.004	0.835	0.990	0.914
	$\hat{\rho}_{t,h}$	0.549	0.451	0.799	0.730	0.785	0.835	0.876	0.879	0.899	0.927
	$\hat{\alpha}_{t,h}$	-0.620	-0.468	-0.322	-0.440	-0.131	-0.224	-0.155	-0.272	-0.165	-0.076

Table 4.2 presents results for 41 sets of mispriced portfolios $\mathcal{I}_t$ for $t = 311 : 351$, as blocks of three rows ($\bar{g}_{t,h}$, $\hat{\rho}_{t,h}$ and $\hat{\alpha}_{t,h}$) for $h = 1 : 10$ (in columns). Mispriced portfolios generally yield positive gains on average for investment horizons longer that four weeks, which confirms BHW's findings. However, the probability that gains are positive is almost always below 50%. Moreover, potential losses can be important, as is revealed by the 5^{th} empirical percentile ($\hat{\alpha}_{t,h}$), which can be above 2 bpbas. Figure 4.1 shows average gains for the first ten (for visual clarity) mispriced portfolio sets. While Table 4.2 highlights the risk associated with mispriced portfolios, this figure does suggest that there might be interesting portfolios in those sets. It also reveals that longer investment horizons could be at least as profitable as a 4-week horizon.

Figure 4.1: Average gains from mispriced portfolios (t=311:321).



Average gains $\bar{g}_{t,h}$ for $t = 311 : 321$ (10 lines) over $h = 1 : 10$ weeks.

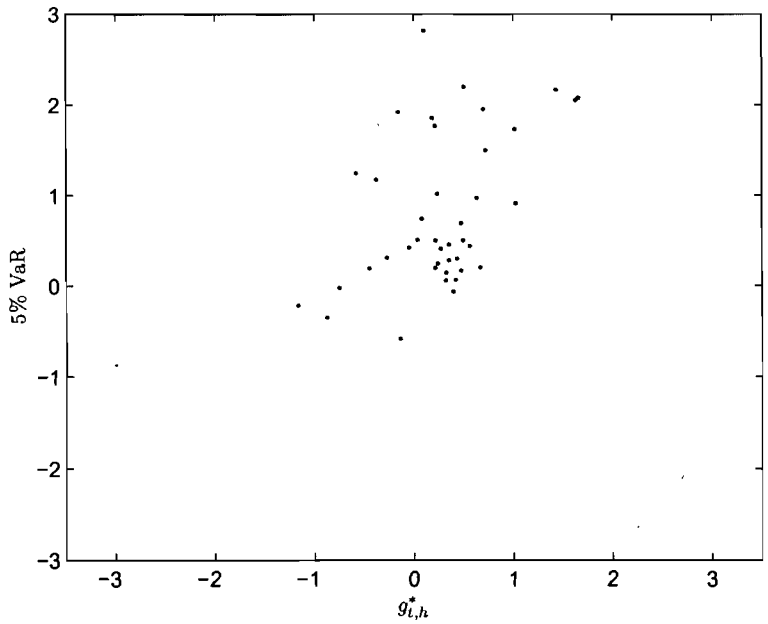
Table 4.3: Optimal portfolio statistics

	h=	1	2	3	4	5	6	7	8	9	10
$t = 311$	$g^*_{t,h}$	0.072	-0.011	0.162	0.078	0.393	0.488	0.577	0.802	0.894	0.816
	$p^*_{t,h}$	0.001	0.002	0.010	0.006	0.011	0.019	0.030	0.056	0.073	0.069
	$\alpha^*_{t,h}$	0.808	0.763	0.810	0.745	0.838	0.793	0.763	0.752	0.714	0.638
$t = 312$	$g^*_{t,h}$	-0.058	0.113	-0.030	0.356	0.424	0.530	0.637	0.710	0.753	0.905
	$p^*_{t,h}$	0.004	0.010	0.011	0.058	0.080	0.099	0.135	0.152	0.129	0.164
	$\alpha^*_{t,h}$	0.618	0.508	0.390	0.283	0.186	0.277	0.200	0.111	0.334	0.285
$t = 313$	$g^*_{t,h}$	0.151	0.053	0.403	0.499	0.587	0.812	0.905	0.827	0.979	0.819
	$p^*_{t,h}$	0.001	0.005	0.032	0.050	0.058	0.110	0.164	0.133	0.185	0.147
	$\alpha^*_{t,h}$	0.782	0.646	0.564	0.508	0.508	0.389	0.361	0.410	0.338	0.347
$t = 314$	$g^*_{t,h}$	-0.158	0.233	0.323	0.422	0.539	0.621	0.509	0.660	0.602	0.480
	$p^*_{t,h}$	0.006	0.056	0.092	0.119	0.147	0.165	0.139	0.174	0.170	0.141
	$\alpha^*_{t,h}$	0.386	0.172	0.099	0.064	0.035	0.037	0.024	-0.044	-0.052	-0.103
$t = 315$	$g^*_{t,h}$	0.363	0.510	0.533	0.671	0.743	0.640	0.773	0.766	0.637	0.754
	$p^*_{t,h}$	0.012	0.061	0.065	0.150	0.151	0.143	0.159	0.159	0.146	0.170
	$\alpha^*_{t,h}$	0.591	0.434	0.438	0.208	0.192	0.160	0.174	0.188	0.107	0.076
$t = 316$	$g^*_{t,h}$	0.182	0.204	0.443	0.566	0.393	0.549	0.668	0.373	0.637	0.860
	$p^*_{t,h}$	0.000	0.002	0.035	0.090	0.036	0.080	0.313	0.077	0.330	0.407
	$\alpha^*_{t,h}$	0.576	0.527	0.488	0.444	0.457	0.415	-0.535	0.270	-0.720	-0.705
$t = 317$	$g^*_{t,h}$	0.060	0.292	0.418	0.244	0.965	0.909	1.091	1.041	0.765	0.813
	$p^*_{t,h}$	0.003	0.055	0.120	0.049	0.501	0.439	0.484	0.464	0.360	0.404
	$\alpha^*_{t,h}$	0.314	0.283	0.254	0.249	-0.896	-0.976	-0.921	-0.896	-0.910	-1.070
$t = 318$	$g^*_{t,h}$	0.010	0.019	0.021	0.039	0.026	0.103	0.154	0.248	0.207	0.238
	$p^*_{t,h}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	$\alpha^*_{t,h}$	0.552	0.536	0.524	0.517	0.508	0.509	0.503	0.493	0.480	0.473
$t = 319$	$g^*_{t,h}$	-0.159	0.060	0.084	0.322	0.212	0.521	0.045	0.374	0.511	0.164
	$p^*_{t,h}$	0.001	0.013	0.042	0.127	0.108	0.253	0.081	0.199	0.271	0.135
	$\alpha^*_{t,h}$	0.349	0.232	0.125	0.056	-0.036	-0.050	-0.103	-0.151	-0.222	-0.264
$t = 320$	$g^*_{t,h}$	0.157	0.120	0.417	0.274	0.620	0.203	0.548	0.675	0.326	0.428
	$p^*_{t,h}$	0.001	0.004	0.041	0.027	0.121	0.033	0.136	0.192	0.090	0.124
	$\alpha^*_{t,h}$	0.575	0.548	0.469	0.416	0.329	0.328	0.263	0.158	0.143	0.027
$t = 321$	$g^*_{t,h}$	-0.037	-0.207	-0.155	-0.140	0.127	0.064	0.057	0.047	-0.104	-0.238
	$p^*_{t,h}$	0.001	0.056	0.076	0.103	0.171	0.166	0.162	0.164	0.139	0.150
	$\alpha^*_{t,h}$	0.471	-0.252	-0.415	-0.582	-0.611	-0.687	-0.751	-0.703	-0.819	-0.979
$t = 322$	$g^*_{t,h}$	0.038	0.000	0.079	0.221	0.250	0.311	0.450	0.216	-0.056	-0.140
	$p^*_{t,h}$	0.001	0.004	0.009	0.017	0.032	0.029	0.049	0.037	0.024	0.016
	$\alpha^*_{t,h}$	0.683	0.697	0.604	0.509	0.458	0.489	0.467	0.378	0.343	0.352
$t = 323$	$g^*_{t,h}$	-0.039	0.041	0.187	0.217	0.281	0.422	0.187	-0.088	-0.169	0.415
	$p^*_{t,h}$	0.004	0.019	0.045	0.053	0.067	0.116	0.087	0.045	0.046	0.128
	$\alpha^*_{t,h}$	0.417	0.299	0.220	0.198	0.184	0.044	-0.071	-0.058	-0.090	-0.099
$t = 324$	$g^*_{t,h}$	0.079	0.229	0.260	0.326	0.469	0.234	-0.044	-0.125	0.463	0.548
	$p^*_{t,h}$	0.014	0.039	0.054	0.075	0.118	0.072	0.044	0.035	0.116	0.124
	$\alpha^*_{t,h}$	0.368	0.303	0.223	0.145	0.105	0.046	0.065	0.067	0.063	-0.012
$t = 325$	$g^*_{t,h}$	0.151	0.181	0.249	0.396	0.155	-0.131	-0.226	0.361	0.439	-0.107
	$p^*_{t,h}$	0.011	0.053	0.095	0.141	0.109	0.076	0.054	0.149	0.179	0.062
	$\alpha^*_{t,h}$	0.471	0.163	0.022	-0.066	-0.163	-0.315	-0.279	-0.247	-0.350	-0.232
$t = 326$	$g^*_{t,h}$	0.044	0.140	0.221	0.355	0.240	0.524	-0.076	0.102	0.715	0.524
	$p^*_{t,h}$	0.002	0.006	0.027	0.040	0.029	0.069	0.023	0.030	0.122	0.081
	$\alpha^*_{t,h}$	0.655	0.546	0.400	0.459	0.404	0.389	0.317	0.296	0.318	0.284
$t = 327$	$g^*_{t,h}$	0.097	0.218	0.009	-0.268	-0.723	0.107	-1.292	-0.005	-0.139	0.529
	$p^*_{t,h}$	0.001	0.010	0.006	0.006	0.011	0.048	0.008	0.030	0.024	0.127
	$\alpha^*_{t,h}$	0.823	0.583	0.405	0.312	-0.097	0.111	-0.322	0.121	0.103	0.086
$t = 328$	$g^*_{t,h}$	0.146	-0.117	-0.412	-0.876	-1.466	-1.444	-1.374	-0.934	-0.589	0.393
	$p^*_{t,h}$	0.023	0.034	0.029	0.007	0.001	0.005	0.010	0.037	0.068	0.276
	$\alpha^*_{t,h}$	0.286	0.066	-0.156	-0.343	-0.422	-0.577	-0.627	-0.715	-0.783	-0.770
$t = 329$	$g^*_{t,h}$	0.071	-0.178	-0.574	-1.165	-1.144	-1.074	-0.633	-0.286	-0.406	-0.654
	$p^*_{t,h}$	0.050	0.026	0.011	0.002	0.004	0.013	0.029	0.069	0.068	0.048
	$\alpha^*_{t,h}$	0.061	0.013	0.001	-0.211	-0.240	-0.263	-0.373	-0.509	-0.570	-0.635
$t = 330$	$g^*_{t,h}$	0.074	0.102	-0.437	-0.047	1.238	0.990	1.088	1.428	0.955	0.582
	$p^*_{t,h}$	0.000	0.007	0.004	0.019	0.262	0.188	0.221	0.331	0.201	0.134
	$\alpha^*_{t,h}$	0.823	0.665	0.505	0.430	0.306	0.247	0.197	0.003	0.018	-0.029

	h=	1	2	3	4	5	6	7	8	9	10
$t = 331$	$g^*_{t,h}$	-0.088	-0.713	-0.328	0.634	0.626	0.843	1.067	0.611	0.131	0.234
	$p^*_{t,h}$	0.000	0.000	0.000	0.021	0.029	0.052	0.095	0.035	0.014	0.024
	$\alpha^*_{t,h}$	1.221	1.087	0.986	0.974	0.842	0.835	0.750	0.758	0.719	0.616
$t = 332$	$g^*_{t,h}$	-0.634	-0.320	0.709	0.718	0.897	1.087	0.659	0.132	0.324	0.887
	$p^*_{t,h}$	0.000	0.000	0.004	0.007	0.016	0.025	0.012	0.001	0.004	0.021
	$\alpha^*_{t,h}$	1.755	1.584	1.535	1.500	1.388	1.365	1.250	1.326	1.287	1.195
$t = 333$	$g^*_{t,h}$	0.499	1.435	1.444	1.628	1.820	1.387	0.853	0.950	1.605	1.676
	$p^*_{t,h}$	0.000	0.010	0.010	0.015	0.022	0.005	0.004	0.004	0.017	0.024
	$\alpha^*_{t,h}$	1.920	1.919	1.953	2.053	2.127	2.087	2.152	2.057	2.013	2.047
$t = 334$	$g^*_{t,h}$	1.440	1.178	1.278	1.433	0.998	0.460	0.559	1.216	1.289	2.071
	$p^*_{t,h}$	0.004	0.005	0.009	0.010	0.004	0.001	0.001	0.001	0.003	0.029
	$\alpha^*_{t,h}$	1.945	1.905	1.923	2.170	2.240	2.251	2.354	2.296	2.358	2.371
$t = 335$	$g^*_{t,h}$	-0.102	0.837	0.742	0.474	0.045	0.107	1.191	0.963	1.852	0.863
	$p^*_{t,h}$	0.001	0.159	0.081	0.030	0.004	0.005	0.049	0.025	0.111	0.014
	$\alpha^*_{t,h}$	0.542	0.498	0.478	0.698	1.030	1.195	1.195	1.311	1.424	1.419
$t = 336$	$g^*_{t,h}$	0.426	0.852	-0.024	-0.446	-0.389	0.679	0.450	1.339	0.372	0.618
	$p^*_{t,h}$	0.020	0.216	0.051	0.013	0.010	0.082	0.052	0.192	0.035	0.052
	$\alpha^*_{t,h}$	0.556	0.412	-0.032	0.191	0.246	0.350	0.413	0.480	0.542	0.599
$t = 337$	$g^*_{t,h}$	-0.115	-0.384	-0.809	-0.750	0.324	0.097	0.981	0.002	0.255	-0.394
	$p^*_{t,h}$	0.054	0.044	0.013	0.008	0.093	0.036	0.174	0.028	0.044	0.014
	$\alpha^*_{t,h}$	-0.127	-0.307	-0.272	-0.021	0.071	0.214	0.288	0.261	0.323	0.256
$t = 338$	$g^*_{t,h}$	-0.264	-0.684	-0.627	0.436	0.208	1.091	0.127	0.373	-0.269	-0.019
	$p^*_{t,h}$	0.073	0.002	0.007	0.073	0.040	0.183	0.036	0.052	0.011	0.016
	$\alpha^*_{t,h}$	-0.379	-0.070	0.113	0.301	0.314	0.302	0.383	0.358	0.505	0.453
$t = 339$	$g^*_{t,h}$	-0.609	-0.272	0.706	0.479	1.361	0.389	0.640	-0.005	0.246	0.187
	$p^*_{t,h}$	0.000	0.000	0.201	0.092	0.259	0.042	0.056	0.008	0.016	0.013
	$\alpha^*_{t,h}$	0.845	0.767	0.017	0.166	0.386	0.470	0.577	0.680	0.829	0.769
$t = 340$	$g^*_{t,h}$	0.096	0.769	0.836	1.658	1.112	1.296	0.836	0.845	1.062	0.899
	$p^*_{t,h}$	0.000	0.000	0.001	0.011	0.003	0.004	0.000	0.001	0.001	0.001
	$\alpha^*_{t,h}$	1.988	2.036	2.061	2.079	2.135	2.159	2.258	2.363	2.323	2.370
$t = 341$	$g^*_{t,h}$	0.671	0.740	1.558	1.012	1.197	0.738	0.747	0.962	0.798	0.632
	$p^*_{t,h}$	0.000	0.000	0.032	0.007	0.007	0.001	0.001	0.006	0.004	0.003
	$\alpha^*_{t,h}$	1.713	1.686	1.748	1.730	1.776	1.959	1.937	1.933	1.957	2.046
$t = 342$	$g^*_{t,h}$	0.026	1.287	1.042	1.024	0.261	0.559	0.816	0.767	0.787	-0.076
	$p^*_{t,h}$	0.000	0.098	0.047	0.064	0.009	0.033	0.067	0.073	0.073	0.007
	$\alpha^*_{t,h}$	1.394	1.083	1.055	0.914	0.801	0.696	0.668	0.599	0.596	0.601
$t = 343$	$g^*_{t,h}$	1.260	1.013	0.996	0.234	0.531	0.790	0.740	0.761	-0.102	0.176
	$p^*_{t,h}$	0.023	0.018	0.031	0.001	0.011	0.033	0.033	0.046	0.004	0.013
	$\alpha^*_{t,h}$	1.442	1.269	1.157	1.014	1.013	0.981	0.856	0.802	0.778	0.747
$t = 344$	$g^*_{t,h}$	0.467	-0.799	0.078	-0.582	0.050	-0.397	-0.306	-0.674	-0.224	-0.249
	$p^*_{t,h}$	0.000	0.000	0.001	0.000	0.004	0.004	0.012	0.007	0.022	0.024
	$\alpha^*_{t,h}$	1.535	1.371	1.376	1.242	1.030	0.719	0.672	0.467	0.300	0.200
$t = 345$	$g^*_{t,h}$	-1.210	-0.814	-1.055	-0.379	-0.734	-0.636	-0.997	-0.553	-0.568	0.056
	$p^*_{t,h}$	0.000	0.000	0.000	0.001	0.000	0.003	0.001	0.005	0.007	0.021
	$\alpha^*_{t,h}$	1.347	1.286	1.267	1.168	1.149	1.019	0.909	0.738	0.629	0.565
$t = 346$	$g^*_{t,h}$	0.448	0.165	0.858	0.504	0.600	0.244	0.692	0.674	1.260	1.772
	$p^*_{t,h}$	0.000	0.000	0.001	0.001	0.001	0.000	0.001	0.002	0.016	0.051
	$\alpha^*_{t,h}$	2.352	2.245	2.281	2.195	2.159	2.231	2.149	2.074	1.881	1.747
$t = 347$	$g^*_{t,h}$	-0.274	0.445	0.084	0.182	-0.168	0.281	0.261	0.851	1.417	1.889
	$p^*_{t,h}$	0.000	0.000	0.001	0.001	0.001	0.005	0.005	0.021	0.050	0.096
	$\alpha^*_{t,h}$	2.216	2.030	1.847	1.855	1.779	1.781	1.633	1.333	1.415	1.311
$t = 348$	$g^*_{t,h}$	0.723	0.340	0.440	0.098	0.544	0.527	1.115	1.635	2.125	3.419
	$p^*_{t,h}$	0.000	0.000	0.000	0.000	0.000	0.001	0.003	0.017	0.037	0.176
	$\alpha^*_{t,h}$	2.869	2.754	2.846	2.815	2.764	2.750	2.580	2.484	2.369	2.234
$t = 349$	$g^*_{t,h}$	-0.422	-0.291	-0.601	-0.158	-0.177	0.414	0.938	1.425	2.714	1.611
	$p^*_{t,h}$	0.000	0.000	0.000	0.000	0.001	0.005	0.017	0.040	0.193	0.069
	$\alpha^*_{t,h}$	2.135	2.019	1.924	1.915	1.863	1.840	1.664	1.554	1.412	1.333
$t = 350$	$g^*_{t,h}$	0.154	0.525	0.641	0.211	0.801	1.327	1.813	3.098	1.999	1.971
	$p^*_{t,h}$	0.000	0.000	0.001	0.000	0.013	0.043	0.098	0.327	0.128	0.140
	$\alpha^*_{t,h}$	2.143	2.116	2.080	1.766	1.578	1.416	1.333	1.302	1.214	0.906
$t = 351$	$g^*_{t,h}$	-0.311	0.128	0.109	0.698	1.228	1.714	3.000	1.898	1.859	1.901
	$p^*_{t,h}$	0.000	0.000	0.000	0.006	0.019	0.049	0.245	0.098	0.112	0.123
	$\alpha^*_{t,h}$	2.217	2.124	2.037	1.954	1.795	1.716	1.563	1.404	1.277	1.115

Table 4.3 presents results for 41×10 optimal portfolios, as 41 blocks of three rows ($g^*_{t,h}$, $p^*_{t,h}$ and $\alpha^*_{t,h}$) for 10 investment horizons. In contrast to mispriced portfolios sets that are function of pricing differentials only ($d_{l,t}$), optimal portfolios rely on the predictive density of gains ($g_{t,h}$). Optimal portfolios generally yield much larger gains than the average mispriced portfolio. These gains are in the order of one basis point of bid-ask spread and are thus economically significant for large institutional investors.

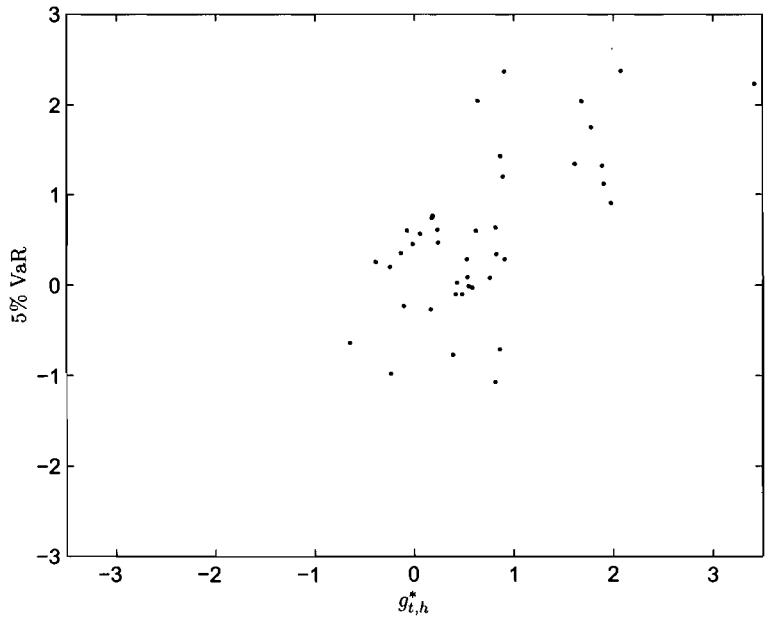
Figure 4.2: Optimal portfolio value at risk (h=4).



Realized optimal portfolio gains $g^*_{t,h}$ after $h = 4$ weeks and posterior 5% value-at-risk (VaR) $\alpha^*_{t,h}$ for $t = 311 : 351$ (41 dots).

Optimal portfolios are not risk-free and lose on occasion. However, the estimated predictive cumulative probability of realized gains $p^*_{t,h}$ are reasonable possible values, which suggest that characteristics of the predictive density other than the expected gains could be useful.

Figure 4.3: Optimal portfolio value at risk (h=10).



Realized optimal portfolio gains $g^*_{t,h}$ after $h = 10$ weeks and posterior 5% value-at-risk (VaR) $\alpha^*_{t,h}$ for $t = 311 : 351$ (41 dots).

One such characteristic of the predictive density is the 5%-value-at-risk. Figures 4.2 and 4.3 respectively show the relationship between realized gains ($g^*_{t,h}$) and value-at-risk ($\alpha^*_{t,h}$) for investment horizons $h = 4$ and $h = 10$. The risk assessment is rather poor in absolute terms: 5% of portfolios should be above the 45^o-line. However, there seems to be a relationship between gains and value-at-risk, which appears more precisely at longer horizons. For 10-week investment horizons, considering portfolios with positive value-at-risk would significantly reduce the range and probability of loses. This suggests that formalizing one's preferences over expected gains and risk through the specification of a utility function of the form (4.8) can prove profitable.

4.6 Appendix A - Prior distributions and hyper-parameters

The following priors, for a re-parameterization of the model, are permutation- and reflection-invariant. The re-parameterization involves two parameters, $\kappa^{\mathbb{Q}}$ and $\mathbf{B_1}$. The first relies on an eigendecomposition of $\mathcal{I} - \kappa^{\mathbb{Q}'}$, i.e.

$$\mathcal{I} - \kappa^{\mathbb{Q}'} = \delta\gamma\delta^{-1}.$$

As eigenvector are defined up to a scalar multiplication, I parameterize them in polar coordinates. Let $\phi \equiv [\phi_1 \dots \phi_K]$ denote the matrix of angles, where the vector ϕ_j, $j = 1, \dots, K$ contains the angles associated with the eigenvector δ_j

$$\phi_{k,j} \equiv \arctan\left(\frac{\delta_{k+1,j}}{\sqrt{\sum_{i=1}^{k}\delta_{i,j}^2}}\right) \quad \text{for} \quad k = 1, \dots, K-1.$$

For $K > 1$, I use a parameterization $(\zeta_1, \dots, \zeta_{K-1}, \sigma)$ of the short rate factor loadings $\mathbf{B_1}$ in polar coordinates. I define ζ to be the $K-1$-dimensional vector of angles with elements

$$\zeta_k \equiv \arctan\left(\frac{B_{1,k+1}}{\sqrt{\sum_{i=1}^{k}B_{1,i}^2}}\right) \quad \text{for} \quad k = 1, \dots, K-1,$$

and σ to be the logarithm of the Euclidean norm of $\mathbf{B_1}$,

$$\sigma \equiv \log\left(\sqrt{\mathbf{B}_1'\mathbf{B}_1}\right).$$

Prior distribution of σ

The logarithm of the Euclidean norm of $\mathbf{B}_1$ is normally distributed

$$f(\sigma|\mu_\sigma^0, \Sigma_\sigma^0) \equiv \mathcal{N}(\sigma|\mu_\sigma^0, \Sigma_\sigma^0).$$

Prior distribution of ζ

The short-rate vector of factor loadings is a priori uniformly distributed on a K-dimensional hyper-sphere with radius e^σ

$$f(\zeta) \equiv \frac{1}{2\pi}\prod_{k=2}^{K}\frac{1}{4\pi}\cos(\zeta_k).$$

Prior distribution of $\kappa^{\mathbb{P}}$

I do not impose stationarity and use a i.i.d. normal distribution

$$f(\kappa_{k,k}^{\mathbb{P}}|\mu_{\kappa^{\mathbb{P}}}^0, \Sigma_{\kappa^{\mathbb{P}}}^0) = \mathcal{N}(\kappa_{k,k}^{\mathbb{P}}|\mu_{\kappa^{\mathbb{P}}}^0, \Sigma_{\kappa^{\mathbb{P}}}^0),$$

for $k = 1, \ldots, K$.

Prior distribution of Σ

As Σ is a correlation matrix, I use a prior distribution proposed by Barnard, McCulloch, and Meng (2000). Defining the one-to-one mapping $g(\mathbf{Q}, \mathbf{D}) = \mathbf{DQD}' = \Sigma$, which decomposes a covariance matrix Σ into a diagonal matrix of standard deviations D and a correlation matrix $\mathbf{Q}$, the distribution is

$$f(\mathbf{Q}|\tau) = |\mathbf{Q}|^{\frac{1}{2}(\tau-1)(K-1)-1}\left(\prod_{i=1}^{K}|\mathbf{Q}_{(ii)}|\right)^{-\frac{\tau}{2}}.$$

It has the property that individual correlations have scaled Beta marginal distributions $\mathcal{B}eta(\frac{\tau-K+1}{2}, \frac{\tau-K+1}{2})$, which is uniform over $[-1, 1]$ for $\tau = K + 1$.

Prior distribution of γ

The eigenvalues of $\mathcal{I} - \kappa^{\mathbb{Q}'}$ are a priori i.i.d. normally distributed

$$f(\gamma_k|\mu_\gamma^0, \Sigma_\gamma^0) \equiv \mathcal{N}(\gamma_k|\mu_\gamma^0, \Sigma_\gamma^0).$$

Prior distribution of ϕ

Eigenvectors are defined up to a scaler multiplication so I consider the space of unit eigenvectors with positive first-element. The K unit eigenvectors of $\mathcal{I} - \kappa^{\mathbb{Q}'}$ are a priori uniformly distributed on the unit half-sphere, which implies the following prior distribution for the angles:

$$f(\phi_k) \equiv \frac{1}{\pi} \prod_{j=2}^{K} \frac{1}{4\pi} \cos(\phi_{k,j}),$$

for $k = 1, \ldots, K$.

Prior distribution of A_1

The prior distribution of A_1 is the following truncated normal distribution:

$$f(A_1|\mu_{A_1}^0, \Sigma_{A_1}^0) = \mathcal{N}(A_1|\mu_{A_1}^0, \Sigma_{A_1}^0)\mathbf{1}_{A_1>0}.$$

Prior distribution of λ_0

The prior distribution of λ_0 is the following truncated normal distribution:

$$f(\lambda_{0,k}|\mu^0_{\lambda_0}, \Sigma^0_{\lambda_0}) \quad = \quad \mathcal{N}\left(\lambda_0|\mu^0_{\lambda_0}, \Sigma^0_{\lambda_0}\right),$$

for $k = 1, \ldots, K$.

Hyper-parameters

Table 4.4: Prior distribution hyper-parameter values.

Parameter	Value
$\mu^0_{A_1}$	1e-003
$\Sigma^0_{A_1}$	2.5e-006
$\mu^0_{\sigma_0}$	-5
$\Sigma^0_{\sigma_0}$	1e+003
$\mu^0_{\lambda_0}$	0
$\Sigma^0_{\lambda_0}$	100
μ^0_{γ}	0
Σ^0_{γ}	5
$\mu^0_{\kappa^{\mathbb{P}}}$	0
$\Sigma^0_{\kappa^{\mathbb{P}}}$	5
τ	10
ν^0_{Ω}	2
γ^0_{Ω}	250
β^0_{Ω}	1e-010

CHAPITRE 5

CONCLUSION

La fonction de vraisemblance des modèles espace-état linéaires est invariante par rapport à un certain ensemble de transformations du vecteur de paramètres, ce qui implique, en particulier, que l'estimateur du maximum de vraisemblance des paramètres n'est pas unique. Au meilleur de mes connaissance, l'approche de l'ensemble de la littérature est de normaliser l'espace des paramètres afin de briser l'invariance de la fonction de vraisemblance et assurer l'unicité de l'estimateur.

Cette thèse propose une nouvelle approche et considère la problématique d'une manière systématique. Elle identifie d'abord un ensemble de situations où il est possible de faire de l'inférence statistique sans briser l'invariance de la fonction de vraisemblance. Il est possible, par exemple, de faire des prévisions en utilisant un estimateur ensembliste des paramètres ou une distribution *a posteriori*. Dès lors, normaliser l'espace des paramètres ou non devient un choix de modélisation statistique. Comme toute décision de modélisation, ce choix doit être fait à la lumière d'une analyse coût-bénéfice. Je présente d'abord, d'un point de vue théorique, les avantages et désavantages anticipés d'une normalisation. Il va sans dire que l'importance relative attribuée à ces avantages et désavantages anticipés dépendra du contexte d'inférence et des préférences de l'économètre. Cependant, pour certains échantillons, il est possible qu'un avantage théorique anticipé soit hors de portée en pratique. Ces situations peuvent être qualifiées de problèmes d'*identification faible*, et se traduisent notamment par des estimateurs ponctuels des paramétres dont la distribution est multimodale, ou par des distributions *a posteriori* multimodales. Ces difficultés doivent être prises en considération dans la

décision de normaliser ou non l'espace des paramètres.

Il existe une multitude de manières de normaliser l'espace des paramètres et, lorsqu'il appert avantageux de normaliser, je propose un critère pour comparer entre elles différentes normalisations. Ce critère, appelé *principe d'identification*, vise à réduire l'ampleur du problème d'identification faible pour un échantillon arbitraire de l'espace échantillonal. Pour satisfaire le principe d'identification, une normalisation doit rencontrer trois critères : elle doit être non restrictive observationellement, apporter l'identification globale des paramètres, et être connexe. Au meilleur de mes connaissance, cette thèse présente la première famille de normalisations des modèles espace-état linéaire satisfaisant ce principe d'identification.

En présence de problèmes d'identification faible, l'approche bayésienne offre deux avantages par rapport à la méthode du maximum de vraisemblance. D'abord, elle fournit de meilleures prédictions parce qu'elle ne repose sur aucun estimateur ponctuel des paramètres. Les deux approches ne sont équivalentes qu'asymptotiquement, et les résultats empiriques que je présente ne constituent qu'un exemple de situation où considérer l'incertitude entourant les paramètres porte fruit. Ceci suggère qu'il est possible que considérer cette incertitude en utilisant un estimateur du maximum de vraisemblance ensembiliste puisse aussi s'avérer utile.

Le second avantage est de nature computationelle. Le principe d'identification ne permet pas d'isoler une normalisation optimale et, dans une situation pratique donnée, on voudra comparer entre elles plusieurs normalisation satisfaisant le principe d'identification. On peut, par exemple, comparer visuellement les distributions des estimateurs définies par différentes normalisations. Puisque l'estimateur du maximum

de vraisemblance des paramètres n'a pas d'expression analytique, sa distribution échantillonale doit être obtenue par méthodes de simulation, ce qui constitue un problème computationel de taille. En revanche, il est possible de comparer un grand nombre de normalisations dans le cadre bayésien quasi-instantanément.

En ce qui a trait à l'inférence pour les modèles affines de la structure à terme, je démontre que la normalisation cannonique proposée par Dai et Singleton (2000) ne satisfait pas le principe d'identification. De plus, je propose une telle normalisation. J'expose aussi le rôle joué par la spécification de la matrice de covariance des erreurs observationnelles. En particulier, spécifier une matrice de plein rang permet de réduire l'autocorrélation des résidus ainsi que les corrélations croisées entre les résidus associés à différentes maturités. Plus généralement, considérant qu'un modèle affine décompose la dynamique de la structure à terme en facteurs communs et idiosyncrasiques, je montre que la modélisation des erreurs joue un rôle important dans cette décomposition. Ce rôle est aisément décrit en considérant les cas extrêmes. Une modélisation très restrictive introduira des problèmes de mauvaise spécification. Par ailleurs, une modélisation trop générale introduira des problèmes d'identification, e.g. lorsque les facteurs et les erreurs appartiennent à la même famille de processus stochastiques. Afin de permettre à l'économètre de choisir sur continuum entre ces modélisations extrêmes, je propose une loi a priori qui permet de spécifier des restrictions souples sur les corrélations croisées et l'hétéroscédasticité des erreurs.

Le dernier article de cette thèse illustre la mise-en-oeuvre des résultats obtenus dans les articles qui le précèdent. Je considère le problème auquel fait face un arbitragiste sur le marché des titres à revenu fixe. Ce problème consiste à maximiser l'espérance d'utilité de l'investisseur en choisissant une stratégie et cette espérance est

calculée sous la densité prédictive estimée à l'aide d'un modèle affine. Bien que ce modèle économétrique présente certaines faiblesses, le cadre proposé semble permettre d'identifier des stratégies d'investissement profitables. Outre le modèle utilisé pour calculer la densité prédictive, les autres aspects du problème d'optimisation sont très simples. En particulier, l'utilité est linéaire et les stratégies considérées sont statiques et relativement simples. Ceci laisse donc place à un important ensemble d'améliorations.

BIBLIOGRAPHIE

Anderson, B. D., Moore, J. B., 1979. Optimal Filtering. Prentice-Hall, Englewood Cliffs, N.J.

Ang, A., Bekaert, G., 2002. Regime switches in interest rates. Journal of Business and Economic Statistics 20, 163–182.

Ang, A., Dong, S., Piazzesi, M., 2007. No-arbitrage Taylor rules, Working paper, Columbia University and University of Chicago.

Ang, A., Piazzesi, M., 2003. A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. Journal of Monetary Economics 50(4), 745–787.

Aoki, M., 1987. State space modeling of time series. Springer-Verlag, New York.

Babbs, S., Nowman, K., 1999. Kalman filtering of generalized Vasicek term structure models. Journal of Financial and Quantitative Analysis 34 (1), 115–130.

Balı, T., Heidari, M., Wu, L., June 2006. Predictability of interest rates and interest-rate portfolios, Working paper, Zicklin School of Business and Caspian Capital Management.

Ball, C., Torous, W., 1996. Unit roots and the estimation of interest rate dynamics. Journal of Empirical Finance 3, 215–238.

Barnard, J., McCulloch, R., Meng, X.-L., 2000. Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. Statistica Sinica 10, 1281–1311.

Bekaert, G., Cho, S., Moreno, A., June 2006. New-Keynesian macroeconomics and the term structure, Working paper, Graduate School of Business, Columbia University.

Bertholon, H., Monfort, A., Pegoraro, F., 2007. Econometric asset pricing modelling, CREST DP.

Blais, S., 2008a. A Bayesian analysis of affine term structure models, Working paper, Université de Montréal.

Blais, S., 2008b. Forecasting with wealkly identified linear state space models, Working paper, Université de Montréal.

Bliss, R. R., 1997. Testing term structure estimation methods. Advances in Futures and Options Research 9, 197–231.

Bound, J., Jaeger, D., Baker, R., 1995. Problems with instrumental variables estimation when the correlations betweeen the instruments and the endogenous explanatory variables is weak. Journal of the American Statistical Association 90, 443–450.

Box, G., Jenkins, G., 1976. Time series analysis : Forecasting and applications. Holden-Day, San Fransisco.

Box, G., Jenkins, G., Reinsel, G., 1994. Time Series Analysis, 3rd Edition. Prentice Hall.

Brockwell, P. J., Davis, R. A., 1991. Time Series : Theory and Methods, 2nd Edition. Springer.

Brooks, S. P., 1998. Quantitative convergence assessment for Markov chain Monte Carlo via cusums. Statistics and Computing 8, 267–274.

Buse, A., 1992. The bias of instrumental variables estimators. Econometrica 60, 173–180.

Campbell, J., Lo, A., MacKinlay, A., 1997. The Econometrics of Financial Markets. Princeton University Press.

Carter, C., Kohn, P., 1994. On the Gibbs sampling for state space models. Biometrika 81, 541–553.

Celeux, G., Hurn, M., Robert, C., 2000. Computational and inferential difficulties with mixture prosterior distribution. Journal of the American Statistical Association 95 (451), 957–970.

Chen, R., Scott, L., 1993. Maximum likelihood estimation for a multifactor equilibrium model of the term structure of interest rates. Journal of Fixed Income 3, 14–31.

Chen, R., Scott, L., 1995. Multi-factor Cox-Ingersoll-Ross models of the term-structure : estimates and tests from a Kalman filter model., Working paper, University of Georgia.

Cheridito, P., Filipović, D., Kimmel, R., December 2003. Market price of risk specifications for affine models : Theory and evidence, Princeton University.

Chernozhukov, V., Hong, H., Tamer, E., September 2007. Estimation and confidence regions for parameter sets in econometric models. Econometrica 75 (5), 1234–1284.

Chib, S., Ergashev, B., September 2008. Analysis of multi-factor affine yield curve models, wroking paper, Washington University in St. Louis and the Federal Reserve Bank of Richmond.

Christensen, J. H. E., Diebold, F. X., Rudebusch, G. D., November 2007. The affine arbitrage-free class of Nelson-Siegel term structure models, nBER Working Paper 13611.

Cochrane, J., 2005. Asset Pricing, Revised Edition. Princeton University Press.

Dai, Q., Le, A., Singleton, K., March 2006. Discrete-time dynamic term structure models with generalized market prices of risk, Working paper, Graduate School of Business, Stanford University.

Dai, Q., Singleton, K., 2000. Specification analysis of affine term structure models. 55 Journal of Finance, 1943–1978.

Dai, Q., Singleton, K., 2002. Expectation puzzles, time-varying risk premia, and affine models of the term structure. Journal of Financial Economics 63, 415–441.

Dai, Q., Singleton, K., 2003. Term structure dynamics in theory and reality. Review of Financial Studies 16, 361–678.

Dai, Q., Singleton, K., Yang, W., 2005. Are regime shifts priced in U.S. Treasury markets ?, Working paper, New York University.

Dewachter, H., Lyrio, M., 2006. Learning, macroeconomic dynamics and the term structure of interest rates, Working paper, Catholic University of Leuven.

Dick, N. P., Bowden, D. C., 1973. Maximum likelihood estimation for mixtures of two normal distributions. Biometrics 29, 781–790.

Duarte, J., 2003. Evaluating an alternative risk preference in affine term structure models, financeLab, Working paper FLWP-2003-2.

Duffee, G., 2002. Term premia and interest rate forecasts in affine models. Journal of Finance 57, 405–443.

Dufour, J.-M., 1997. Some impossibility theorems in econometrics, with applications to structural and dynamic models. Econometrica 65, 1365–1389.

Dufour, J.-M., Hsiao, C., 2008. "Identification", The New Palgrave Dictionary of Economics, 2nd Edition. Palgrave Macmillan.

Evans, M., 2003. Real risk, inflation risk, and the term structure. Economic Journal 113(487), 345–389.

Frühwirth-Schnatter, S., 1994. Data augmentation and dynamic linear models. Journal of Time Series Analysis 15, 183–202.

Frühwirth-Schnatter, S., 2001. Markov Chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association 96 (453), 194–205.

Frühwirth-Schnatter, S., 2004. Efficient Bayesian parameter estimation. In : Harvey, A., Koopman, S. J., Shephard, N. (Eds.), State Space and Unobserved Component Models : Theory and Applications. Cambridge University Press, pp. 123–151.

Frühwirth-Schnatter, S., Geyer, A., 1998. Bayesian estimation of econometric multifactor Cox-Ingersoll-Ross models of the term structure of interest rates via MCMC methods, Working paper, Vienna University of Economics and Business Administration.

Frühwirth-Schnatter, S., Wagner, H., 2008. Stochastic model specification search for Gaussian and non-Gaussian state space models, iFAS Research Paper Series 2008-36.

Galichon, A., Henry, M., 2009. A test of non-identifying restrictions and confidence regions for partially identified parameters. Journal of Econometrics, forthcoming.

Garcia, R., Luger, R., February 2007. Risk aversion, intertemporal substitution, and the term structure of interest rates., Working paper, Université de Montréal and Emory University.

Geweke, J., 2007. Interpretation and inference in mixture models : Simple MCMC works. Computational Statistics & Data Analysis 51, 3529–3550.

Geyer, A., Pichler, S., 1999. A state-space approach to estimate and test multifactor Cox-Ingersoll-Ross models of the term structure. Journal of Financial Research 22 (1), 107–130.

Gouriéroux, C., Monfort, A., Polimenis, V., 2002. Affine term structure models, Working paper, CREST.

Hamilton, J., Waggoner, D., Zha, T., 2007. Normalization in econometrics. Econometric Reviews 26, 221 – 252.

Hamilton, J. D., 1994. Time Series Analysis. Princeton University Press.

Harvey, A. C., 1989. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press.

Hill, B., 1963. Information for estimating the proportions in mixtures of exponential and normal distributions. Journal of the American Statistical Association 58 (304), 918–932.

Hiller, G. H., 1990. On the normalization of structural equations : properties of direction estimators. Econometrica 58 (5), 1181–1194.

Hördahl, P., Tristani, O., Vestin, D., 2006. A joint econometric model of macroeconomic and term-structure dynamics. Journal of Econometrics 131, 405–444.

Jacquier, E., Johannes, M., Polson., N., 2007. MCMC maximum likelihood for latent state models. Journal of Econometrics 137.

Jegadeesh, N., Pennacchi, G., 1996. The behavior of interest rates implied by the term structure of eurodollar futures. Journal of Money, Credit and Banking 28 (3), 426–446.

Jennrich, R. I., 1978. Rotational equivalence of factor loading matrices with specified values. Psychometrika 43 (3), 421–426.

Kim, C., Nelson, C., 1998. Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching. The Review of Economics and Statistics 80, 188–201.

Kleibergen, F., Hoek, H., March 2000. Bayesian analisys of ARMA models, Tinbergen Institute Discussion Paper TI 2000-027/4.

Lamoureux, C., Witte, H., 2002. Empirical analysis of the yield curve : The information in the data viewed through the window of Cox, Ingersol and Ross. The Journal of Finance 57, 1479–1520.

Leamer, E. E., 1973. Multicollinearity : A Bayesian perspective. The Review of Economics and Statistics 55, 371–380.

Litterman, R., Scheinkman, J., 1991. Common factors affecting bond returns. Journal of Fixed Income, 54–61.

Loken, E., 2004. Multimodality in mixture models and factor models. In : Gelman, A., Meng, X.-L. (Eds.), Applied Bayesian Modeling and causal inference from incomplete-data perspectives. John Wiley & Son, pp. 203–213.

Manski, C., 2003. Partial Identification of Probability Distributions. Springer-Verlag, New York.

Mas-Colell, A., Whinston, M. D., Green, J. R., 1995. Microeconomic Theory. Oxford University Press.

McCulloch, J., 1975. The tax-adjusted yield curve. Journal of Finance 30, 811–830.

Modigliani, F., Sutch, R., 1966. Innovations in interest rate policy. American Economic Review 56, 178–197.

Modigliani, F., Sutch, R., 1967. Debt management and the term structure of interest rates : an empirical analysis of interest rates. Journal of Political Economy 75, 569–589.

Monfort, A., Pegoraro, F., 2007. Switching VARMA term structure models. Journal of Financial Econometrics 51 (1), 105–153.

Müller, P., Polson, N., Stroud, J., 2003. Nonlinear state-space models with state-dependent variances. Journal of the American Statistical Association 98, 377–386.

Nelson, C. R., Startz, R., 1990. The distribution of the instrumental variable estimator and its t-ratio when the instrument is a poor one. Journal of Business 63 (S125-S140).

Redner, R. A., Walker, H. F., 1984. Mixture densities, maximum likelihood and the EM algorithm. SIAM Review 26 (2), 195–239.

Robert, C. P., 2001. The Bayesian choice, 2nd Edition. Springer.

Robert, C. P., Casella, G., 2004. Monte Carlo Statistical Methods, 2nd Edition. Springer.

Rothenberg, T. J., May 1971. Identification in parametric models. Econometrica 39 (3), 577–591.

Royden, H. L., 1988. Real analysis, 3rd Edition. Prentice Hall.

Sanford, A., Martin, G., 2005. Simulation-based Bayesian estimation of affine term structure models. Computational Statistics and Data Analysis, Special Issue on Computational Econometrics 2 49, 527–554.

Shumway, R., Stoffer, D., 1983. An approach to time series smoothing and forecasting using the EM algorithm. Journal of Time Series Analysis 3 (4), 253–264.

Stephens, M., 1997. Bayesian methods for mixtures of normal distributions. Ph.D. thesis, University of Oxford.

Stephens, M., 2000. Dealing with label switching in mixture models. Journal of the Royal Statistical Society B 62, 795–809, part 4.

Stoffer, D. S., Wall, K. D., 1991. Bootstrapping state-space models : Gaussian maximum likelihood estimation and the Kalman filter. Journal of the American Statistical Association 86 (416), 1024–1033.

Watson, M., Engle, R., 1983. Altenative algorithms for the estimation of dynamic factor, MIMIC and varying coefficient regression models. Journal of Econometrics 23, 385–400.