Normalizing Empirically Underidentified Linear State-Space Models.[☆]

Sébastien Blais

Abstract

The likelihood function of LSSMs is invariant under a group of parameter transformations associated to affine transformations of the state vector. Thus, certain parameters are not identified. Identification is usually obtained through normalization: one restricts attention to a particular parameter subspace in order to ensure that parameter point estimators are well defined. Normalizing LSSMs in empirical work has more influence on statistical inference than is commonly appreciated as the invariance property of the likelihood function has at least three implications for Bayesian inference. First, certain parameters have no substantive interpretation. Specifying prior beliefs on such quantities is, at best, conceptually difficult to justify. Second, the affine group provides a natural structure for augmenting the parameter space and improving the numerical efficiency of posterior sampling. Finally, if certain parameter elements are close to being unidentified empirically, the parameter posterior can be multimodal or have extremely large high-order moments, which complicates its interpretation. I propose invariant prior distributions for the parameters of LSSMs that ensure that predictive densities do not depend on parameterization choice. I also present a structural parameter expansion data augmentation (SPX-DA) algorithm for computing the parameter posterior in a simple and efficient manner. Using artificial data, I show how one popular normalization can define an almost invariant parameter posterior when a model is empirically underidentified.

[☆]This draft: May 24, 2017.
Email address: sblais@sebastienblais.com (Sébastien Blais)

1. Introduction

Let $\mathbf{y}_t$ be a N-dimensional vector of observables at time t and ξ_t a K-dimensional vector of unobserved state variables. A Markovian Gaussian linear state-space model (LSSM) is defined by the system of equations

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

$$\mathbf{y}_t = \mathbf{B} + \mathbf{H}\xi_t + \mathbf{w}_t, \tag{2}$$

where $\mathbf{v}_t$ and $\mathbf{w}_t$ are independent Gaussian white noises with covariance matrices $\mathbf{Q}$ and $\mathbf{R}$, respectively. Let also $\Sigma = (\mathcal{I} \otimes \mathcal{I} - \mathbf{F} \otimes \mathbf{F})^{-1} \operatorname{vec}(\mathbf{Q})$ denote the covariance matrix of ξ_t. Equation (1) is referred to as the *state equation* and equation (2) as the *observation equation*. State variables are also known as *factors* and $\mathbf{H}$ as the matrix of *factor loadings*. Factor analysis corresponds to the case $\mathbf{E} = \mathbf{0}$ and $\mathbf{F} = \mathbf{0}$, and structural vector autoregressions correspond to the case $\mathbf{E} = \mathbf{0}$ and $\mathbf{R} = \mathbf{0}$ with $K = N$. I use the generic notation p to denote a density function. The function's support is indicated as a subscript when this precision is required. For instance, $p_\Theta(\theta)$ is a density function with respect to a measure ν on Θ.

The likelihood function of LSSMs is invariant under the group of affine transformations of the latent variables. If $\mathcal{L}$ denotes the space of K-dimensional vectors and $\mathcal{G}$ denotes the space of $K \times K$ invertible matrices, the system (1-2) can be written in terms of the transformed state variable $\eta_t = \mathbf{G}\xi_t + \mathbf{L}$ for any $(\mathbf{L}, \mathbf{G}) \in \mathcal{L} \times \mathcal{G}$ and the log-likelihood function satisfies[1]

$$l(\mathbf{M}_{\mathbf{L},\mathbf{G}}(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q}) | \mathbf{y}) = l(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q} | \mathbf{y}), \tag{3}$$

where

$$\mathbf{M}_{\mathbf{L},\mathbf{G}}(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q}) = \left(\mathbf{B} - \mathbf{H}\mathbf{G}^{-1}\mathbf{L}, \mathbf{H}\mathbf{G}^{-1}, \mathbf{R}, \mathbf{G}\mathbf{E} + \left(\mathcal{I} - \mathbf{G}\mathbf{F}\mathbf{G}^{-1}\right)\mathbf{L}, \mathbf{G}\mathbf{F}\mathbf{G}^{-1}, \mathbf{G}\mathbf{Q}\mathbf{G}^{\top}\right) \tag{4}$$

and $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_T)$. For future reference, the Jacobian of this transformation is $\mathbf{J}_{\mathbf{M}_{\mathbf{L},\mathbf{G}}} =$

[1] The density function of the initial state vector, ξ_1, must satisfy $p(\mathbf{G}\xi_1 + \mathbf{L} | \mathbf{M}_{\mathbf{L},\mathbf{G}}(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q})) = p(\xi_1 | \mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q})$. Certain candidate densities are presented in Section 3.

$\det(\mathbf{G})^{K+2-N}$. The solution to the likelihood-maximization problem is thus not unique. Non-identification is typically broken through normalization: one restricts attention to a particular parameter subspace in order to ensure that parameter point estimators are well defined. A *normalization* is thus a parameter subspace, say $\Theta^N \subseteq \Theta$, that is not observationally restrictive, meaning that it does not contain any information about the observables. In particular, for every $\theta \in \Theta$ there exists a $\theta' \in \Theta^N$ such that $p(\mathbf{y}|\,\theta) = p(\mathbf{y}|\,\theta')$. For instance, under one popular normalization (Harvey, 1989; Geweke and Zhou, 1996), the system (1-2) can be equivalently written as

$$\zeta_t = \mathbf{0} + \tilde{\mathbf{F}}\zeta_{t-1} + \tilde{\mathbf{v}}_t, \tag{5}$$

$$\mathbf{y}_t = \tilde{\mathbf{B}} + \tilde{\mathbf{H}}\zeta_t + \mathbf{w}_t. \tag{6}$$

where $\tilde{\mathbf{H}}$ is $N \times K$ matrix with $\tilde{\mathbf{H}}_{n,k} = 0$ for $k > n$ and $\tilde{\mathbf{H}}_{n,k} > 0$ for $k = n$, and $\text{Cov}[\tilde{\mathbf{v}}_t] = \mathcal{I}$, a $K \times K$ identity matrix.

Let $\theta = \{\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q}\}$ and Θ denote the unnormalized parameter space. The normalized system (5-6) results from imposing normalization

$$\tilde{\Theta} = \{\theta \in \Theta |\, \mathbf{E} = \mathbf{0}; \mathbf{H}_{n,k} = 0 \text{ for } k > n; \mathbf{H}_{n,k} > 0 \text{ for } k = n; \mathbf{Q} = \mathcal{I}\} \tag{7}$$

on the system (1-2). I use the generic notation $\Theta^N = \left\{(\theta_1, \theta_2) \middle|\, (\theta_1, \theta_2) \in \Theta = \Theta_1^N \times \Theta_2^N; \theta_2 = \bar{\theta}_2\right\}$. For a normalization Θ^N the transformation (4) implicitly defines a function $\mathbf{M}_{\Theta^N} : \Theta_1^N \times \mathcal{L} \times \mathcal{G} \to \Theta$. The unnormalized parameter space Θ can thus be generated by any normalization and the affine group. In that sense, (1-2) is the structural overparameterization of a LSSM that is induced by the invariance property of its likelihood function. For $\tilde{\Theta}$, the transformation is

$$\begin{aligned}\mathbf{M}_{\tilde{\Theta}}&\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{L}, \mathbf{G}\right) \\ &= \left(\tilde{\mathbf{B}} - \tilde{\mathbf{H}}\mathbf{G}^{-1}\mathbf{L}, \tilde{\mathbf{H}}\mathbf{G}^{-1}, \mathbf{R}, \left(\mathcal{I} - \mathbf{G}\tilde{\mathbf{F}}\mathbf{G}^{-1}\right)\mathbf{L}, \mathbf{G}\tilde{\mathbf{F}}\mathbf{G}^{-1}, \mathbf{G}\mathbf{G}^{\top}\right).\end{aligned} \tag{8}$$

Where it is well defined[2], its inverse $\mathbf{M}_{\tilde{\Theta}}^{-1} : \Theta \to \tilde{\Theta}_1 \times \mathcal{L} \times \mathcal{G}$ does not have a closed-form solution

[2]The parameter subspace on which the inverse is not well defined is a property of a normalization that influences the shape of the parameter posterior. This relationship is examined in more detail in Section 2.2

but it can be implemented with standard matrix decomposition routines. The expressions *normalization* and *parameterization* are sometimes used interchangeably because there is often a natural parameterization associated to a particular normalization. For instance, $\mathbf{E}$ and $\mathbf{Q}$ are not parameters under (7). In this paper, a change in parameterization always results from a change in normalization and this should not be a source of confusion[3].

One standard way of normalizing a model is restricting the support of the parameter prior. For example, one would approximate the parameter posterior under $\tilde{\Theta}$ by generating a sample from $p_{\tilde{\Theta}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}} \middle| \mathbf{y}\right) \propto p\left(\mathbf{y} \middle| \tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}\right) p_{\tilde{\Theta}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}\right)$. Alternatively, normalization can be operationalized as a mapping (McCulloch and Rossi, 1994; Stephens, 1997; Frühwirth-Schnatter, 2001), that is by generating a sample from $p_{\Theta}\left(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q} | \mathbf{y}\right) \propto p\left(\mathbf{y} | \mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q}\right) p_{\Theta}\left(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q}\right)$ and computing $\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{L}, \mathbf{G}\right) = \mathbf{M}_{\tilde{\Theta}}^{-1}\left(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q}\right)$. Both approaches to normalization are inferentially valid, but they define the same parameter posterior only if

$$p_{\tilde{\Theta}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}\right) = \int_{\mathcal{L}\times\mathcal{G}} p_{\Theta}\left(\mathbf{M}_{\tilde{\Theta}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{L}, \mathbf{G}\right)\right) \left|\mathbf{J}_{\mathbf{M}_{\tilde{\Theta}}}\right| \nu\left(\mathrm{d}\mathbf{L}\, \mathrm{d}\mathbf{G}\right),$$

where $\mathbf{J}_{\mathbf{M}_{\tilde{\Theta}}}$ denotes the Jacobian of the transformation (8). Therefore, parameter priors can be chosen so that both approaches define the same parameter posterior. The main contribution of this paper is showing that operationalizing normalization as a mapping is simpler to implement and numerically more efficient than restricting the parameter prior's support, for any normalization.

Outline

The invariance property of the likelihood function influences statistical inference in at least three ways, which are the subject of the next three sections. First, normalization determines the shape of the parameter posterior and thus its interpretation. The influence of normal-

[3] A simple example of reparameterization that involves more than restricting the parameter space is modeling a standard deviation as $\delta = \ln(\sigma)$. I do not consider this type of reparameterization in this paper.

ization on finite-sample parameter inference has been documented in a number of settings[4]. Normalization does not merely ensure that parameter estimators are well defined, it also has broader implications for inference for it defines the sampling distribution of the parameter MLE and the parameter posterior. In general, Hamilton, Waggoner, and Zha (2007) state that "poor normalizations can lead to multimodal distributions, disjoint confidence intervals, and very misleading characterizations of the true statistical uncertainty." Such problems arise when a model is *empirically underidentified* (or *locally almost unidentified* or *weakly identified*). The first expression was introduced by Kenny (1979), which he defines as "zero or near-zero denominators in the estimates of structural parameters." Dufour and Hsiao (2008) define weak identification in the following terms: "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*." In this paper, I adopt a definition similar to Dufour's (1997) definition of a locally almost unidentified model and I say that a parametric model is *empirically underidentified* by a data sample if the likelihood-maximizing parameter value is close, in terms of statistical uncertainty, to the parameter subspace where the model is unidentified. Thus, empirical underidentification is a joint property of both the (normalized) model and the data sample. In Section 2, I show how the influence of normalization on the parameter posterior in LSSMs is related to inefficient normalization of invariance under certain subgroups of the affine group. In particular, I explain how one popular normalization could define a parameter posterior that is multimodal, almost rotation invariant, or even almost improper.

Second, invariance has consequences for prior specification. LSSMs can be interpreted as hierarchical models in which the parameter elements governing the dynamics of the state vector are hyperparameters. Hyperparameters are often modeled with noninformative pri-

[4]For instance, Hillier (1990) and Millsap (2001) examine the influence of normalization in structural equation models. McMillin (2001) and Rubio-Ramìrez, Waggoner, and Zha (2010) consider structural vector autoregressions. See Koop, Strachan, van Dijk, and Villani (2006) for a recent discussion of the influence of normalization on inference in cointegrated models.

ors, because of either convenience or a lack of prior information (Hobert and Casella, 1996), although this is not standard practice for the hyperparameters of LSSMs. Prior specification can adversely interact with normalization and distort inferences about parameters (Hamilton, Waggoner, and Zha, 2007). This interaction can also have material consequences for model selection (Frühwirth-Schnatter and Lopes, 2010; Frühwirth-Schnatter and Wagner, 2010). Such difficulties arise if $p(\mathbf{y})$ is not independent of normalization choice. They are avoided if the following condition is satisfied for every pair of normalizations (Θ^1, Θ^2):

$$\int_{\Theta^1} p(\mathbf{y}|\,\theta)\, p_\Theta(\theta)\ \nu(\mathrm{d}\theta) = \int_{\Theta^2} p(\mathbf{y}|\,\theta)\, p_\Theta(\theta)\ \nu(\mathrm{d}\theta)\,, \tag{9}$$

where $p_\Theta(\theta)$ is a density with respect to a measure ν on Θ. Requiring that normalization contains no information about the observables thus severely constrains prior specification. In Section 3, I propose invariant parameter priors that do not express prior beliefs over the relative plausibility of observationally equivalent parameter values. These priors satisfy (9) and ensure that statistical inference is not unduly influenced by normalization.

Normalization also affects the computational efficiency of a data augmentation (DA) algorithm for LSSMs (Pitt and Shephard, 1999; Papaspiliopoulos, Roberts, and Sköld, 2003). However, the question of which parameterization performs best has been somewhat subsided by recent techniques that do not rest on the choice of one particular parameterization (See Papaspiliopoulos, Roberts, and Sköld, 2007, for a discussion.). For instance, Yu and Meng (2011) propose an interweaving strategy (IS) that combines two parameterizations within a single sweep of a Gibbs sampler. In a nutshell, their algorithm is assimilable to a reversible Gibbs sampler (Robert and Casella, 2004, algorithm A.41) with a symmetric scan under different parameterizations. Alternatively, Liu and Wu (1999) propose a parameter expansion DA (PX-DA) algorithm. They augment the parameter space with an artificial expansion parameter α such that $\int p(\mathbf{y}, \zeta|\,\theta, \alpha)\ \mathrm{d}\zeta = p(\mathbf{y}|\,\theta)$ for the purpose of improving the mixing properties of the DA algorithm (See Meng and van Dyk, 1999; Liu and Wu, 1999, for related approaches.). IS and PX-DA perform impressively well for one-factor LSSMs, but implementation is dif-

ficult when $K > 1$ and $K \neq N$[5]. State-space models have a structural interpretation that provides a basis for augmenting the parameter space in a more natural manner than the artificial overparameterization of PX-DA. Respecting the likelihood function's structure simplifies implementation. If drawing from a conditional parameter posterior with a support that is restricted by an identifying restriction can be challenging from an analytical or computational point of view, removing that restriction often simplifies this task dramatically[6]. For instance, McCulloch and Rossi (1994) invoke such arguments for motivating a structural overparameterization of the multinomial probit model in which the matrix of regression coefficients and the error covariance matrix are not restricted. Structural overparameterization is thus motivated by analytical or computational convenience rather than numerical efficiency. In Section 4, I propose an implementation of the DA algorithm for the structural overparameterization (1-2) of a LSSM. This structural parameter expansion DA (SPX-DA) algorithm converges at least as fast a standard DA algorithm under any normalization $\Theta^N \subseteq \Theta$. In addition to its numerical efficiency, one key advantage of the SPX-DA algorithm is that it can be easily implemented under any normalization.

Section 5 presents empirical evidence on the numerical efficiency of the SPX-DA algorithm for artificial and real data. In section 6, I illustrate the influence of normalization on the shape of the parameter posterior. I also show that one popular normalization can define an almost invariant parameter posterior and I propose a novel normalization that defines a parameter posterior with more desirable properties.

Elements of group theory

I conclude this introduction with a brief presentation of certain elementary definitions and results from group theory (See Eaton, 1989, for a detailed presentation of group invariance

[5] Simpson, Niemi, and Roy (2017) consider an application of IS in which $\mathbf{H}$ and $\mathbf{F}$ are considered known parameters.

[6] Ruud (1991) remarks that structural overparameterization can simplify implementation of the EM algorithm by replacing certain constrained optimization problems with unconstrained ones.

applications in statistics.). A group is a nonempty set Γ, together with an associative binary operation, that (a) is closed under that operation, (b) has an identity element, and (c) in which every element has an inverse in Γ. The affine group is closed under the operation $(\mathbf{L}_1, \mathbf{G}_1)(\mathbf{L}_2, \mathbf{G}_2) = (\mathbf{G}_1\mathbf{L}_2 + \mathbf{L}_1, \mathbf{G}_1\mathbf{G}_2)$, the identity transformation is $(\mathbf{0}, \mathcal{I})$, and $(\mathbf{L}, \mathbf{G})^{-1} = (-\mathbf{G}^{-1}\mathbf{L}, \mathbf{G}^{-1})$.

For an element $\omega \in \Gamma$ and a measurable subset $\Omega \subseteq \Gamma$, the notation $\omega\Omega \subseteq \Gamma$ denotes the subset resulting from the left-action of ω on every element of Ω through the group's operation. A measure μ_l satisfying $\int_\Omega \mu_l(\mathrm{d}\gamma) = \int_{\omega\Omega} \mu_l(\mathrm{d}\gamma)$ is a left Haar measure on Γ. It is unique up to a multiplicative constant. A right Haar μ_r measure is defined similarly. If Γ is commutative or compact, $\mu_l = \mu_r$ and the Haar measure is said to be unimodular. One important result is that a Haar measure is finite ($\mu(\Gamma) < \infty$) if and only if Γ is compact. A Haar measure on a compact group can therefore be normalized to be a probability measure. Grossly speaking, a Haar measure can be assimilated to a uniform or noninformative measure on Γ. The left Haar measure on the affine group is proportional to $|\det(G)|^{K+1}$.

A parameter subspace in which elements are observationally equivalent to one another can be described by a group of transformations, which can be constructed as follows. For a given group Γ, a function $\mathbf{f} : \Gamma \times \Theta \to \Theta$ satisfying (a) $\mathbf{f}(\mathbf{e}, \theta) = \theta$, for all $\theta \in \Theta$ and where $\mathbf{e}$ is the identity element of Γ; and (b) $\mathbf{f}(\gamma_1\gamma_2, \theta) = \mathbf{f}(\gamma_1, \mathbf{f}(\gamma_2, \theta))$, for all $\gamma_1, \gamma_2 \in \Gamma$ and $\theta \in \Theta$ is said to specify Γ acting on the left of Θ. For each $\gamma \in \Gamma$, define the transformation $\mathbf{M}_\gamma(\theta) = \mathbf{f}(\gamma, \theta)$. Then $\mathcal{M}_\Gamma(\Theta) = \{\mathbf{M}_\gamma | \, \gamma \in \Gamma\}$ is a group under function composition. The notation $\mathcal{M}_\Gamma(\Theta)$ makes dependence on the space Θ explicit: $\mathcal{M}_\Gamma(\Theta)$ is a group of transformations on Θ onto Θ. I will omit this dependence and write $\mathcal{M}_\Gamma$ when this causes no confusion. Notice that the commutativity, compactness and countability of $\mathcal{M}_\Gamma$ correspond to those of Γ.

A function $\psi(\theta)$ is invariant under $\mathcal{M}_\Gamma(\Theta)$ if $\psi(\theta) = \psi(\mathbf{M}_\gamma(\theta))$ for all $\theta \in \Theta$ and all $\mathbf{M}_\gamma \in \mathcal{M}_\Gamma(\Theta)$ (Eaton, 1989, definition 2.4). I will say that a statistical model is invariant under $\mathcal{M}_\Gamma$ if its likelihood function is invariant under that group, implying that $\mathbf{M}_\gamma(\theta)$ observationally equivalent to θ. LSSMs are invariant under $\mathcal{M}_{\mathcal{L}\times\mathcal{G}}(\Theta) = \{\mathbf{M}_{\mathbf{L},\mathbf{G}}(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q}) | \, (\mathbf{L}, \mathbf{G}) \in \mathcal{L} \times \mathcal{G}\}$.

Finally, a density function $p(\theta)$ (which might not be proper) is invariant under $\mathcal{M}_\Gamma(\Theta)$ if $p(\mathbf{M}_\gamma(\theta))\left|\mathbf{J}_{\mathbf{M}_\gamma}\right| = p(\theta)$, for all $\mathbf{M}_\gamma \in \mathcal{M}_\Gamma(\Theta)$, where $\mathbf{J}_{\mathbf{M}_\gamma}$ is the Jacobian of $\mathbf{M}_\gamma$ evaluated at θ.

2. Empirical underidentification

Empirical underidentification has severe consequences for maximum-likelihood 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; Hillier, 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)"

Empirical underidentification has consequences for Bayesian inference as well. In particular, the parameter posterior can be multimodal and credibility regions can be disjoint. In such situations, a parameter point estimator is an inappropriate summary of the parameter posterior distribution for most purposes. Computing a credibility interval as the region between two posterior quantiles is similarly misleading.

As there are many ways to normalizing any given model, it is natural to ask if certain normalizations define parameter posteriors with more desirable properties than alternatives. Symmetric, unimodal parameter posteriors facilitate communication of empirical results for instance. Hamilton, Waggoner, and Zha (2007) propose a theoretical framework for guiding the

choice of normalization according to an *identification principle*[7]. In some models, this principle yields a unique normalization that defines a unimodal parameter posterior. In LSSMs with more than two state variables, it is less straightforward to apply. Although it usefully defines a preorder on normalizations by ruling out uncountably many poor normalizations, it falls short of recommending a unique optimal normalization. Moreover, it does not guarantee that any particular normalization defines a unimodal parameter posterior. The practical guidance that the identification principle offers is thus incomplete and comparing several normalizations is recommended in empirical work (Frühwirth-Schnatter, 2001; Hamilton et al., 2007). In this section, I show how decomposing the affine group into subgroups and operationalizing normalization as a mapping help to understand the influence of normalization on the parameter posterior's shape.

2.1. Decomposing the affine group

If the likelihood function and the parameter prior are invariant under a group of transformations, inefficient normalization defines an almost invariant parameter posterior. The practical implications for the shape of the posterior depend on two properties of the group: countability[8] and compactness. Thus, decomposing the affine group into subgroups will prove useful for better understanding how normalization determines the shape of the parameter posterior.

As a first step, one can decompose an affine transformation into a translation, $(\mathbf{L}, \mathcal{I})\,\xi = \xi + \mathbf{L}$ and a linear transformation $(\mathbf{0}, \mathbf{G})\,\xi = \mathbf{G}\xi$. Finding useful subgroups of the linear group is accomplished by the use of three standard decompositions of square matrices. The QR decomposition factorizes a square matrix $\mathbf{X} = \mathbf{UT}$ into the product of an orthogonal matrix $\mathbf{U}$ and a upper-triangular matrix $\mathbf{T}$. A matrix is orthogonal if its transpose is equal to its inverse. The group of orthogonal matrices, denoted by $\mathcal{U}$, is known as the orthogonal

[7] The framework of Hamilton, Waggoner, and Zha (2007) is presented and extended in Appendix A.

[8] Although every example of a countable group in this paper is finite, countability rather than finiteness is the relevant property. An example of an infinitely countable group is that of translating an angle $\rho \in \Re$ by a multiple of 2π.

group. Orthogonal matrices are sometimes referred to as rotation matrices, but rotation matrices (which I denote by $\mathbf{O} \in \mathcal{O}$) are more properly defined as orthogonal matrices with a determinant equal to 1, *i.e.* $\mathcal{O} = \{\mathbf{O} \in \mathcal{U} | \det(\mathbf{O}) = 1\}$. Two subgroups of the orthogonal group that are not subgroups of the rotation group are the permutation and the reflection groups. Permutation matrices (which I denote by $\mathbf{P} \in \mathcal{P}$) are obtained by permuting the rows of an identity matrix and satisfy $\mathbf{P} = \mathbf{P}^{-1}$. Reflections matrices are diagonal matrices (denoted by $\mathbf{S} \in \mathcal{S}$) with diagonal elements equal to 1 or -1 and satisfy $\mathbf{S} = \mathbf{S}^\top = \mathbf{S}^{-1}$. The QR decomposition of a real square matrix $\mathbf{X}$ always exists and it is unique if $\mathbf{X}$ is invertible and if the diagonal elements of $\mathbf{T}$ are positive. In other words, if $\mathbf{X}$ is invertible, the decomposition is unique up to reflection because $\mathbf{UT} = \left(\mathbf{US}^\top\right)(\mathbf{ST})$ as $\mathbf{US}^\top \in \mathcal{U}$ and $\mathbf{ST}$ is upper triangular.

The singular value decomposition factorizes a symmetric positive definite matrix $\Sigma = \mathbf{UDU}^\top$ as the product of an orthogonal matrix and a scaling matrix $\mathbf{D}$. The group of scaling transformations is the group of diagonal, positive-definite matrices, which I denote by $\mathcal{D}$. The decomposition always exists. It is unique up to reflection and permutation because it satisfies $\mathbf{UDU}^\top = \mathbf{US}^\top \left(\mathbf{SDS}^\top\right) \mathbf{SU}^\top$ and $\mathbf{UDU}^\top = \mathbf{UP}^\top \left(\mathbf{PDP}^\top\right) \mathbf{PU}^\top$ as $\mathbf{SDS}^\top, \mathbf{PDP}^\top \in \mathcal{D}$ and $\mathbf{UP}^\top \in \mathcal{U}$. Finally, the Cholesky decomposition factorizes a symmetric positive definite matrix as $\Sigma = \mathbf{T}^\top\mathbf{T}$. It always exists and it is unique up to reflection because $\mathbf{T}^\top\mathbf{T} = \left(\mathbf{T}^\top\mathbf{S}^\top\right)(\mathbf{ST})$.

The likelihood function of LSSMs is invariant under the translation, scaling, rotation, permutation and reflection groups. The reflection, permutation and rotation groups are subgroups of the orthogonal group, which is compact. In contrast, the translation and scaling groups are not compact. Because a proper parameter posterior cannot be invariant under a group that is not compact, inefficient normalization of translation and scale invariance can have severe consequences: the parameter posterior could be almost improper. This relationship between normalization compactness and posterior propriety is related to the demonstration of Dufour (1997) that, in a classical setting, no valid confidence set which is almost surely bounded does exist for an empirically underidentified parameter with an unbounded range. An example of an almost improper parameter posterior on $\tilde{\Theta}$ will be presented in Section 6.2.

Inefficient normalization of invariance under a compact group has consequences as well. The permutation and reflection subgroups are countable: $\mathcal{M}_{\mathbf{0},\mathcal{S}}$ and $\mathcal{M}_{\mathbf{0},\mathcal{P}}$ respectively contain 2^K and $K!$ transformations. For example, the likelihood function of a LSSM with $K = 3$ latent variables has 48 equivalent lobes. Inefficient normalization of reflection and permutation invariance can thus define a multimodal parameter posterior. Inefficient normalization of invariance under rotation can result in a parameter posterior that exhibits a circular shape. Section 6.1 presents an example of a parameter posterior that is almost invariant under the reflection, permutation and rotation groups.

2.2. Operationalizing normalization as a mapping

In addition to being a central building block of the SPX-DA algorithm, operationalizing normalization as a mapping is useful for practical and theoretical reasons. From a practical perspective, it can be used for reparameterizing a model by postprocessing a posterior sample. This allows the investigator to analyze the influence of an alternative normalization on the posterior posterior at very little computational cost when empirical underidentification difficulties are suspected. It also simplifies the implementation of certain normalizations. For example, in order to implement a DA algorithm under the restriction that $\mathbf{H}^\top\mathbf{H}$ is a diagonal matrix, one would have to parameterize $\mathbf{H}$ in terms of a $(N-1)\times K$ lower triangular matrix of angles ϕ (Heiss and Sannino, 1990; Heiss, 1994) as

$$\mathbf{H} = \mathbf{C}_1\mathbf{C}_2\ldots\mathbf{C}_K\mathbf{W},$$

where

$$\begin{aligned}
\mathbf{C}_k &= \rho_{k,k+1}\rho_{k,k+2}\cdots\rho_{k,N},\\
\mathbf{W}_{(N\times K)} &= \left[\begin{smallmatrix}\mathcal{I}\\ \mathbf{0}\end{smallmatrix}\right],\\
\rho_{i,j} &= \begin{bmatrix}\mathcal{I} & & & \\ & \cos\phi_{i,j} & & -\sin\phi_{i,j} \\ & & \mathcal{I} & \\ & \sin\phi_{i,j} & & \cos\phi_{i,j} \\ & & & & \mathcal{I}\end{bmatrix}_{(N\times N)},
\end{aligned}$$

and specify prior information about ϕ. Operationalizing this normalization as a restriction of the posterior's support requires additional programming and foregoing the convenience of standard distributions. In contrast, operationalizing it as a mapping requires little analytical,

programming or computational effort (Kaufmann and Schumacher, 2013).

From a theoretical point of view, the properties of the mapping $\mathbf{M}_{\Theta^N}$ are informative about the influence of Θ^N on the shape of the parameter posterior. In particular, Θ^N performs inefficiently in a neighborhood of the parameter subspace on which $\mathbf{M}_{\Theta^N}^{-1}$ does not exist. An example best illustrates how the properties of the mapping influence statistical inference. The popular normalization (7) can be written as

$$\tilde{\Theta} = \Theta^{\mathbf{E}} \cap \Theta^{\mathbf{Q}_{\mathcal{I}}} \cap \Theta^{\mathbf{H}_{\triangle}} \cap \Theta^{\mathbf{H}_{d1}},$$

where

$$\Theta^{\mathbf{E}} = \{\theta \in \Theta | \, \mathbf{E} = \mathbf{0}\}, \tag{10}$$

$$\Theta^{\mathbf{Q}_{\mathcal{I}}} = \{\theta \in \Theta | \, \mathbf{Q} = \mathcal{I}\}, \tag{11}$$

$$\Theta^{\mathbf{H}_{\triangle}} = \left\{\theta \in \Theta \middle| \, \tilde{\mathbf{H}}_{n,k} = 0, k > n\right\}, \tag{12}$$

$$\Theta^{\mathbf{H}_{d1}} = \left\{\theta \in \Theta \middle| \, \tilde{\mathbf{H}}_{n,k} > 0, k = n\right\}. \tag{13}$$

In order to operationalize $\tilde{\Theta}$ as a mapping, one computes the inverse of (8). In practice, this is accomplished by finding the functions $\mathbf{L} : \Theta \to \mathcal{L}$ and $\mathbf{G} : \Theta \to \mathcal{G}$ such that $\mathbf{M}_{\mathbf{L}(\theta),\mathbf{G}(\theta)}(\theta) \in \tilde{\Theta}$ for all $\theta \in \Theta$. From (4), $\mathbf{L}(\theta)$ is given by the solution of $\mathbf{0} = \mathbf{GE} + (\mathcal{I} - \mathbf{GFG}^{-1})\mathbf{L}$, which is $\mathbf{L}(\theta) = \mathbf{G}(\mathbf{F} - \mathcal{I})^{-1}\mathbf{E}$. Substituting this solution into (4) reveals the consequences of inefficient normalization of translation invariance for the shape of the parameter posterior: $\tilde{\mathbf{B}} = \mathbf{B} + \mathbf{H}(\mathcal{I} - \mathbf{F})^{-1}\mathbf{E}$ can take extremely large values if $\mathcal{I} - \mathbf{F}$ is close to being singular. Clearly, (8) is not onto Θ because the inverse of $\mathcal{I} - \mathbf{F}$ does not exist everywhere on Θ. In empirical work, $p\left(\tilde{\mathbf{B}} \middle| \mathbf{y}\right)$ could have very fat tails. Heuristically, (10) performs inefficiently because the unconditional expectation of the state vector, $\mathbb{E}[\xi_t] = (\mathcal{I} - \mathbf{F})^{-1}\mathbf{E}$, does not exit if $\mathcal{I} - \mathbf{F}$ is singular. If the state vector is not stationary, the model is translation invariant even if $\mathbf{E}$ is restricted to being equal to a constant vector. This argument is holds independently of the manner that normalization is operationalized. However, it seems likely that operationalizing normalization as a mapping will facilitate exploration of the posterior's fat tails. I present a simulation experiment that supports this intuition in Section 6.2.

The function $\mathbf{G}(\theta)$, implicitly defined by the conditions (11-13), does not have a closed-form expression. But it can be computed using standard matrix decomposition routines. If $\mathbf{H}_{1:K,1:K}$ denotes the first $K \times K$ block of $\mathbf{H}$, one finds the function $\mathbf{G}(\theta)$ by computing[9] the QR decomposition $\mathbf{U}\mathbf{T}_1 = \mathbf{H}^{\top}_{1:K,1:K}$, the Cholesky decomposition $\mathbf{T}_2\mathbf{T}_2^{\top} = \mathbf{U}^{\top}\mathbf{Q}\mathbf{U}$ and the product $\mathbf{G}_1 = {\mathbf{T}_2^{\top}}^{-1}\mathbf{U}^{\top}$. The Cholesky decomposition of $\mathbf{U}^{\top}\mathbf{Q}\mathbf{U}$ and the QR decomposition of $\mathbf{H}_{1:K,1:K}$ are defined up to reflection. Therefore, imposing (11-12) preserves reflection invariance[10]. One imposes (13) by finding the reflection matrix $\mathbf{S}$ such that the diagonal elements of $\mathbf{H}_{1:K,1:K}\mathbf{G}_1^{-1}\mathbf{S}$ are positive. Thus, the solution is $\mathbf{G} = \mathbf{S}\mathbf{G}_1$.

The QR decomposition of $\mathbf{H}^{\top}_{1:K,1:K}$ is not unique if $\mathbf{H}_{1:K,1:K}$ is not invertible. In that case, (11-12) does not break permutation and rotation invariance. The model is locally unidentified on the parameter subspace where $\mathbf{H}_{1:K,1:K}$ is not invertible and the parameter posterior is almost rotation invariant if the likelihood-maximizing value of $\mathbf{H}_{1:K,1:K}$ is close to being singular. This possibility is illustrated with artificial data in Section 6.1. In this example, normalization fails to break permutation and rotation invariance because then chosen block of $\mathbf{H}$ does not have full rank (Geweke and Singleton, 1980), which can result from modeling too many factors (Lopes and West, 2004). In empirical work, the investigator might try several other blocks (Carvalho et al., 2008; Frühwirth-Schnatter and Lopes, 2010) of $\mathbf{H}$ if he suspects that the first block is rank-deficient or select K rows of $\mathbf{H}$ following an *ad hoc* procedure (Forni et al., 2000). A more robust normalization might be desirable however. In the empirical section of this paper I break rotation invariance by imposing $\Theta^{\mathbf{Q}_{\mathcal{D}}} \cap \Theta^{\xi_{\mathcal{I}}}$, where

$$\Theta^{\mathbf{Q}_{\mathcal{D}}} = \{\theta \in \Theta | \, \mathbf{Q} \in \mathcal{D}\}, \tag{14}$$

$$\Theta^{\xi_{\mathcal{I}}} = \{\theta \in \Theta | \, \Sigma = \mathcal{I}\}, \tag{15}$$

[9] An alternative—computing the Cholesky decomposition $\mathbf{T}_1\mathbf{T}_1^{\top} = \mathbf{Q}$, the QR decomposition $\mathbf{U}\mathbf{T}_2 = \mathbf{T}_1^{\top}\mathbf{H}^{\top}_{1:K,1:K}$, and the product $\mathbf{G}_1 = \mathbf{U}^{\top}{\mathbf{T}_1}^{-1}$—defines the same mapping.

[10] Implementations of the QR and Cholesky decompositions typically do not preserve reflection invariance and produce only one of the 2^K solutions. One of these solutions should be selected randomly for operationalizing normalization as a mapping. Similarly, typical implementations of the singular value decomposition produce only one of the $K!2^K$ solutions.

and Σ is the unconditional covariance matrix of the state vector. In this parameter subspace the state vector is unconditionally uncorrelated and has unitary unconditional variance. State innovations are uncorrelated as well. This normalization thus gives simple interpretation to the matrix of factor loadings. It also has advantage that one does not have to arbitrarily choose a particular block of $\mathbf{H}$. Like (11-12), (14-15) preserves reflection invariance, but it preserves permutation invariance too. Permutation normalization can thus be chosen freely. Notice that if the elements of $\mathbf{H}$ are *a priori* uncorrelated they will be so *a posteriori* as well if (15) is imposed and if $\mathbf{R}$ is proportional to an identity matrix. Reflection and permutation normalization based on elements of $\mathbf{H}$ is then often straightforward, as is illustrated in Section 6.1.

3. Prior specification

It is conceptually inconsistent to express prior beliefs over the relative plausibility of observationally equivalent parameter values. For finite mixture distributions, Geweke (2007) argues that "If the state labels have no substantive interpretation, then the prior density must also be permutation invariant." Indeed, specifying prior beliefs on quantities that have no substantive interpretation is, at best, conceptually difficult to justify. Prior information should reflect the invariance property of the likelihood function. If the likelihood function is invariant under a group of transformations, the prior density should be invariant under that group as well. Invariant priors are noninformative about dimensions that have no substantive interpretation. This ensures that that (9) is satisfied and that inference for invariant quantities is not influenced by normalization. If the likelihood function and the parameter prior are invariant, so is the parameter posterior. If the parameter posterior is invariant, predictive densities are not influenced by normalization.

Specifying a proper prior density that is noninformative about observationally equivalent parameter values is possible only if $\mathcal{M}_\Gamma$ is a compact group. Therefore, there exists no proper invariant prior under the affine group. A prior (on Θ) that is invariant under an affine trans-

formation of the state variables must satisfy

$$p_{\Theta}\left(\mathbf{M}_{\mathbf{L},\mathbf{G}}\left(\mathbf{B},\mathbf{H},\mathbf{R},\mathbf{E},\mathbf{F},\mathbf{Q}\right)\right)\left|\det\left(\mathbf{G}\right)\right|^{K+2-N}=p_{\Theta}\left(\mathbf{B},\mathbf{H},\mathbf{R},\mathbf{E},\mathbf{F},\mathbf{Q}\right). \tag{16}$$

For example, a prior proportional to

$$p_{\Theta}^{\mathbf{Q}}\left(\mathbf{B},\mathbf{H},\mathbf{R},\mathbf{E},\mathbf{F},\mathbf{Q}\right)=p\left(\mathbf{R}\right)\det\left(\mathbf{Q}\right)^{-\frac{1-N+(K+1)}{2}} \tag{17}$$

is invariant under affine transformation of the state variables, for any marginal density $p\left(\mathbf{R}\right)$. Conditionally on the state variable, the posterior of $\mathbf{Q}$ is an inverse Wishart distribution with $T-N$ degrees of freedom. There exists other invariant priors. A prior proportional to

$$p_{\Theta}^{\mathbf{H}}\left(\mathbf{B},\mathbf{H},\mathbf{R},\mathbf{E},\mathbf{F},\mathbf{Q}\right)=p\left(\mathbf{R}\right)\left|\det\left(\mathbf{H}^{\top}\mathbf{H}\right)\right|^{1-N+(K+1)} \tag{18}$$

will be useful when $\mathbf{Q}$ is not a parameter element of the model. For example, one can obtain an invariant prior on $\tilde{\Theta}$ by computing the density of $\mathbf{M}_{\tilde{\Theta}}^{-1}\left(\mathbf{B},\mathbf{H},\mathbf{R},\mathbf{E},\mathbf{F},\mathbf{Q}\right)$, the inverse of (8), as I do in Section 4.2.

Invariance criteria have been used for obtaining noninformative parameter prior distributions[11] and other definitions of invariance can be found in the literature. Berger (1985) and Eaton (1989), for instance, define invariance in terms of groups of transformations on the sample space. Alternatively, George and McCulloch (1993) refer to the latter concept as *sample space invariance* and they define *parametrization invariance* in terms of invertible parameter transformations. They show how a sample-space invariant or a parameterization invariant prior can be constructed through the choice of a discrepancy measure. A parameterization invariant prior can also be specified through the choice of a group of transformations. This approach can be criticized on the basis that the group can be chosen in several ways. A prior density satisfying (16) is a parameterization invariant prior where the group of parameter transformations is induced by the invariance property of the likelihood function. In that sense, the choice of the transformation group is not arbitrary[12].

[11] See George and McCulloch (1993) and Kass and Wasserman (1996), for a discussion and Bayarri, Berger, Forte, and García-Donato (2012), for an application to model choice.

[12] Writing the system (1-2) in terms of the alternative transformed state variable $\upsilon_t=\mathbf{G}\left(\xi_t+\mathbf{L}\right)$ implies that an invariant prior must satisfy (16) as well.

I will refer to $p(\xi_1)$ as a prior density even if ξ_1 is not an element of the model's parameter. The likelihood function of a LSSM satisfies (3) only if $p(\mathbf{G}\xi + \mathbf{L}|\, \mathbf{M}_{\mathbf{L},\mathbf{G}}(\mathbf{B},\mathbf{H},\mathbf{R},\mathbf{E},\mathbf{F},\mathbf{Q})) = p(\xi|\, \mathbf{B},\mathbf{H},\mathbf{R},\mathbf{E},\mathbf{F},\mathbf{Q})$. There are at least three ways of specifying $p(\xi_1)$ so that this conditions holds. A first possibility is specifying an hierarchical prior and treating its hyperparameters as elements of the model's parameter. For instance, if ones specifies a multivariate normal density with hyperparameters μ_{ξ_1} and Σ_{ξ_1} then the model's parameter is $\{\mathbf{B},\mathbf{H},\mathbf{R},\mathbf{E},\mathbf{F},\mathbf{Q},\mu_{\xi_1},\Sigma_{\xi_1}\}$, the density satisfies $p(\mathbf{G}\xi_1 + \mathbf{L}|\, \mathbf{G}\mu_{\xi_1} + \mathbf{L}, \mathbf{G}\Sigma_{\xi_1}\mathbf{G}) = p(\xi_1|\, \mu_{\xi_1}, \Sigma_{\xi_1})$, and a prior proportional to $p(\mathbf{R})\det(\mathbf{Q})^{-\frac{2-N+2(K+1)}{2}}$ is invariant. Notice that (3) is not satisfied if $p(\xi_1)$ is a multivariate normal density with fixed parameters. Another possibility is assuming stationarity and specifying a normal prior with mean and variance equal to $(\mathcal{I} - \mathbf{F})^{-1}\mathbf{E}$ and $(\mathcal{I}\otimes\mathcal{I} - \mathbf{F}\otimes\mathbf{F})^{-1}\mathrm{vec}(\mathbf{Q})$, which does not involve additional parameter elements. A third possibility is specifying a flat prior on the initial state vector.

Computational or other considerations might lead one to specifying diffuse conditionally conjugate priors. But doing so can have severe consequences for parameter estimation and model selection because conditionally conjugate priors do not satisfy (9). For example, Hamilton et al. (2007) warn that prior specification can adversely interact with normalization and distort inferences about parameters. In particular, they show how specifying a diffuse normal prior for the location parameters of a mixture of two normal distributions can lead to severe bias in parameter estimation; the more diffuse the prior, the larger the bias. Carvalho et al. (2008) remark that the ordering of the observed variable is an important modeling decision under $\tilde{\Theta}$ for model selection. In order to make statistical inference independent of the ordering choice, Frühwirth-Schnatter and Lopes (2010) and Kaufmann and Schumacher (2013) propose methods for specifying priors that respect the invariance of the likelihood function under orthogonal transformation.

For a dynamic linear trend model, Frühwirth-Schnatter and Wagner (2010) argue that a normal prior on $\tilde{\mathbf{H}}$ is more appropriate for model selection applications than an inverse gamma prior on $\mathbf{Q}$ under an alternative scale normalization. In order to better understand how scale normalization affects inference with conditionally conjugate priors, consider a simple LSSM

with $N = K = 1$. Under the scale normalization that imposes $\mathbf{Q} = 1$, a zero-mean normal prior on the factor loading, $\tilde{\mathbf{H}} \sim \mathcal{N}(0, \sigma^2)$, is conditionally conjugate. This distributional assumption implies that $\tilde{\mathbf{H}}^2$ is gamma-distributed, $\tilde{\mathbf{H}}^2 \sim \mathcal{G}(\frac{1}{2}, 2\sigma)$. By scale invariance, this prior is equivalent to $\mathbf{Q} \sim \mathcal{G}(\frac{1}{2}, 2\sigma)$ under the scale normalization that imposes $\mathbf{H} = 1$, which is not conditionally conjugate. A standard conditionally conjugate prior for a variance parameter is an inverse gamma distribution, which attributes much less weight to neighborhoods of zero than a gamma distribution. Indeed, it is well known that the hyperparameters of the inverse gamma prior have a strong influence on the posterior of a variance parameter if its value is close to zero, where the model is locally unidentified.

4. Posterior sampling

In this section I describe the SPX-DA algorithm and I explore its relation to the DA and PX-DA algorithms[13]. In particular, I elicit the conditions under which these algorithms define the same parameter posterior on $\tilde{\Theta}$. Then I show that the SPX-DA algorithm corresponds to a particular implementation of the PX-DA algorithm with certain optimality properties. Using $\tilde{\Theta}$ for this discussion simplifies exposition but is not restrictive, although implementing the DA and PX-DA algorithms under alternative normalizations could be analytically or computationally challenging. In contrast, the simplicity and the numerical efficiency of the SPX-DA algorithm are independent of the particular choice of normalization.

4.1. The posterior sampling algorithms

For (5-6), one simple DA algorithm that operationalizes normalization as a restriction of the posterior's support would proceed as follows:

[13]The posterior samplers that I describe in this section involve a naive accept-reject step that operationalizes reflection normalization or imposes stationarity. They are not, therefore, pure Gibbs samplers. For notational clarity however, I omit this dependence on the current state of the Markov chain and I write, for example, $p_{\tilde{\Theta}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}} \middle| \zeta'\right)$ instead of $p_{\tilde{\Theta}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}} \middle| \zeta', \tilde{\mathbf{B}}', \tilde{\mathbf{H}}', \mathbf{R}', \tilde{\mathbf{F}}'\right)$.

DA Algorithm 1.

1. Draw ζ' from $p\left(\zeta\middle|\tilde{\mathbf{B}},\tilde{\mathbf{H}},\mathbf{R},\tilde{\mathbf{F}},\mathbf{y}\right)$;
2. Draw $\left(\tilde{\mathbf{B}}',\tilde{\mathbf{H}}',\mathbf{R}',\tilde{\mathbf{F}}'\right)$ from

$$p_{\tilde{\Theta}}^{DA}\left(\tilde{\mathbf{B}},\tilde{\mathbf{H}},\mathbf{R},\tilde{\mathbf{F}}\middle|\zeta',\mathbf{y}\right)\propto p\left(\mathbf{y},\zeta'\middle|\tilde{\mathbf{B}},\tilde{\mathbf{H}},\mathbf{R},\tilde{\mathbf{F}}\right)p_{\tilde{\Theta}}^{DA}\left(\tilde{\mathbf{B}},\tilde{\mathbf{H}},\mathbf{R},\tilde{\mathbf{F}}\right),$$

where the support of $p_{\tilde{\Theta}}^{DA}\left(\tilde{\mathbf{B}},\tilde{\mathbf{H}},\mathbf{R},\tilde{\mathbf{F}}\right)$ is restricted to $\tilde{\Theta}$ and $\zeta=(\zeta_1,\ldots,\zeta_T)$. Notice that if the parameter prior is proportional to

$$p_{\tilde{\Theta}}^{\mathbf{R}}=p\left(\mathbf{R}\right), \tag{19}$$

for any prior conditionally conjugate $p\left(\mathbf{R}\right)$, the conditional parameter posterior can be factorized as

$$p_{\tilde{\Theta}}^{DA}\left(\tilde{\mathbf{B}},\tilde{\mathbf{H}},\mathbf{R},\tilde{\mathbf{F}}\middle|\zeta,\mathbf{y}\right)=p^{DA}\left(\mathbf{R}\middle|\zeta,\mathbf{y}\right)p^{DA}\left(\tilde{\mathbf{B}},\tilde{\mathbf{H}}\middle|\mathbf{R},\zeta,\mathbf{y}\right)p^{DA}\left(\tilde{\mathbf{F}}\middle|\zeta\right), \tag{20}$$

in which each factor is a standard distribution, up to a naive accept-reject step that operationalizes reflection normalization and imposes stationarity. Without this accept-reject step, the DA Algorithm 1 would be a two-stage Gibbs sampler.

For a LSSM, if $\mathcal{A}$ is a group and an expansion parameter $\alpha\in\mathcal{A}$ indexes a differentiable mapping $\mathbf{M}_\alpha\left(\zeta\right)$ such that $p\left(\mathbf{y}\middle|\mathbf{M}_\alpha\left(\zeta\right),\theta,\alpha\right)=p\left(\mathbf{y}\middle|\mathbf{M}_\alpha\left(\zeta\right),\theta\right)$, PX-DA sampling (Liu and Wu, 1999) could proceed as follows[14]:

PX-DA Algorithm 2.

1. Draw ζ' from $p\left(\zeta\middle|\tilde{\mathbf{B}},\tilde{\mathbf{H}},\mathbf{R},\tilde{\mathbf{F}},\mathbf{y}\right)$;
2. Draw $\left(\tilde{\mathbf{B}}',\tilde{\mathbf{H}}',\mathbf{R}',\tilde{\mathbf{F}}',\alpha^*\right)$ from

$$\begin{aligned}&p_{\tilde{\Theta}\times\mathcal{A}}^{PX-DA}\left(\tilde{\mathbf{B}},\tilde{\mathbf{H}},\mathbf{R},\tilde{\mathbf{F}},\alpha\middle|\zeta',\mathbf{y}\right)\\&\quad\propto p\left(\mathbf{y},\mathbf{M}_\alpha\left(\zeta'\right)\middle|\tilde{\mathbf{B}},\tilde{\mathbf{H}},\mathbf{R},\tilde{\mathbf{F}}\right)\left|\mathbf{J}_{\mathbf{M}_\alpha}\right|p_{\tilde{\Theta}}^{DA}\left(\tilde{\mathbf{B}},\tilde{\mathbf{H}},\mathbf{R},\tilde{\mathbf{F}}\right)p_{\mathcal{A}}^{PX-DA}\left(\alpha\middle|\tilde{\mathbf{B}},\tilde{\mathbf{H}},\mathbf{R},\tilde{\mathbf{F}}\right),\end{aligned}$$

where $\mathbf{J}_{\mathbf{M}_\alpha}$ denotes the Jacobian of $\mathbf{M}_\alpha\left(\zeta\right)$ evaluated at ζ and $p_{\mathcal{A}}^{PX-DA}\left(\alpha\middle|\tilde{\mathbf{B}},\tilde{\mathbf{H}},\mathbf{R},\tilde{\mathbf{F}}\right)$ is (a) a proper density function, (b) the improper limit of a sequence of proper priors, or (c) propor-

[14] In this and the other algorithms described in this paper, starred symbols denote intermediate quantities.

tional to the left Haar measure on $\mathcal{A}$. If $p_{\mathcal{A}}^{PX-DA}\left(\alpha\middle|\tilde{\mathbf{B}},\tilde{\mathbf{H}},\mathbf{R},\tilde{\mathbf{F}}\right)=p_{\mathcal{A}}^{PX-DA}(\alpha)$, Liu and Wu (1999) show that the PX-DA Algorithm 2 converges as least as fast the DA Algorithm 1. If $p_{\mathcal{A}}^{PX-DA}(\alpha)$ is proportional to the left Haar measure on $\mathcal{A}$, it converges at least as fast as a DA algorithm under any normalization $\Theta^N \subseteq \tilde{\Theta}\times\mathcal{A}$.

The expansion parameter α in PX-DA is not identified by the data, *i.e.* $p(\mathbf{y}|\theta,\alpha)=p(\mathbf{y}|\theta)$. Identification is generally considered to be a property of the likelihood function that is not of particular interest to Bayesian econometricians. As long as the parameter posterior is proper, a Gibbs sampler converges under fairly mild conditions (Roberts and Smith, 1994; Hobert, Robert, and Goutis, 1997). If the parameter posterior on the unnormalized parameter space Θ is proper, it can be recovered by a standard DA algorithm. If a sample from the parameter posterior on a particular normalized parameter space Θ^N is needed, it can be obtained by operationalizing normalization as a mapping of each element of a sample from the posterior on Θ to an observationally equivalent parameter value in Θ^N. Stephens (1997) and Frühwirth-Schnatter (2001) use such a strategy for computing the parameter posterior in a finite mixture of normal distributions, and McCulloch and Rossi (1994) do so in a multinomial probit model. For finite mixture distributions, this transformation corresponds to permuting certain elements of the parameter vector and its Jacobian is identically equal to one. If the parameter prior is invariant under permutation, operationalizing normalization as a mapping is therefore equivalent to operationalizing it as a restriction of the prior's support. For the multinomial probit model, the mapping corresponds to scaling certain elements of the parameter vector by another element. The Jacobian of this transformation is not identically equal to one. The two approaches to operationalizing normalization are not equivalent and define different parameter posteriors unless the parameter priors are explicitly chosen for this equivalence to hold. Clearly, both approaches to normalization are inferentially valid and any difference in the posterior they define can be attributed to prior specification.

The SPX-DA algorithm takes the unnormalized parameter space, Θ, as an expanded parameter space. This parameter space is induced by the invariance property of the likelihood function under $\mathcal{M}_{\mathcal{L}\times\mathcal{G}}$. Because Θ in not compact, the parameter posterior is improper if

the parameter prior is improper. Thus, $p(\xi|\,\theta)$ is not well defined if the parameter prior is invariant. After each sweep of the sampler, normalization is operationalized by mapping the parameter vector to $\Theta^N \times \mathcal{L} \times \mathcal{G}$ so that the conditional posterior of the state vector is well defined. Under $\tilde{\Theta}$, the SPX-DA algorithm proceeds as follows:

SPX-DA Algorithm 3.

1. Draw ξ^* from $p\left(\xi \middle| \tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \mathbf{0}, \tilde{\mathbf{F}}, \mathcal{I}, \mathbf{y}\right)$;
2. Draw $(\mathbf{B}^*, \mathbf{H}^*, \mathbf{R}^*, \mathbf{E}^*, \mathbf{F}^*, \mathbf{Q}^*)$ from
$$p_{\Theta}^{SPX-DA}(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q}|\,\xi^*, \mathbf{y}) \propto p(\mathbf{y}, \xi^*|\,\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q}) \times p_{\Theta}^{SPX-DA}(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q});$$
3. Compute $\left(\tilde{\mathbf{B}}', \tilde{\mathbf{H}}', \mathbf{R}', \tilde{\mathbf{F}}', \mathbf{L}^*, \mathbf{G}^*\right) = \mathbf{M}_{\tilde{\Theta}}^{-1}(\mathbf{B}^*, \mathbf{H}^*, \mathbf{R}^*, \mathbf{E}^*, \mathbf{F}^*, \mathbf{Q}^*)$,

where $\mathbf{M}_{\tilde{\Theta}}^{-1}$ is the inverse of (8). The inverse transformation does not have a explicit solution but its computation requires only simple linear algebra operations that are described in Section 2.2. Compared to PX-DA and IS, implementing an SPX-DA algorithm is extremely simple and requires little programming. In particular, if the parameter prior is proportional to (17), the conditional parameter posterior can be factorized as

$$p_{\Theta}^{SPX-DA}(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q}|\,\xi^*, \mathbf{y}) = p^{SPX-DA}(\mathbf{R}|\,\xi^*, \mathbf{y})\, p^{SPX-DA}(\mathbf{B}, \mathbf{H}|\,\mathbf{R}, \xi^*, \mathbf{y}) \times p^{SPX-DA}(\mathbf{Q}|\,\xi^*)\, p^{SPX-DA}(\mathbf{E}, \mathbf{F}|\,\mathbf{Q}, \xi^*)$$

in which each factor is a standard distribution, up to an accept-reject step that operationalize imposes stationarity. Implementing alternative normalizations, *e.g.* imposing $\mathbf{H}^\top\mathbf{H} = \mathcal{I}$ or $\mathrm{Cov}[\xi_t] = \mathcal{I}$, requires only simple algebra.

4.2. Statistical equivalence

In general, the DA Algorithm 1 and SPX-DA Algorithm 3 define different parameter posteriors. Because the likelihood function is invariant under $\mathcal{M}_{\mathcal{L}\times\mathcal{G}}$, any difference can be interpreted as the result of different parameter prior specifications. As a consequence, both algorithm define the same posterior for a particular choice of the parameter priors. Liu and Wu

(1999) show that the DA Algorithm 1 and PX-DA Algorithm 2 define the same parameter posterior. If the PX-DA Algorithm 2 and SPX-DA Algorithm 3 define the same posterior for a particular choice of priors, so does the DA Algorithm 1.

For this analysis, I define the expansion $\alpha \in \mathcal{A}$ so that $\tilde{\Theta} \times \mathcal{A}$ and Θ are equinumerous. A natural choice is $\alpha = (\mathbf{L}, \mathbf{G})$. Let $p^{PX-DA}_{\tilde{\Theta}\times\mathcal{L}\times\mathcal{G}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{L}, \mathbf{G} \middle| \mathbf{y}\right)$ denote the parameter posterior defined by the PX-DA Algorithm 2 corresponding to this choice. As $(\mathbf{L}, \mathbf{G})$ is not identified by the data,

$$p^{PX-DA}_{\tilde{\Theta}\times\mathcal{L}\times\mathcal{G}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{L}, \mathbf{G} \middle| \mathbf{y}\right) = p^{DA}_{\tilde{\Theta}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}} \middle| \mathbf{y}\right) p^{PX-DA}_{\mathcal{L}\times\mathcal{G}}\left(\mathbf{L}, \mathbf{G} \middle| \tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}\right).$$

The parameter posterior defined by the SPX-DA Algorithm 3 is

$$\begin{aligned} &p^{SPX-DA}_{\tilde{\Theta}\times\mathcal{L}\times\mathcal{G}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{L}, \mathbf{G} \middle| \mathbf{y}\right) \\ &\qquad = p^{SPX-DA}_{\Theta}\left(\mathbf{M}_{\tilde{\Theta}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{L}, \mathbf{G}\right) \middle| \mathbf{y}\right) \left|\mathbf{J}_{\mathbf{M}_{\tilde{\Theta}}}\right| \\ &\qquad \propto p\left(\mathbf{y} \middle| \mathbf{M}_{\tilde{\Theta}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{L}, \mathbf{G}\right)\right) p^{SPX-DA}_{\tilde{\Theta}}\left(\mathbf{M}_{\tilde{\Theta}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{L}, \mathbf{G}\right)\right) \left|\mathbf{J}_{\mathbf{M}_{\tilde{\Theta}}}\right|, \end{aligned}$$

where $\mathbf{J}_{\mathbf{M}_{\tilde{\Theta}}}$ denotes the Jacobian of the transformation (8). If $K = 1$, $\mathbf{J}_{\mathbf{M}_{\tilde{\Theta}}} = \det(\mathbf{G})^{1-N} \det\left(1 - \tilde{\mathbf{F}}\right)$[15]. By the invariance property of the likelihood function,

$$p\left(\mathbf{y} \middle| \mathbf{M}_{\tilde{\Theta}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{L}, \mathbf{G}\right)\right) = p\left(\mathbf{y} \middle| \tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \mathbf{0}, \tilde{\mathbf{F}}, \mathcal{I}\right).$$

Therefore, the PX-DA and SPX-DA algorithms define the same parameter posterior distribution if

$$p^{DA}_{\tilde{\Theta}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}\right) p^{PX-DA}_{\mathcal{L}\times\mathcal{G}}\left(\mathbf{L}, \mathbf{G} \middle| \tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}\right) = p^{SPX-DA}_{\Theta}\left(\mathbf{M}_{\tilde{\Theta}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{L}, \mathbf{G}\right)\right) \left|\mathbf{J}_{\mathbf{M}_{\tilde{\Theta}}}\right|, \tag{21}$$

or equivalently if

$$p^{SPX-DA}_{\Theta}(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q}) = p^{PX-DA}_{\tilde{\Theta}\times\mathcal{L}\times\mathcal{G}}\left(\mathbf{M}^{-1}_{\tilde{\Theta}}(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q})\right) \left|\mathbf{J}_{\mathbf{M}^{-1}_{\tilde{\Theta}}}\right|. \tag{22}$$

For example, $\mathbf{J}_{\mathbf{M}^{-1}_{\tilde{\Theta}}} = \det(\mathbf{Q})^{\frac{N-1}{2}} \det(1 - \mathbf{F})^{-1}$ if $K = 1$. If $p^{DA}_{\tilde{\Theta}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}\right)$ is proportional to (19) and $p^{PX-DA}_{\mathcal{L}\times\mathcal{G}}\left(\mathbf{L}, \mathbf{G} \middle| \tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}\right) \propto |\det(\mathbf{G})|^2$ (*i.e.* the left Haar measure on the affine

[15] Computing $\mathbf{J}_{\mathbf{M}_{\tilde{\Theta}}}$ when $K > 1$ is complicated by the fact that $\tilde{\mathbf{H}}$ has a lower triangular block and that $\mathbf{G}\mathbf{G}^{\top}$ is a symmetric matrix.

group when $K = 1$), specifying a prior proportional to

$$p_{\Theta}^{SPX-DA}(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q}) = p(\mathbf{R}) \left| \det(\mathbf{Q})^{\frac{N+1}{2}} \det(1 - \mathbf{F})^{-1} \right| \tag{23}$$

ensures that the SPX-DA algorithm recovers $p_{\tilde{\Theta}}^{DA}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}} \middle| \mathbf{y}\right)$. Using the invariant prior (18), one can obtain a prior on $\tilde{\Theta}$ that is invariant under $\mathcal{M}_{\mathcal{L}\times\mathcal{G}}$. If $K = 1$, specifying a prior proportional to

$$p_{\tilde{\Theta}\times\mathcal{L}\times\mathcal{G}}^{PX-DA}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{L}, \mathbf{G}\right) = p(\mathbf{R}) \left| \det\left(\tilde{\mathbf{H}}^{\top}\tilde{\mathbf{H}}\right)^{3-N} \det(\mathbf{G})^{N-5} \det\left(1 - \tilde{\mathbf{F}}\right) \right| \tag{24}$$

expresses no prior beliefs over the relative plausibility of observationally equivalent parameter values.

The PX-DA and SPX-DA algorithms define the same parameter posterior if parameter priors are chosen appropriately. However, $p_{\tilde{\Theta}\times\mathcal{L}\times\mathcal{G}}^{PX-DA}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{L}, \mathbf{G} \middle| \zeta, \mathbf{y}\right)$ is not a standard distribution if the prior is proportional to (24). Sampling could be implemented by a Metropolis-Hastings step but numerical efficiency would be impaired. Similarly, sampling from $p_{\Theta}^{SPX-DA}(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q} | \xi, \mathbf{y})$ is computationally more demanding if the prior is proportional to (23). For this reason, comparing the numerical efficiency of the DA and SPX-DA algorithm for a particular parameter posterior is futile. In Section 5, I examine the mixing properties of the DA and SPX-DA algorithms with parameter priors proportional to (19) and (17), respectively.

Deriving the conditions under which the PX-DA and SPX-DA algorithms define the same parameter posterior sheds light on one way in which prior specification can distort parameter inference in LSSMs. Condition (23) implies that an invariant prior attributes a lower probability to neighborhoods of $\tilde{\mathbf{F}} = 1$ than the flat prior (19). As a consequence, the mode of the posterior density of $\tilde{\mathbf{F}}$ will larger than the maximum-likelihood estimate when the prior is proportional to (19). An invariant prior, in contrast, does not distort inference in this manner, as is illustrated in Section 6.2. Moreover, because posterior sampling can be numerically unstable when $\mathcal{I} - \tilde{\mathbf{F}}$ is close to being singular, specifying a prior proportional to (19) could impair a posterior sampler's numerical stability when the observables are highly persistent and the sample size is relatively small.

4.3. SPX-DA as an implementation of PX-DA

Defining $\mathbf{M}_{\mathbf{L},\mathbf{G}}(\zeta) = \mathbf{G}^{-1}(\zeta - \mathbf{L})$, $p\left(\mathbf{y}, \mathbf{M}_{\mathbf{L},\mathbf{G}}(\zeta) \middle| \tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}\right)$ is the density associated to the system

$$\mathbf{G}^{-1}(\zeta_t - \mathbf{L}) = \mathbf{0} + \tilde{\mathbf{F}}\mathbf{G}^{-1}(\zeta_{t-1} - \mathbf{L}) + \tilde{\mathbf{v}}_t, \tag{25}$$

$$\mathbf{y}_t = \tilde{\mathbf{B}} + \tilde{\mathbf{H}}\mathbf{G}^{-1}(\zeta_t - \mathbf{L}) + \mathbf{w}_t. \tag{26}$$

The Jacobian of the transformation is $\det(\mathbf{G})^{-T}$ and Algorithm 2 can be written as follows:

PX-DA Algorithm 2.1

1. Draw ζ' from $p\left(\zeta \middle| \tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{y}\right)$;
2. Draw $\left(\tilde{\mathbf{B}}', \tilde{\mathbf{H}}', \mathbf{R}', \tilde{\mathbf{F}}', \mathbf{L}^*, \mathbf{G}^*\right)$ from

$$p^{PX-DA}_{\tilde{\Theta}\times\mathcal{L}\times\mathcal{G}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{L}, \mathbf{G} \middle| \zeta', \mathbf{y}\right)$$
$$\propto p\left(\mathbf{y}, \mathbf{M}_{\mathbf{L},\mathbf{G}}(\zeta') \middle| \tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}\right) \left|\det(\mathbf{G})^{-T}\right| p^{PX-DA}_{\tilde{\Theta}\times\mathcal{L}\times\mathcal{G}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{L}, \mathbf{G}\right).$$

Drawing from $p^{PX-DA}_{\tilde{\Theta}\times\mathcal{L}\times\mathcal{G}}\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{L}, \mathbf{G} \middle| \zeta', \mathbf{y}\right)$ can be simplified by writing the system (26-25) as

$$\zeta_t = \left(\mathcal{I} - \mathbf{G}\tilde{\mathbf{F}}\mathbf{G}^{-1}\right)\mathbf{L} + \mathbf{G}\tilde{\mathbf{F}}\mathbf{G}^{-1}\zeta_{t-1} + \mathbf{G}\tilde{\mathbf{v}}_t,$$
$$\mathbf{y}_t = \tilde{\mathbf{B}} - \tilde{\mathbf{H}}\mathbf{G}^{-1}\mathbf{L} + \tilde{\mathbf{H}}\mathbf{G}^{-1}\zeta_t + \mathbf{w}_t$$

and making the change of variable defined by (8). Doing so produces the following algorithm:

PX-DA Algorithm 2.3.

1. Draw ζ' from $p\left(\zeta \middle| \tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}, \mathbf{y}\right)$;
2. Draw $(\mathbf{B}^*, \mathbf{H}^*, \mathbf{R}^*, \mathbf{E}^*, \mathbf{F}^*, \mathbf{Q}^*)$ from

$$p^{PX-DA}_{\Theta}(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q} | \zeta', \mathbf{y})$$
$$\propto p(\mathbf{y}, \zeta' | \mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q})\, p^{PX-DA}_{\tilde{\Theta}\times\mathcal{L}\times\mathcal{G}}\left(\mathbf{M}^{-1}_{\tilde{\Theta}}(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{E}, \mathbf{F}, \mathbf{Q})\right) \left|\mathbf{J}_{\mathbf{M}^{-1}_{\tilde{\Theta}}}\right|.$$

3. Compute $\left(\tilde{\mathbf{B}}', \tilde{\mathbf{H}}', \mathbf{R}', \tilde{\mathbf{F}}', \mathbf{L}^*, \mathbf{G}^*\right) = \mathbf{M}^{-1}_{\tilde{\Theta}}(\mathbf{B}^*, \mathbf{H}^*, \mathbf{R}^*, \mathbf{E}^*, \mathbf{F}^*, \mathbf{Q}^*)$.

This algorithm is identical to the SPX-DA Algorithm 3 if prior specification satisfies (21) or

(22). If the prior is proportional to (17), $\left(\tilde{\mathbf{B}}, \tilde{\mathbf{H}}, \mathbf{R}, \tilde{\mathbf{F}}\right)$ and $(\mathbf{L}, \mathbf{G})$ are *a priori* independent. When this is the case, the SPX-DA algorithm corresponds to an implementation of the PX-DA algorithm in which the invariance property of the likelihood function is exploited for choosing an expansion scheme with two properties. First, the expanded parameter space respects the model's structure and simplifies sampling from the condition parameter posterior. Second, because Θ corresponds to an expansion of *any* normalization, it converges at least as fast as a DA algorithm under any normalization $\Theta^N \subset \Theta$. Liu and Wu (1999) show that the PX-DA Algorithm 2 converges at least as fast as a DA algorithm under any normalization $\Theta^N \subseteq \tilde{\Theta} \times \mathcal{A}$ if $p_{\mathcal{A}}^{PX-DA}(\alpha)$ is proportional to the left Haar measure on $\mathcal{A}$. This condition is incompatible with (16) and it is unnecessary when $\mathcal{A} = \mathcal{L} \times \mathcal{G}$.

5. Numerical efficiency

In this section[16], I use artificial as well as real data for analyzing the mixing properties of the SPX-DA Algorithm 3 with a parameter prior proportional to (17). The mixing properties of the DA Algorithm 1 with a prior proportional to (19) are presented as a benchmark. I use a forward-filtering-backward-simulation (FSBS) algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994) for drawing the state variables in both algorithms. For simplicity, I specify a flat prior on ξ_1, which requires an information-filtering version of the algorithm (See Grewal and Andrews, 2008, for an detailed presentation.). More efficient algorithms exist for drawing the state variable (See Kim et al., 1998; McCausland, 2012, for example), but the results presented in this section suggest that little performance improvement could be obtained by improving the sampling of the state variable.

[16]Because parameter priors are improper, the method proposed by Geweke (2004) cannot be used for validating the algorithms implemented in this section. However, they have been thoroughly validated with proper parameter priors and an alternative blocking scheme. That the algorithms define the same parameter posterior has been validated as well with priors proportional to (19) and (23).

5.1. Measuring numerical efficiency

There is no universal definition of numerical efficiency, but it is often presented as a property of the posterior sampler. Measuring numerical efficiency, by contrast, requires reference to an inference problem. For instance, efficiency can be defined as the computing time—the number of iterations times the computing time per iteration—required for estimating a quantity with a certain precision. The computing time per iteration for the DA and SPX-DA algorithms is dominated by the drawing of the state variable and are thus almost identical. For MCMC algorithms, the number of iterations required for estimating a scalar-valued function of the parameter vector, say $h(\theta)$, with a certain precision is influenced by the autocorrelation time of posterior sample, $\tau = 1 + 2\sum_{q=1}^{\infty} \rho(q)$, where $\rho(q)$ is the autocorrelation of $h(\theta)$ at lag q. For a posterior sample of M iterations, the variance of the posterior mean of $h(\theta)$ is equal to $\frac{\mathrm{Var}(h(\theta))}{M} \times \tau$. In other words, an MCMC sampler requires a simulation size τ times larger than an i.i.d. sampler for estimating $h(\theta)$ with the same precision. The inefficiency factor of a quantity is an estimator of its autocorrelation time. In this paper, it is computed as

$$1 + 2\sum_{q=1}^{500} \left(1 - \frac{q}{500}\right) \hat{\rho}(q), \tag{27}$$

where $\hat{\rho}(q)$ is the sample autocorrelation of the quantity at lag q. Parameter inefficiency factors are relevant measures of numerical efficiency if the inference objective is parameter estimation, but their usefulness is less apparent for other objectives. For instance, inefficient estimation of the parameter vector would not be a concern in forecasting applications if predictive densities were computed efficiently.

In empirically underidentified finite mixture models, Geweke (2007) shows that, contrary to assessments in earlier literature, a standard DA algorithm could reliably recover the informational content of the parameter posterior with a reasonable number of iterations. Finite mixture models are permutation invariant. He stresses the distinction between quantities of interest that are permutation invariant (*e.g.* predictions) and those that are not (*e.g.* certain component distribution parameters). He finds that, although estimating the latter precisely can be difficult, estimation of invariant quantities poses no particular problem. In finite mix-

ture models, a standard DA algorithm's ability to compute the posterior of invariant quantities efficiently is not impaired by empirical underidentification. I will assess whether this holds true in LSSMs as well by computing the inefficiency factors associated to certain invariant quantities.

5.2. Simulation setup

I use artificial data for examining the performance of the DA and SPX-DA algorithms for LSSMs. The data is generated as

$$\zeta_{t+1} = \mathbf{0} + \alpha_F \begin{bmatrix} 1 & & & \\ & 0.75^1 & & \\ & & \ddots & \\ & & & 0.75^{K-1} \end{bmatrix} \zeta_t + \tilde{\mathbf{v}}_t \tag{28}$$

$$\mathbf{Y}_t = \mathbf{0} + \begin{bmatrix} \mathcal{I} \\ \mathbf{1} \end{bmatrix} \zeta_t + \sqrt{\alpha_R}\mathbf{w}_t, \tag{29}$$

where $\tilde{\mathbf{v}}_t$ and $\mathbf{w}_t$ are vectors of independent standard normal variables, $\mathbf{1}$ is a $(N-K)\times K$ matrix of ones, α_F and α_R are scalars. In this section, $K = 2$, $N = 4$ and $T = 200$. My results are based on $50,000$ iterations after a burn-in of $5,000$ iterations.

In an application of IS (Yu and Meng, 2011) to inference in stochastic volatility models, Kastner and Frühwirth-Schnatter (2014) remark that inefficiency factors may depend substantially on the actual realization of the artificial data. In order to mitigate the effect of this variability, they generate several artificial data sets and report the median inefficiency factor for each parameter element. The sampling distribution of the inefficient factor depends on the Markov chain's transition kernel and on the number of iterations. Table 1 gives descriptive statistics of the distribution of the inefficiency factor (27) for a sample of 50,000 observations from a Gaussian first-order auto-regressive process with auto-regressive coefficient $\rho \in \{0, 0.7, 0.9, 0.99, 1\}$. For such processes, the autocorrelation time is equal to $1+2\frac{\rho}{1-\rho}$. These statistics provide some guidance for appreciating the simulations results reported in this section. For example, the inefficiency factor lies outside the interval [0.78, 1.23] with a

statistic \ ρ	0	0.7	0.9	0.99	1
mean	0.99	5.58	18.51	158.10	487.77
standard error	0.01	0.06	0.21	1.38	0.72
median	0.99	5.54	18.43	158.00	489.68
skewness	0.31	0.22	0.22	0.04	−1.64
excess kurtosis	0.32	−0.03	0.10	0.05	3.67
95 % prob. interval	[0.78, 1.23]	[4.38, 6.90]	[14.63, 22.79]	[130.69, 185.58]	[468.68, 496.21]
autocorrelation time	1.00	5.67	19.00	199.00	∞

Table 1: Descriptive statistics of the sampling distribution of the inefficiency factor (27) for a sample of 50,000 observations from a Gaussian first-order auto-regressive process with auto-regressive coefficient $\rho \in \{0, 0.7, 0.9, 0.99, 1\}$. Based on 5,000 artificial samples. Standard error is for 101 samples.

probability of 0.05 if the autocorrelation time is equal to one, and the standard error of the mean over 101 simulations is 0.01.

The model defined by the system (28-29) is globally identified on $\tilde{\Theta}$ if $\alpha_F < 1$. As α_F approaches one, $\tilde{\mathbf{B}}$ becomes unidentifiable and the model becomes translation invariant. Also, as α_R decreases, $\mathbf{Y}$ becomes more correlated in the cross-section and time dimensions. In both cases, the DA algorithm (under $\tilde{\Theta}$) becomes less efficient at estimating $\tilde{\mathbf{B}}$ and the state vector (Pitt and Shephard, 1999). As the convergence properties of a DA algorithm depend on the data-generating parameter values (See Papaspiliopoulos et al., 2007, for a discussion), I compute inefficiency factors on a coarse grid by setting $(\alpha_F, \alpha_R) \in \{0.7, 0.9, 0.95\} \times \{0.01, 0.1, 0.5\}$. Using (28-29), 101 artificial data samples were generated for each point of the grid. I impose the restriction $\mathbf{R} = r\mathcal{I}$[17] and I specify an inverse gamma distribution on r with shape parameter equal to 2 and scale parameter equal to α_R.

[17]Error specification has no material effect on numerical efficiency: imposing that $\mathbf{R}$ is equal to a diagonal matrix produces qualitatively similar results with artificial data, which are not reported in this paper. This alternative restriction is used for describing the numerical efficiency of the SPX-DA with actual data in Section 5.4.

5.3. Simulation results

For $\alpha_F = 0.9$ and $\alpha_R = 0.1$, Table 2 reports the minimum, median, maximum and mean inefficiency factors of each parameter element for the DA algorithm (columns 1 to 4) and for the SPX-DA algorithm (columns 5 to 8). It also reports these statistics for the inefficiency factors of the SPX-DA algorithm as a proportion of the those of the DA algorithm (columns 9 to 12). The DA algorithm is very inefficient at estimating $\tilde{\mathbf{B}}$ and requires several hundred times more iterations than an independent sampler for achieving the same precision. Estimation of the state vector ζ_T is as inefficient as that of $\tilde{\mathbf{B}}$, which is not surprising because they are related by translation invariance. With respect to the estimation of the matrix of factor loading, the picture is somewhat less dramatic on average but the DA algorithm is several hundred times less efficient than an independent sampler for certain elements too. That the elements of the second column of $\tilde{\mathbf{H}}$ are estimated more precisely than those of the first is attributable to the lower persistence of the second state variable. A similar pattern is observed for the estimation of the matrix of auto-regressive coefficients. In contrast, estimation of $\mathbf{R}$ is quite efficient, being only a few times less efficient than an independent sampler. The DA algorithm estimates $\mathbf{R}$ efficiently because this parameter element is drawn from $p(\mathbf{R}|\,\zeta, \mathbf{y})$ when the parameter prior is proportional to (19) and the conditional posterior is factorized as (20). Thus, it is isolated from any inefficiency in the estimation of other parameter elements. Also, this density is invariant under affine transformation of the state vector: $p(\mathbf{R}|\,\zeta, \mathbf{y}) = p(\mathbf{R}|\,\mathbf{G}\zeta + \mathbf{L}, \mathbf{y})$, for all $(\mathbf{L}, \mathbf{G}) \in \mathcal{L} \times \mathcal{G}$.

Inefficiency, as it is being quantified by (27), might therefore be influenced by the invariance property of the likelihood function in a way that generalizes the results of Geweke (2007) for finite mixture distributions, which are permutation invariant. Because the permutation group is countable, inefficient normalization defines a multimodal parameter posterior. Depending on the severity of the empirical underidentification problem, a standard DA algorithm could visit certain lobes of the parameter posterior only infrequently or never at all. LSSMs are invariant under the affine group, which is not countable. Because inefficient normalization does not necessary defines a multimodal posterior, poor mixing over observationally equivalent parameter

values is not always easily visible in LSSMs. In order to better characterize the relationship between numerical efficiency and invariance, Table 2 also reports the inefficiency factors of two additional invariant quantities. The first is the vector of one-step-ahead predictions for the vector of observables, $\hat{\mathbf{y}}_{T+1}$, which I define as[18]

$$\hat{\mathbf{y}}_{T+1} = \tilde{\mathbf{B}} + \tilde{\mathbf{H}}\tilde{\mathbf{F}}\zeta_T. \tag{30}$$

The second is the magnitude of the eigenvalues of the matrix of auto-regressive coefficients, in descending order, which I denote by λ. Considering the inefficient estimation of certain elements of the parameter vector, the DA algorithm is surprisingly efficient at computing predictions. It is only about 50 times less efficient than an independent sampler. A significant component of the low numerical efficiency of the DA algorithm, as quantified by the inefficiency factor of factor loadings for instance, could be attributable to poor mixing over almost observationally equivalent parameter values. This interpretation is confirmed by the inefficiency factors of the eigenvalues of $\tilde{\mathbf{F}}$.

For $\alpha_F = 0.9$ and $\alpha_R = 0.1$, the SPX-DA algorithm is almost as efficient as an independent sampler for estimation every parameter element, with the notable exception of $\mathbf{R}$, for which it provides no material improvement (Table 2, columns 5 to 8). Predictions and eigenvalue are estimated with impressive precision as well. As anticipated, the inefficiency factors of the DA algorithm (under $\tilde{\Theta}$) depend on the parameter values of the data generating process[19]. For parameter elements of the observation equation (6), inefficiency increases with α_F (Table 3, Panels a, b and c). Inefficiency increases with α_R for $\mathbf{R}$, but decreases for $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{H}}$. For the matrix of autoregressive coefficients and its eigenvalues, inefficiency increases with α_R when α_F is large but decreases when α_F is small (Panels d and f). The inefficiency factors of predictions

[18] Computing the inefficiency factor of $\mathbf{y}_{T+1} = \tilde{\mathbf{B}} + \tilde{\mathbf{H}}\left(\tilde{\mathbf{F}}\zeta_T + \tilde{\mathbf{v}}_{T+1}\right) + \mathbf{w}_{T+1}$ would be misleading because the innovations in the observation and state equations reduce the posterior sample's autocorrelation.

[19] Reflection normalization is implemented as an accept-reject step in the DA algorithm. In all, 909 artificial data samples were generated. For each sample, the rejection rate associated with reflection normalization was identically equal to zero. Therefore, reflection normalization had no influence on the numerical efficiency of the DA algorithm.

(Panel e) follow a pattern opposite to that of $\mathbf{R}$: they decrease as α_R or α_F increase. For every quantity and parameter value considered, SPX-DA is at least as efficient as DA, and often much more so (Table 3). Also, inefficiency factors fluctuate much less with the parameter values of the data generating process, if at all. For instance, the inefficiency factors of $\tilde{\mathbf{B}}$ and $\mathbf{R}$ do not seem to vary with α_F (Panels a and c). Those of $\tilde{\mathbf{F}}$ do not vary with α_R (Panel d). Although the DA and SPX-DA algorithms estimate $\mathbf{R}$ equally precisely when α_R are α_F small, the SPX-DA algorithm is more efficient when they are large.

5.4. Term structure data

The empirical results that follow pertain to the estimation of a LSSM with three state variables for a panel of zero-coupon bond yields. I use the data of Joslin, Singleton, and Zhu (2011), which was obtained by bootstrapping[20] Constant Maturity Treasury yields assuming constant forward rates between maturities. Seven maturities (6 months, and 1, 2, 3, 5, 7, and 10 years) have been observed between January 1990 to December 2007, for a total of 216 monthly observations. Yields are measured in basis points. Correlation between interest rates of various maturities is high and factor models are thus attractive for modeling interest rate panels (Table 5). But serial correlations are equally high, which makes inference challenging. Average inefficiency factors are reported in Table 6 when $\mathbf{R}$ is restricted to being a diagonal matrix (Panel b) or proportional to an identity matrix (Panel a). They are consistent with the main conclusions of the simulations experiments of Section 5.3:

1. The DA algorithm estimates invariant quantities ($\mathbf{R}$, $\hat{\mathbf{y}}_{T+1}$ and λ) more efficiently than quantities that are not invariant ($\tilde{\mathbf{B}}$, $\tilde{\mathbf{H}}$, $\tilde{\mathbf{F}}$ and ζ_T);
2. The SPX-DA is almost as efficient as an independent sampler for every parameter element but $\mathbf{R}$, for which it is only a few times less efficient.

[20] Bootstrapping is an iterative method for extracting zero-coupon bond rates from coupon-paying bond yields. It has no relation to the resampling statistical procedure of the same name.

	DA				SPX-DA				DA / SPX-DA			
	min	median	max	mean	min	median	max	mean	min	median	max	mean
Element	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$\tilde{\mathbf{B}}_1$	290.4	443.7	488.5	436.2	0.6	1.0	1.4	1.0	312.6	451.4	700.4	453.8
$\tilde{\mathbf{B}}_2$	106.3	253.1	391.8	256.7	0.6	1.0	1.3	1.0	94.7	251.8	418.6	264.0
$\tilde{\mathbf{B}}_3$	300.9	453.4	489.4	447.7	0.6	1.0	1.4	1.0	325.6	460.2	737.4	468.0
$\tilde{\mathbf{B}}_4$	303.1	453.5	489.4	447.8	0.6	1.0	1.4	1.0	325.0	459.5	743.6	467.4
$\tilde{\mathbf{H}}_{1,1}$	61.4	106.0	218.2	108.1	0.9	1.4	2.1	1.4	34.7	77.7	175.6	80.2
$\tilde{\mathbf{H}}_{2,1}$	122.0	195.3	303.4	201.7	0.8	1.2	2.0	1.3	77.2	154.0	332.6	163.0
$\tilde{\mathbf{H}}_{3,1}$	155.2	241.3	370.6	247.2	0.7	1.1	2.1	1.2	122.6	215.0	383.6	221.1
$\tilde{\mathbf{H}}_{4,1}$	154.9	242.0	370.4	247.2	0.8	1.1	2.1	1.2	122.7	215.5	379.0	221.0
$\tilde{\mathbf{H}}_{2,2}$	26.9	54.9	93.9	56.0	0.8	1.2	1.9	1.2	22.4	46.3	84.3	46.7
$\tilde{\mathbf{H}}_{3,2}$	31.6	58.7	101.4	60.1	0.9	1.2	1.9	1.2	25.6	51.1	86.0	50.8
$\tilde{\mathbf{H}}_{4,2}$	31.5	58.9	102.2	60.1	0.7	1.2	1.6	1.2	25.6	50.3	85.3	51.0
$\mathbf{R}_{1,1}$	2.4	3.6	5.4	3.7	1.9	2.9	4.2	2.9	0.7	1.3	2.4	1.3
$\tilde{\mathbf{F}}_{1,1}$	7.8	40.7	351.7	63.7	0.6	1.1	1.5	1.1	6.3	38.2	368.9	61.6
$\tilde{\mathbf{F}}_{2,1}$	10.2	62.4	206.2	78.9	0.8	1.2	1.7	1.2	9.3	56.2	245.4	69.2
$\tilde{\mathbf{F}}_{1,2}$	2.9	8.2	72.1	11.8	0.7	1.1	1.6	1.1	2.0	8.1	57.6	10.8
$\tilde{\mathbf{F}}_{2,2}$	4.1	11.1	51.6	13.2	0.7	1.1	1.7	1.2	3.2	9.5	52.2	11.9
$\zeta_{T,1}$	160.3	375.7	482.1	369.2	0.6	1.0	1.4	1.0	183.3	385.6	595.3	384.3
$\zeta_{T,2}$	44.7	145.4	364.1	157.3	0.6	1.0	1.4	1.0	38.9	146.0	332.1	157.1
$\hat{\mathbf{y}}_{T+1,1}$	4.7	31.7	148.4	36.2	0.6	1.0	1.4	1.0	5.9	32.7	122.0	36.9
$\hat{\mathbf{y}}_{T+1,2}$	11.8	32.0	123.2	37.0	0.7	1.1	1.5	1.0	9.2	32.3	110.3	36.5
$\hat{\mathbf{y}}_{T+1,3}$	12.1	61.0	158.3	65.2	0.7	1.0	1.4	1.0	10.2	63.0	170.3	66.1
$\hat{\mathbf{y}}_{T+1,4}$	11.6	60.9	157.7	65.3	0.7	1.0	1.4	1.0	10.0	61.4	164.2	66.5
λ_1	6.9	40.9	389.1	62.2	0.6	1.0	1.5	1.0	6.4	41.1	419.5	62.0
λ_2	2.7	7.8	29.8	9.3	0.8	1.1	1.7	1.1	2.2	7.1	25.7	8.5

Table 2: Minimum, median, maximum and mean of the inefficiency factors of the DA algorithm (columns 1 to 4), of the inefficient factors of the SPX-DA algorithm (columns 5 to 8), and of the inefficiency factors of the DA algorithm as a proportion of the inefficient factors of the SPX-DA algorithm (columns 9 to 12) for 101 artificial data samples of $T = 200$ observations generated from (28-29) with $\alpha_F = 0.9$ and $\alpha_R = 0.1$. $\mathbf{R}$ is proportional to an identity matrix. Inefficiency factors are computed by (27).

α_R \ α_F	0.70	0.90	0.99
0.01	440.2	462.2	467.2
0.10	290.0	400.9	430.4
0.50	118.3	322.2	390.2

Panel a - $\tilde{\mathbf{B}}$

α_R \ α_F	0.70	0.90	0.99
0.01	305.3	355.2	393.9
0.10	79.8	136.7	231.3
0.50	27.5	59.1	150.0

Panel b - $\tilde{\mathbf{H}}$

α_R \ α_F	0.70	0.90	0.99
0.01	2.8	2.9	3.1
0.10	3.1	3.6	4.7
0.50	4.1	4.8	8.6

Panel c - $\mathbf{R}$

α_R \ α_F	0.70	0.90	0.99
0.01	16.7	30.3	46.5
0.10	9.8	30.6	70.0
0.50	7.3	28.8	76.8

Panel d - $\tilde{\mathbf{F}}$

α_R \ α_F	0.70	0.90	0.99
0.01	139.8	69.2	101.1
0.10	56.8	46.4	33.8
0.50	10.6	14.6	10.8

Panel e - $\hat{\mathbf{y}}_{T+1}$

α_R \ α_F	0.70	0.90	0.99
0.01	10.3	10.8	9.5
0.10	8.3	24.3	14.0
0.50	5.8	29.6	29.7

Panel f - λ

Table 3: Median inefficiency factors of the DA algorithm for 101 artificial data sets of 200 observations generated from (28-29) with $(\alpha_F, \alpha_R) \in \{0.7, 0.9, 0.99\} \times \{0.01, 0.1, 0.5\}$. $\mathbf{R}$ is proportional to an identity matrix. Inefficiency factors are computed by (27). The average over the elements of each parameter is reported.

α_R \ α_F	0.70	0.90	0.99
0.01	1.0	1.0	1.1
0.10	1.0	1.0	1.1
0.50	1.0	1.0	1.1

Panel a - $\tilde{\mathbf{B}}$

α_R \ α_F	0.70	0.90	0.99
0.01	1.0	1.0	1.1
0.10	1.1	1.2	1.4
0.50	1.9	2.1	2.5

Panel b - $\tilde{\mathbf{H}}$

α_R \ α_F	0.70	0.90	0.99
0.01	2.9	2.9	3.3
0.10	2.9	2.9	3.2
0.50	3.1	3.0	3.3

Panel c - $\mathbf{R}$

α_R \ α_F	0.70	0.90	0.99
0.01	1.0	1.0	1.1
0.10	1.1	1.1	1.3
0.50	2.1	1.8	2.0

Panel d - $\tilde{\mathbf{F}}$

α_R \ α_F	0.70	0.90	0.99
0.01	1.0	1.0	1.1
0.10	1.0	1.0	1.1
0.50	1.1	1.1	1.1

Panel e - $\hat{\mathbf{y}}_{T+1}$

α_R \ α_F	0.70	0.90	0.99
0.01	1.0	1.0	1.1
0.10	1.1	1.1	1.2
0.50	1.9	1.6	1.6

Panel f - λ

Table 4: Median inefficiency factors of the SPX-DA algorithm for 101 artificial data sets of 200 observations generated from (28-29) with $(\alpha_F, \alpha_R) \in \{0.7, 0.9, 0.99\} \times \{0.01, 0.1, 0.5\}$. $\mathbf{R}$ is proportional to an identity matrix. Inefficiency factors are computed by (27). The average over the elements of each parameter is reported.

	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$	$\mathbf{y}_6$	$\mathbf{y}_7$
$\mathbf{y}_1$	1.000	-	-	-	-	-	-
$\mathbf{y}_2$	0.994	1.000	-	-	-	-	-
$\mathbf{y}_3$	0.963	0.983	1.000	-	-	-	-
$\mathbf{y}_4$	0.928	0.957	0.993	1.000	-	-	-
$\mathbf{y}_5$	0.840	0.879	0.946	0.977	1.000	-	-
$\mathbf{y}_6$	0.774	0.818	0.901	0.944	0.992	1.000	-
$\mathbf{y}_7$	0.687	0.736	0.833	0.889	0.965	0.988	1.000
Serial	0.982	0.981	0.978	0.976	0.974	0.975	0.975

Table 5: First-order serial correlation and cross-correlations of zero-coupon bond rates.

	DA	SPX-DA
Element	(1)	(2)
$\tilde{\mathbf{B}}$	481.7	1.1
$\tilde{\mathbf{H}}$	344.8	1.9
$\mathbf{R}$	8.6	2.7
$\tilde{\mathbf{F}}$	99.3	1.4
ζ_T	437.2	1.0
$\hat{\mathbf{y}}_{T+1}$	5.5	1.1
λ	12.9	1.0

Panel a

	DA	SPX-DA
Element	(1)	(2)
$\tilde{\mathbf{B}}$	483.8	1.3
$\tilde{\mathbf{H}}$	399.5	2.0
$\mathbf{R}$	8.4	5.8
$\tilde{\mathbf{F}}$	113.8	1.4
ζ_T	447.4	1.3
$\hat{\mathbf{y}}_{T+1}$	11.7	1.3
λ	6.2	1.2

Panel b

Table 6: Term structure data. Mean inefficiency factors of the DA (column 1) and the SPX-DA (column 2) algorithms. $\mathbf{R}$ is proportional to an identity matrix (Panel a) or is a diagonal matrix (Panel b). Inefficiency factors are computed by (27).

6. Almost invariant posteriors

I use artificial data for illustrating the influence of normalization on parameter inference when the model is empirically underidentified. In particular, I show how $\tilde{\Theta}$ can define an almost invariant parameter posterior. Such a posterior has an irregular shape that the DA algorithm does not always fully recover even after a fairly large number of iterations. I analyze the empirical consequences of invariance under orthogonal transformation and translation in two separate simulation experiments. The DA and SPX-DA algorithms are implemented as in Section 5. They define different parameter posteriors because they integrate different priors. The influence of prior specification is also examined in the present section.

6.1. Invariance under orthogonal transformation

In order to isolate the influence of invariance under orthogonal transformation on statistical inference, the data generating process is chosen so that there is no empirical underidentification difficulty associated with translation or scale invariance. As translation invariance can cause empirical underidentification problems when that data is strongly persistent, the diagonal elements of the matrix of autoregressive coefficients are set to relatively low values. Difficulties associated to scale invariance are avoided by assuming that the number of state variables is known. For this simulation experiment, the data is generated as

$$\zeta_{t+1} = \mathbf{0} + \begin{bmatrix} 0.4 & 0 \\ 0 & 0.8 \end{bmatrix} \zeta_t + \tilde{\mathbf{v}}_t \tag{31}$$

$$\mathbf{Y}_t = \mathbf{0} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \zeta_t + \mathbf{w}_t, \tag{32}$$

where $\tilde{\mathbf{v}}_t$ and $\mathbf{w}_t$ are vectors of independent standard normal variables. Imposing (11-12) does not break invariance under permutation and rotation: post-multiplying $\tilde{\mathbf{H}}_{1:2,1:2} = \left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right]$, by a permutation or a rotation matrix matrix results in a lower triangular matrix. Permutation invariance interferes with reflection normalization under $\tilde{\Theta}$ because imposing (13) is ineffective if

Parameter	Mode 1	2	3	4	5	6
$\tilde{\mathbf{H}}$	$\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ -1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & -1 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix}$
$\tilde{\mathbf{F}}$	$\begin{bmatrix} 0.5 & 0 \\ 0 & 0.9 \end{bmatrix}$	$\begin{bmatrix} 0.5 & 0 \\ 0 & 0.9 \end{bmatrix}$	$\begin{bmatrix} 0.5 & 0 \\ 0 & 0.9 \end{bmatrix}$	$\begin{bmatrix} 0.5 & 0 \\ 0 & 0.9 \end{bmatrix}$	$\begin{bmatrix} 0.9 & 0 \\ 0 & 0.5 \end{bmatrix}$	$\begin{bmatrix} 0.9 & 0 \\ 0 & 0.5 \end{bmatrix}$

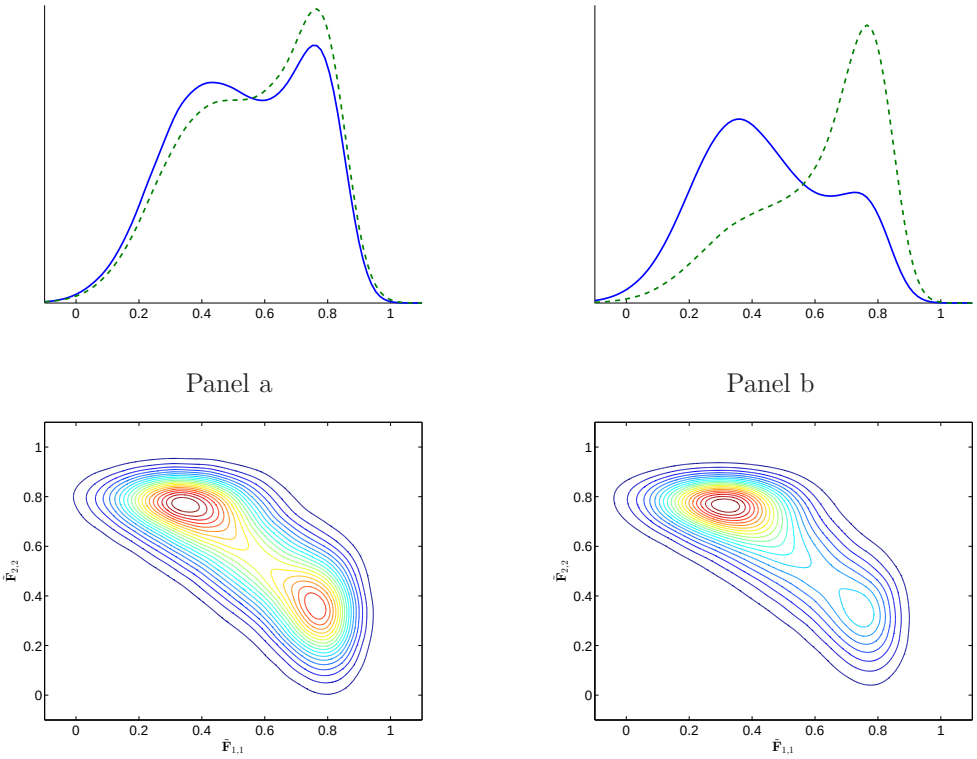
Table 7: The six pairs of parameter values (which I refer to as modes of the parameter posterior) that are observationally equivalent to the system (31-31).

the diagonal elements of $\tilde{\mathbf{H}}_{1:2,1:2}$ are equal to zero. Thus, permutation and reflection invariance implies that there are six observationally equivalent parameter values. For future reference, they are described and labeled in Table 7. In finite samples, this observational equivalence does not hold exactly because the first row of $\tilde{\mathbf{H}}$ will not be identically zero, but the parameter posterior could be multimodal and almost invariant under rotation. For clarity of exposition however, I will refer to the modes of the parameter posterior through the intermediary of the parameter values reported in Table 7.

By construction, the SPX-DA algorithms explores every lobe of the parameter posterior[21]. The DA algorithm does not always do so. Figure 1 (Panels a and b) shows the posterior density[22] of the diagonal elements of $\tilde{\mathbf{F}}$ for one artificial data set of 200 observations, based on $500,000$ iterations after a burn-in of $5,000$ iterations. The marginal posterior densities recovered by the DA (Panel a) and the SPX-DA (Panel b) algorithms have surprisingly different shapes. The contour plots of the joint densities (Panels c and d), however, show that they

[21] Matrix decomposition algorithms must implemented in order to randomly select one element of their solution set. For instance, there are $K!$ solutions to the QR decomposition of an invertible $K \times K$ matrix.

[22] Joint densities are computed using the kernel density estimator proposed by Botev, Grotowski, and Kroese (2010). For consistency, marginal densities are computed by numerical integration of joint densities.



Panel a Panel b

Panel c Panel d

Figure 1: Posterior marginal (Panels a and b) and joint (Panels c and d) densities of $\tilde{\mathbf{F}}_{1,1}$ (solid blue line) and $\tilde{\mathbf{F}}_{2,2}$ (dashed green line). Panels a and c correspond to the posterior sample from the DA Algorithm 1. Panels b and d correspond to the sample from the SPX-DA Algorithm 3.

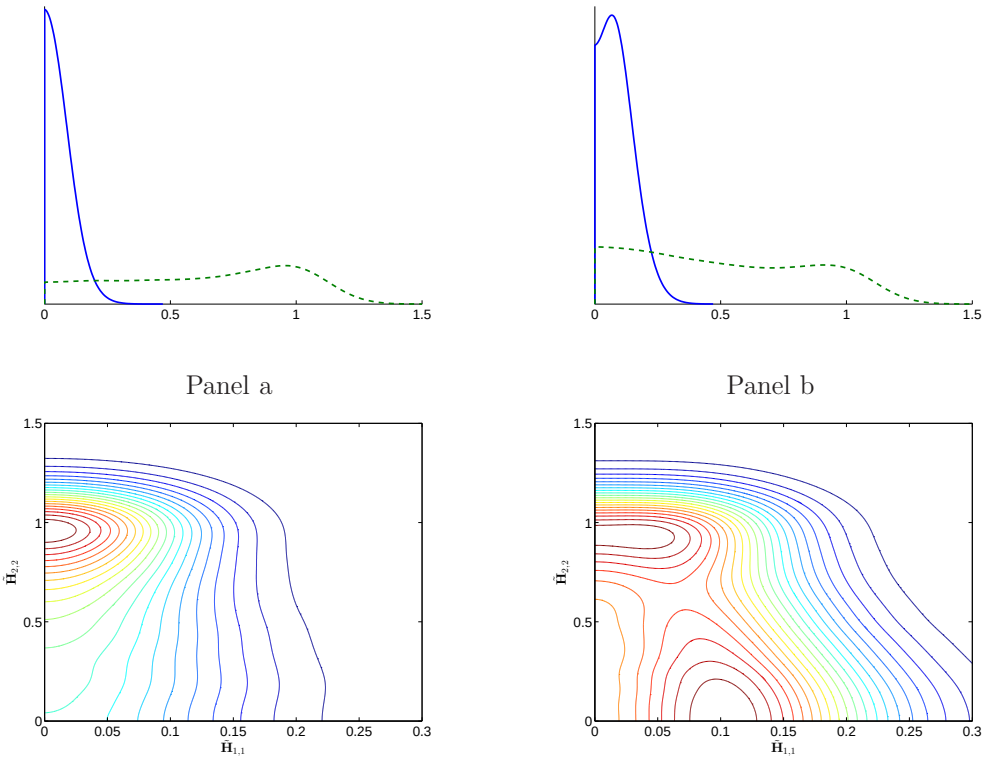


Figure 2: Posterior marginal (Panels a and b) and joint (Panels c and d) densities of $\mathbf{H}_{1,1}$ (solid blue line) and $\mathbf{H}_{2,2}$ (dashed green line) under $\tilde{\Theta}$. Panels a and c correspond to the posterior sample from the DA Algorithm 1. Panels b and d correspond to the sample from the SPX-DA Algorithm 3.

are characterized by the same modes but that the algorithms do not visit them with the same frequency. In particular, the DA algorithm visits modes 5 or 6 (or both) more often than the SPX-DA algorithm does.

The posterior distribution of other parameter elements is irregular as well. Figure 2 shows the posterior distribution of $\tilde{\mathbf{H}}_{1,1}$ and $\tilde{\mathbf{H}}_{2,2}$, the elements upon which reflection normalization rests under $\tilde{\Theta}$. Because the marginal densities are not negligible at zero, imposing $\tilde{\mathbf{H}}_{1,1} > 0$ and $\tilde{\mathbf{H}}_{2,2} > 0$ does not break reflection invariance efficiently. The contour plots (Panels c and d) confirm that the DA algorithm spends more time around mode 5 or 6 than the SPX-DA algorithm does. In order to distinguish between modes 5 and 6, Figure 3 shows the posterior of $\tilde{\mathbf{H}}_{3,1}$ and $\tilde{\mathbf{H}}_{2,2}$ and reveals that the DA algorithm never visits mode 5. Figure 4 (Panels c and

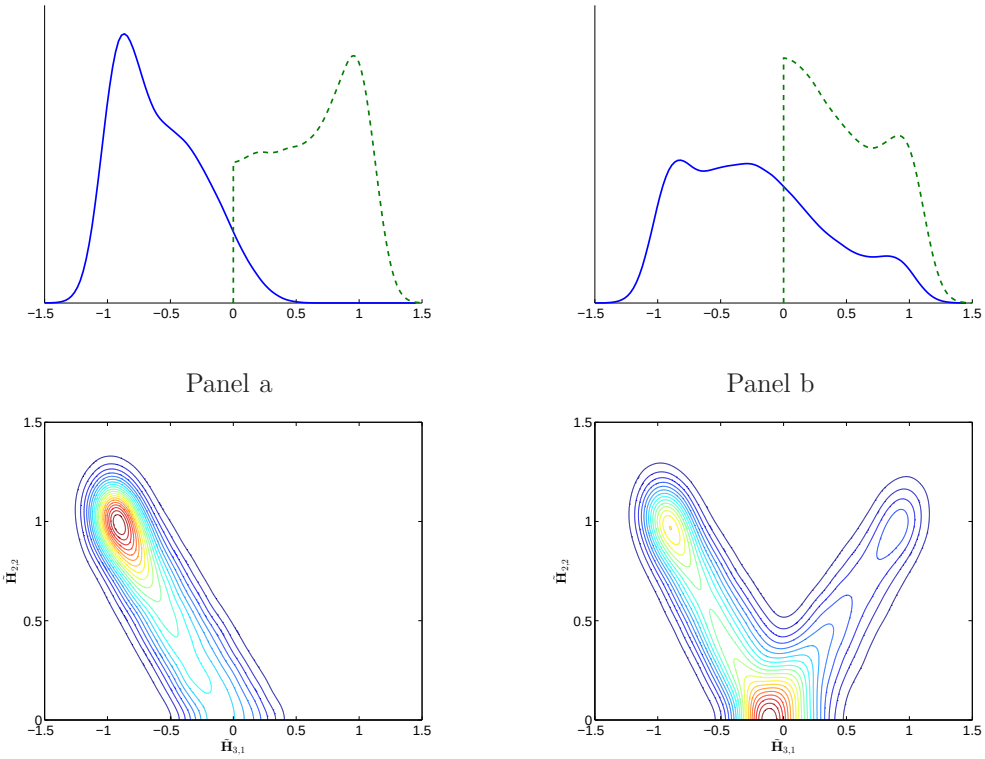


Figure 3: Posterior marginal (Panels a and b) and joint (Panels c and d) densities of $\tilde{\mathbf{H}}_{3,1}$ (solid blue line) and $\tilde{\mathbf{H}}_{2,2}$ (dashed green line). Panels a and c correspond to the posterior sample from the DA Algorithm 1. Panels b end d correspond to the sample from the SPX-DA Algorithm 3.

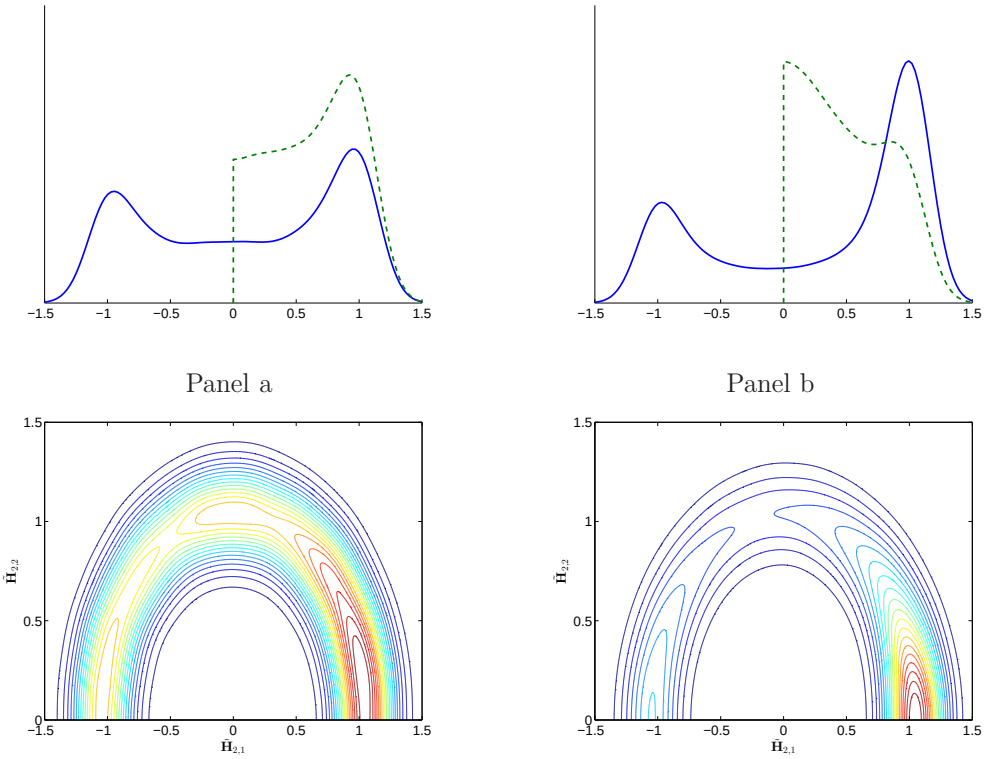


Figure 4: Posterior marginal (Panels a and b) and joint (Panels c and d) densities of $\tilde{\mathbf{H}}_{2,1}$ (solid blue line) and $\tilde{\mathbf{H}}_{2,2}$ (dashed green line). Panels a and c correspond to the posterior sample from the DA Algorithm 1. Panels b and d correspond to the sample from the SPX-DA Algorithm 3.

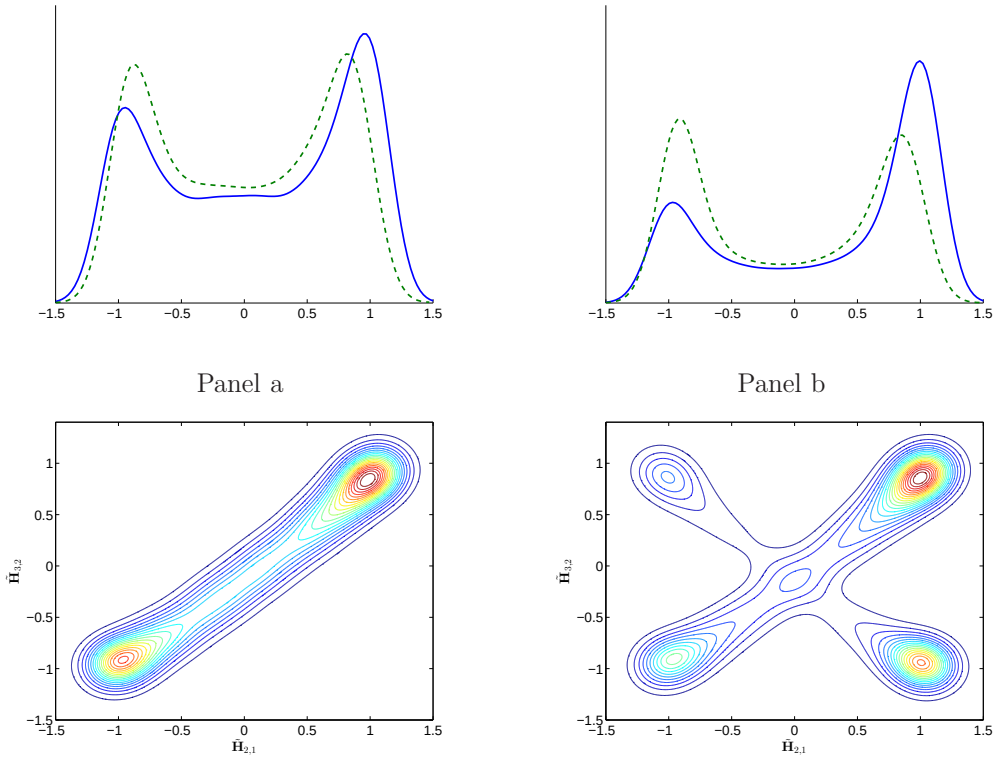


Figure 5: Posterior marginal (Panels a and b) and joint (Panels c and d) densities of $\tilde{\mathbf{H}}_{2,1}$ (solid blue line) and $\tilde{\mathbf{H}}_{3,2}$ (dashed green line). Panels a and c correspond to the posterior sample from the DA Algorithm 1. Panels b and d correspond to the sample from the SPX-DA Algorithm 3.

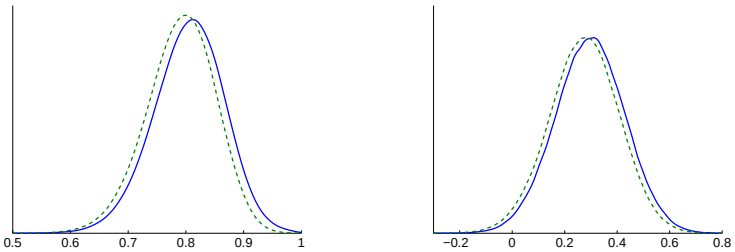
d) shows that the joint posterior of $\left(\tilde{\mathbf{H}}_{2,1}, \tilde{\mathbf{H}}_{2,2}\right)$ is almost rotation invariant. The posterior has a circular shape due to inefficient breaking of rotation invariance. Both algorithms spend more time around modes 1 or 3 than around other modes. Finally, the DA algorithm never visits modes 2 and 3 (Figure 5, Panels c and d). In sum, the DA algorithm switches between mode 1 and mode 4, visiting mode 6 in the transition, but never visits modes 2, 3 and 5.

For this artificial data set, $\tilde{\Theta}$ defines a parameter posterior with an irregular shape, which complicates communication of empirical results about certain parameters. Reporting parameter point estimators would be inadequate for must purposes for instance. Breaking invariance under permutation and reflection would be more effective if the elements of the factor loadings matrix were uncorrelated *a posteriori*. When the parameter prior is proportional to (17), $\text{Cov}\left[\mathbf{H}|\,\mathbf{R}, \xi, \mathbf{y}\right]$ is a diagonal matrix if both $\mathbf{R}$ and Σ are. I thus break rotation invariance by imposing (14-15). Under this normalization, the lobes of the joint posterior of $(\mathbf{F}_{1,1}, \mathbf{F}_{2,2})$ are well separated and imposing $\mathbf{F}_{1,1} > \mathbf{F}_{2,2}$ efficiently breaks permutation invariance. In addition, the marginal posterior density of $\mathbf{H}_{3,1}$ and $\mathbf{H}_{2,2}$ attribute negligible probability to neighborhoods of zero and imposing $\mathbf{H}_{3,1} > 0$ and $\mathbf{H}_{2,2} > 0$ efficiently breaks reflection invariance. For future reference, I denote this normalization by

$$\Theta^{O} = \left\{\theta \in \Theta |\, \Sigma = \mathcal{I}; \mathbf{Q} \in \mathcal{D}; \mathbf{F}_{1,1} > \mathbf{F}_{2,2}; \mathbf{H}_{3,1} > 0; \mathbf{H}_{2,2} > 0\right\}. \tag{33}$$

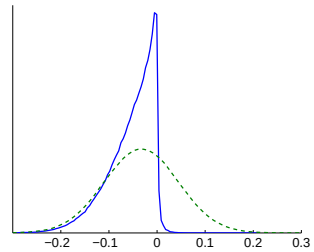
I operationalize normalization by postprocessing the posterior samples. For the diagonal elements of $\mathbf{F}$, the DA and SPX-DA algorithms define very similar parameter posteriors under (33). The marginal posterior densities (Figure 6, Panels a and b) are unimodal and have a regular shape. If the investigator wanted to communicate empirical results about the persistence of the state vector, this normalization would serve him well.

As implemented in this paper, the DA and SPX-DA algorithm define different parameter posteriors because they integrate different priors. For the diagonal elements of $\mathbf{F}$, the difference is relatively small (Figure 6, Panels a and b) but it is consistent with the observation made in Section 4.2: the modes of the marginal posteriors defined that the DA algorithm are larger

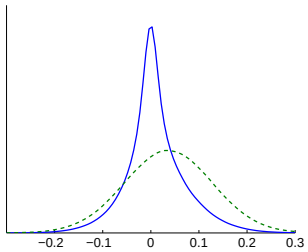


Panel a - $\mathbf{F}_{1,1}$

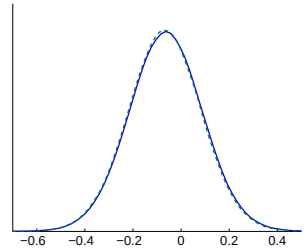
Panel b - $\mathbf{F}_{2,2}$



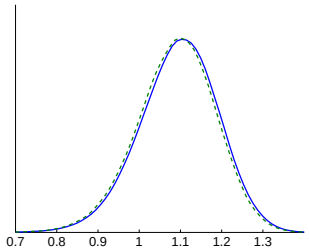
Panel c - $\mathbf{H}_{1,1}$



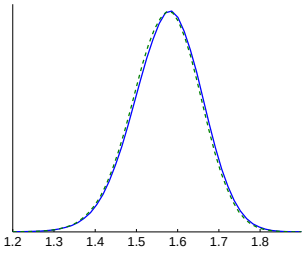
Panel d - $\mathbf{H}_{1,2}$



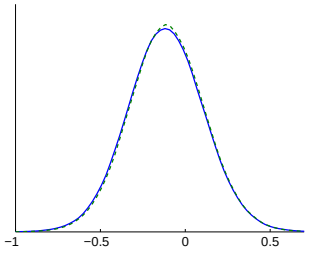
Panel e - $\mathbf{H}_{2,1}$



Panel f - $\mathbf{H}_{2,2}$



Panel g - $\mathbf{H}_{3,1}$



Panel h - $\mathbf{H}_{3,2}$

Figure 6: Posterior marginal densities of the diagonal elements of $\mathbf{F}$ and of the elements of $\mathbf{H}$ recovered by the DA (solid blue line) and SPX-DA (dashed green line) algorithms under normalization (33).

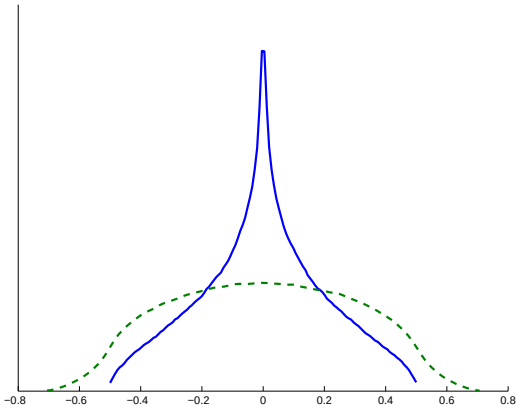
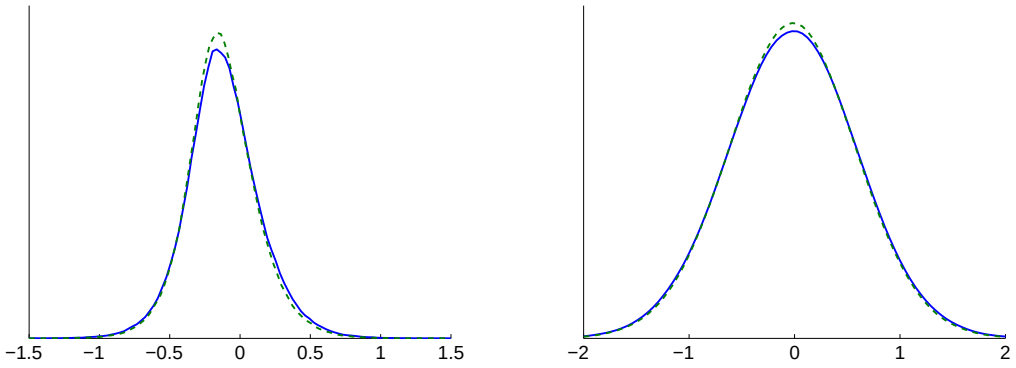


Figure 7: Marginal densities of the top left element of $\mathbf{HU}$ when $\mathbf{U}$ is distributed according to the Haar measure on the orthogonal group and $\mathbf{H}_{1,1}$ is uniformly distributed on $\left[-\frac{1}{2}, \frac{1}{2}\right]$ ($\mathbf{H}_{1,2} = 0$) (solid blue line) and $(\mathbf{H}_{1,1}, \mathbf{H}_{1,2})$ is uniformly distributed on $\left[-\frac{1}{2}, \frac{1}{2}\right]^2$ (dashed green line).

than those defined by the SPX-DA algorithm[23]. The influence of prior specification on the elements on the first row (first $K-1$ rows, when $K > 2$) of $\mathbf{H}$ is more apparent (Panels c and d). This is because specifying a flat prior on the elements of a lower triangular matrix is not equivalent to specifying a flat prior on the elements of a square matrix (equivalent priors must satisfy equation (21)). In order to better understand the influence of prior specification, suppose that a uniform prior on the interval $\left[-\frac{1}{2}, \frac{1}{2}\right]$ is specified instead of a flat prior and consider the distribution of the top left element of $\mathbf{HU}$ when $\mathbf{U}$ is distributed according to the Haar measure on the orthogonal group. If the first $K \times K$ block of $\mathbf{H}$ is lower triangular, the density of the top left element of $\mathbf{HU}$ has a sharp spike at zero (Figure 7, solid blue line). In contrast, the density is not so different from that of a uniform random variable if the first $K \times K$ block of $\mathbf{H}$ is a square matrix (Figure 7, dashed green line). Notice that the difference between the marginal posterior of the elements of the second and third rows is not material (Figure 6, Panels e to h), which confirms that the observed differences in other marginal posteriors are indeed due to prior specification.

[23]The influence of prior specification on the posterior of the diagonal elements of $\mathbf{F}$ is explored further in Section 6.2.



Panel a - $\hat{\mathbf{y}}_{2,T+1}$ Panel b - $\hat{\mathbf{y}}_{3,T+1}$

Figure 8: Posterior marginal densities of predictions recovered by the DA (solid blue line) and SPX-DA (dashed green line) algorithms under $\tilde{\Theta}$. Predictions are defined by (30).

That the DA and SPX-DA algorithms recover so similar parameter posteriors under (33) is somewhat puzzling for it suggests that the DA algorithm fully captures the informational content of the parameter posterior in spite of its failure to explore every lobe. Geweke (2007) argues that efficiently exploring every lobes of a permutation invariant posterior distribution is not necessary for the posterior sampler to fully capture its informational content. He states that "Simple MCMC works" unless "there are mixing problems beyond those arising from permutation invariance." Because each lobe contains the same information, exploring any single lobe is sufficient. Heuristically, not visiting observationally equivalent parameter values at the right frequency is inconsequential. The posterior densities of predictions under $\tilde{\Theta}$ recovered by the DA (Figure 8, solid blue line) and SPX-DA (dashed green line) algorithms are very similar. This suggests that not visiting almost observationally equivalent parameter values at the right frequency is almost inconsequential.

6.2. Invariance under translation

In order to better understand the influence of translation invariance on statistical inference, the data generating process is chosen so that there is no underidentification difficulty associated with invariance under scaling and orthogonal transformations. The data is generated according

to (28-29) with $K = N = 1$, $T = 200$, $\alpha_F = 0.99$ and $\alpha_R = 1$. There is no permutation or rotation invariance in a one-factor model. For the artificial data that I analyze in this section[24], the posterior distribution of the factor loading is unimodal and attributes negligible probability to neighborhoods of zero so there is no empirical underidentification difficulty associated with reflection and scale invariance. Breaking translation invariance by imposing (10) performs poorly on the parameter subspace on which $\mathcal{I} - \tilde{\mathbf{F}}$ is close to being singular. In this simulation exercise, $\tilde{\mathbf{F}} = 0.99$ is a sufficiently high value for illustrating the implications of translation empirical underidentification.

Because the translation group is not compact, a translation invariant density function is improper. Inefficiently breaking translation invariance could therefore result in an almost improper parameter posterior. If the mathematical definition of impropriety is unambiguous, anticipating the observable characteristics of a sample from an almost improper distribution proves challenging. It requires defining some sort of sequence of which an improper density is the limit, in a way that gives meaning to *being close to impropriety*. One such sequence is the order of the lowest central moment that is not finite. For an improper density, moments of order higher than or equal to zero are not finite. Arguably, an almost improper density should not have a finite variance, and perhaps not a finite mean either.

Figure 9 shows the value of $\tilde{\mathbf{B}}$ for every 1000th iteration of the DA algorithm (Panel a) and a kernel estimation of the posterior density (Panel c, solid blue line). Even after 1,000,000 iterations, the DA algorithm failed to recover the marginal posterior of $\tilde{\mathbf{B}}$ with a precision that would allow one to draw any meaningful conclusion about its shape. In contrast, the SPX-DA algorithm shows no particular sign of poor mixing. Panel b of Figure 9 shows the value of $\tilde{\mathbf{B}}$ for every 1000th iteration of the SPX-DA algorithm, *truncated to the* $[-50, 50]$ *interval*, thus excluding 3,2% of the sample. The DA algorithm never explores outside this interval. The posterior density recovered by the SPX-DA algorithm has very fat tails, which are difficult to estimate. Iterations for which the value of $\tilde{\mathbf{B}}$ lies outside the $[-50, 50]$ interval are excluded

[24]Qualitatively similar results were obtained with other artificial data sets.

as well in the kernel estimate shown in Panel c (dashed green line). The sample posterior standard deviation and excess kurtosis are respectively 1,295 and 299,639. The largest value is 932,334. Assuming that the sample is t-distributed, the maximum-likelihood estimate of the number of degrees of freedom is $\nu = 1.23$. A t-distributed random variable with $1 < \nu \leq 2$ as infinite variance and undefined higher-order moments. Its mean is undefined if $\nu \leq 1$. An alternative way of assessing whether the variance is finite is analyzing the distribution of the sample mean. I divided the posterior sample randomly into 1,000 sub-samples of 1,000 observations. The distribution of the sample mean is undoubtedly not normal: assuming the sample mean is t-distributed, the maximum-likelihood estimate of the number of degrees of freedom is 1.03. These observations suggest that the parameter posterior recovered by the SPX-DA algorithm might not have a finite mean, which is arguably as close to being improper as a proper distribution can be.

As in the previous simulation exercises, the DA and SPX-DA algorithms define different parameter posteriors. But the influence of the prior specification is more transparent in the present one. Specifying a flat prior proportional to (19) attributes a higher probability to neighbourhoods of $\tilde{\mathbf{F}} = \mathcal{I}$ than specifying an invariant prior proportional to (17). When the observables are highly persistent and the sample size is relatively small, the DA and SPX-DA algorithms can define quite different parameter posteriors. Panel d of Figure 9 shows a kernel estimation of the posterior of $\tilde{\mathbf{F}}$ recovered by the DA (solid blue line) and SPX-DA algorithm (dashed green line). In contrast to the posterior of $\tilde{\mathbf{B}}$, the DA algorithm has no difficulty recovering the posterior of $\tilde{\mathbf{F}}$. But it defines a parameter posterior with a higher mode (0.998) than that defined by the SPX-DA algorithm (0.988), which is equal to the maximum-likelihood estimate. For this reason, the invariant prior integrated in the SPX-DA algorithm is arguably more compatible with the common notion of a *noninformative* prior than the flat prior integrated in the DA algorithm.

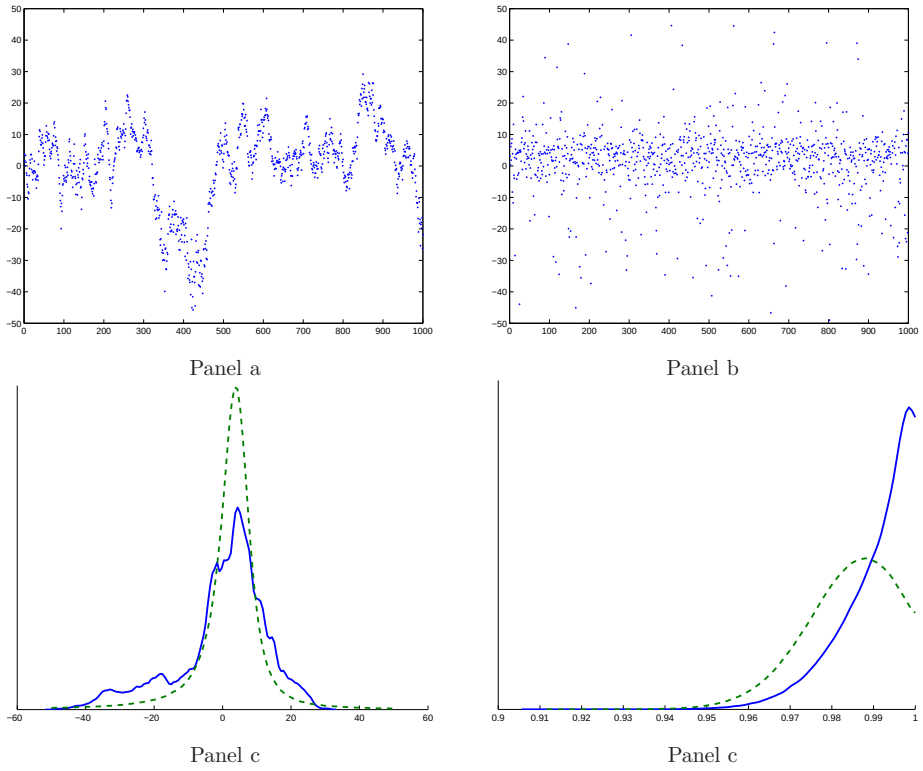


Figure 9: Translation empirical underidentification. Panels a and b respectively show the value of $\tilde{\mathbf{B}}$ for every 1000th iteration of the DA and SPX-DA algorithm. The artificial data is generated as described in section 6.2. Also shown is a kernel estimation of the marginal posterior of $\tilde{\mathbf{B}}$ (Panel c) and $\tilde{\mathbf{F}}$ (Panel d) recovered by the DA algorithm (solid blue line) and the SPX-DA algorithm (dashed green line).

7. Discussion

Empirical work in econometrics often begins with specification of a structural model involving a parameter that has more elements than can be estimated because they are not identified. Identification is usually obtained by restricting the support of the parameter prior density to a particular subspace in which the parameter is identified. When a DA algorithm is used for computing the parameter posterior, operationalizing normalization in this manner is computationally inefficient. The main contribution of this paper is a novel posterior sampler for state-space models. It exploits the invariance property of the model's likelihood function for simplifying implementation and improving numerical efficiency of posterior sampling. In particular, the SPX-DA algorithm outperforms a standard DA algorithm under any parameterization that can be expressed as a restriction of the unnormalized parameter space induced by the invariance group. From a practitioner's perspective, SPX-DA provides substantial computational efficiency gains with only minor modifications to a standard DA algorithm. In addition, the SPX-DA's simplicity and efficiency do not depend on normalization choice. In fact, operationalizing normalization as a mapping allows the investigator to consider parameterizations that would be impracticable with a standard DA algorithm. Such parameterizations could prove useful in empirical work when standard normalizations define parameter posteriors with undesirable properties.

One apparent drawback of SPX-DA is that the parameter posterior it defines is generally different from that defined by a standard DA algorithm, unless the Jacobian of the transformation that operationalizes normalization is identically equal to one. While both approaches are inferentially valid, they define different posteriors because the integrate, implicitly, different priors. With respect to the parameter posteriors that they define, comparing the SPX-DA algorithm to a standard DA algorithm therefore amounts to comparing the prior that they typically integrate. Another contribution of this paper is proposing parameter priors that do not express prior beliefs over the relative plausibility of observationally equivalent parameter values. With such priors the invariance property of the likelihood function carries over

to the parameter posterior and predictive densities. For a one-factor LSSM, the invariant prior (23) ensures that the mode of the autoregression coefficient's posterior is equal to its maximum-likelihood estimate. But the empirical performance of invariant priors for parameter estimation, forecasting and model selection merits further examination.

I have provided only a few examples of invariant prior densities, which are noninformative about dimensions that have no substantive interpretation but, necessarily, informative about other dimensions. In certain situations, an invariant prior that is more informative about invariant dimensions might be desirable. Invariant quantities are arguably the only quantities that one can reasonably have prior beliefs about. For instance, the investigator might have prior beliefs about the magnitude of the matrix of autoregression coefficients' largest eigenvalue, say λ_{max}. Clearly, a prior proportional to $p(\mathbf{R})\, p(\lambda_{max}) \det(\mathbf{Q})^{-\frac{1-N+(K+1)}{2}}$ satisfies (16) for any marginal density $p(\lambda_{max})$. Another invariant quantity is the proportion of the observables' variance that is attributable to the state vector, *i.e.* the diagonal elements of $\mathbf{H}\Sigma\mathbf{H}'$ as a proportion of those of $\mathbf{H}\Sigma\mathbf{H}' + \mathbf{R}$. Thus, invariant priors should not be assimilated to noninformative priors and can be used when prior information is available.

Autoregressive-moving-average (ARMA) models have LSSM representations[25] that are not invariant under the affine group because certain parameter elements are restricted to being equal to zero or one. However, these models are not immune to an empirical underidentification difficulty that is known as *root cancelation* or *redundant parameter*. It is well known (Box and Jenkins, 1976) that parameter point estimators become unreliable when an autoregressive root is close to a moving-average root and this difficulty is the object ongoing research. Kleibergen and Hoek (2000) propose priors for a reparameterization of ARMA models that penalizes regions of the parameter space where roots are close to canceling out. Alternatively, Cogley and Startz (2013) propose a framework for attributing a specified probability to the parameter subspace where roots cancel out. Describing the invariance property of ARMA

[25] See Aoki (1987), Brockwell and Davis (1991) and Hamilton (1994) for LSSM representations of ARMA models.

models and its relationship to its LSSM representations would possibly help finding other solutions to the near root cancellation problem. The tools that I present in this paper could also help addressing empirical underidentification difficulties in the mixtures models, structural vector autoregressions, and cointegration models considered by Hamilton, Waggoner, and Zha (2007).

Measuring the numerical efficiency of a posterior sampler is more complicated than is commonly appreciated. An essential first step in this assessment is being precise about the inference objective. Efficiency cannot be defined in the abstract. A standard DA algorithm is surprisingly efficient for computing predictions in LSSMs. This contrasts with extremely inefficient computation of certain parameter posteriors. Although the high autocorrelation of the parameter posterior sample generated by a standard DA algorithm suggests poor mixing properties, simulation results indicate that it reliably recovers the informational content of the parameter posterior after a reasonable number of iterations. Heuristically, slow exploration of almost observationally equivalent parameter values is almost inconsequential. One implication of this observation is that efficiency gains brought by a particular algorithm for computing the parameter posterior do not necessarily translate into more efficient computation of predictive densities. From that perspective, relating the numerical efficiency of a MCMC sampler to the inefficiency factors of its parameter could be misleading when parameter estimation is not the inference objective.

References

Andrews, D. W. K., Cheng, X., 2012. Estimation and inference with weak, semi-strong, and strong identification. Econometrica 80 (5), 2153–2211.
URL http://dx.doi.org/10.3982/ECTA9456

Aoki, M., 1987. State space modeling of time series. Springer-Verlag, New York.

Bayarri, M. J., Berger, J., Forte, A., García-Donato, G., 2012. Criteria for bayesian model choice with application to variable selection. The Annals of Statistics 40 (3), 1550–1577.
URL http://dx.doi.org/10.1214/12-AOS1013

Berger, J., 1985. Statistical Decision Theory and Bayesian Analysis, 2nd Edition. Springer.

Botev, Z., Grotowski, J., Kroese, D., 2010. Kernel density estimation via diffusion. The Annals of Statistics 38 (5), 2916–2957.
URL http://dx.doi.org/10.1214/10-AOS799

Box, G., Jenkins, G., 1976. Time series analysis: Forecasting and applications. Holden-Day, San Fransisco.

Brockwell, P. J., Davis, R. A., 1991. Time Series: Theory and Methods, 2nd Edition. Springer.

Buse, A., 1992. The bias of instrumental variables estimators. Econometrica 60, 173–180.
URL http://dx.doi.org/10.2307/2951682

Carter, C., Kohn, P., 1994. On the Gibbs sampling for state space models. Biometrika 81, 541–553.
URL http://www.jstor.org/stable/2337125

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Quanli, W., West, M., 2008. High-dimensional sparse factor modeling: Applications in gene expression genomics. Journal of the American Statistical Association 103 (484), 1438–1456.
URL http://dx.doi.org/10.1198/016214508000000869

Cogley, T., Startz, R., June 2013. Robust estimation or ARMA models with near root cancellation, working paper, University of California at Santa Barbara.

Dufour, J.-M., 1997. Some impossibility theorems in econometrics, with applications to structural and dynamic models. Econometrica 65 (6), 1365–1389.
URL http://dx.doi.org/10.2307/2171740

Dufour, J.-M., Hsiao, C., 2008. Identification. In: Durlauf, S. N., Blume, L. E. (Eds.), The New Palgrave Dictionary of Economics, 2nd Edition. Palgrave Macmillan.
URL http://dx.doi.org/10.1057/9780230226203.0762

Eaton, M. L., 1989. Group invariance application in statistics. Regional Conference Series in Probability and Statistics. Institute of Mathematical Statistics.
URL http://www.jstor.org/stable/4153172

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2000. The generalized dynamic-factor model: Identification and estimation. The Review of Economics and Statistics 82 (4), 540–554.
URL http://www.jstor.org/stable/2646650

Frühwirth-Schnatter, S., 1994. Data augmentation and dynamic linear models. Journal of Time Series Analysis 15 (2), 183–202.
URL http://dx.doi.org/10.1111/j.1467-9892.1994.tb00184.x

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.
URL http://www.jstor.org/stable/2670359

Frühwirth-Schnatter, S., Lopes, H. F., 2010. Parsimonious Bayesian factor analysis when the number of factors is unknown, Technical report, University of Chicago, Booth School of Business.

Frühwirth-Schnatter, S., Wagner, H., 2010. Stochastic model specification search for Gaussian and non-Gaussian state space models. Journal of Econometrics 154 (1), 85–100.
URL http://dx.doi.org/10.1016/j.jeconom.2009.07.003

George, E., McCulloch, R., 1993. On obtaining invariant prior distributions. Journal of Statistical Planning and Inference 37 (2), 169–179.
URL http://www.sciencedirect.com/science/article/pii/037837589390086L

Geweke, J., 2004. Getting it right: Joint distribution tests of posterior simulators. Journal of the American Statistical Association 99 (467), 799–804.
URL http://dx.doi.org/10.1198/016214504000001132

Geweke, J., 2007. Interpretation and inference in mixture models: Simple MCMC works. Computational Statistics & Data Analysis 51, 3529–3550.
URL http://dx.doi.org/10.1016/j.csda.2006.11.026

Geweke, J., Zhou, G., 1996. Measuring pricing error of the arbitrage pricing theory. The Review of Financial Studies 9 (2), 557–587.
URL http://www.jstor.org/stable/2962214

Geweke, J. F., Singleton, K. J., 1980. Interpreting the likelihood ratio statistic in factor models when sample size is small. Journal of the American Statistical Association 75 (369), 133–137.
URL http://dx.doi.org/10.1080/01621459.1980.10477442

Grewal, M. S., Andrews, A. P., 2008. Kalman Filtering: Theory and Practice Using MATLAB, 3rd Edition. Wiley-IEEE Press.

Hamilton, J., Waggoner, D., Zha, T., 2007. Normalization in econometrics. Econometric Reviews 26 (2-4), 221 – 252.
URL http://dx.doi.org/10.1080/07474930701220329

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.

Heiss, W. D., 1994. Distributions of angles of a random unit vector and random orthogonal matrices. Zeitschrift für Physik A 349, 9–12.
URL http://dx.doi.org/10.1007/BF01296327

Heiss, W. D., Sannino, A. L., 1990. Avoided level crossing and exceptional points. Journal of Physics A 23, 1167–1178.
URL http://dx.doi.org/10.1088/0305-4470/23/7/022

Hillier, G. H., 1990. On the normalization of structural equations: Properties of direction estimators. Econometrica 58 (5), 1181–1194.
URL http://www.jstor.org/stable/2938305

Hobert, J., Robert, C., Goutis, C., 1997. Connectedness conditions for the convergence of the Gibbs sampler. Statistics & Probability Letters 33 (3), 235–240.
URL http://dx.doi.org/10.1016/S0167-7152(96)00132-0

Hobert, J. P., Casella, G., 1996. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. Journal of the American Statistical Association 91 (436), 1461–1473.
URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476714

Joslin, S., Singleton, K. J., Zhu, H., 2011. A new perspective on Gaussian dynamic term structure models. The Review of Financial Studies 24 (3), 926–970.
URL http://dx.doi.org/10.1093/rfs/hhq128

Kass, R. E., Wasserman, L., 1996. The selection of prior distributions by formal rules. Journal of the American Statistical Association 91 (435), 1343–1370.
URL http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1996.10477003

Kastner, G., Frühwirth-Schnatter, S., 2014. Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. Computational Statis-

tics & Data Analysis 76, 408–423.
URL http://dx.doi.org/10.1016/j.csda.2013.01.002

Kaufmann, S., Schumacher, C., April 2013. Bayesian estimation of sparse dynamic factor models with order-independent identification, Working Paper 13.04, Study Center Gerzensee.

Kenny, D. A., 1979. Correlation and causality. Wiley.

Kim, S., Shephard, N., Chib, S., 1998. Stochastic volatility: likelihood inference and comparison with ARCH models. The Review of Economic Studies 65 (3), 361–393.

Kleibergen, F., Hoek, H., March 2000. Bayesian analisys of ARMA models, Tinbergen Institute Discussion Paper TI 2000-027/4.

Koop, G., Strachan, R., van Dijk, H., Villani, M., 2006. Bayesian approaches to cointegration. In: Mills, T., Patterson, T. (Eds.), The Pelgrave Handbook of Econometrics: Theoretical Econometrics. Vol. 1. Pelgrave Macmillan, Ch. 25, pp. 871–898.

Liu, J., Wu, Y., 1999. Parameter expansion for data augmentation. Journal of the American Statistical Association 94 (448), 1264–1274.
URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10473879

Lopes, H. F., West, M., 2004. Bayesian model assessment in factor analysis. Statistica Sinica 14 (1), 41–67.
URL http://www.jstor.org/stable/24307179

McCausland, W. J., 2012. The HESSIAN method: Highly efficient simulation smoothing, in a nutshell. Journal of Econometrics 168 (2), 189–206.
URL http://dx.doi.org/10.1016/j.jeconom.2011.12.003

McCulloch, R., Rossi, P. E., 1994. An exact likelihood analysis of the multinomial probit model. Journal of Econometrics 64 (1), 207–240.
URL http://dx.doi.org/10.1016/0304-4076(94)90064-7

McMillin, W. D., 2001. The effect of monetary policy shocks: Comparing contemporaneous versus long-run identifying restrictions. Southern Economic Journal 67 (3), 618–636.
URL http://dx.doi.org/10.2307/1061454

Meng, X.-L., van Dyk, D. A., 1999. Seeking efficient data augmentation schemes via conditional and marginal augmentation. Biometrika 86 (2), 301–320.
URL http://www.jstor.org/stable/2673513

Millsap, R. E., 2001. When trivial constraints are not trivial: The choice of uniqueness constraints in confirmatory factor analysis. Structural Equation Modeling: A Multidisciplinary Journal 8 (1), 1–17.
URL http://dx.doi.org/10.1207/S15328007SEM0801_1

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 (1), S125–S140.
URL http://www.jstor.org/stable/2353264

Papaspiliopoulos, O., Roberts, G. O., Sköld, M., 2003. Non-centered parameterizations for hierarchical models and data augmentation. In: Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M. (Eds.), Bayesian Statistics. Vol. 7. Oxford University Press, pp. 307–326.

Papaspiliopoulos, O., Roberts, G. O., Sköld, M., 2007. A general framework for the parametrization of hierarchical models. Statistical Science 22 (1), 59–73.
URL http://www.jstor.org/stable/27645805

Pitt, M. K., Shephard, N., 1999. Analytic convergence rates and parameterization issues for the Gibbs sampler applied to state space models. Journal of Time Series Analysis 20 (1), 63–85.
URL http://dx.doi.org/10.1111/1467-9892.00126

Robert, C. P., Casella, G., 2004. Monte Carlo Statistical Methods, 2nd Edition. Springer.

Roberts, G. O., Smith, A. F. M., 1994. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. Stochastic Processes and their Applications 49 (2), 207–216.
URL http://dx.doi.org/10.1016/0304-4149(94)90134-1

Rubio-Ramìrez, J. F., Waggoner, D. F., Zha, T., 2010. Structural vector autoregressions: Theory of identification and algorithms for inference. The Review of Economic Studies 77 (2), 665–696.
URL http://dx.doi.org/10.1111/j.1467-937X.2009.00578.x

Ruud, P. A., 1991. Extensions of estimation methods using the EM algorithm. Journal of Econometrics 49 (3), 305–341.
URL http://dx.doi.org/10.1016/0304-4076(91)90001-T

Simpson, M., Niemi, J., Roy, V., 2017. Interweaving Markov Chain Monte Carlo strategies for efficient estimation of dynamic linear models. Journal of Computational and Graphical Statistics 26 (1), 152–159.
URL http://dx.doi.org/10.1080/10618600.2015.1105748

Stephens, M., 1997. Bayesian methods for mixtures of normal distributions. Ph.D. thesis, University of Oxford.

Yu, Y., Meng, X.-L., 2011. To center or not to center: That is not the question - An ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. Journal of Computational and Graphical Statistics 20 (3), 531–570.
URL http://dx.doi.org/10.1198/jcgs.2011.203main

Appendix A. The identification principle

Because there are many ways to normalizing a model, it is natural to ask whether certain normalizations define parameter posteriors with more desirable finite-sample properties than alternatives. In the following discussion, unimodal, symmetric and low kurtosis parameter posterior and parameter estimator sampling distributions are assumed to be desirable. I will say that a distribution with these attributes is *regular*.

Because empirical underidentification difficulties arise when the model is not globally identified, Hamilton et al. (2007) propose an *identification principle*[26] as a general guideline for the choice of normalization. In this appendix, I briefly describe their innovative approach (which they illustrate with numerous examples) and I generalize some of its elements. This description is presented as four conditions that normalization should satisfy. I also provide examples for which the approach falls short of providing a unique normalization or ensuring that the parameter posterior has a regular shape.

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

$$\Theta^N = \bigcap_{i=1}^{I} \left\{ \theta \in \Theta \middle| \, \mathbf{g}_i^\top \theta \geq 0 \right\} \cap \bigcap_{j=1}^{J} \left\{ \theta \in \Theta \middle| \, \mathbf{h}_j^\top \theta = \tau_j \right\},$$

for some real scalar $\tau_1, \ldots, \tau_J$ and sets of conformable linearly independent real vectors $\{\mathbf{g}_1, \ldots, \mathbf{g}_I\}$ and $\{\mathbf{h}_1, \ldots, \mathbf{h}_J\}$. Such normalizations impose $I + J$ identifying restrictions. Half-spaces are useful for normalizing countable groups of transformations while hyperplanes can normalize uncountable groups of transformations. For example, one would break invariance under a group of $(I+1)!$ permutations with a normalization consisting in the intersection of I half-spaces. In contrast, the intersection of J hyperplanes would break invariance under a group

[26] Andrews and Cheng (2012) implicitly consider a similar notion by introducing a canonical parametrization under which the parameter vector is partitioned into three subvectors, *i.e.* $\theta = (\theta_1, \theta_2, \theta_3)$, in such a manner that θ_3 is identified if and only if $\theta_1 \neq 0$, θ_2 is not related to the identification of θ_3, and θ_1 and θ_2 are always identified. Unfortunately, LSSMs cannot be parameterized in this manner.

of transformations that is equinumerous to $\Re^J$. Considering intersections of half-spaces and hyperplanes is not as restrictive as it might seem. In particular, one can specify half-spaces and hyperplanes in any space that is homeomorphic to Θ (Hobert et al., 1997).

IP1 - Convexity

Multimodality issues are more likely if the normalization is not convex. A minimal, if perhaps trivial, criterion is thus that a *normalization should be convex*. As intersections of half-spaces and hyperplanes are convex, a normalization consisting in intersections of half-spaces and hyperplanes satisfies this criterion.

IP2 - The boundary of an half-space that break invariance under a countable group

Hamilton, Waggoner, and Zha (2007) advise that "... the boundaries of $[\Theta^N]$ should correspond with the loci in $[\Theta]$ along which the [model] is locally unidentified or the log likelihood diverges to $-\infty$.". The expression "corresponds to" is not technically precise. In the light of the next example, conditions on the boundaries of normalizations should more precisely be expressed as follows, which I will refer to as **condition IP2**:

The hyperplane defining the half-space $\left\{\theta \in \Theta \middle| \, \mathbf{g}^\top \theta \geq 0\right\}$ *should either (a) include the parameter subspace on which the model is locally unidentified or (b) be included in the parameter subspace on which the log likelihood function diverges to* $-\infty$.

Example 1. *Consider the location-and-scale mixture of two normal distributions*

$$l\left(\mu_1, \mu_2, \pi, \sigma_1, \sigma_2 \middle| \, \mathbf{y}\right) = \pi \phi\left(\mathbf{y} \middle| \, \mu_1, \sigma_1\right) + \left(1 - \pi\right) \phi\left(\mathbf{y} \middle| \, \mu_2, \sigma_2\right).$$

The parameter subspace on which the model is locally unidentified is

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

and the log likelihood diverges to $-\infty$ *as* θ *gets closer to*

$$\Theta^{-\infty} = \left\{\theta \in \Theta \middle| \, \sigma_1 = 0\right\} \cup \left\{\theta \in \Theta \middle| \, \sigma_2 = 0\right\}.$$

Condition IP2 rules out identifying restrictions based on π such as

$$\Theta^{\pi} = \{\theta \in \Theta | \, \pi \geq 0.5\}$$

because the boundary of Θ^{π} does not include Θ^{u}. Intuitively, this normalization would perform poorly if the data came from a mixture distribution with $\pi = 0.5$ or with $\theta \in \Theta^{u}$, in which case π is not identified.

Condition IP2 also rules out identifying restrictions like

$$\Theta^{\sigma^*} = \{\theta \in \Theta | \, \sigma_1 \geq 0, \sigma_2 \in (-\infty, -2) \cup [0, 2]\}$$

because the boundary of Θ^{σ^*} is not included in $\Theta^{-\infty}$. However, condition IP2 does not yield a unique solution as it is satisfied by the normalizations

$$\Theta^{\mu\sigma_\alpha} = \{\theta \in \Theta | \, \alpha_\mu (\mu_1 - \mu_2) + \alpha_\sigma (\sigma_1 - \sigma_2) \geq 0\}$$

for $(\alpha_\mu, \alpha_\sigma) \in \Re^2$. Moreover, none of these normalizations ensures unimodality because the model is permutation invariant on the normalizations' boundary. For instance, the special case $\Theta^{\mu} = \{\theta \in \Theta | \, \mu_1 \geq \mu_2\}$ would perform poorly if the data came from a mixture distribution with $\mu_1 = \mu_2$ and the posterior distribution of σ_1 and σ_2 would be multimodal if $\sigma_1 \neq \sigma_2$[27]*.* □

In some models, condition IP2 yields a unique normalization, which provides global identification on its interior and ensures unimodal distributions. In slightly more general models (*e.g.* example 1), it is less straightforward to apply, as it may yield uncountably many normalizations, none of which ensures unimodal distributions because the model is invariant under a countable group of transformations on the normalization's boundary.

IP3 - Local identification at all interior points

Hamilton, Waggoner, and Zha (2007) suggest that specifying conditions on the hyperplane defining the boundary of a half-space ensures "that the model is locally identified at all interior points". This is not the case, however, because specifying the boundary of a half-space does

[27] See Geweke (2007) for an empirical illustration.

not uniquely defines that half-space: there are two halves. As the following example illustrates, this lack of uniqueness can be a problem when the invariance group contains more that two transformations.

Example 2. *Consider the scale mixture of three normal distributions*

$$l\left(\mu, \pi_1, \pi_2, \sigma_1, \sigma_2, \sigma_3 | \, \mathbf{y}\right) = \pi_1 \phi\left(\mathbf{y} | \, \mu, \sigma_1\right) + \pi_2 \phi\left(\mathbf{y} | \, \mu, \sigma_2\right) + \left(1 - \pi_1 - \pi_2\right) \phi\left(\mathbf{y} | \, \mu, \sigma_3\right).$$

The parameter subspace on which the model is locally unidentified is

$$\Theta^u = \left\{\theta \in \Theta | \, \sigma_1 = \sigma_2\right\} \cup \left\{\theta \in \Theta | \, \sigma_1 = \sigma_3\right\} \cup \left\{\theta \in \Theta | \, \sigma_2 = \sigma_3\right\}.$$

Normalizations

$$\Theta^{\sigma_a} = \left\{\theta \in \Theta | \, \sigma_1 \geq \sigma_2, \sigma_1 \geq \sigma_3\right\},$$

$$\Theta^{\sigma_b} = \left\{\theta \in \Theta | \, \sigma_1 \geq \sigma_2, \sigma_1 \leq \sigma_3\right\},$$

$$\textit{and} \quad \Theta^{\sigma_c} = \left\{\theta \in \Theta | \, \sigma_1 \geq \sigma_2, \sigma_2 \geq \sigma_3\right\}$$

satisfy condition IP2, but only Θ^{σ_b} and Θ^{σ_c} ensure local identification at all interior points: the model is locally unidentified on $\left\{\theta \in \Theta | \, \sigma_2 = \sigma_3\right\} \subset \text{interior}\left(\Theta^{\sigma_a}\right)$. □

Because condition IP2 does not imply local identification at all interior points, the latter condition must be verified separately.

IP4 - Commutativity of mapping composition

When a model is invariant under several groups of transformations and normalization is operationalized as a mapping, this mapping is obtained sequentially by composing the mappings that operationalize each normalization. Certain normalizations can only be operationalized by composing mappings in a particular order. For example, if permutation invariance were to be broken by sorting the elements of a parameter vector in ascending order and reflection invariance by ensuring that the elements of that vector be positive, reflection invariance must be broken before permutation invariance. For certain other normalizations, composition order has no consequence.

Example 3 (Example 1, continued)**.** *The model is permutation invariant, but it is also reflection invariant because*

$$\begin{aligned} l\left(\mu_1, \mu_2, \pi, \sigma_1, \sigma_2 \middle| \mathbf{y}\right) &= l\left(\mu_1, \mu_2, \pi, -\sigma_1, \sigma_2 \middle| \mathbf{y}\right) \\ &= l\left(\mu_1, \mu_2, \pi, \sigma_1, -\sigma_2 \middle| \mathbf{y}\right) \\ &= l\left(\mu_1, \mu_2, \pi, -\sigma_1, -\sigma_2 \middle| \mathbf{y}\right). \end{aligned}$$

Consider a parameter value $(\sigma_1, \sigma_2) = (-1, -2)$*. Operationalizing* $\Theta^{\sigma_1 \geq 0, \sigma_2 \geq 0, \sigma_1 \geq \sigma_2}$ *as a mapping and applying permutation normalization first would result in the parameter value*

$$(-1, -2) \rightarrow (-2, -1) \rightarrow (2, 1) \notin \Theta^{\sigma_1 \geq 0, \sigma_2 \geq 0, \sigma_1 \geq \sigma_2}.$$

In contrast, applying reflection normalization first results in

$$(-1, -2) \rightarrow (1, 2) \rightarrow (1, 2) \in \Theta^{\sigma_1 \geq 0, \sigma_2 \geq 0, \sigma_1 \geq \sigma_2}.$$

□

This property is more than an implementation detail. For if mappings must be composed in a particular order and the first normalization that is operationalized fails empirically, this failure might propagate to the following normalizations. Normalizations that can be operationalized by composing mappings in any order are more robust to empirical underidentification and thus preferable.