Forecasting with Weakly Identified Linear State-Space Models.

Sébastien Blais [a,1]

[a] *Bank of Canada.*
First draft: December 17, 2007.
This draft: 20 September 2009.

Abstract

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

JEL classification: C11; C5; C52

Keywords: State-space models; Out-of-sample forecasts; Bayesian methods; Kalman filter

Email address: sblais@sebastienblais.com.

[1] The latest version of this working paper is available at http://www.sebastienblais.com/research/fwilssm.pdf. Financial support from CDP Capital is gratefully acknowledged.

Introduction

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

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

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

As the term suggests and is generally understood, a normalization is a restriction of the parameter space (*i.e.* a parameter subspace) that does not contain any information about the observables or the parameters. From that perspective, normalization

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

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

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

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

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

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

This paper is organized as follows.

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

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

In the third section, I propose permutation- and reflection-invariant prior distributions for the parameters of Gaussian LSSMs. Some of these priors rely on a reparameterization of the model that is easier to interpret. While normalization and parameterization do not change the informational content of the likelihood function, they might affect the interpretation of the parameters and, consequently, the specification of prior information. I highlight situations in which one can inadvertently penalize reasonable regions of the parameter space.

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

The fifth section considers the relationship between forecasting performance and weak identification. Because Bayesian predictive densities are reflection- and permutation-invariant, they are not affected by permutation and reflection normalization. Using simulations, I compare the performance of Bayesian and ML forecasts, on the basis of

out-of-sample root mean square errors, and I find that the advantage of taking parameter uncertainty into account increases as reflection identification becomes weaker. I conclude with a research agenda for future research on these matters.

1 Weak identification

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

Definition 1 *A **parametric statistical model** is a triplet $(\mathcal{Y}, \mathcal{F}, \Theta)$, where $\mathcal{Y}$ is the sample space, $\mathcal{F} \equiv \left\{ f(y|\,\theta) \middle|\, y \in \mathcal{Y},\, \theta \in \Theta \right\}$ is a set of parametric probability density functions on $\mathcal{Y}$ and Θ is the parameter set. The **likelihood function** of the model is the function $l(\theta|\,y) = f(y|\,\theta)$.*

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

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

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

Example 1 (Normal mean) *Consider*

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

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

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

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

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

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

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

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

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

$$f(\mathbf{y}|\, \mu_1, \mu_2, \pi, \sigma) = f(\mathbf{y}|\, \mu_2, \mu_1, 1 - \pi, \sigma), \tag{1.1}$$

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

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

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

1.1 What is normalization?

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

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

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

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

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

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

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

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

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

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

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

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

where

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

is a singleton: the identity transformation.

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

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

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

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

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

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

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

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

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

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

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

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

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

1.2 Is normalization necessary?

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

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

Example 7 (Location mixture, continued) *Permutation invariance implies that the likelihood function admits two equivalent global maxima, sitting at the summit*

***symmetric lobes**: if $\left(\hat{\mu}, \hat{\Pi}, \hat{\sigma}\right)$ is a global maximum, so is $\left(\mathbf{P}\hat{\mu}, \mathbf{P}\hat{\Pi}, \hat{\sigma}\right)$, for $\mathbf{P} = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$.*

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

Definition 5 *A set of transformations*

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

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

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

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

I will argue in Section 3 that it is conceptually inconsistent to express prior beliefs over the relative plausibility of observationally equivalent parameter values. In other words, if the likelihood function is invariant with respect to a given transformation set, prior distributions should also be invariant with respect to that set.

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

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

Example 10 (Normal mean, continued) *$f(y|\, b, \sigma^2)$ is invariant with respect to $\mathcal{T}_S(\Theta)$ and $\mathcal{T}_D(\Theta)$. Finding a proper joint prior distribution that is invariant with respect to $\mathcal{T}_S(\Theta)$ is straightforward. For example, a joint prior is invariant with respect to $\mathcal{T}_S(\Theta)$ as soon as its marginal prior $f(b)$ is symmetric and centered on zero. In contrast, there exists no proper joint prior $f(b, \sigma^2)$ such that $f(b, \sigma^2) = f(ab, \sigma^2/a^2|\, a)$*

for all $a \in \mathcal{A} = (0, \infty)$ *as this would require that the prior be constant over unbounded sets.*

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

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

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

1.3 What are the costs and benefits of normalization?

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

In the Bayesian framework, if prior distributions and the likelihood function are invariant with respect to a set of transformations $\mathcal{T}$, then so are posterior distributions. Invariant *proper* prior and posterior distributions are perfectly valid characterization of uncertainty. In cases where $\mathcal{T}$ is finite (but not a singleton), some posterior distributions are multimodal and thus cause interpretational difficulties. For example, if the bimodal posteriors of μ_1 and μ_2 in (1.1) are symmetric with respect to zero, the posterior means of these parameters are both equal to zero, $\mathbb{E}[\mu_1|\mathbf{y}] = \mathbb{E}[\mu_2|\mathbf{y}] = 0$ which is not particularly informative about the mixture components.

The model should therefore be normalized if the investigator uses mixtures or LSSMs as classification tools where the interpretation of the components or state variables is of interest [2]. When the parameters are not of direct interest however, such as when one uses latent-variable models as flexible parameterizations of the observables's distribution, transformation invariance introduces no interpretational difficulty and normalization, beyond what is required in order to obtain well defined set estimators, is

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

unnecessary.

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

1.4 Are the potential benefits of normalization always achievable?

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

These problems arise when some elements of θ are weakly identified. Except in the context of instrumental variables (IV) and the generalized method of moments (GMM), weak identification has not been defined precisely. Dufour and Hsiao (2008) write: "More generally, any situation where a parameter may be difficult to determine because we are close to a case where a parameter ceases to be identifiable may be called *weak identification*." Many common situations fit this description. For example, multicollinearity issues arise in linear regression models when the sample covariance matrix of the regressors is "close" to being singular. If one restricts attention to ML inference, weak identification problems occur when the Fisher information matrix is close to being singular at the pseudo-true parameter values. Thus, weak identification is a joint property of both the model $\left(\mathcal{Y}, \mathcal{F}, \Theta^N\right)$ and the data y, as the term "empirical underidentification" used in psychometrics emphasizes. In this paper, I say that a parametric model is **weakly identified** if $\hat{\Theta}\left(\Theta^N\right)$ is close to Θ^l, where $\Theta^l \subseteq \Theta$ is the **singularity parameter subspace** where the Fisher information matrix is singular.

Example 12 (Location mixture, continued) *The information matrix is singular on* $\{\theta \in \Theta | \, \mu_1 = \mu_2\} \subset \Theta^\pi$*, where the probability* π *is unidentified. Thus we say that the model is weakly identified if the pseudo-true parameters* μ_1 *and* μ_2 *are too close to each other. In such situations, the lobes of the likelihood function are not well separated and they are not symmetric with respect to their respective mode. A symmetric normal approximation of the ML estimators's sampling distribution is thus unlikely to be accurate. Indeed, Dick and Bowden (1973) compare a Monte Carlo approximation of the parameter sampling variances to their asymptotic counterparts,*

which are approximated by a power series expansion of the information matrix developed by Hill (1963). They report that [Summary] "the sample variance of the estimates can be as much as three times greater than the estimated asymptotic variances".

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

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

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

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

1.5 How best to normalize?

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

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

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

The two other criteria pertain to the shape of point estimator sampling or parameter posterior distributions. Obviously, multimodality issues will arise more often if the normalization is disconnected. A second criteria is thus that the normalization should be connected [3]. Note that intersections of half spaces and hyper-plans are connected spaces. Also, continuous bijections preserve connectedness (Royden, 1988).

Global identification does not only ensure ML estimator's uniqueness, it also affects its sampling distribution. If global identification is achieved through normalization, then normalization has implications for estimator sampling and parameter posterior distributions. Hiller (1990) shows how normalization in structural equations models affects the finite-sample distribution of ordinary least squares and two-stage least squares estimators. Unfortunately, as Hamilton, Waggoner, and Zha (2007) note, "the fact that normalization can materially affect the conclusions one draws from likelihood-based methods is not widely recognized."

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

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

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

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

a) is observationally unrestrictive;
b) is connected;
c) provides global identification.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.6 Summary

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

2 Normalization of LSSMs

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

$$\xi_{t+1} = \mathbf{F}\xi_t + \mathbf{v}_t, \tag{2.1}$$
$$\mathbf{y}_t = \mathbf{B} + \mathbf{A}'\mathbf{x}_t + \mathbf{H}'\xi_t + \mathbf{w}_t, \tag{2.2}$$

where $\mathbf{v}_t$ and $\mathbf{w}_t$ are Gaussian white noises with covariance matrices $\mathbf{Q}$ and $\mathbf{R}$, respectively [4]. Equation (2.1) is referred to as the *state equation* and equation (2.2) as the *observation equation*. State variables are also known as *factors* and $\mathbf{H}$ as the matrix of *factor loadings*. For expositional clarity I consider only the case $\mathbf{A} = \mathbf{0}$ and $N \geq K$, but this is not a substantive restriction.

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

$$l(\mathbf{B}, \mathbf{M}'^{-1}\mathbf{H}, \mathbf{R}, \mathbf{MFM}^{-1}, \mathbf{MQM}', \mathbf{M}\xi_1|y_t) \equiv l(\mathbf{B}, \mathbf{H}, \mathbf{R}, \mathbf{F}, \mathbf{Q}, \xi_1|y_t). \tag{2.3}$$

Thus, the system (2.1-2.2) can be written as

$$\tilde{\xi}_{t+1} = \tilde{\mathbf{F}}\tilde{\xi}_t + \tilde{\mathbf{v}}_t, \tag{2.4}$$
$$\mathbf{y}_t = \mathbf{B} + \mathbf{A}'\mathbf{x}_t + \tilde{\mathbf{H}}'\tilde{\xi}_t + \mathbf{w}_t, \tag{2.5}$$

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

2.1 Primitive transformations

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

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

Let $\mathcal{T}_\mathbf{D}$, $\mathcal{T}_\mathbf{O}$, $\mathcal{T}_\mathbf{P}$ and $\mathcal{T}_\mathbf{S}$ denote these four sets of primitive transformations. Permutation and reflection matrices are orthogonal matrices, *i.e.* $\mathbf{P}'\mathbf{P} = \mathbf{S}'\mathbf{S} = \mathcal{I}$; rotation matrices are special orthogonal matrices, *i.e.* $\mathbf{O}'\mathbf{O} = \mathcal{I}$ and $|\mathbf{O}| = 1$. Any linear

[4] Appendix A shows how to generalize my results to LSSMs with correlated errors.

transformation can be decomposed into these primitive transformations in order to help clarifying invariance issues.

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

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

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

> "In order for the model to be identifiable, restrictions must be placed on [$\mathbf{Q}$] and [$\mathbf{H}$]. In classical factor analysis, the covariance matrix of the common factors is taken to be an identity matrix. However, this is not sufficient to make the model identifiable since if [$\mathbf{M}$] is an orthogonal matrix, [(2.4-2.5)] still satisfies all the restrictions of the original model because [$\mathbf{Var}(\mathbf{M}\mathbf{v}_t) = \mathbf{M}\mathbf{M}' = \mathcal{I}$]. Some restrictions are needed on [$\mathbf{M}$], and one way of imposing them is to require that the ij-th element of [$\mathbf{M}$], [$\mathbf{M}_{ij}$], be zero for $j > i$, $i = 1, \ldots, K-1$. Alternatively, [$\mathbf{Q}$] can be set equal to a diagonal matrix while [$\mathbf{M}_{ij}$] $= 0$ for $j > i$ and [$\mathbf{M}_{ii}$] $= 1$ for $i = 1, \ldots, K$."

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

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

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

Both of Harvey's normalizations are observationally restrictive because they involve K^2 parameter restrictions while breaking scale invariance requires K restrictions and breaking rotation invariance requires $K(K-1)/2$ restrictions. Thus, $K(K-1)/2$ additional parameter restrictions reduce the model's flexibility. I will present several normalizations in this section, but consider a simple one here in order to illustrate how one can normalize scales and rotations with $K(K+1)/2$ restrictions. The cen-

tral issue is that identity matrices are diagonal matrices with diagonal elements all set to 1, which implies that $\mathbf{M}\mathcal{I}\mathbf{M}' = \mathcal{I}$ if $\mathbf{M}$ is an orthogonal matrix. In contrast, consider diagonal matrices with diagonal elements set to distinct values, say $\mathbf{Q}_{kk} = k$ for $k = 1, \ldots, K$. Because $\mathbf{MQM}' \neq \mathbf{Q}$ when $\mathbf{M}$ is orthogonal, these $K(K+1)/2$ restrictions break rotation and scale invariance.

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

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

2.2 Breaking rotation invariance

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

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

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

For the reasons given above, $\Theta^{\mathbf{H}_{lt}}$ does not provide global identification. Nor does $\Theta^{\mathbf{F}_{lt}}$, by similar arguments. This could lead one to consider $\Theta^{\mathbf{F}_{sym}}$ as simultaneous

row-and-column permutation is a bijective transformation for symmetric matrices. However, this is observationally restrictive because the eigenvalue of real symmetric matrices are real. In particular, this would precludes latent variables with sinusoidal dynamics.

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

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

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

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

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

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

where

$$\mathbf{B}_k = \rho_{k,k+1} \rho_{k,k+2} \cdots \rho_{k,N},$$
$$\rho_{i,j} = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & \cos\gamma_{i,j} & & & -\sin\gamma_{i,j} & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & \sin\gamma_{i,j} & & & \cos\gamma_{i,j} & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{bmatrix}_{N \times N},$$
$$\mathbf{U}_{N \times K} = \begin{bmatrix} \delta_1 & & \\ & \ddots & \\ & & \delta_K \\ \mathbf{0}_{N-K \times K} & & \end{bmatrix}$$

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

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

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

2.3 Breaking scale invariance

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

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

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

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

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

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

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

breaks scale invariance and provides global identification.

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

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

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

2.4 Breaking permutation invariance

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

$$\begin{array}{ccccccccc}
\alpha_1 f_1(\gamma_{1,1}-\gamma_{2,1}) & +\ldots+ & \alpha_{N-1} f_{N-1}(\gamma_{1,N-1}-\gamma_{2,N-1}) & + & \alpha_N f_N(\mathbf{F}_{1,1}-\mathbf{F}_{2,2}) & + & \alpha_{N+1} f_{N+1}(\mathbf{Q}_{1,1}-\mathbf{Q}_{2,2}) & > 0 \\
\alpha_1 f_1(\gamma_{2,1}-\gamma_{3,1}) & +\ldots+ & \alpha_{N-1} f_{N-1}(\gamma_{2,N-1}-\gamma_{3,N-1}) & + & \alpha_N f_N(\mathbf{F}_{2,2}-\mathbf{F}_{3,3}) & + & \alpha_{N+1} f_{N+1}(\mathbf{Q}_{2,2}-\mathbf{Q}_{3,3}) & > 0 \\
 & & & & \vdots & & & \\
\multicolumn{9}{c}{\alpha_1 f_1(\gamma_{K-1,1}-\gamma_{K,1}) + \ldots + \alpha_{N-1} f_{N-1}(\gamma_{K-1,N-1}-\gamma_{K,N-1}) + \alpha_N f_N(\mathbf{F}_{K-1,K-1}-\mathbf{F}_{K,K}) + \alpha_{N+1} f_{N+1}(\mathbf{Q}_{K-1,K-1}-\mathbf{Q}_{K,K}) > 0}
\end{array}$$

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

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

$$\begin{array}{ccccccccc}
\alpha_1 f_1(\mathbf{H}_{1,1}-\mathbf{H}_{2,1}) & +\ldots+ & \alpha_N f_N(\mathbf{H}_{1,N}-\mathbf{H}_{2,N}) & + & \alpha_{N+1} f_{N+1}(\mathbf{F}_{1,1}-\mathbf{F}_{2,2}) & > 0 \\
\alpha_1 f_1(\mathbf{H}_{2,1}-\mathbf{H}_{3,1}) & +\ldots+ & \alpha_N f_N(\mathbf{H}_{2,N}-\mathbf{H}_{3,N}) & + & \alpha_{N+1} f_{N+1}(\mathbf{F}_{2,2}-\mathbf{F}_{3,3}) & > 0 \\
 & & \vdots & & & \\
\multicolumn{5}{c}{\alpha_1 f_1(\mathbf{H}_{K-1,1}-\mathbf{H}_{K,1}) + \ldots + \alpha_N f_N(\mathbf{H}_{K-1,N}-\mathbf{H}_{K,N}) + \alpha_{N+1} f_{N+1}(\mathbf{F}_{K-1,K-1}-\mathbf{F}_{K,K}) > 0}
\end{array}$$

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

2.5 Breaking reflection invariance

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

$$\begin{array}{ccccc}
\alpha_{1,1} f_{1,1}(\mathbf{H}_{1,1}) & +\ldots+ & \alpha_{1,N} f_{1,N}(\mathbf{H}_{1,N}) & > 0 \\
\alpha_{2,1} f_{2,1}(\mathbf{H}_{2,1}) & +\ldots+ & \alpha_{2,N} f_{2,N}(\mathbf{H}_{2,N}) & > 0 \\
 & \vdots & & \\
\multicolumn{4}{c}{\alpha_{K,1} f_{K,1}(\mathbf{H}_{K,1}) + \ldots + \alpha_{K,N} f_{K,N}(\mathbf{H}_{K,N}) > 0,}
\end{array}$$

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

to reflections through

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

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

2.6 Root cancelation in the ARMA representation

Weak identification has an interesting interpretation in terms of root cancelation or redundant parameter issues. It is well-known that root cancelation can make parameter point estimators unreliable (Box and Jenkins, 1976). I first show that weak reflection identification in a one-factor LSSM implies root cancelation in its ARMA representation. Because this particular LSSM is not a common representation of ARMA processes, I next show that root cancelation in an ARMA(1,1) process implies weak reflection identification for Aoki's canonical LSSM representation of the stochastic process.

The 1-factor LSSM for an univariate process

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

has the ARMA(1,1) representation

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

where

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

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

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

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

$$\xi_t = \begin{bmatrix} \rho & 0 \\ 1 & 0 \end{bmatrix} \xi_{t-1} + \begin{bmatrix} v_{1,t} \\ 0 \end{bmatrix}, \tag{2.6}$$

$$y_t = \alpha + \begin{bmatrix} 1 & \theta \end{bmatrix} \xi_t. \tag{2.7}$$

Setting one factor loading to 1 in (2.7) breaks reflection invariance but do not provide global identification because the model is locally unidentified on the line $\rho = -\theta$. This is easily seen by substituting (2.6) into (2.7):

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

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

3 Prior Distributions

In this section, I propose permutation- and reflection-invariant prior distributions for the parameters of the LSSM (2.1-2.2). For finite mixture distributions, Geweke (2007, p. 3537) argues that "If the state labels have no substantive interpretation, then the prior density must also be permutation invariant." More generally, prior information should reflect the invariance property of the likelihood function and specifying prior beliefs on quantities that have no substantive interpretation is, at best, conceptually difficult to justify. Moreover, inference might be sensitive to prior specification if priors are informative with respect to reflection or permutation and some parameters are weakly identified.

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

3.1 Permutation- and reflection-invariant priors

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

Alternative approaches require some analysis in order to see how permutation or reflection affects each element of each parameter. Reparameterization sometimes helps in this analysis. Some parameters are naturally reflection invariant, *e.g.* $\mathbf{Q}_{kk}$ or $\mathbf{F}_{kk}$, and permutation invariance is obtained by any exchangeable prior distribution on the diagonal elements of $\mathbf{Q}$ or $\mathbf{F}$ [5]. An exchangeable normal distribution has the form $\mathcal{N}(\mu\iota, \sigma^2((1-\rho)\mathcal{I} + \rho\iota\iota'))$. As another special case, i.i.d. univariate priors are permutation invariant. Priors that are symmetric with respect to 0 are reflection invariant. They are equivalently specified as priors on the absolute values of the parameters.

3.2 Normalization, parameterization, conditional conjugacy and prior information

Permutation- and reflection-invariant, proper priors provide no information with respect to permutation and reflection, but are informative in other dimensions. When computational or other considerations leads one to specifying conditionally conjugate priors on the model's parameters, normalization and parameterization can have unexpected consequences for the resulting inference.

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

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

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

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

3.3 Priors for $\mathbf{F}$, ξ_1 *and* $\mathbf{Q}$

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

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

3.4 Priors for γ

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

4 Posterior Simulation

This section describes posterior simulation for the LSSM (2.1-2.2). I propose a Metropolis-within-Gibbs sampler. In this sampler, parameters without standard conditional posteriors are drawn with the factors as a single block. Next, I propose an extension of Frühwirth-Schnatter's (2001) random permutation sampler to LSSMs and I and discusses the implementation of permutation and reflection normalizations.

4.1 Posterior simulator

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

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

(1) Generate $\mathbf{B}^{(m)} \sim p(\mathbf{B}|y, \gamma^{(m-1)}, \mathbf{R}^{(m-1)}, \mathbf{F}^{(m-1)}, \mathbf{Q}^{(m-1)}, \xi_1^{(m-1)}, \xi^{(m-1)})$
(2) Generate $\mathbf{Q}^* \sim p(\mathbf{Q}|y, , \mathbf{B}^{(m)}, \gamma^{(m-1)}, \mathbf{R}^{(m-1)}, \mathbf{F}^{(m-1)}, \xi_1^{(m-1)}, \xi^{(m-1)})$

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

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

where

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

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

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

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

4.2 Metropolis-Hastings-within-Gibbs

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

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

the proposal is

$$q\left(\gamma', \xi' \middle| \mathbf{y}, \gamma, \Phi, \Sigma_\gamma\right) = p\left(\xi' \middle| \mathbf{y}, \gamma', \Phi\right) \phi\left(\gamma' \middle| \gamma, \Sigma_\gamma\right), \tag{4.1}$$

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

Note that the joint proposal (4.1) does not depend on ξ, the current state of the factors. The Markov chain is less autocorrelated and therefore more efficient than if it did. Because $p\left(\xi'\middle|\mathbf{y},\gamma',\Phi\right)$ is exact, the proposal can be close to its target for a relatively large Σ_γ.

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

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

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

4.3 Mixture sampler

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

Multimodal posteriors constitute a computational challenge for which tempering methods have proved useful (Robert and Casella, 2004, p. 540). However, the mixture sampler I describe in this paper takes advantage of the symmetry of the joint posterior in order to efficiently explore all of its $K!2^K$ lobes with high numerical efficiency. It generalizes Frühwirth-Schnatter's (2001) random permutation sampler in two ways. First, (2.4-2.5) reveals that permutation invariance in LSSMs does not correspond to simple permutations of parameter indices. My mixture sampler deals with more general parameter transformations. Second, it addresses reflection invariance. The invariance property of the mixture sampler follows directly from that of the permutation sampler (See Frühwirth-Schnatter, 2001, Appendix, for a proof).

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

10 are accepted with probability one. Otherwise, one computes the acceptance probability.

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

If it is inferentially redundant, why then would anyone use the random permutation sampler? One answer is a practical one. Assessing the mixing properties of a MCMC sampler is no simple task. In LSSMs, standard methods of assessing convergence must take into account permutation and reflection invariance. For example, methods based on cumulative sums, like Brook's (1998), must consider permutation- and reflection-invariant quantities. Plotting the output of an MCMC sampler is another common way of doing a quick diagnosis of the generated chain. This exercise is complicated by reflection and permutation invariance. For example, for the state-space model (2.1-2.2) with three factors, there are six lobes any element of $\mathbf{B}$ can visit. A trend in the path of a parameter (which could indicate that the effect of initial conditions has not died out) can be difficult to see graphically when the chain keeps switching between lobes. Indeed, this is possibly what led Celeux et al. (2000) [p. 957] to assert that "we consider that almost the entirety of MCMC samplers implemented for mixture models has failed to converge!" The random mixture sampler ensures that all modes are visited. Geweke (2007) proposes to build permuted copies of the parameter vector as a post-simulation step. This approach is inferentially equivalent to the random permutation sampler.

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

5 Simulations Results

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

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

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

I limit my empirical investigation to the impact of weak reflection identification, the nature of the weak permutation identification problem being similar. I set $B = 0$, $R = 1$, $Q = 1$, $F = 0.95$ and $\xi_1 = 0$ and look at the impact of varying H. A smaller factor loading H makes the reflection weak identification problem more severe because the Fisher information matrix is singular at $H = 0$. I consider $H = \{0.005, 0.01, 0.05, 0.1\}$. Note the sample size is irrelevant in itself as one would obtain similar results with a larger T and smaller H's.

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

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

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

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

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

Table 1

Out-of-sample relative performances and weak identification

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

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

Table 1 presents the relative RMSEs of out-of-sample forecasts ($\text{RMSE}_{\text{Bayes}}/\text{RMSE}_{\text{MLE}}$). For a significant proportion of samples, the EM algorithm converges to an AR(1) model. This would not be of particular concern in practice but I present these cases separately because a proper prior on R prevents this from happening in the Bayesian framework. Whether all data sets are considered together, or separated according to where the EM algorithm converged, the improvement increases as H approaches 0. Furthermore, the improvement stabilizes when H is large enough and the lobes are well separated.

6 Concluding Remarks

Inference for linear state-space models is complicated by a weak identification problem; if latent variables are too similar or if factor loadings are too small, the Fisher information matrix is close to being singular and the factors's reflection and permutation are weakly identified. I argue that a connected normalization providing global parameter identification is more likely to produce unimodal posterior distributions or

a maximum-likelihood estimator with unimodal sampling distribution in finite sample, and I propose an observationally unrestrictive normalization of LSSM satisfying these conditions. However, I stress that unimodal distribution cannot be ensured by an observationally unrestrictive normalization.

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

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

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

References

Anderson, B. D., Moore, J. B., 1979. Optimal Filtering. Prentice-Hall, Englewood Cliffs, N.J.

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

Bound, J., Jaeger, D., Baker, R., 1995. Problems with instrumental variables estimation when the correlations betweeen the instruments and the endogenous explanatory variables is weak. Journal of the American Statistical Association 90, 443–450.

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

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

Brooks, S. P., 1998. Quantitative convergence assessment for Markov chain Monte Carlo via cusums. Statistics and Computing 8, 267–274.

Buse, A., 1992. The bias of instrumental variables estimators. Econometrica 60, 173–180.

Carter, C., Kohn, P., 1994. On the Gibbs sampling for state space models. Biometrika 81, 541–553.

Celeux, G., Hurn, M., Robert, C., 2000. Computational and inferential difficulties with mixture prosterior distribution. Journal of the American Statistical Association 95 (451), 957–970.

Chernozhukov, V., Hong, H., Tamer, E., September 2007. Estimation and confidence regions for parameter sets in econometric models. Econometrica 75 (5), 1234–1284.

Chib, S., Ergashev, B., September 2008. Analysis of multi-factor affine yield curve models, wroking paper, Washington University in St. Louis and the Federal Reserve Bank of Richmond.

Dick, N. P., Bowden, D. C., 1973. Maximum likelihood estimation for mixtures of two normal distributions. Biometrics 29, 781–790.

Dufour, J.-M., 1997. Some impossibility theorems in econometrics, with applications to structural and dynamic models. Econometrica 65, 1365–1389.

Dufour, J.-M., Hsiao, C., 2008. "Identification", The New Palgrave Dictionary of Economics, 2nd Edition. Palgrave Macmillan.

Frühwirth-Schnatter, S., 1994. Data augmentation and dynamic linear models. Journal of Time Series Analysis 15, 183–202.

Frühwirth-Schnatter, S., 2001. Markov Chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association 96 (453), 194–205.

Frühwirth-Schnatter, S., 2004. Efficient Bayesian parameter estimation. In: Harvey, A., Koopman, S. J., Shephard, N. (Eds.), State Space and Unobserved Component Models: Theory and Applications. Cambridge University Press, pp. 123–151.

Frühwirth-Schnatter, S., Wagner, H., 2008. Stochastic model specification search for Gaussian and non-Gaussian state space models, iFAS Research Paper Series 2008-36.

Galichon, A., Henry, M., 2009. A test of non-identifying restrictions and confidence regions for partially identified parameters. Journal of Econometrics, forthcoming.

Geweke, J., 2007. Interpretation and inference in mixture models: Simple MCMC works. Computational Statistics & Data Analysis 51, 3529–3550.

Hamilton, J., Waggoner, D., Zha, T., 2007. Normalization in econometrics. Econometric Reviews 26, 221 – 252.

Hamilton, J. D., 1994. Time Series Analysis. Princeton University Press.

Harvey, A. C., 1989. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press.

Hill, B., 1963. Information for estimating the proportions in mixtures of exponential

and normal distributions. Journal of the American Statistical Association 58 (304), 918–932.

Hiller, G. H., 1990. On the normalization of structural equations: properties of direction estimators. Econometrica 58 (5), 1181–1194.

Jacquier, E., Johannes, M., Polson., N., 2007. MCMC maximum likelihood for latent state models. Journal of Econometrics 137.

Jennrich, R. I., 1978. Rotational equivalence of factor loading matrices with specified values. Psychometrika 43 (3), 421–426.

Kim, C., Nelson, C., 1998. Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching. The Review of Economics and Statistics 80, 188–201.

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

Leamer, E. E., 1973. Multicollinearity: A Bayesian perspective. The Review of Economics and Statistics 55, 371–380.

Loken, E., 2004. Multimodality in mixture models and factor models. In: Gelman, A., Meng, X.-L. (Eds.), Applied Bayesian Modeling and causal inference from incomplete-data perspectives. John Wiley & Son, pp. 203–213.

Manski, C., 2003. Partial Identification of Probability Distributions. Springer-Verlag, New York.

Nelson, C. R., Startz, R., 1990. The distribution of the instrumental variable estimator and its t-ratio when the instrument is a poor one. Journal of Business 63 (S125-S140).

Redner, R. A., Walker, H. F., 1984. Mixture densities, maximum likelihood and the EM algorithm. SIAM Review 26 (2), 195–239.

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

Royden, H. L., 1988. Real analysis, 3rd Edition. Prentice Hall.

Shumway, R., Stoffer, D., 1983. An approach to time series smoothing and forecasting using the EM algorithm. Journal of Time Series Analysis 3 (4), 253–264.

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

Stephens, M., 2000. Dealing with label switching in mixture models. Journal of the Royal Statistical Society B 62, 795–809, part 4.

Stoffer, D. S., Wall, K. D., 1991. Bootstrapping state-space models: Gaussian maximum likelihood estimation and the Kalman filter. Journal of the American Statistical Association 86 (416), 1024–1033.

Watson, M., Engle, R., 1983. Altenative algorithms for the estimation of dynamic factor, MIMIC and varying coefficient regression models. Journal of Econometrics 23, 385–400.

A Invariance to linear transformations with correlated errors

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

LSSMs with correlated errors can be represented by the system of equations (2.1-2.2) with

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

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

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

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

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

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

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

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

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

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

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

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

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