Basturk, Nalan; Hoogerheide, Lennart; van Dijk, Herman K.

Working Paper

Bayesian Analysis of Boundary and Near-Boundary Evidence in Econometric Models with Reduced Rank

Working Paper, No. 11/2017

Provided in Cooperation with:
Norges Bank, Oslo

Suggested Citation: Basturk, Nalan; Hoogerheide, Lennart; van Dijk, Herman K. (2017) : Bayesian Analysis of Boundary and Near-Boundary Evidence in Econometric Models with Reduced Rank, Working Paper, No. 11/2017, ISBN 978-82-7553-988-3, Norges Bank, Oslo, https://hdl.handle.net/11250/2495574

This Version is available at:
https://hdl.handle.net/10419/210121

WORKING PAPER

Bayesian analysis of boundary and near-boundary evidence in econometric models with reduced rank

NORGES BANK RESEARCH

11 | 2017

NALAN BASTURK,
LENNART HOOGERHEIDE
AND
HERMAN K. VAN DIJK

NORGES BANK

Working papers fra Norges Bank, fra 1992/1 til 2009/2 kan bestilles over e-post:
FacilityServices@norges-bank.no

Fra 1999 og senere er publikasjonene tilgjengelige på www.norges-bank.no

Working papers inneholder forskningsarbeider og utredninger som vanligvis ikke har fått sin endelige form. Hensikten er blant annet at forfatteren kan motta kommentarer fra kolleger og andre interesserte. Synspunkter og konklusjoner i arbeidene står for forfatternes regning.

Working papers from Norges Bank, from 1992/1 to 2009/2 can be ordered by e-mail
FacilityServices@norges-bank.no

Working papers from 1999 onwards are available on www.norges-bank.no

Norges Bank's working papers present research projects and reports (not usually in their final form) and are intended inter alia to enable the author to benefit from the comments of colleagues and other interested parties. Views and conclusions expressed in working papers are the responsibility of the authors alone.

ISSN 1502-819-0 (online)
ISBN 978-82-7553-988-3 (online)

Bayesian Analysis of Boundary and Near-Boundary Evidence in Econometric Models with Reduced Rank*

Nalan Baştürk[1], **Lennart Hoogerheide**[2] **and Herman K. van Dijk**[3,†]

[1] *Maastricht University e-mail:* n.basturk@maastrichtuniversity.nl

[2] *VU University Amsterdam e-mail:* l.f.hoogerheide@vu.nl

[3] *Econometric Institute, Erasmus University Rotterdam and Norges Bank e-mail:* hkvandijk@ese.eur.nl

"Why econometrics should always and everywhere be Bayesian"
— C. Sims (2007)

Abstract:

Weak empirical evidence near and at the boundary of the parameter region is a predominant feature in econometric models. Examples are macroeconometric models with weak information on the number of stable relations, microeconometric models measuring connectivity between variables with weak instruments, financial econometric models like the random walk with weak evidence on the efficient market hypothesis and factor models for investment policies with weak information on the number of unobserved factors. A Bayesian analysis is presented of the common issue in these models, which refers to the topic of a reduced rank. Reduced rank is a boundary issue and its effect on the shape of the posteriors of the equation system parameters with a reduced rank is explored systematically. These shapes refer to ridges due to weak identification, fat tails and multimodality. Discussing several alternative routes to construct regularization priors, we show that flat posterior surfaces are integrable even though the marginal posterior tends to infinity if the parameters tend to the values corresponding to local non-identification. We introduce a lasso type shrinkage prior combined with

*This paper should not be reported as representing the views of Norges Bank. The views expressed are those of the authors and do not necessarily reflect those of Norges Bank. The authors are indebted to Lukasz Gatarek and Richard Kleijn for expert research assistance. We also thank the editor, two anonymous referees and Mattias Villani for their comments on an earlier version of this paper.

†Corresponding author: hkvandijk@ese.eur.nl.

orthogonal normalization which restricts the range of the parameters in a plausible way. This can be combined with other shrinkage, smoothness and data based priors using training samples or dummy observations. Using such classes of priors, it is shown how conditional probabilities of evidence near and at the boundary can be evaluated effectively. These results allow for Bayesian inference using mixtures of posteriors under the boundary state and the near-boundary state. The approach is applied to the estimation of education-income effect in all states of the US economy. The empirical results indicate that there exist substantial differences of this effect between almost all states. This may affect important national and state-wise policies on required length of education. The use of the proposed approach may, in general, lead to more accurate forecasting and decision analysis in other problems in economics, finance and marketing.

1. Introduction

Inference near and at the boundary of the parameter space of a probability model is occurring frequently in the field of econometrics. We list three economic and financial topics where (near-)boundary evidence became empirically relevant in the second halve of the twentieth century and it led to important econometric research. In micro-econometrics the estimation of the effect of length of education on earned income encountered the (near-)boundary of weak or no endogeneity and/or weak or no identification. In macro-econometrics investigating which and how many stable relations exist between macroeconomic time series has been extensively explored in order to estimate forecast and policy uncertainty. Here moving to the boundary refers to going from near-nonstationarity to unit roots. In financial econometrics efficient data reduction using large cross sectional data on stocks was investigated using a certain number of unobserved factors which affect, for instance, equity momentum strategies. Weak information on the number of factors is a near-boundary issue. To motivate our analysis, we provide in Section 2 several illustrative examples also for more general model structures.
The literature dealing with these issues is substantial and an extensive overview is outside the scope of this paper. In the frequentist econometric literature the focus has been largely on testing whether one's view is at the boundary and on assessing what is the sensitivity of the test when one is near the boundary. We restrict ourselves to listing three classic tests: the Anderson-Rubin test for (over-)identification which is regularly used in the literature on the education-income analysis (Anderson and Rubin, 1950); the Johansen test used for determining the number of stable relations in macro-economic time series (Johansen, 1991); and the Anderson-Rubin test for determining the number of factors (Anderson and Rubin, 1956)
The major message of the present paper is that many modeling, forecasting and policy problems in non-experimental empirical econometrics are not about asymptotically valid parameter estimation and testing near or at a boundary. Given several different sources of information on features of economic processes, the relevant issue is to use this information and average over the available evidence on the different states of the economy, near and at the boundary, where

the evidence on these states is measured using posterior probability weights. The Bayesian approach is eminently suitable for this. We take the viewpoint that the scientific evidence should be reported in such a way that the information specified in the likelihood dominates with respect to other sources of information, see Baştürk et al. (2014a) for a historical background. Thus our approach to specifying prior information is one where relatively weak information is used compared to that of the likelihood.

In order to back-up the general message, this paper makes four points. The first is to show that there exists a common structure in the three issues mentioned and that the effect of the boundary issue on the shape of the posterior densities of the model parameters can be studied within the context of a standard *reduced rank* regression model under different restrictions on the parametric structure and alternative choices of weak priors. It is well-known that the shape of the likelihood, and therefore the shape of the posterior with a flat prior, in the standard multivariate regression model is bell-shaped or elliptical. As a consequence, credibility regions of parameters can be simply determined using second order moments. However, the posterior density of the matrix of equation system parameters in a reduced rank model is non-elliptical. We provide in Section 2 several motivating examples. This nonstandard shape refers to several typical features. We focus on two features that have an effect on the existence of posterior moments: a ridge or, more generally, flat parts in the surface and heavy tails. A ridge refers to weak or non-identification of parameters and it makes a marginal posterior density unbounded, while very heavy tails make the use of first and higher order moments unsuitable for all inference. We will show in Section 3 that the posterior in a standard or *workhorse* reduced rank model, which in our case is a cointegration model, is locally integrable even in the case of a flat prior with flat parts in the posterior surface and the tails are heavy but also integrable. Therefore, the search for plausible restrictions on the parameter space has become an important topic of research. Apart from this research line, we also show that using triangular restrictions on the parameters modify the workhorse model into an instrumental variable regression model and that a normal prior on some equation parameters together with a diagonal covariance matrix on the disturbances modify the workhorse model into a static factor model. We will show that these typical restrictions help in making a posterior with a flat prior more regular with existence of first and higher order moments. We note that, given the structure of our three types of reduced rank models, multi-modality and skewness (of multiple parameters) are more computational problems about numerical evaluation of the posterior but not about the existence. More complex mixture models may give existence problems due to weak empirical identification of a component of the mixture, see for instance Frühwirth-Schnatter (2006).

A second purpose of the paper is to discuss alternative ways that appeared in the literature of specifying *prior regularization* information. This is helpful for determining model weights. One way is to use a more technical econometric approach. That is, construct priors that are based on information or reference theory concepts connected to the identification issue. However, we shall argue

that these priors are in many cases not sufficient for making posteriors proper. We add in Section 4 a new result on the existence of the posterior distribution of model parameters with a reduced rank where the regularizing prior information is based on weak and plausible restrictions on the *range* of the parameters of interest. We introduce a *lasso type shrinkage prior combined with orthogonal normalization*. We also, briefly, explore several other routes that deal with regularizing prior information. The focus is then more on prior information that makes *economic models behave more reasonably*, see Sims (2008). That is, one may be more interested in regular behavior of a nonlinear function of the equation system parameters like the impulse response function of a model after a shock. Here the implications of prior information for posterior and predictive analysis are important. Other examples are the effect of prior information on multipliers of an econometric model, which is prior-predictive analysis and such an effect on posterior estimates of stability of a model, which refers to posterior-predictive analysis.

A third purpose of this paper is to show how the evaluation of conditional probabilities on the evidence of different states of an econometric model can be made operational when the prior information is weak. That is, although the issue of weak identification is not an impediment for obtaining a proper probability, weak prior information and a nearly flat posterior do play a major role in the evaluation of posterior and predictive probabilities of evidence near and at a boundary of non-identification and irrelevant instruments. Given the bounded regions of integration, the Bartlett/Jeffreys/Lindley paradox, see Jeffreys (1939), Lindley (1957) and Bartlett (1957) does not show up as a mathematical statistical result, but it appears as a serious practical problem for model evaluation when prior probabilities are assumed over regions where there is weak or no data information. Here the use of a training sample and weak economic information is recommended. Second, a sensitivity analysis is recommended in order to obtain more robustness in the results. We explore several routes that are described in Section 5. Once a model weight is obtained, Bayesian inference can proceed with model averaging in order to estimate mixtures of models suitable for forecasting and policy analysis.

As a final contribution, in Section 6 we explore the regional differences between all states of the US with respect to the effect of length of education on earned income using an instrumental variables model and a mixture of endogenous and weakly exogenous states of the model. We obtain strong empirical evidence that the financial income returns of education vary substantially between almost all states in the USA. This may affect important state and national policies on the requires length of education.

We emphasize that there is much more done on the topic of model averaging in Bayesian econometrics, a recent example in the field of macroeconomics is given in Strachan and Van Dijk (2013). We refer to the Handbook of Bayesian Econometrics, Geweke et al. (2011), and to the Supplementary Material in the Online Appendix for more examples in the fields of economics, finance and marketing. In Section 7 several perspectives for further research are presented.

Remark 1: Given the length of this paper which is due to a combination of its

survey character as well as presentation of new results, the material is divided into a main text and Supplementary Material which is in the Online Appendix.
Remark 2: The development of efficient computational procedures using simulation-based methods has been essential and an active area of research in Bayesian econometrics but it is a topic beyond the scope of this paper. For a historical analysis of the development of this topic since the early nineteen-seventies we refer to Baştürk et al. (2014a). Modern hardware and software including parallel computation allow detailed analysis of many of the issues listed in this paper.
Remark 3: Bayesian inference of mixture processes is extensively studied in the statistical literature, see *e.g.* Frühwirth-Schnatter (2006) and Mengersen et al. (2011). In this paper we focus on the issues that refer to the evidence near and at the boundary of *econometric* models and how to average over these states.

2. Motivating examples

In this section we provide several motivating examples of the boundary and near-boundary issues and the irregular likelihoods resulting in these examples. One econometric model, the cointegration model, serves as workhorse model for reduced rank analysis in this paper. Two other models, the instrumental variable and the factor model, are special cases of the workhorse model. We illustrate the boundary and near-boundary issue for the cointegration and instrumental variable models using simulated and real data. In addition to these motivating examples, we provide three other empirical applications where the boundary issue is evident in the Supplementary Material.

Posteriors of an instrumental variables (IV) model: The restricted reduced form of an IV model for data y_i with one explanatory variable x_i and two instruments (z_{1i}, z_{2i}) can be written as follows:

$$\begin{pmatrix} y_i \\ x_i \end{pmatrix} = \begin{pmatrix} \beta \\ 1 \end{pmatrix} \begin{pmatrix} \pi_1 & \pi_2 \end{pmatrix} \begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix} + \begin{pmatrix} u_i \\ v_i \end{pmatrix}, \tag{2.1}$$

where β, π_1 and π_2 are scalar model parameters, and disturbances $\begin{pmatrix} u_i & v_i \end{pmatrix}'$ have e.g. an iid normal distribution. This restricted reduced formulation of the model clearly shows the reduced rank structure within this class of models.
Under flat priors, the posterior distribution of the model parameters for the above IV model has a ridge at the region implying 'a move from weak to irrelevant instruments', where $\pi_1 = \pi_2 = 0$. We illustrate this issue in Figure 1. More details are given in the Supplementary material, in Hoogerheide et al. (2007b) and Zellner et al. (2014).

Posteriors of a cointegration model: The second model we consider is a cointegration model, specifically a Vector Error Correction Model (VECM), with data $y_{1,t}, y_{2,t}$:

$$\begin{pmatrix} \Delta y_{1,t} \\ \Delta y_{2,t} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \begin{pmatrix} 1 & -\beta \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix}, \tag{2.2}$$

Fig 1: 95% HPD credible set for π_1, π_2, β for simulated data from the IV model

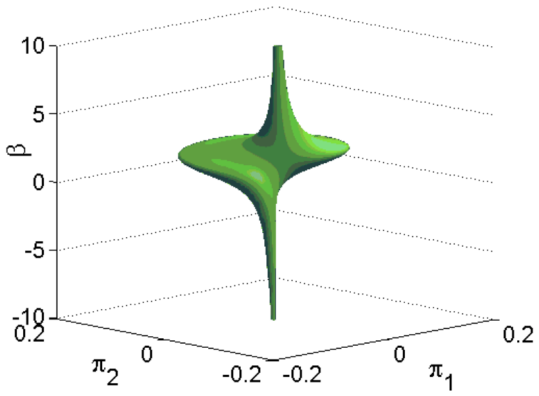
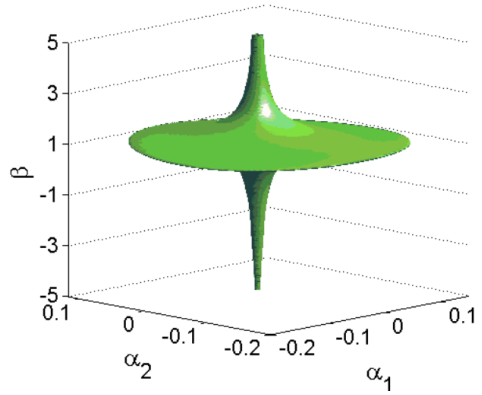


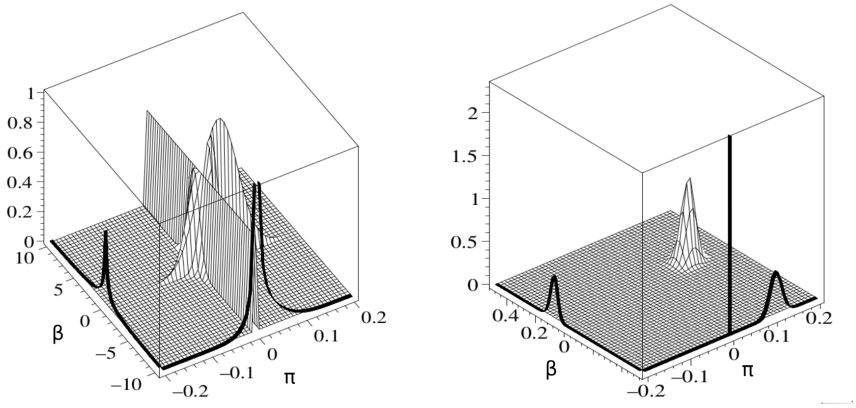
Fig 2: 95% HPD credible set for $\alpha_1, \alpha_2, \beta$ for simulated data from the VECM



where $(\alpha_1, \alpha_2, \beta)$ are the model parameters, the disturbances $(\varepsilon_{1,t}\, \varepsilon_{2,t})'$ have iid normal distributions. Similar to the earlier IV model formulation, the reduced rank issue is evident in the matrix multiplication on the right hand side of this model.
The boundary issue for the posterior distributions for the cointegration model under diffuse priors is illustrated in Figure 2. In this case, the 'boundary' corresponds to the case where there is no dynamic adjustment in the model towards an equilibrium, i.e. $\alpha_1 = \alpha_2 = 0$.
In the Supplementary material the set-up of the experiments for Figures 1 and 2 is given.

Education-income analysis using the IV model: As a first empirical motivating example, we present the posterior density of the parameters of an instrumental variables model for education and income data from individuals living in the US, which are analyzed in Angrist and Krueger (1991) and Hoogerheide and Van Dijk (2008) among others. The fundamental issue is that years of education in these data are instrumented with a dummy variable for individuals born in quarters 2-4 of a year. Quarter of birth had an effect on the years of compulsory schooling, due to the compulsory schooling laws. These data represent a typical 'weak instrument' case since the explanatory power of quarter of birth on education is expected to be present only for individuals whose years of education were affected by the compulsory schooling requirement. We refer to the Supplementary Material in Appendix A.1 for an introduction and more explanations of the instrumental variable model.
Figure 3 illustrates the boundary issue which refers to local non-identification of the posteriors under flat priors for the income-education data of the state of New York and the whole US. The two figures of the joint posterior kernels in the model with the effect of education on income (β) and the effect of quarter of birth differences on education (Π) show a substantial 'ridge' in the posterior. For New York data, this ridge is visible at $\Pi = 0$, which dominates the marginal posterior of Π. On the other hand, for the US data, the shapes are nearly elliptical, which reflects that in this case the quarter-of-birth instrument is less weak. The peak around the posterior mode is high compared with the ridge around $\pi = 0$, so that the latter is not visible in the joint posterior density kernel (even though the marginal posterior of π tends to ∞ for $\pi \to 0$). We will show in Section 3 and the Supplementary material A.3.2 that the ridge is integrable but the bimodality is a serious issue for simple inference using only a second moment to measure estimation uncertainty. We refer here also to the Supplementary Material for more empirical examples.
We end this section by summarizing the issue: our motivation for more methodological analysis is that non-elliptical shapes appear in much of the non-experimental empirical econometric analysis. Possible causes of typical shapes need to be studied.
As an important note we emphasize that is it not easy and probably not a good strategy to perform a conjugate analysis when the likelihood is not regular. Since conjugacy would involve some prior irregularity in this context.

Fig 3: Posterior density kernels for simple instrumental variables models for the effects of education on income (β) using the difference in mean education between men born in quarters 2-4 and quarter 1 (π). The model is applied to Angrist and Krueger (1991) data on income and education.



data of New York (29015 observations) data of USA (329509 observations)

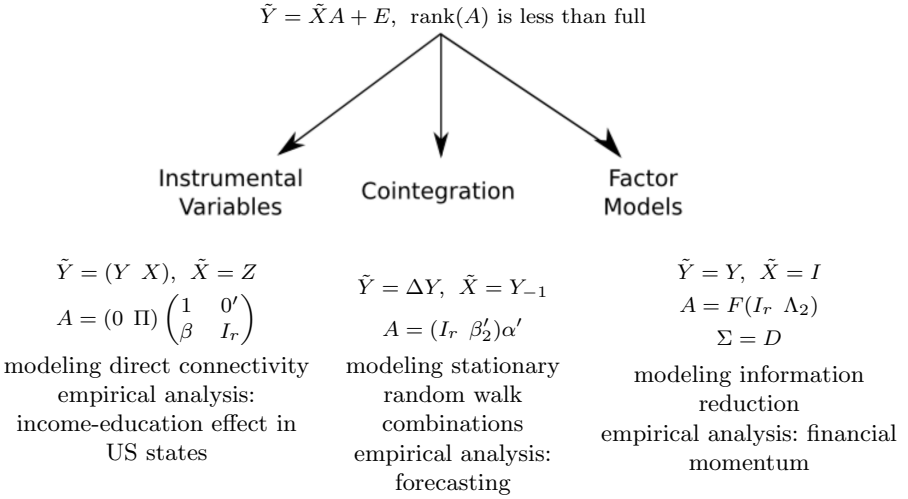
3. Basic model structures, nonstandard likelihood shapes and posterior existence

3.1. Common structure of three reduced rank regression models and summary of posterior existence results

In this section we start to investigate the effect of a reduced rank on the likelihood shape and existence of a posterior within the context of a cointegration model. This model serves as our workhorse model since it can be interpreted as a multivariate regression model where the matrix of equation parameters has reduced rank, see the middle of Figure 4. Using an improper flat prior and linear normalization, it is clear from the cointegrated equation system that a value of $\alpha = 0$ results in a ridge in the parameter space. We will show that this feature leads to an unbounded marginal posterior that is however integrable on a finite region around $\alpha = 0$. We further show that the posterior of α has heavy tails but the density is proper. We note that all conditional distributions are proper with first and higher order moments. We emphasize that the posterior of this cointegration model has the same features as the posterior of a full system Simultaneous Equations Model, an Error in Variables model, and a Static Factor model with no prior information on the factors.

We investigate in the Supplementary Material A.3.3 the effect of imposing a lower triangular structure on the equation system parameters. It is interesting to observe that we can then move from the workhorse model to the so-called Instrumental Variable (IV) regression model, see the left side of Figure 4. Given this triangular structure, we show that the posterior with a flat prior, which

Fig 4: Common structure of three reduced rank econometric models: General structure of reduced rank regression models with linear normalization/identification



leads to a ridge in the posterior surface when the matrix $\Pi = 0$, is a proper density for the case of enough instrumental variables. A large number of instruments makes the tail behavior of the posterior more regular with existence of first and higher order moments. Thus an improper prior yields in this situation a much more regular posterior. The case of many instruments and that of weak endogeneity versus strong endogeneity together with weak and strong identification are all analyzed. We note that there exists an analogy with a triangular cointegraton system, see Martin and Martin (2000).

Thirdly, we explore, also in the Supplementary Material A.3.4, the case where the covariance matrix of the disturbances is diagonal together with the assumption of a standard normal prior on the matrix β. Now, we can move from the workhorse model to a static factor model, see the model on the right of Figure 4. Here the matrix of the unobserved factors F plays the same role as the matrix β in the cointegration model. Similarly the matrix Λ in the factor model has the same role as the matrix α in the cointegration model. When one adds the normal assumption and the one of a diagonal covariance of the disturbances then the posterior with a flat prior is proper. We emphasize that the effect of a diagonal covariance matrix within an IV model yields well behaved student t posterior densities.[1]

There exist several lines of criticisms on our use of flat priors and linear normal-

[1] We note that due to the similarity of three model structures, one can prove the equivalence of the Anderson-Rubin test for overidentification and the Johansen test for cointegration. For details, see Hoogerheide and Van Dijk (2001).

ization. It is well-known that the posterior results using a linear normalization may, in an empirical analysis, be sensitive for the ordering of the variables. In the case of IV this ordering is natural since one is mainly interested in the effect that a possibly endogenous explanatory variable may have on the left hand side endogenous variable (years of education on earned income). But in cointegration and factor models one is often symmetric between variables or factors. Then orthogonal or orthonormal normalization is interesting to explore. We investigate that in Section 4. Second, a uniform prior on parameters is not invariant to a transformation. It is very important that one specifies the prior information on the parameter that reflects the issue of interest. We will also explore this issue more in Section 4 and in the Supplementary material.

3.2. Likelihood shape and existence of posterior in a workhorse reduced rank model: the case of cointegration

A cointegration model constitutes a general class of a reduced rank regression model. Special cases with different restrictions on the parametric structure are covered in the Supplementary Material for the instrumental variable regression model and the static factor model.

3.3. Posterior of a standard cointegration model under linear normalization and a diffuse prior

A Vector AutoRegressive (VAR) model of lag order 1 is usually specified as

$$y_t = \Phi y_{t-1} + \varepsilon_t, \;\; \varepsilon_t \sim NID(0, \Sigma), \;\; \text{for } t = 1, \ldots, T, \tag{3.1}$$

where y_t is $k \times 1$ dimensional vector of observations on economic variables (in deviation from their mean) at time t; Φ is a $k \times k$ matrix of parameters belonging to the observations on the lagged endogenous variables; the disturbances ε_t for $t = 1, \ldots, T$ have independent Gaussian distributions with Σ a positive definite symmetric (PDS) parameter matrix. Observations on y_0 are given as initial values. A basic paper on this VAR model is Sims (1980). For a general introduction to the class of models we refer also to Johansen (1995).
The VAR model equation (3) can be cast into the Vector Error Correction Model (VECM) as follows:

$$\Delta y_t = \Pi' y_{t-1} + \varepsilon_t, \;\; \varepsilon_t \sim NID(0, \Sigma), \;\; \text{for } t = 1, \ldots, T, \tag{3.2}$$

where $\Pi' = \Phi - I_k$. In matrix notation, this error correction model can be specified as:

$$\Delta Y = Y_{-1}\Pi + E, \tag{3.3}$$

where ΔY is a $T \times k$ matrix of observations Δy_1 to Δy_T in its rows and similarly, Y_{-1} is a $T \times k$ matrix of observations containing y_0 to y_{T-1} in its rows. The $T \times k$ random matrix E has a matric-variate distribution, $E \sim MN\,(0, I_T, \Sigma)$.

Stationarity of the process corresponds to Π having full rank. Then all series converge to a finite long run mean and have a bounded variance in the long run. When Π has rank 0, a k-dimensional random walk occurs. The long run mean is equal to the next period mean and long run variance tends to infinity. The more interesting case is where the process $\{y_t\}$ has a so-called *cointegrating* rank r, that is, when Π has rank $r < k$. In this case one has r cointegrating or otherwise stated r stable relations between k economic variables and the matrix Π can be specified as the product of two $k \times r$ matrices α and β with full column rank and $\Pi = \beta\alpha'$.

The resulting model is called a *cointegrating* VECM, which in matrix notation takes the following form:

$$\Delta Y = Y_{-1}\beta\alpha' + E. \tag{3.4}$$

The number of parameters in α and β together may be larger than the number of free parameters in Π under a rank restriction. For the case of k variables and $r \leq k$ cointegrating relations, it holds for any $(r \times r)$ non-singular matrix R that:

$$\Pi = \beta\alpha' = (\beta R)(\alpha R^{-1})',$$

with $\text{rank}(\beta) = \text{rank}(\beta R)$ and $\text{rank}(\alpha) = \text{rank}(\alpha R^{-1})$. That is, the parameters β and α are non-identified. A straightforward way of identifying the parameters is by using a linear normalization on β as restriction:

$$\beta = \begin{pmatrix} I_r \\ \beta_2 \end{pmatrix}, \tag{3.5}$$

where β_2 is a $(k-r) \times r$ matrix, see Kleibergen and Van Dijk (1994); Kleibergen and Paap (2002) among others. We will consider as an alternative in Section 4.2 the case of orthogonal normalization.

Consider a diffuse class of priors defined on the space of (α, β_2) and on the space of positive definite matrices Σ given as $p(\alpha, \beta_2, \Sigma) \propto |\Sigma|^{-h/2}, h > 1$. We make use of the prior value $h = k + 1$, which gives an equivalence between the marginal posterior of (α, β_2) and their, so-called, concentrated likelihood function. We discuss the effect of a more general choice of h later.

The posterior density (apart from the integrating constant) under the normalization is obtained by multiplying the likelihood and the diffuse prior which yields:

$$p\left(\alpha, \beta_2, \Sigma \mid Y\right) \propto |\Sigma|^{-(T+k+1)/2} \exp\left[-\frac{1}{2}\text{tr}\left\{\Sigma^{-1}\left(\Delta Y - Y_{-1}\beta\alpha'\right)'\left(\Delta Y - Y_{-1}\beta\alpha'\right)\right\}\right]. \tag{3.6}$$

We note that for notational convenience, we make use of only the symbol Y to denote the data $(\Delta Y, Y_{-1})$.

In the previous section it is shown empirically that the shape of such a posterior (more precisely the marginal one after integrating out Σ) is such that there exists a ridge in the surface when $\alpha = 0$. We will show analytically that this feature leads to an unbounded marginal posterior that is however integrable and, further, that the tails are heavy but the posterior remains integrable. It

is noteworthy that all conditionals are proper density function with first and higher order moments.

Marginal and conditional posterior densities We consider marginal and conditional posterior density functions of the parameters under a diffuse prior and discuss existence conditions for the posterior distributions and their first and higher order moments. A summary of the derivations and results is presented in Figure 5. For details on the derivation we refer to the online Appendix A.3.2. We note that our results are quite general and several are, to best of our knowledge, novel.

Marginal densities of α and β_2 after integrating out Σ Application of the inverse-Wishart integration step yields the joint posterior distribution of (α, β_2) with density:

$$p\left(\alpha, \beta_2 \mid Y\right) \propto \left|\left(\Delta Y - Y_{-1}\beta\alpha'\right)'\left(\Delta Y - Y_{-1}\beta\alpha'\right)\right|^{-T/2}. \tag{3.7}$$

Exact expressions of the conditional densities which are of the *matrix-t class* are presented in Appendix A.3.2.

Marginal posterior of β_2 and existence of moments From (9), using a matrix-t density step on α and applying a matrix decomposition and properties of the projection matrix, as presented in Appendix A.3.1 and A.3.2, one can obtain the following result:

Proposition *Given the standard form of a cointegration model under linear normalization and using a diffuse class of priors, the marginal posterior of the cointegration parameters β_2 is proportional to a matrix-t density times a polynomial in β_2. This density is proper, independent of the cointegrating rank r, but no first or higher order moments exist.*

It is noteworthy that this result is also independent of the difference $k - r$. We come back in the case of the IV model, presented in the Online Appendix. This result extends the analysis and results of Kleibergen and Van Dijk (1994). We further note that the choice of the prior parameter h does not play a role in the existence condition for the distribution function.

Marginal posterior of α and existence of moments It is shown in Appendix A.3.1 and A.3.2 that using a matrix-t density step on β_2 and applying a matrix decomposition and properties of the projection matrix presented in that appendix, one can obtain the marginal posterior density of α.

Proposition *Given the standard form of a cointegration model under linear normalization and using a diffuse class of priors, the marginal posterior density of the adjustment parameters α is a rational function in α and this density is not proportional to a known form of densities.*

Fig 5: Derivation Scheme for Posterior Densities of a Cointegration model with k variables and $r < k$ cointegrating relations under a diffuse prior.

Model and posterior

$$\Delta Y = Y_{-1}\beta\alpha' + E,\ \ E \sim N(0, \Sigma \otimes I_T)$$

Identification restriction is linear normalization on β

$$\Delta Y = Y_{-1}\begin{pmatrix} I_r \\ \beta_2 \end{pmatrix}\alpha' + E,\ \beta_2 \text{ is } (k-r)\times r,\ \alpha \text{ is } k \times r$$

posterior has ridge at $\alpha = 0$*, but joint density is proper*

Conditional posteriors

$p(\alpha, \beta_2, \Sigma \mid Y)$, data $= \{\Delta Y, Y_{-1}\}$ *is summarized as* Y

↓	↓	↓
complete sum of squares in α	complete sum of squares in β_2	use Inverse-Wishart dist.
↓	↓	↓
$p(\alpha \mid \beta_2, \Sigma, Y) \propto$ matrix Normal density	$p(\beta_2 \mid \alpha, \Sigma, Y) \propto$ matrix Normal density	$p(\Sigma \mid \beta_2, \alpha, Y) \propto$ inverse-Wishart density
↓	↓	↓

Conditional moments of $p(\beta_2 \mid \alpha, \Sigma, data)$*,* $p(\alpha \mid \beta_2, \Sigma, Y)$ *and* $p(\Sigma \mid \beta_2, \alpha, Y)$ *exist for all values of the conditioning parameters in their domains and for all finite* k *and* r*.*

Marginal posteriors of α and β_2

$p(\alpha, \beta_2, \Sigma \mid Y)$

↓

Inverse-Wishart step on Σ

↓

$$p(\alpha, \beta_2 \mid Y) \propto \left|(\Delta Y - Y_{-1}\beta\alpha')'(\Delta Y - Y_{-1}\beta\alpha')\right|^{-T/2}$$

↓	↓
complete sum of squares on α	complete sum of squares on β_2 in three steps
↓	↓
$p(\alpha \mid \beta_2, Y) \propto$ a matrix t density	$p(\beta_2 \mid \alpha, Y) \propto$ a matrix t density
conditional moment exist for all values of β_2 *in its domain*	*conditional moments exist for all values of* α *in its domain*
↓	↓
matrix t-density step on α	matrix t-density step on β_2
↓	↓
use matrix decomposition and properties of the projection matrix:	use matrix decomposition and properties of the projection matrix:
↓	↓
$p(\beta_2 \mid Y)$ *is proportional to a matrix t density times a polynomial in* β_2*. It is a proper density independent of the cointegrating rank r, but no first or higher order moments exist.*	$p(\alpha \mid Y)$ *is a rational polynomial function in* α *and not a member of a known class of densities. It is integrable despite having an asymptote at* $\alpha = 0$*. The tails are heavy but integrable.*

Existence of the marginal posterior of $\alpha|Y$ It is shown in Appendix A.3.2 that a sufficient condition for the existence of the posterior of α at $\alpha = 0_{(k\times r)}$ is:

$$\int \left|\alpha' D^{-1}\alpha\right|^{-(k-r)/2} d\alpha < \infty, \tag{3.8}$$

where D is a matrix which only depends on data.

We next analyze two shape features: the asymptote in the interior when $\alpha = 0_{(k\times r)}$ and the tail behavior when α tends to infinity. We show that the determinant in (100) is integrable around $\alpha = 0$ despite the asymptote at $\alpha = 0_{(k\times r)}$ and we show that the tails are heavy but integrable.

2-dimensional vector case $r = 1, k = 2$ For simplicity, consider the integral on a ball A_k with radius R for the special case, $k = 2$, $r = 1$ where for ease of exposition we assume that the data matrices have been scaled and rotated such that $Y'_{-1}Y_{-1} = I_k$:

$$\int_{A_k} |\alpha'\alpha|^{-(k-r)/2} d\alpha. = \iint_{\alpha_1^2+\alpha_2^2\leq R^2} (\alpha_1^2+\alpha_2^2)^{-1/2} d\alpha_1 d\alpha_2. \tag{3.9}$$

We perform a polar coordinate transformation of α_1, α_2 to show that the above integral is finite but depends on the value of R. Consider the change of variables:

$$\begin{aligned} \alpha_1 &= \lambda\cos\theta, \;\; \alpha_2 = \lambda\sin\theta \\ \lambda^2 &= \alpha_1^2+\alpha_2^2, \;\; \theta = \tan^{-1}(\alpha_2/\alpha_1), \end{aligned}$$

where $\theta \in (0, 2\pi]$, $\lambda > 0$ and the determinant of the Jacobian for this change of variables is

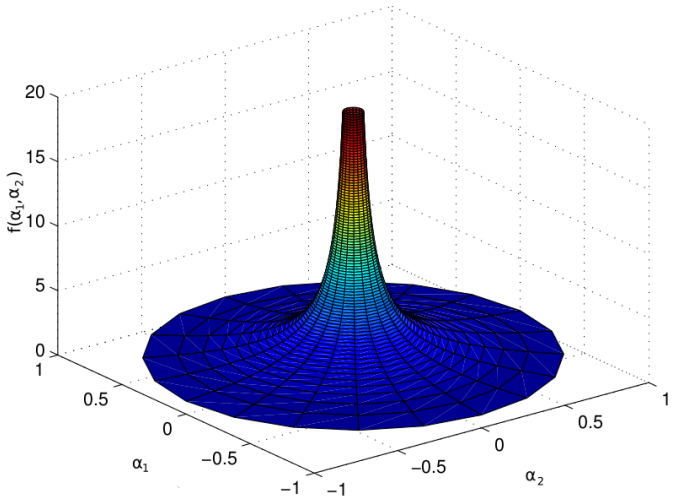
$$|J| = \begin{vmatrix} \cos\theta & -\lambda\sin\theta \\ \sin\theta & \lambda\cos\theta \end{vmatrix} = \lambda(\cos^2\theta + \sin^2\theta) = \lambda. \tag{3.10}$$

With the change of variables, the integral in (11) becomes:

$$\int_{\theta=0}^{2\pi}\int_{\lambda=0}^{R} (\lambda^2)^{-1/2}\lambda d\lambda d\theta = \int_{\theta=0}^{2\pi}\int_{\lambda=0}^{R} 1 d\lambda d\theta = 2\pi R, \tag{3.11}$$

The integral corresponds to the volume under the graph of $f(\alpha) = (\alpha'\alpha)^{-1/2}$. The volume over the region $\{\alpha|\alpha'\alpha \leq 1\}$ can be computed by integrating the surfaces of circles with radius $f(\alpha)$ for $1 \leq f(\alpha) < \infty$ and the surfaces α of circles with radius 1 for $0 \leq f(\alpha) < 1$. Figure 6 illustrates this: for each function value $f(\alpha) = (\alpha'\alpha)^{-1/2}$ with $f(\alpha)$ as the horizontal 'slice' through the graph is a circle with radius $1/f(\alpha)$. For any finite R the integral is bounded from which we conclude that the asymptote poses no problems. A proof that the asymptote poses no problem for the general vector and the matrix case is presented in the online Appendix A.3.2.
If however R tends to ∞ the integral in equation eq:polar2 also goes to ∞ at a

Fig 6: $f(\alpha) = (\alpha'\alpha)^{-1/2}$ for $\alpha'\alpha \leq 1$, where $\alpha = (\alpha_1, \alpha_2)'$.



rate R, so that the *sufficient* condition is not satisfied then. However, the tails are integrable and the marginal posterior of α is proper. The easiest way to see this is as follows. We show in Appendix 3.2 that the marginal posterior of β_2 is proper but it has no first or higher order moments, see equation (A.66). Further, the conditional posterior of α given β_2 is proper for each value of β_2, see (A.39) and (A.58). Therefore, the joint posterior of (α, β_2) is proper. We could simulate α from its (marginal) posterior by simulating β_2 from its marginal posterior and simulating α given the draw of β_2. We emphasize that the line of reasoning to show that the tails are integrable is a general one. That is, it holds for the bivariate case, the general vector case and the matrix case.

All this leads to the following proposition:

Proposition *Given the standard form of a cointegration model under linear normalization and using a diffuse class of priors, the marginal posterior density of α, given in Appendix A.3.2, equation (A.72), is integrable despite the fact that it has an asymptote at $\alpha = 0$. The tails are heavy but integrable, so that the marginal posterior density of α is proper.*

This result also holds for the Simultaneous Equations Model when there exist only a few restrictions on the structure, the Errors-in-Variable model and the Static Factor Model with no information on the factors.

General conclusion of Section 3 In this section we have shown that, using a flat prior, Bayesian analysis of a general reduced rank model yields non-elliptical

shapes of posteriors that can be classified as: flatness and unboundedness due to weak or non-identification and weak or irrelevant instruments. We further showed that unbounded posteriors are locally integrable under weak conditions and posterior tails are heavy but integrable. These results are to the best of our knowledge new. We will show in the Supplementary Material that by making use of extra restrictions such as a lower triangular matrix of β one can obtain proper posteriors with more desired properties (existence of higher order posterior moments). This is shown in the Supplementary material for the Instrumental Variable Model. Alternatively, one may use a weakly informative prior such as a normal prior $N(0, cI)$ with c a large constant on α which makes the tails of the posterior of α more regular. This can be seen in the class of factor models, see for instance Geweke (1996).
We note that, given the structure of our three types of models, multi-modality and skewness (of multiple parameters) are more a computational problem about numerical evaluation. More complex mixture models may give an existence problem due to weak empirical identification of a component of the mixture but this is a topic beyond the scope of this paper. In the next section we investigate how regularization priors deal with the two issues of flat regions (unbounded marginals) and heavy tails.

4. Regularization priors

Since the early nineteen-seventies there has been a strong tradition in Bayesian econometrics of studying the shape and integrability of posteriors of parameters of models with a reduced rank under flat priors. The first class of models studied was the Simultaneous Equations Model (SEM) where the issue of endogeneity of explanatory variables was analyzed. One of the early important papers is Drèze (1976) where a posterior density is presented of the parameters of a single SEM equation, marginalized with respect to all parameters in the remaining part of the SEM where no restrictions were imposed. For a detailed explanation of the shape of the likelihood of the full model and of one single equation we refer to Bauwens and Van Dijk (1990). Next, the so-called Incomplete Simultaneous Equations (INSEM) model, see Zellner et al. (1988), was studied from a Bayesian point of view. This model was shown to be a triangular SEM model and to be identical to an IV model. Bauwens and Van Dijk (1990) present a derivation of the marginal posterior of the single equation parameters but do not discuss in detail under what conditions this is a proper density.
In the present section we present a set of priors that are potentially suitable for making posterior densities proper. First, in Section 4.1 we follow an econometric methodological or statistical approach to specifying weak prior information that is intended to make an unbounded posterior more regular by using the information matrix and an other reference approach. In Section 4.2 we present a new result on a lasso type shrinkage prior combined with orthogonal normalization that serves this purpose well. Furthermore, in Section 4.3 we specify prior information that is meant to make economic models behave 'reasonably'.

A motivation for the latter property was given by Sims (2008) for the case of macroeconomic models. This can be applied more generally to all economic models.
A final point of this section is that in order to obtain robust results for posterior and predictive analysis with weak prior information, it is recommended to use a sequence of priors with increasing amounts of information starting from very weak prior information. Therefore the contents of this section are organized with listing regularization priors in increasing amount of information.

4.1. Information matrix, subspace and reference priors

Information Matrix and Embedding priors: An alternative to using a flat prior on the parameters of a cointegration model (as workhorse model for a reduced rank) is provided by the Information Matrix prior, also known as Jeffreys prior. It is proportional to the square root of the determinant of the information matrix and it can be specified as:

$$p(\Sigma) \propto |\Sigma|^{-(k+1)/2} \tag{4.1}$$

$$\begin{aligned} p(\alpha,\beta_2|\Sigma) \propto & |\mathcal{I}(\alpha,\beta_2|\Sigma)|^{\frac{1}{2}} \\ = & \left| \left(\frac{\partial \operatorname{vec}(\Pi)}{\partial (\operatorname{vec}(\alpha)' \ \operatorname{vec}(\beta_2)')} \right)' \mathcal{I}(\Pi|\Sigma) \left(\frac{\partial \operatorname{vec}(\Pi)}{\partial (\operatorname{vec}(\alpha)' \ \operatorname{vec}(\beta_2)')} \right) \right|^{\frac{1}{2}} \\ = & \left| \left(I_n \otimes \beta \quad \alpha' \otimes \begin{pmatrix} 0 \\ -I_{n-r} \end{pmatrix} \right)' (\Sigma^{-1} \otimes Y_{-1}'Y_{-1}) \right. \\ & \left. \times \left(I_n \otimes \beta \quad \alpha' \otimes \begin{pmatrix} 0 \\ -I_{n-r} \end{pmatrix} \right) \right|^{\frac{1}{2}} \\ \propto & |\beta' Y_{-1}' Y_{-1} \beta|^{\frac{1}{2}(k-r)} |\alpha \Sigma^{-1} \alpha'|^{\frac{1}{2}(k-r)} |\Sigma|^{-\frac{1}{2}(k+1)}, \end{aligned} \tag{4.2}$$

where k is the dimensionality. For a derivation and more details on Jeffreys prior see, Kleibergen and Van Dijk (1994), Uhlig (1994), Kleibergen and Van Dijk (1998), Martin and Martin (2000) and Martin (2001). Both $\mathcal{I}(\alpha,\beta_2|\Sigma)$ and $\mathcal{I}(\Pi|\Sigma)$ denote the conditional information matrices. The distinctive feature of this prior is its ability to annihilate probability mass at points where the identification problem occurs. This result also holds for the instrumental variable model, see the example in Figure 3 in Section 2. To visualize the effects of applying the Information Matrix prior to the likelihood of the cointegration model we present the shape of this prior and the shape of credible sets and the posterior distribution in Figure 7. In the Figures of the prior and posterior density of (α_1,α_2) the activity of Information matrix prior is evident around point $(0,0)$. It is clear from the equations and from the figure that Jeffreys prior relates to strength of information on β (long term equilibrium) and α (speed of adjustment). This prior gives no weight to the state where the model is not identified (where the likelihood exhibits a ridge) and it gives more weight to values of the

parameters α and β when the likelihood also has some weight. More formally, the Information Matrix or Jeffreys prior is a polynomial in these parameters and the prior density kernel tends to infinity when the parameters tend to infinity. Therefore this class of priors is not suitable as regularization prior in the general case of a reduced rank model where the problem is with the tail behavior. However, this class of priors can be used for the case of the Instrumental Variable regression model where the tail behavior of the likelihood is very regular for a large number of instrumental variables, see the analysis in the Online Appendix A.2.3.

We emphasize that there exists an equivalence between the Jeffreys prior and the prior that stems from the **embedding approach**, see, for instance, Kleibergen and Zivot (2003). In the embedding approach one specifies a flat prior on the unrestricted reduced form and makes use of a transformation of random variables to the parameter of the structural form. This approach has been used to specify priors for a simultaneous equations model and a co-integration model, see Kleibergen and Van Dijk (1998) and Kleibergen and Paap (2002). For the embedding approach the same conclusion holds as for Jeffreys prior approach. We present an empirical analysis in the Supplementary Material, Appendix A.1.3. Another interesting analysis is presented for this IV model comparing Bayes and GMM by Sims (2007). We refer to that paper for details.

Subspace/Reference based priors Villani (1998), see also Villani (2000), proposed a prior on the subspace spanned by the columns of the matrix with reduced rank using the concept of a Grassmann manifold. This prior was then transformed to a prior on the parameters α and β in the linear normalization case, treated in Section 3, in order to perform Gibbs sampling. Villani (2005) continued this line of work, now labeled as a reference approach but still based on the subspace approach. It gave proper posteriors that are invariant to the ordering of variables.

Strachan and Inder (2004) and Strachan and Van Dijk (2004) applied the subspace approach to the case of orthonormal normalization. This led to a prior of the parameters β defined on a bounded region. These authors developed a sampling algorithm that allowed to sample from the orthonormal normalization. We refer to the survey by Koop et al. (2006) for a more detailed analysis of the subspace/reference approach.

Conclusion Although the technical approaches listed so far are elegant and 'repair' some or all anomalies of the likelihood function of a reduced rank regression model, we take a different direction in the present paper. The reason being that we intend to work with several states of the econometric model, that is, near the boundary of a reduced rank as well as at the boundary. We want to specify a convenient class of priors that yield proper posteriors which can be used to effectively evaluate posterior and/or predictive probabilities at and near a boundary. Further, we discuss priors that explore implications for posterior and predictive probabilities that may be used for prediction and decision analysis, that is, prior- and posterior-predictive and -decision analysis.

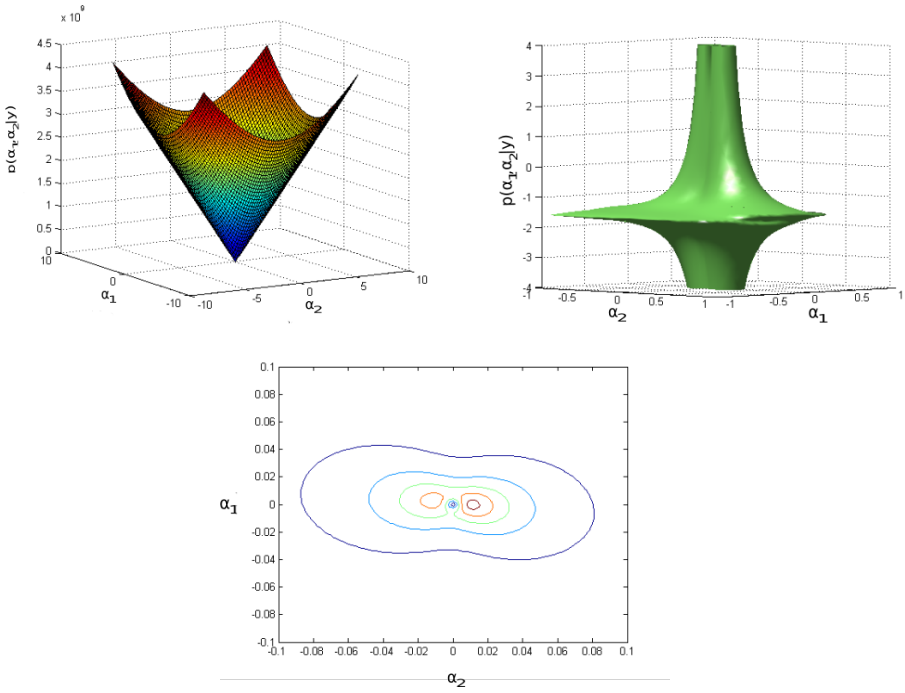


Fig 7: Shape of the Information Matrix or Jeffreys prior, credible sets and posterior distributions under this prior. Data generated from one unit root cointegration model (eigenvalues $\lambda = (0.6074, 1.0)$) with $\alpha = (0.5, -0.0561)'$, $\beta = (-0.6640, 1.0799)'$; $\Pi_1 = \Pi + I = (0.6680,\ 0.5399;\ 0.0373,\ 0.9394)$.

4.2. Orthogonal normalization and lasso type shrinkage prior

Given a diffuse prior and under linear normalization we have shown that the marginal posteriors of the parameters of interest of a workhorse reduced rank regression model are not regular in the sense that they do not belong to a known class of densities like the matrix-t densities. We took the cointegration model as an example. We note that in the case of such a model, when the parameter matrix has everywhere full rank the posterior is regular. That occurs when the data in the cointegration model are all stationary. Also in the case when the rank is zero, that is, when all data series are random walks one encounters regular posteriors. We now explore an approach where weak regularizing prior information is introduced that makes use of restrictions, in particular, plausible restrictions on the range of the parameters. For expository purposes we continue with the cointegration model but emphasize that our results hold also for the instrumental variable and factor model with sometimes slight modifications.

Identification and orthogonal normalization In general an $n \times k$ matrix of rank r has $(n+k)r - r^2$ free elements, that is $(n-r)(k-r)$ restrictions. In our case, the $k \times k$ matrix Π has rank r and therefore it has $2kr - r^2$ independent free elements and $(k-r)^2$ restrictions. The matrices α and β in the parametrization $\Pi = \beta\alpha'$ with $\text{rank}(\Pi) = r$ together have $2kr$ elements, which are r^2 too many to identify α and β. The normalization $\beta_1 = I_r$ that we used in the previous sections exactly accounts for the additional r^2 required restrictions. The parametrization $\Pi = \beta\alpha'$ can be linked to the singular value decomposition $\Pi = USV'$, where the rectangular $k \times r$ matrix U is an element of the Stiefel manifold $U'U = I_r$ and the square $r \times r$ matrix V is an element of the manifold of orthogonal matrices $V'V = I_r$. S is a diagonal $r \times r$ matrix with positive diagonal entries equal to the singular values of Π. We denote the vector of these diagonal elements as $\lambda = (\lambda_1, \ldots, \lambda_r)'$. Note that the manifolds on which U and V are defined have finite volume. The manifold on which λ is defined is not bounded and we shall come back to that later.
E.g. Kleibergen and Van Dijk (1998) and Kleibergen and Paap (2002) explicitly link their parametrization to the singular value decomposition and they combine it with the linear restriction $\beta_1 = I_r$. This linear normalization subsequently implies a mapping from these manifolds to Cartesian coordinates in Euclidean space, that is $\alpha \in \mathbb{R}^{k \times r}$ and $\beta_2 \in \mathbb{R}^{(k-r) \times r}$. This mapping thus transforms from manifolds with finite volume (except λ) to unbounded spaces.
Another common normalization of β used in the literature is $\beta'\beta = I_r$. A major motivation for the choice of this orthogonal normalization of the matrix β is that in this case no preferred ordering of the variables is imposed and the region of integration for β is bounded. In the case of a VAR these may be reasonable assumptions in several situations, in particular, when one considers a set of similar price indices or quantity series.

We emphasize that this normalization alone is not sufficient to identify both α and β. This normalization imposes only $r(r+1)/2$ unique restrictions, because of the symmetry of $\beta'\beta$, so an additional $r(r-1)/2$ restrictions are required. One could impose these on β but this should be done with caution in order to avoid the issue of imposing too much structure through the combination of ordering, restricting and assigning a flat prior. For a more information on normalization and identification, we refer to Hamilton et al. (2007).

Lasso type shrinkage prior under orthogonal normalization We propose an approach that more directly uses the structure of the singular value decomposition and also makes use of the concept of lasso type shrinkage priors, see Tibshirani (1996).

As specified above, the singular value decomposition is not uniquely defined. Any simultaneous permutation of the columns of U, S and V also constitutes a singular value decomposition. A common way to avoid this ambiguity is by ordering the singular values that occur on the diagonal of S as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$. We shall use this ordering. Ordering the singular values is also more straightforward than devising an ordering of the columns of U and V directly (or the columns of α and β for that matter).

Because of this ordering each element λ_{i+1} for $i = 1, \ldots, r-1$ is bounded by λ_i. Only λ_1 remains unbounded towards $+\infty$. Integrability is thus determined by the behaviour of λ_1.

Having fixed the ordering of the singular values the uniqueness of the singular value decomposition when all λ_i's are different is up to simultaneous sign changes of corresponding columns of U and V which could be mitigated for instance by imposing a positive sign for the first non-zero entry in each column of U. Finally, if a singular value occurs more than once, then the columns of U and V corresponding to these singular values are not uniquely defined. Any other orthonormal basis that spans the same space will also do. Although in this particular case the transformation between the matrix Π and its singular value decomposition (U, S, V) is still not invertible everywhere, this is however an event with zero measure and we observe that the Jacobian of this transformation equals 0 whenever a repeated singular value occurs because then the factor $\lambda_i^2 - \lambda_j^2$ will be 0 for some $i < j$.

We analyse the specification in which we combine $\beta'\beta = I_r$ with $\alpha'\alpha = I_r$ in the parametrization $\Pi = \beta\Lambda\alpha'$ with Λ diagonal. This corresponds directly to the singular value decomposition $\Pi = USV'$ with $\beta = U$, $\alpha = V$ and $\Lambda = S = \mathrm{diag}(\lambda)$. The restriction $\alpha'\alpha = I_r$ imposes $r(r+1)/2$ restrictions which amount to r restrictions more than required, but λ subsequently provides these extra r degrees of freedom.

Λ and α in this parametrization combine into α in the usual parametrization $\Pi = \beta\alpha'$ as in the previous bullet.

The advantage of this specification is that now both α and β have finite support. If the issue of non-integrability arises it will be in the parameter λ, and if so it is also clear they will also have to be repaired in λ.

Regarding the econometric interpretation of the parametrization $\Pi = \beta\Lambda\alpha'$ we

may think of $\beta' y_t$ as the deviation from the r cointegrating relations $\beta' y_t = 0$ between the k variables y_t, which is similar to the role of β in the more usual parametrization $\Pi = \beta\alpha'$. The interpretation of λ is that of the rate of adjustment of the system towards each of the r cointegrating relations. α in the parametrization $\Pi = \beta\Lambda\alpha'$ describes the contribution of each of the k variables y_t to the adjustment towards each of these r cointegrating relations. This has advantages over the more usual parametrization $\Pi = \beta\alpha'$ in which the speed of adjustment towards the cointegrating relations is amalgamated with the distribution of these adjustments over the variables into one single parameter matrix (also denoted α).
Each data vector y_t defines a vector in k-dimensional space. The geometric interpretation is that β defines r directions in the space of the data. Λ scales in these directions and α rotates the result to a r dimensional subspace of the data. To distinguish the parameter matrix α in $\Pi = \beta\Lambda\alpha'$ from the parameter matrix α in the usual parametrization we shall denote the latter by α^* such that $\Pi = \beta\alpha^{*\prime}$ in the remainder of this section. In order to translate results on α and $\Pi = \beta\Lambda\alpha'$ back and forth to α^* and $\Pi = \beta\alpha^{*\prime}$ we now briefly describe how they are related. Both parametrizations are linked by the relation $\alpha^* = \alpha\Lambda$. This can be seen when we combine $\beta'\beta = I_r$ with $\alpha^{*\prime}\alpha^* = S$ in the parametrization $\Pi = \beta\alpha^{*\prime}$ where S is a $r \times r$ diagonal matrix with λ_i, $i = 1, \ldots, r$, as diagonal elements. The relation with the singular value decomposition $\Pi = USV'$ is $\beta = U$, $\alpha^* = VS = \alpha\Lambda$. This also gives exactly the number of required restrictions: all off-diagonal elements of $\alpha^{*\prime}\alpha^*$ are constrained to 0 and because of the symmetry of $\alpha^{*\prime}\alpha^*$ each off-diagonal element occurs twice which results in $r(r-1)/2$ unique restrictions. In terms of the columns α_i^* of α^*: $\alpha_i^{*\prime}\alpha_i^* = \lambda_i^2$ for $i = 1, \ldots, r$ and $\alpha_i^{*\prime}\alpha_j^* = 0$ for $i \neq j$.

Prior choice and existence of posterior moments In Appendix A.3.2 we present a derivation where given that diffuse priors are specified for α and β on their respective Stiefel manifolds and a usual diffuse prior on Σ one can derive proper posteriors and existence of first and higher order moments.
For convenience we present here the reasoning, which proceeds as follows. Using the parametrization $\Pi = \beta\Lambda\alpha'$ and the normalizing restrictions $\alpha'\alpha = I_r$, $\beta'\beta = I_r$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$ all parameters except λ_1 are defined on bounded sets (conditionally upon the (finite) value of λ_1). A natural choice for an uninformative prior is the uniform prior over these sets. Only λ_1 is defined on an infinite interval. A natural choice for λ_1 that is consistent with the uniform prior on the simplex for $\lambda_2, \ldots, \lambda_r | \lambda_1$ is the exponential distribution. Another way to look at this, is that although $\lambda \in [0, \infty)$ has infinite support, it can also be transformed to the unit interval on which a uniform prior can be specified. By doing so, all model parameters (except the covariance matrix Σ) are bounded to finite areas. Specifically, when either the transformation $\lambda^\flat = \exp(-\lambda) \in (0, 1]$ or $\lambda^\sharp = 1 - \exp(-\lambda) \in [0, 1)$ is used and a standard uniform density is specified on $\lambda^\flat$ or $\lambda^\sharp$ then λ also has a standard exponential distribution. Using a similar argument the rate parameter θ could be included by specifying a uniform prior on e.g. $\exp(-\theta\lambda)$. A note refers to the rate θ of the exponential distribution. By

choosing θ to a value close to 0, the exponential distribution tends towards a flat distribution over the positive real numbers.
We can summarize the results from this section as follows.

Proposition *Given the standard form of a cointegration model and using a lasso type shrinkage prior under orthogonal normalization on the parameters of the matrix with reduced rank, the marginal posteriors of these parameters are proper with finite first and higher order moments.*

We emphasize that the cointegration model serves as an example of a general reduced rank model but our result holds generally for this class of models. That is, one may also apply it to the instrumental variable model and the factor model when in these latter models one does not want to impose specific restrictions like triangularity and/or diagonality.

4.3. Short survey of other regularization priors

Inequality conditions where data and economic information matters: As explained in the previous subsection area restrictions play a useful role in formulating prior information. Baumeister and Hamilton (2015) have carried this issue further. These authors explore the effect of sign restrictions, coming from broad economic considerations, on vector autoregressive models under different identification conditions. They also explore the effect of weak prior information on implied impulse response functions. Apart from restrictions based on economic relationships and characteristics, there exist data based inequality conditions that can also be relevant as prior information. A simple example of this is the restriction that autoregressive parameters in a dynamic model should not be taken outside the unit interval since explosive time series are highly unlikely for the long run because the occurrence of a regime change is then very likely. An analogous point can be made for values of the autoregressive parameters close to zero. From stylized facts of macroeconomic and financial time series it is well-known that the relevant range of the autoregressive parameters is a subinterval of the unit interval close to the unit root. For more details of the locally uniform prior where the data play a role, we refer to the next section and to Schotman and Van Dijk (1991b).

Dummy observations and training sample priors: One popular way to make use of weak data-based prior information is to split the data into two parts: a training set and a 'hold-out' set of data. In the first part the weak prior is transformed to an informative posterior which serves as a prior for the second part of the data and this leads to model validation and forecasting. For an illustrative example we refer to the next section and for background to, e.g. Berger et al. (2004). Another approach is to construct a so-called imaginary sample by introducing a set of dummy observations. It yields a pragmatic class of priors, proposed by Sims (2004, 2005). This approach can be combined with a more informative prior approach, see below.

Dynamic patterns for parameters: Given the dynamic nature of many models in economics, it is very natural to allow not only the variables but also the parameters of such models to move through time. However, simply adding a subscript of time to an equation system parameter yields an intractable likelihood since a T-dimensional integral is added to the Bayesian inferential problem. The well-known Normal or Kalman Filter imposes in such a case a structure on the dynamic parameters and it forms a pattern which yields a tractable likelihood and posterior. The literature on this topic is abundant and we refer only to two basic textbooks for more background: Pole et al. (1994) and Harvey (1990). A related approach is the Minnesota prior, see Doan et al. (1992), which may be characterized as a smoothness priors. This class of priors is meant to improve forecasting properties by making use of stylized facts of macroeconomic time series.
One may also explore the predictive implications of a prior. For instance, does a weak prior on the equation system parameters give plausible prior values of multipliers, impulse response function and/or periods of oscillations from an implied business cycle. For an early reference we refer to Van Dijk and Kloek (1980). The literature on this *prior-predictive approach* is substantial and a more detailed analysis is outside the scope of the present paper.
We also mention an approach where the priors are anchored to some long run plausible values. A basic approach was taken by Schotman and Van Dijk (1991a,b) for the unit root case. It was extended by Villani (2009) to refer to long run plausible values and recently again extended to be combined with a dummy variable prior by Giannone et al. (2015). A similar idea is to connect the prior to a plausible posterior-predictive analysis, see Gelman et al. (1996) and Baştürk et al. (2014b).

Economic structural information: We end this brief survey by mentioning the approach to add economic structural information like so-called DSGE priors due to Del Negro and Schorfheide (2004), while Strachan and Van Dijk (2013) combine economic information and technical econometric information.
We conclude that there are many useful approaches to explore the sensitivity of the posterior and predictive results with respect to a sequence of weak priors where the amount of prior information is gradually increasing. This will be illustrated in the next section. Finally, we note that the issue of sensitivity of weak priors and also prior choice is very much studied in the Bayesian literature, see for instance Tuyl et al. (2008).

5. Model probabilities under regularization priors and possibly irregular likelihoods

This section forms a bridge between the more theoretical analysis of the shape of posterior densities for the reduced rank regression model with possibly irregular likelihoods and the empirical analysis of a micro-econometric problem on the education-income effect where we make use of mixtures of models. An

important concern is how to give probabilistic weights to evidence that is near and at the boundary of the parameter region of reduced rank models using Bayesian methods. We show that, although the issue of weak identification is not an impediment for showing posterior existence of distributions, very weak prior information does play a major role for the evaluation of posterior and predictive probabilities of evidence near and at a boundary of identification and relevance of instruments. We illustrate that the Bartlett/Jeffreys/Lindley paradox is not only a mathematical or statistical result but it shows up as a problem when flat prior density kernels are assumed over regions where there is little empirical evidence like near a boundary with weak instruments. This issue was pointed out by Hoogerheide and Van Dijk (2013). Here a training sample and weak economic information on area restrictions is recommended together with a sensitivity analysis in order to obtain more robustness in the results. We present two examples. One refers to a basic time series model where the likelihood is regular but the prior interval contains many many irrelevant parameter values. In order to save space, the issue of model evaluation without and with regularization priors is discussed for this class of models in the Supplementary material in Online Appendix, section A.4. The second example studies the effect that an irregular likelihood due to a lack of identification and the presence of weak instruments has on model probability evaluation within the context of an IV model. These results are reported in Section 5.1. Armed with these results, we continue in the next section with an empirical analysis using a mixture of models with mixing probabilities coming from evidence near and at a boundary.

5.1. IV Model probabilities under alternative identification and endogeneity structures using training sample priors

In this subsection we apply the predictive likelihood approach, see Gelfand and Dey (1994) and Eklund and Karlsson (2007), to simulated data from the IV model. Our purpose is show that, although the posterior densities in an IV model with diffuse type priors and weak instruments/identification are very non-regular and require special simulation based procedures to evaluate their shape, it is relatively easy to evaluate posterior/predictive probabilities near and at the boundary using reasonable area restrictions and training sample priors. In the next section a mixture of posteriors under endogeneity and exogeneity for the IV model is estimated using US data.

In the present subsection we investigate the robustness of the results on estimating predictive probabilities for the case of no endogeneity for different levels of endogeneity, different levels of empirical identification and different lengths of training samples, where the total number of observations is 1000 for each simulated dataset. We will use the basic structural IV model from Section 3, see also Appendix A.3.3. For simplicity and for computational convenience we take the case of one endogenous variable and one instrument, where $\beta = 0$ and $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ and the parameter ρ indicates the degree of endogeneity with $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$. We restrict the parameters to a plausible finite region.

Table 1: Simulation experiments for the IV model: $Pr(\rho = 0 \mid \text{data})$ for different levels of endogeneity (ρ), different instrument strength (π) and different prior data percentages (m). Standard deviations and numerical standard errors of $Pr(\rho = 0 \mid \text{data})$ based on 10 sets of simulated data are reported in parentheses and square brackets, respectively.

	percentage m of observations that are used as training sample											
	$m = 50\%$				$m = 10\%$				$m = 0.5\%$			
	level of endogeneity				level of endogeneity				level of endogeneity			
identification level /	strong	medium	weak	no	strong	medium	weak	no	strong	medium	weak	no
instrument strength	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.1$	$\rho = 0$	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.1$	$\rho = 0$	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.1$	$\rho = 0$
strong π =1	0.00	0.00	0.27	0.57	0.00	0.00	0.32	0.75	0.00	0.00	0.66	0.97
	(0.00)	(0.00)	(0.18)	(0.04)	(0.00)	(0.00)	(0.26)	(0.03)	(0.00)	(0.00)	(0.31)	(0.00)
	[0.00]	[0.00]	[0.06]	[0.01]	[0.00]	[0.00]	[0.08]	[0.01]	[0.00]	[0.00]	[0.10]	[0.00]
medium π =0.5	0.00	0.00	0.44	0.57	0.00	0.00	0.53	0.73	0.00	0.00	0.89	0.97
	(0.00)	(0.00)	(0.14)	(0.04)	(0.00)	(0.00)	(0.23)	(0.04)	(0.00)	(0.00)	(0.09)	(0.01)
	[0.00]	[0.00]	[0.04]	[0.01]	[0.00]	[0.00]	[0.07]	[0.01]	[0.00]	[0.00]	[0.03]	[0.00]
weak π =0.1	0.00	0.41	0.53	0.54	0.00	0.39	0.71	0.73	0.00	0.58	0.88	0.88
	(0.00)	(0.18)	(0.05)	(0.04)	(0.00)	(0.20)	(0.16)	(0.11)	(0.00)	(0.27)	(0.05)	(0.05)
	[0.00]	[0.06]	[0.02]	[0.01]	[0.00]	[0.06]	[0.05]	[0.03]	[0.00]	[0.09]	[0.01]	[0.02]
irrelevant π =0	0.50	0.48	0.49	0.48	0.49	0.49	0.52	0.53	0.72	0.70	0.70	0.70
	(0.05)	(0.05)	(0.06)	(0.05)	(0.09)	(0.14)	(0.15)	(0.13)	(0.09)	(0.11)	(0.14)	(0.14)
	[0.02]	[0.02]	[0.02]	[0.02]	[0.03]	[0.04]	-0.14	[0.04]	[0.03]	[0.03]	[0.04]	[0.04]

Left panel of Table 1: For the cases of strong and medium instruments/identification and strong and medium level of endogeneity the posterior probability $Pr(\rho = 0 \mid \text{data})$ is correctly chosen as zero, given the 50% training sample. That is, let $y^\star$ be the training sample and $\tilde{y}$ be the validation sample, then $Pr(\rho = 0 \mid y^\star, \tilde{y})$ is much smaller than $Pr(\rho = 0 \mid y^\star)$, since the data $\tilde{y}$ contain much evidence about ρ being not equal to zero. For the bottom row, it holds that $\pi = 0$ implies that β, Σ, ρ are not identified. That is, the data contain no information on ρ and thus the posterior probability $Pr(\rho = 0 \mid \text{data})$ is equal to the prior probability $Pr(\rho = 0) = 50\%$.
For the right hand column one would expect that $Pr(\rho = 0 \mid \text{data}) = 1$. However, the situation is as follows. Given $y^\star$ and $\tilde{y}$, $Pr(\rho = 0 \mid y^\star)$ using 50% of data is already rather precisely located around $\rho = 0$, with a standard deviation only about $\sqrt{2}\times$ larger than for $Pr(\rho = 0 \mid y^\star, \tilde{y})$. This implies $Pr(\rho = 0 \mid y^\star, \tilde{y}) = \sqrt{2} Pr(\rho = 0 \mid y^\star)$ which leads to $Pr(\rho = 0 \mid \text{data}) = \sqrt{2}/(1 + \sqrt{2}) \approx 0.586$. [2]

Middle panel of Table 1: The results in the upper left corner are as expected: $\approx 0\%$. Similarly, the results in the bottom row are: $\approx 50\%$. The results in the right column follow from $Pr(\rho = 0 \mid \text{data}) \approx \sqrt{1/m}/(1 + \sqrt{1/m}) = \sqrt{1/10}/(1 + \sqrt{1/10}) \approx 0.760$.

Right panel of Table 1: Again the results in the upper left corner are as expected: $\approx 0\%$. Next, the advantage of very small training sample m is shown at the top of the right column: $Pr(\rho = 0 \mid \text{data})$ is close to 1, which is the true value given that $\rho = 0$. The disadvantage of a very small m is recognized as a case of Bartlett/Jeffreys/Lindley paradox. That is, the false null hypothesis $\rho = 0$ is wrongly favored in the bottom row and in the third column of results. The reason is that $Pr(\rho = 0 \mid y^\star)$ after only 5 of 1000 observations is still very diffuse. That is, more diffuse than $Pr(\rho = 0 \mid y^\star, \tilde{y})$ after 1000 observations.
The conclusions of Table 1 may be summarized as follows.

In the interior of the parameter region. For the cases of strong and medium instruments and strong and medium level of endogeneity the posterior probability $Pr(\rho = 0 \mid \text{data})$ is correctly chosen as zero for several values of the length of the training sample.

At the boundaries of the parameter region. For the bottom row, which refers to the case of no identification/irrelevant instruments, the estimated posterior probability $Pr(\rho = 0 \mid \text{data})$ is sensitive for the length of the training sample. A training sample of less than 10 percent should not be selected. For the right hand columns in all three panels, which refers to the case of no endogeneity, the estimated posterior probability $Pr(\rho = 0 \mid \text{data})$ is also very sensitive to the length of the training sample. A small training sample and a large validation sample are to be recommended in this case.

[2] Given a training sample fraction equal to m and given a normal distribution with mean $\rho = 0$, stdev = const / $\sqrt{\#data}$, one has $Pr(\rho = 0 \mid y^\star, \tilde{y}) = \sqrt{1/m} \times Pr(\rho = 0 \mid y^\star)$ and $Pr(\rho = 0 \mid \text{data}) = \sqrt{1/m}/(1 + \sqrt{1/m}) = 1/(1 + \sqrt{1/m})$.

Near the boundaries of the parameter region. This refers to the third column and third row in each of the three panels. Here there also exists a trade-off between the case of weak instruments/identification and the case of weak and no endogeneity. In the case of weak instruments/identification one would prefer a large training sample to get informative priors while for the case of no endogeneity a small training sample so that most of the data can be used for validation.
It is clear evidence from the results of the Table that the choice of a 'prior data' percentage m is important. The advantage of the predictive likelihood approach is that m is a scalar. This may be easier to choose than specifying an entire not 'too non-informative or not too informative' prior density. The problem of predictive likelihood remains: How to choose this scalar m? A practical sensitivity analysis is: simply show results for multiple values of m and find the interval of m values where results are 'similar'.

General conclusion of Section 5. The evaluation of predictive model probabilities under weak prior information and near a boundary of the parameter region gives correct results which are relatively robust under the condition of choosing the right training sample. A sensitivity analysis is recommended for the length of the training sample. In extreme cases very near and at the boundary with weak identification one should be very careful with strong conclusions. More informative priors are then to be recommended.

6. Bayesian mixtures to analyze the effect of length-of-education on earned income in US states

In this section, we present and apply a predictive likelihood approach for model comparison or model combination to the Angrist and Krueger (1991) data on income and education, which are also analyzed in Hoogerheide and Van Dijk (2008). Angrist and Krueger (1991) data consist of men born in the US during the periods 1920-1929, 1930-1939 and 1940-1949, where the data for the first group are collected in 1970, and the data for the last two groups are collected in 1980.[3] We use a subset of their data, consisting of men born during the period 1930-1939, including the data on weekly wages, number of completed years of education and instruments consisting of quarter of birth dummies. The data include 51 states and 329.509 observations.[4] The IV model applied to data from

[3] For an introduction to a Bayesian analysis of an IV model using real and simulated data, we refer to the Supplementary Material in the Online Appendix A.1.

[4] The source of the data is the 1980 Census, 5 percent public sample, also available from `econ-www.mit.edu/faculty/angrist/data1/data/angkru1991`. We refer to the online appendix for a summary of these data.

each state is[5]:

$$\tilde{y}_i = \alpha_1 + \tilde{x}_i\beta + \sum_{t=1}^{9} D_{t,i}\delta_t + \tilde{\epsilon}_i \tag{6.1}$$

$$\tilde{x}_i = \alpha_2 + \sum_{q=2}^{4} D_{q,i}\Pi_q + \sum_{t=1}^{9} D_{t,i}\delta_t + \tilde{\nu}_i \tag{6.2}$$

where $\tilde{y}_i$ and $\tilde{x}_i$ are the natural logarithm of the weekly wage and the number of completed years of education for the person i in 1979, respectively.
In (16) and (17), $D_{t,i}$ for $t = \{1, \ldots, 9\}$ are the dummy variables for year of birth which take the value 1 if individual i was born in year $1929+t$, and 0 otherwise. $D_{q,i}$ for $q = \{2, 3, 4\}$ are the quarter of birth dummy variables which take the value 1 if individual i was born in quarter q, and 0 otherwise. α_1 and α_2 are constants, and $\tilde{\epsilon}_i$ and $\tilde{\nu}_i$ are disturbances assumed to be normally distributed, and independent across individuals.
The model in (16) and (17) is similar to the model of Hoogerheide and Van Dijk (2006). For simplicity, we do not consider interactions of year dummies and quarter of birth dummies as instruments. Furthermore, the model employed here does not include state dummies, as each state is analyzed separately. We simplify the IV model in (16) and (17) correcting for the constant terms and exogenous year of birth dummies. Using this simplification, the IV model becomes:

$$y_i = x_i\beta + \epsilon_i, \tag{6.3}$$

$$x_i = Z_i\Pi + \nu_i, \tag{6.4}$$

where y_i, x_i are the residuals from regressing the log weekly wage and years of education on a constant and year of birth dummies, respectively. Z_i is the 3×1 vector of instruments, obtained from regressing quarter of birth dummies on a constant and the year of birth dummies. ϵ_i and ν_i are the error terms that have a joint normal distribution, and are independent across individuals.

6.1. Bayesian model mixtures using predictive model probabilities

In order to calculate the predictive model probabilities, we define two models M_0 and M_1, where M_0 is a nested model compared to M_1. In the IV model example, M_1 corresponds to the IV model while the nested model M_0 corresponds to M_1 with a parameter restriction: $\rho = 0$. The posterior odds ratio K_{01} for comparing M_0 with model M_1 is the product of the Bayes factor and the prior odds ratio:

$$K_{01} = \frac{p\left(Y \mid M_0\right)}{p\left(Y \mid M_1\right)} \times \frac{p\left(M_0\right)}{p\left(M_1\right)}, \tag{6.5}$$

[5] In order to keep the notation simple, we do not define an index for each state, but note that the described IV model is applied to each US state separately.

where Y is the observed data, and the prior model probabilities $(p(M_1), p(M_0)) \in (0,1) \times (0,1)$ and $p(M_1) + p(M_0) = 1$.
It is often difficult to compute K_{01} since the marginal likelihoods are given by the following integrals: $p(Y \mid M_0) = \int_{\theta_{-\rho}} \ell(\rho = 0, \theta_{-\rho})\, p_0(\theta_{-\rho})\, d(\theta_{-\rho})$ and $p(Y \mid M_1) = \int_{\theta_{-\rho},\rho} \ell(\rho, \theta_{-\rho})\, p(\rho, \theta_{-\rho})\, d(\rho)\, d(\theta_{-\rho})$, where $\theta_{-\rho}$ are the model parameters apart from ρ. We therefore calculate model probabilities using the Savage-Dickey Density Ratio (SDDR). Dickey (1971) shows that the Bayes factor can be calculated using a single model if the alternative models are nested and the prior densities satisfy the condition that the prior for $\theta_{-\rho}$ in the restricted model M_0 equals the conditional prior for $\theta_{-\rho}$ given $\rho = 0$ in the model M_1, i.e. $p_1(\theta_{-\rho} \mid \rho = 0) = p_0(\theta_{-\rho})$[6]. In this case, (20) becomes:

$$K_{01} = \frac{p(\rho = 0 \mid Y, M_1)}{p(\rho = 0 \mid M_1)} \times \frac{p(M_0)}{p(M_1)}, \tag{6.6}$$

where $p(\rho \mid Y) = \int p(\rho, \theta_{-\rho} \mid Y) d\theta_{-\rho}$ and $p(\rho) = \int p(\rho, \theta_{-\rho}) d\theta_{-\rho}$[7]. We perform the model averaging scheme using the model probabilities in Section 5. Specifically, given the posterior odds ratio, it is possible to weight the evidence of alternative models using Bayesian Model Averaging (BMA). We consider the effect of model uncertainty on the estimation of the parameter β, as this parameter is the main focus in most cases. The information about β is summarized by the following posterior:

$$p(\beta \mid Y) = p(\beta \mid Y, M_0)\, p(M_0 \mid Y) + p(\beta \mid Y, M_1)\, p(M_1 \mid Y). \tag{6.7}$$

Furthermore, functions of parameters, i.e. $g(\beta)$ in the IV model are estimated by:

$$E[g(\beta \mid Y)] = E[g(\beta \mid Y, M_0)] p(M_0 \mid Y) + E[g(\beta \mid Y, M_1)] p(M_1 \mid Y). \tag{6.8}$$

Hence both models under consideration should be estimated, and the inference on parameters is simply the weighted average of the results in both models. The weights in averaging the results are the posterior model probabilities.

6.2. Empirical results

The degree of instrument strength (indicated by posterior densities of Π_2, Π_3 and Π_4) differs substantially across states, as reported in Hoogerheide and Van Dijk (2006). A second source of heterogeneity across states is the degree of endogeneity (indicated by posterior ρ). For some states, such as Maine, Minnesota and Texas, 95% intervals for posterior ρ densities do not include point 0,

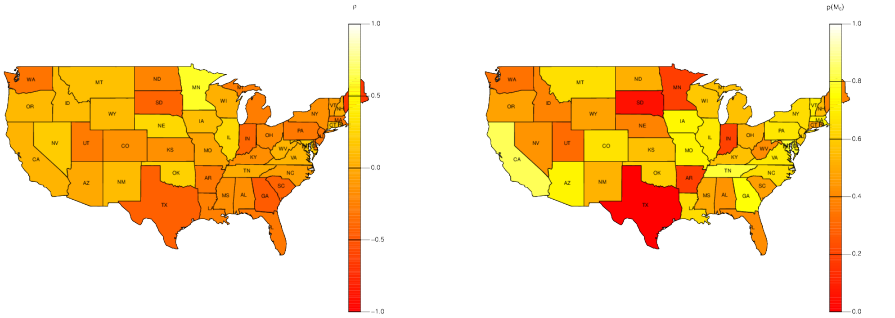
[6] Notice that the condition for SDDR holds if we define the prior for $\theta_{-\rho}$ in the restricted model equal to the conditional prior of $\theta_{-\rho}$ given $\rho = 0$ in the unrestricted model.

[7] As a generalization, Verdinelli and Wasserman (1995) show that K_{01} is equal to the Savage-Dickey density ratio in (21) times a correction factor when the prior condition fails.

while for the rest of the states 95% posterior intervals of ρ include the value 0. Besides the finding of heterogeneity across states, we conclude that the use of instruments may not be necessary for most states. For further details of these estimation results we refer to the online appendix.

Posterior means for the degree of endogeneity of ρ in the IV model, and the predictive model probabilities for model M_0, corresponding to the model with $\rho = 0$ in which the instruments are not used for the estimation of β, are given in Figure 8. For the predictive model probabilities, the training sample consists of

Fig 8: Degree of endogeneity in the US states and predictive model probabilities for model M_0. M_0 denotes the model with $\rho = 0$ in which no instruments are used for the estimation of β.



Posterior mean ρ for US states

Model probabilities for the model M_0 ($\rho = 0$) for US states

a randomly chosen subset of 10% of the observations, prior model probabilities in (21) are chosen to be equal. Furthermore, the effect of the training sample choice is partially eliminated by averaging predictive model probabilities from 20 different random training samples.[8]

Model probabilities are quite close to 0.5 and do not show a clear preference for either model, except for a few states such as Texas and Tennessee. For Texas, model probabilities indicate that the IV model is necessary. For Tennessee on the other hand, we find strong evidence against the need for the IV model. We conclude that choosing one of the alternative models according to these probabilities can be quite inaccurate, and employ model averaging to infer the state-specific effects of income on education.

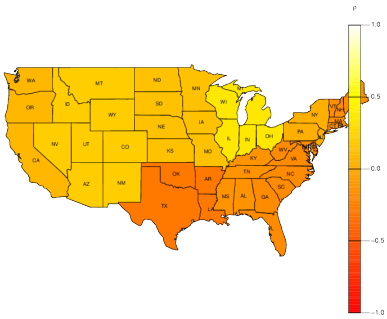
We next present how *average* effects of education on income can be inferred using the model probabilities. The *average* estimated effects of education on

[8]The results with 5% and 25% training sample sizes and a single training sample were similar, except for some states with very small number of observations, such as South Dakota.

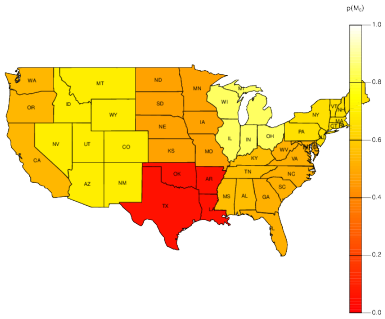
income for the US states, i.e. the posterior distributions resulting from BMA, are summarized in Table 2. Model probabilities are achieved by using training sample with 10% of the observations, averaged over 20 repetitions. The main advantage of model averaging is the improved efficiency of the estimates. Standard deviations of posterior β draws are less than half of those achieved by the IV model only.

Regional patterns in income-education relationship - analysis of US divisions: We further analyze the income-education relationship in US divisions. We apply the IV model in (18) and (19) to 9 divisions for the Angrist and Krueger (1991) data according to the Census Bureau designated areas. The purpose of this analysis is to compare the results in terms of instrument strength with those of Hoogerheide and Van Dijk (2006), who show that quarter of birth dummies are strong instruments mainly in southern states. Furthermore, we document the effect of averaging the data within divisions or regions.
Table 2 reports posterior results of the IV model for US divisions. Similar to the state-specific results, census regions show heterogeneity both in terms of instrument strength and the degree of endogeneity. Posterior results for education effects on income are quite different across divisions. Especially for the West North Central division, the posterior standard deviation is quite high, indicating the relatively weak instruments in this division. Figure 9 presents posterior mean ρ and model probabilities for M_0, the model with $\rho = 0$ in which no instruments are used for the estimation of β. The training sample consists of 25%

Fig 9: Degree of endogeneity in the US divisions and predictive model probabilities for model M_0. M_0 denotes the model with $\rho = 0$ in which no instruments are used for the estimation of β.



Posterior mean ρ for US divisions



Model probabilities for the model M_0 ($\rho = 0$) for US divisions

TABLE 2
Income-education effects in US states

Average effects of education on income								
State	Mean	Std. Dev.	State	Mean	Std. Dev.	State	Mean	Std. Dev.
AL	0.11	0.03	LA	0.09	0.04	OH	0.08	0.04
AZ	0.11	0.03	ME	0.09	0.03	OK	0.04	0.03
AR	0.07	0.02	MD	0.05	0.03	OR	0.05	0.10
CA	0.05	0.01	MA	0.21	0.09	PA	0.11	0.03
CO	0.07	0.02	MI	0.08	0.03	RI	0.07	0.03
CT	0.06	0.04	MN	-0.06	0.08	SC	0.12	0.03
DE	0.02	0.05	MO	0.07	0.03	SD	0.16	0.07
DC	0.10	0.04	MS	0.09	0.04	TN	0.07	0.01
FL	0.13	0.05	MT	0.04	0.04	TX	0.16	0.06
GA	0.12	0.02	NC	0.08	0.02	UT	0.09	0.07
HI	0.08	0.04	NC	0.09	0.04	VT	0.06	0.03
ID	0.05	0.06	NE	0.03	0.09	VA	0.08	0.04
IL	0.05	0.08	NH	0.09	0.04	WA	0.10	0.09
IN	0.04	0.03	NJ	0.09	0.03	WV	0.06	0.03
IA	0.15	0.12	NM	0.05	0.05	WI	0.05	0.03
KS	0.08	0.03	NV	0.03	0.06	WY	0.04	0.06
KY	0.07	0.01	NY	0.08	0.03			

Parameter estimates					
	β	Π_2	Π_3	Π_4	ρ
New England Division	0.11	0.09	0.17	0.21	-0.16
	(0.05)	(0.04)	(0.04)	(0.04)	(0.23)
Middle Atlantic Division	0.07	0.07	0.03	0.03	0.03
	(0.07)	(0.02)	(0.02)	(0.03)	(0.31)
East North Central Division	-0.03	0.07	0.02	0.08	0.36
	(0.08)	(0.02)	(0.02)	(0.02)	(0.25)
West North Central Division	0.02	-0.06	0.01	0.02	0.15
	(0.13)	(0.04)	(0.04)	(0.04)	(0.40)
South Atlantic Division	0.11	-0.01	0.14	0.22	-0.18
	(0.03)	(0.03)	(0.03)	(0.03)	(0.16)
East South Central Division	0.09	0.03	0.27	0.41	-0.13
	(0.02)	(0.04)	(0.04)	(0.04)	(0.12)
West South Central Division	0.12	-0.04	0.20	0.30	-0.29
	(0.02)	(0.04)	(0.04)	(0.03)	(0.11)
Mountain Division	0.01	0.20	0.14	0.18	0.21
	(0.08)	(0.05)	(0.06)	(0.06)	(0.30)
Pacific Division	0.04	0.23	0.21	0.11	0.08
	(0.05)	(0.04)	(0.04)	(0.04)	(0.23)

Note: The top panel reports means and standard deviations of effect of education on income, resulting from BMA, for the US states. The bottom panel reports posterior means for 9 US divisions. All results are based on 30000 draws (3000 burn-in). Estimated posterior standard errors are reported in parentheses.

of the observations[9]. Predictive model probability for the nested model without

[9] We experimented with the model using smaller training samples, and the results are quite insensitive to the training sample size.

instruments is far from 0.5 only for two regions: East North Central Division, and West South Central Division. In East North Central division, the model without instruments is favored by model probabilities. Notice that the states within this region are quite heterogeneous in terms of predictive model probabilities reported in Figure 8. The IV model is clearly necessary only for two states in this region, namely Arkansas (AR) and Texas (TX). Hence 'average' income-education relationship within this region is determined mainly by these two states. This problem is also seen in East North Central division. According to posterior model probabilities in Figure 8, this region consists of states where an IV model is clearly preferred, such as Minnesota (MN) and South Dakota (SD), and also states where the IV model is not necessary, such as Iowa (IA). Hence the 'average' model probability for this region reported in Figure 9 is misleading.

For the US data on the income-education relationship, we conclude that there is substantial heterogeneity in the effect of the length of years of education on earned income. We document that differences between states are characterized by different instrument strengths, as reported by Hoogerheide and Van Dijk (2006). Our results also show that the degree of endogeneity is different across states and regions.

Using this data set we have shown different, and mostly weak power of quarter of birth in explaining education. This finding, in combination with the not so severe problem of endogeneity makes it hard to assess whether the IV model should be preferred over a simpler and more parsimonious linear regression model without instruments. Hence we conclude that averaging over these alternative models is a reasonable way to deal with model uncertainty.

General conclusion of section 6 We have shown that the effect of length of education on earned income differs considerably among almost all US states. This may have important policy implication of determining the length of required schooling. This issue should be investigated in more detail.

7. Conclusions and perspectives

We have sketched in this paper an approach using Bayesian mixtures to average over those states of an econometric model which are known as near a boundary and at a boundary with the purpose to obtain more precise structural inference, accurate forecasting and effective policy analysis. In order to do this several results have been established. There exists a common structure in three well-known econometric models where the matrix of equation system parameters has reduced rank. The case of a reduced rank can be interpreted as a boundary in the parameter space. The econometric models are the cointegration, instrumental variables and factor model. Using a flat prior, the effect that the reduced rank has on the shape of the likelihood/posterior has been studied for a general workhorse model, that is equal to a cointegration model. Marginal posterior densities of equation parameters are of the student-t type times a polynomial or rational function. Their shapes may contain ridges due to weak identification,

be bimodal and have very fat tails. These posteriors can get nicer properties (such as finite higher order moments) when extra restrictions are imposed like triangular ones for the instrumental variable model and a diagonal covariance matrix for the factor model. But their shapes may still be strongly non-elliptical. In order to obtain meaningful posterior and predictive probabilities of states near and at the boundary of a reduced rank, weak regularization priors are discussed and compared. These are dealing with area restrictions, smoothness properties and training samples. As a novel class we introduce a lasso type shrinkage prior combined with orthogonal normalization which restricts the range of the parameters in a plausible way. A sensitivity analysis with respect to a sequence of weak priors is recommended.

The conditional posterior and predictive probabilities of different states of the econometric model near and at the boundary can then be used to estimate Bayesian mixture processes of several relevant economic and econometric issues.

We end this paper with listing some perspectives. The Bayesian approach to econometrics is now dominant in the field of macroeconomics. This is due to the pioneering work by Sims and his co-workers on Bayesian analysis of vector autoregressive models. The basic paper is Sims (1980) and an incomplete list of a few recent references are Sims and Zha (1998), Primiceri (2005) and Del Negro and Schorfheide (2011). An extension is to use more complex economic model structures like Dynamic Stochastic Equilibrium models, see Herbst and Schorfheide (2015). Complex cointegration models and inferential issues of these models have been studied extensively as well. Within this literature, we refer to Strachan (2003) for parameter instability, Jochmann et al. (2013) and Sugita et al. (2016) for regime-switching models, Jochmann and Koop (2015) for structural breaks, Koop et al. (2011) for time-varying parameter models and Chan et al. (2017) for cointegrating rank variations. Also in the fields of finance and marketing the Bayesian approach is becoming the dominant one. More details are presented in the Handbook of Bayesian Econometrics, Geweke et al. (2011), We emphasize another perspective. There exists already much research to extend the analysis of this paper to models with time varying mixtures and to connections with expert systems and machine learning. See, among others, Frühwirth-Schnatter (2006), Chan et al. (2012), Billio et al. (2013), Casarin et al. (2015) and Baştürk et al. (2016a).

All these extensions require a much more algorithmic approach to evaluating posterior probabilities of parameters and unknown unobserved states. Simulation based Bayesian Econometrics (SBBE) should be developed even more than already done so using modern software like parallel algorithms, filtering methods and modern hardware like clusters of machines of GPU processing. Developing operational methods useful for Bayesian empirical econometrics will lead to more insight in structural analysis, more accurate forecasting and more effective policy analysis with implied probabilistic components.

Appendices. Supplementary Material (Online Appendix)

A.1. Appendix for Section 2: Introduction to Bayesian analysis of instrumental variables in order to measure the effect of length-of-education on earned income in the USA

A.1.1. Introduction

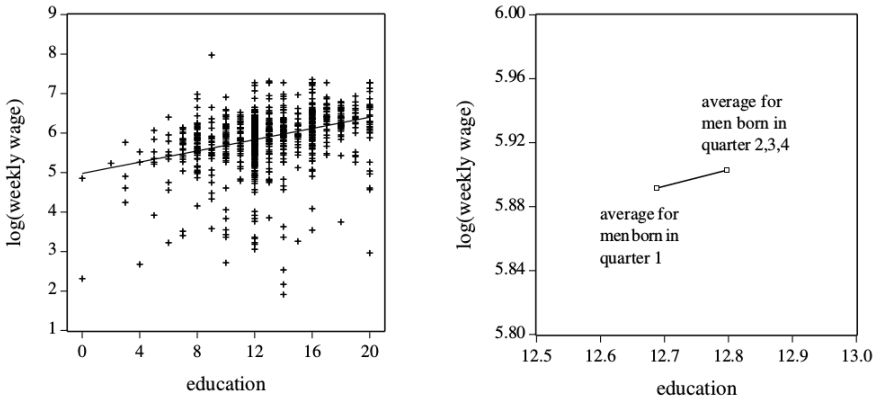
A well-known example of the use of instrumental variables in econometrics is the measurement of the effect of length-of-education on income, the (monetary) return on education.[10] Measuring this effect is a matter of great importance for several decision processes. For example, the results of such analysis are relevant for government agencies responsible for compulsory schooling laws, for school districts considering changes in school entrance policies and also for parents deciding when to enroll their children to school. However, a problem is that intellectual capabilities, which are usually not observed, not only influence education but also directly affect income. Therefore, a simple regression of income on the number of years of education may lead to incorrect conclusions. For example, more intelligent students find school less difficult and may choose to obtain more schooling to signal their high ability. So, even if extra years of education have no effect on income, people with higher education will on average have higher incomes because of their higher abilities. Therefore, one may expect that an ordinary regression of income on the years of education leads to an upward bias, i.e. an overestimated effect of education on income. Further, the (often unobserved) intellectual capabilities, income and education level of the parents may also cause an upward bias, as the parents' characteristics may also influence the education level and have a direct effect on income. For example, it may be the case that children of more intelligent and higher educated parents on average learn more at home. Another problem is the measurement error in reported education. First, usually only the completed (integer) number of years of education is reported. Second, people may misreport their education spell[11]. If the measurement error would be the only problem, one would expect that a simple regression of income on education would result in a downward bias, i.e. an underestimated effect of education on income, as the part of the variation in education that is merely due to measurement error does not lead to variation in income.

A method for solving these problems is the use of instrumental variables. These instrumental variables must be correlated with education but uncorrelated with latent capabilities (and measurement errors). Intuitively, in this way one focuses on the direct effect of education on income, while other effects on income are filtered out. However, it is hard to find variables that are correlated with education but uncorrelated with intellectual capabilities. Angrist and Krueger

[10] This subsection is based on Hoogerheide (2006).

[11] Siegel and Hodge (1968) find that the correlation between individuals' education reported in two surveys is only 0.933.

Fig A.1: Measuring the effect of education on income: simple regression of (logarithm of) income on education (left), or using quarter of birth as an instrumental variable (right).



(1991) use American data and suggest using quarter of birth to form instrumental variables. These instruments exploit that students born in different quarters have different average education spells. This results since most school districts require students to have turned age six by a certain date, a so-called 'birthday cut-off' which is typically near the end of the year, in the year they enter school, whereas compulsory schooling laws compel students to remain at school until their sixteenth, seventeenth or eighteenth birthday. This asymmetry between school-entry requirements and compulsory schooling laws compels students born in certain months to attend school longer than students born in other months: students born earlier in the year enter school at an older age and reach the legal drop-out age after less education. Hence, for students who leave school as soon as the schooling laws allow for it, those born in the first quarter have on average attended school for three quarters less than those born in the fourth quarter.
Angrist and Krueger (1991) use three data sets on men born in three decades, emphasizing results for the data set on 329509 men born in the years 1930-1939. This data set contains the number of completed years of education and the logarithm of weekly earnings in 1979. Figure A.1 illustrates the difference between simply regressing income on education and using quarter of birth to form an instrumental variable. The left panel shows how the effect of education on income is estimated by simple regression. The estimate is the steepness of the regression line, the line which minimizes the sum of squared (vertical) deviations of points from this line. For all data of the US this estimate is 0.0709: each added year of education results on average in a 7.09% increase in income. However, this method may overestimate or underestimate the effect of education on income because of latent intellectual capabilities or measurement errors, respectively. The right panel illustrates how the effect of schooling on earnings can be estimated using quarter of birth as an instrument. The average edu-

cation spell for men born in the first quarter is 12.6881 years, while for men born in other quarters the average education spell is 12.7969. So, the schooling laws imply that men born in the first quarter on average have 0.1088 years less education than men born in the other quarters. Further, for men born in the first quarter the average logarithm of income is 0.0111 ($= 5.9027 - 5.8916$) less than for those born in other quarters. In other words, men born in the first quarter have on average an income that is (approximately) 1.11% lower than men born in the other quarters. The key assumption is now that quarter of birth only influences income because of its effect on education, so that we may interpret the 1.11% difference in average income as a result of the difference in average education spell of 0.1088 years: each added year of education results on average in a 10.20% ($= 0.0111/0.1088$) increase in income. So, at first sight it seems that if any bias exists in the simple regression, then this is a downward bias: measurement errors in reported years of education may have caused an underestimation of the return on education. However, in the above-mentioned approaches we have only obtained estimates of the effect of education spell on income, but we have no measure of the uncertainty on these estimates: we have no lower and upper bounds between which the effect of education on income lies (with a certain probability). In order to obtain such a probability interval we must specify a model; this will be done in the sequel of this section.

First note that if the average education spell is exactly the same for those born in the first quarter and the others, then the approach using quarter of birth as an instrument does not work. In that case one can not identify the difference in income per year of education, as this leads to a division by zero. This illustrates that in instrumental variables models the problem of *local non-identification* may occur: if the instrument (quarter of birth) has no effect on the explanatory variable (education), then one can not identify the effect of the explanatory variable on the variable that is to be explained (income) using this instrument. Furthermore, if the average education spell is almost equal for those born in the first quarter and the others, then there is obviously much uncertainty on the estimated return on education. For in this case some changes of education and/or income for a few persons would result in a quite different estimate of the return on education. This kind of situation in which instruments only explain a small fraction of the variation in (some of) the explanatory variables, is usually referred to as the case of *weak instruments*. In fact, the difference in average education spell of 0.1088 years is small as compared to the variation in education spells across individuals (with education spells varying between 0 and 20 years, having a standard deviation of 3.28 years), so that the uncertainty on the estimate of the return on education is obviously much larger in the approach using quarter of birth as an instrument than in the simple regression. In other words, much information is lost by merely using the averages of education and income for the two quarter-of-birth groups; in the extreme case where the average education spell would be exactly the same for both groups, no information on return on education would be left. So, although the systematic error (of over- or underestimation) due to latent capabilities or measurement errors is avoided by using instruments, this approach may result in probability intervals for the

return on education that are so wide that they are of little practical use.
We consider a simple, illustrative model for the returns to schooling. First define $D_{quarter,i}$ as the following 0/1 variable: $D_{quarter,i} = 1$ if person i is born in quarter 2,3 or 4, and $D_{quarter,i} = 0$ if person i is born in quarter 1. The model is as follows:

$$\log wage_i = \beta\, education_i + \varepsilon_i \tag{A.1}$$
$$education_i = \pi\, D_{quarter,i} + v_i \tag{A.2}$$

for $i = 1, 2, \ldots, T$, where $\log wage$, $education$ and $D_{quarter}$ are taken in deviation from their means, so that no constant terms occur in (24) and (25). The parameter β is the average effect of one extra year of education on income: on average, one more year of schooling results in an increase of income of 100β %. The parameter π is the difference in the mean education spell between men born in quarter 2, 3 or 4 and men born in quarter 1. This is the case of exact identification in which there are as many instruments (only $D_{quarter}$) as explanatory endogenous variables (only $education$). The error terms ε_i and v_i are assumed to be independent across observations and normally distributed:

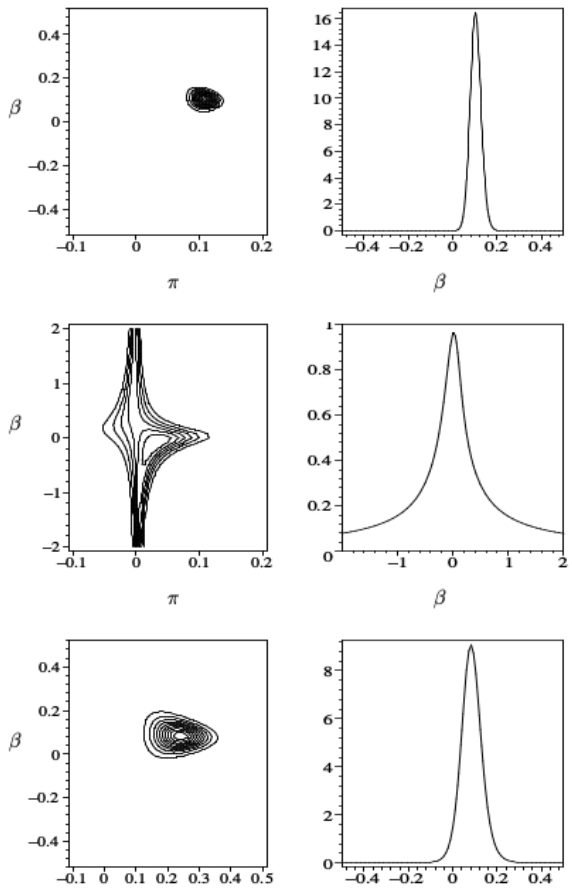
$$\begin{pmatrix} \varepsilon_i \\ v_i \end{pmatrix} \sim N(0, \Sigma), \qquad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

We specify the following non-informative prior density kernel of Drèze (1977):

$$p(\beta, \pi, \Sigma) \propto |\Sigma|^{-h/2} \text{ with } h > 0. \tag{A.3}$$

In the remainder of this Appendix, a derivation is presented of the joint posterior kernel of (π, β) and the marginal posterior kernel of β that follow from the prior specification in (26). Figure A.2 shows the shapes of the joint posterior kernel of (π, β) and the marginal posterior kernel of β (for the choice of $h = 3$) on bounded domains for three data sets: data of all states of the US, data of the state of New York, and joint data of the states of Kentucky, Tennessee and Arkansas.
First it should be noted that in this case of exact identification, both the joint posterior of (π, β) and the marginal posterior of β under the flat prior (26) are improper on $\mathbb{R}^2$ and $\mathbb{R}$, respectively: the integrals of the joint and marginal posterior density kernel are infinite. Although improper on $\mathbb{R}^2$, the joint posterior of (π, β) can be made proper by restricting β and/or π to a certain area. For more details on the existence of posteriors of the IV model, we refer to Appendix A.3.3 and Zellner et al. (2014). It is seen for data of all states of the US and for data of Kentucky, Tennessee and Arkansas that the joint posterior of (π, β) has a clear peak away from $\pi = 0$. This indicates that a sufficiently large difference in average education spell exists between men born in the first quarter and the others, so that valuable results on the returns to schooling can be obtained. In these cases the marginal posterior of β seems to have a bell-shape (with a peak around $\beta = 0.10$). On the other hand, for data of the state of New York the joint posterior kernel (π, β) displays a ridge around $\pi = 0$. In New

Fig A.2: Contour plot of joint posterior density kernel of (π, β) (left) and marginal posterior density kernel of β (right) for data of US (top, $T = 329509$) the state of New York (middle, $T = 29015$), the states of Kentucky, Tennessee and Arkansas (bottom, $T = 23062$).



York there is no or little difference in average education spell between the two quarter-of-birth groups, so that the instrumental variable (IV) approach gives no or little information on returns to schooling. The parameter π can take values close to 0, and for these values of π the parameter β can take a wide range of values; this reflects the local non-identification of β for $\pi = 0$. This leads to a marginal posterior of β with fat tails. Notice that the data set of New York even has somewhat more observations than the data set of Kentucky, Tennessee and Arkansas, so that it is not the size of the data set that causes the difference in the posterior of β. The only reason is the huge difference in strength of the quarter-of-birth instrument between the states.
We refer to section 6 in the paper and Section A.6 for the applications of the IV model to the Angrist and Krueger (1991) data for the case of all US states.

A.1.2. Illustrative posterior evidence near at at the boundary for NY State under a flat and Jeffreys' prior

We further investigate the evidence on a non-elliptical posterior density that may occur in the IV regression model.[12] We consider the simple, illustrative model for the measurement of the effect of education on income for two different data sets of Angrist and Krueger (1991) which were shown in Section 2 of the paper. We illustrate the effect of instrument strength on the posterior shapes, as the strength of the instrument differs considerably between the two data sets. We also will compare the peculiar posterior shape under a flat prior with the shape based on the Information Matrix or Jeffreys' prior. This latter one is a 'regularization prior' that in combination with the likelihood function yields posteriors with more regular properties.
The marginal posterior of β under a flat prior is given by:

$$p(\beta|y,x,Z) \quad \propto \quad \frac{[(y-x\beta)'(y-x\beta)]^{-(T-1)/2}}{[(y-x\beta)'M_Z(y-x\beta)]^{-(T-k-1)/2}} \tag{A.4}$$

with $M_Z = I - Z(Z'Z)^{-1}Z'$, see Zellner et al. (2014), and the marginal posterior of π [13] is given by:

$$\begin{aligned} p(\pi|y,x,Z) \quad \propto \quad & [(x-Z\pi)'(x-Z\pi)]^{-(T-1)/2}\,(\pi'Z'M_xZ\pi)^{-1/2} \times \\ & \times \left(\frac{\pi'Z'M_{[y\ x]}Z\pi}{\pi'Z'M_xZ\pi}\right)^{-(T-1)/2}. \end{aligned} \tag{A.5}$$

These posterior densities have two peculiar properties:

(a) **Local non-identification at $\pi = 0$:** The marginal posterior of π has an asymptote at $\pi = 0$ because of the term $(\pi'Z'M_xZ\pi)^{-1/2}$. In the case

[12] This subsection is based on Hoogerheide and Van Dijk (2008).
[13] For comparison with the earlier section we use here the symbol π where later in the appendix we make use of the capital Π.

of $k = 1$ instrument, the posterior is not integrable except on a bounded domain.

(b) **Regular posterior behavior of β when irrelevant instruments are added:** The marginal posterior of β becomes tighter if (possibly irrelevant) instruments are added. Moments exist up to the order of overidentification $(k-1)$; for $k = 1$, the marginal posterior of β is improper. This result appeared in an informal way in Maddala (1976), commenting on Drèze (1976). For a derivation see, Zellner et al. (2014).

The local non-identification of β when $\pi = 0$ is most easily seen from the restricted reduced form corresponding to the structural form (24)-(25):

$$\begin{pmatrix} y_i \\ x_i \end{pmatrix} = \begin{pmatrix} \beta \\ 1 \end{pmatrix} \pi' z_i + \begin{pmatrix} v_{1i} \\ v_i \end{pmatrix} \tag{A.6}$$

with $v_{1i} = v_i\beta + \varepsilon_i$ and $(v_{1i}, v_i)' \sim N(0, \Omega)$.

The left panel in Figure A.3 illustrates this feature for the data of the state of New York and the whole US. For the joint posterior kernel of β and π for New York state data, a substantial 'ridge' is visible at $\pi = 0$; the marginal posterior of π is completely dominated by the asymptote at $\pi = 0$. On the other hand, for the US data, the shapes are nearly elliptical, which reflects that in this case the quarter-of-birth instrument is less weak. The peak around the posterior mode[14] is high compared with the ridge around $\pi = 0$, so that the latter is not visible in this graph of the kernel of the joint posterior density.

We now consider the Information Matrix or Jeffreys' prior. This Jeffreys prior, the square root of the determinant of the information matrix, is given by:

$$p(\beta, \pi, \Sigma) \propto |\Sigma|^{-2} \, (\pi' Z' Z \pi)^{1/2} \, \sigma_{22.1}^{-(k-1)/2} \tag{A.7}$$

with $\sigma_{22.1} = \sigma_{22} - \sigma_{12}^2/\sigma_{11}$, for the structural form (24)-(25), or equivalently by:

$$p(\beta, \pi, \Omega) \propto |\Omega|^{-2} \, (\pi' Z' Z \pi)^{1/2} \, ((\beta \; 1)\Omega^{-1}(\beta \; 1)')^{(k-1)/2} \tag{A.8}$$

for the corresponding restricted reduced form (29); see Appendix A of Hoogerheide et al. (2007a) for a derivation of this prior.

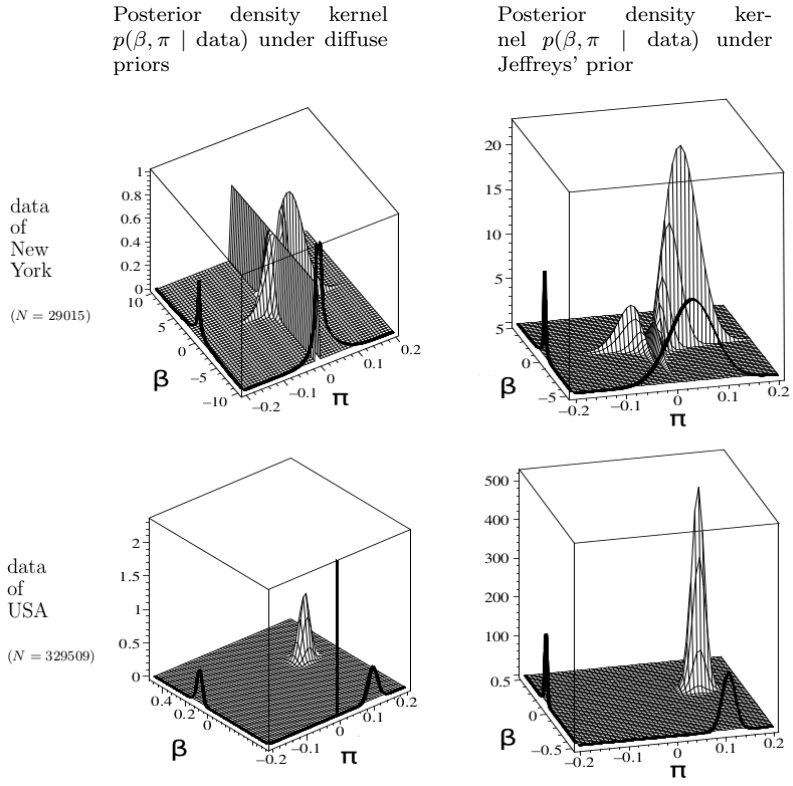
The factor $(\pi' Z' Z \pi)^{1/2}$ is 0 for $\pi = 0$, which reflects that in the restricted reduced form β only occurs in the product $\pi\beta$, so that for $\pi = 0$ the model contains no information on β. Hence for $\pi = 0$ the likelihood is constant over values of β, so that the first and second order derivatives of the log-likelihood with respect to β are zero, and the determinant of the information matrix, minus the expectation of the Hessian of the log-likelihood, is 0 for zero values of π.

Intuitively speaking, the factor $(\pi' Z' Z \pi)^{1/2}$ in the prior 'cancels' the asymptote of the posterior at $\pi = 0$.

The $((\beta \; 1)\Omega^{-1}(\beta \; 1)')^{(k-1)/2}$ factor in the prior influences the tail behavior of

[14] In this simple example, the posterior mode is given by $(\beta, \pi) = (\hat{\beta}_{2SLS}, \hat{\pi}_{OLS}) = (y'z/x'z, x'z/z'z)$.

Fig A.3: Posterior density kernels for IV models for measurement of the effect of income on education (β) using the difference in mean education between men born in quarters 2-4 and quarter 1 (π). The figures show the posterior kernel of (β, π). At the axes, marginal posteriors of β and π are shown.



the marginal posterior of β and makes it independent of the number of instruments k such that it has Cauchy type tails.
Note that for $k = 1$ instrument the Jeffreys prior (30) reduces to

$$p(\beta, \pi, \Sigma) \propto |\Sigma|^{-2}|\pi|, \tag{A.9}$$

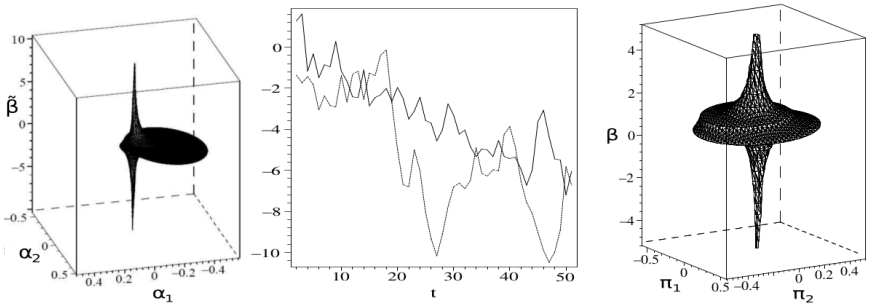
which is simply the diffuse prior in (26) with $h = 4$ multiplied with $|\pi|$. One interpretation of the Information Matrix or Jeffreys' prior is that a priori one prefers a strong instrument; that is, π is preferred to be large (in absolute sense). An intuitively appealing explanation is that this Jeffreys prior is just a *'regularization prior'* that does not immediately reflect prior beliefs, but in combination with the likelihood function yields posteriors with desirable properties (in the sense that the aforementioned peculiar properties resulting from the diffuse prior do not occur).
Note that also for $k > 2$ the factor $(\pi' Z' Z \pi)^{1/2}$ in the prior takes high values (in absolute sense) for large elements of π, while in this case the $((\beta\ 1)\Omega^{-1}(\beta\ 1)')^{(k-1)/2}$ factor takes high values for (in absolute sense) large values of β. In the likelihood of the (restricted reduced form of) the IV model, it is the occurrence of the product $\pi\beta$ that causes points (π, β) with π and β both attaining extremely large values to have small posterior probability.
The right panel of Figure A.3 illustrates the posterior shapes under the Jeffreys prior for the data of New York state and the US. For the US data, the graphs look similar to the graphs under the diffuse prior, except for the disappearance of the asymptote at $\pi = 0$ for the marginal posterior of π. For the New York state data, the differences with the posterior shapes under the diffuse prior are huge. Under the Jeffreys prior, there is no ridge or asymptote at $\pi = 0$, and the tails of the marginal posterior of β are thinner (and integrable). Also note that, although the *joint* posterior kernel of β, π tends to 0 for $\pi \to 0$, the *marginal* posterior of π does not drop in neighborhoods of $\pi = 0$: for $\pi \to 0$ the lower values of the posterior density kernel $p(\beta, \pi|y, x, Z)$ are compensated by the fact that for $\pi \to 0$ the posterior $p(\beta, \pi|y, x, Z)$ becomes less sensitive with respect to changes in β, as β only occurs in the likelihood in the product $\pi\beta$. In other words, the marginal posterior probability mass of π does not decrease for $\pi \to 0$, this posterior probability mass is just spread over a wider range of values for β. Finally, note that although the Jeffreys prior 'cures' some of the peculiar properties under the diffuse prior, the posterior may still display non-elliptical shapes such as bimodality. In the main text of this paper we argue that other regularization priors are more suitable, see Sections 4 and 5 of the main text.

A.2. Appendix for section 2: Other examples about non-elliptical shapes of posteriors and predictive densities

A.2.1. More background on similarity of mathematical structure and posterior shapes in IV model and Vector Error Correction Model

We further illustrate the result in Section 2 that the similar mathematical structure of the instrumental variable model and the vector autoregressive model un-

Fig A.4: A Highest Posterior Density credible set for the parameters α_1, α_2, $\tilde{\beta}$ in the VECM under a diffuse prior for simulated data from a VECM with $\alpha_1 = -0.05$, $\alpha_2 = 0.05$, $\tilde{\beta} = 1$ (left); the simulated data from the VECM (middle); an HPD credible set in an IV model in a similar simulation experiment (right).



der cointegration restrictions leads to similar posterior shapes. Consider again the restricted reduced form of an IV model with 2 instruments z_{1i}, z_{2i} $(i = 1, \ldots, N)$, and a simple vector error correction model (VECM) under a cointegration restriction for 2 variables y_{1t}, y_{2t} $(t = 1, \ldots, T)$:

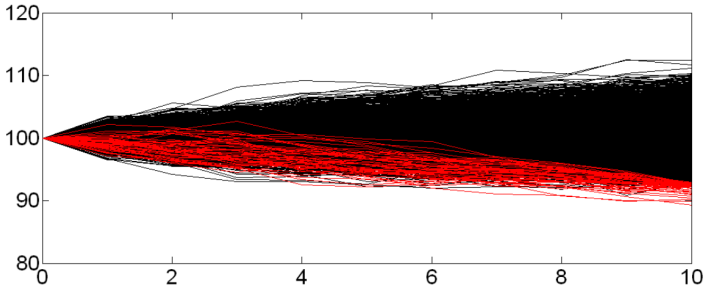
$$\text{IV:} \begin{pmatrix} y_i \\ x_i \end{pmatrix} = \overbrace{\begin{pmatrix} \beta \\ 1 \end{pmatrix} (\pi_1 \;\; \pi_2)}^{\text{reduced rank}} \begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix} + \begin{pmatrix} v_{1i} \\ v_i \end{pmatrix}$$

$$\text{VECM:} \begin{pmatrix} \Delta y_{1t} \\ \Delta y_{2t} \end{pmatrix} = \overbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} (1 \;\; -\tilde{\beta})} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}$$

which have in common that they contain a parameter matrix with reduced rank. In both models, local non-identification plays a role. In the IV model, the parameter β is not identified for $\pi_1 = \pi_2 = 0$, whereas in the VECM the parameter $\tilde{\beta}$ is not identified for $\alpha_1 = \alpha_2 = 0$.

We show a simulation experiment with $\alpha_1 = -0.05$, $\alpha_2 = 0.05$, $\tilde{\beta} = 1$, so that there is slow adjustment towards the cointegration relation $y_1 = y_2$, $(\varepsilon_{1t}, \varepsilon_{2t}) \sim N(0, I)$, for a rather small data set $(T = 50)$. The left panel of Figure A.4 shows a Highest Posterior Density (HPD) credible set for $(\alpha_1, \alpha_2, \beta)$ under a diffuse prior similar to the diffuse prior for the IV model, for $-0.5 < \alpha_j < 0.5$ $(j = 1, 2)$, $-10 < \tilde{\beta} < 10$. The middle panel of Figure A.4 shows the simulated data from the VECM. The right panel shows approximately the same non-elliptical posterior shapes for a similar simulation experiment in the IV model.

Fig A.5: 10-day ahead return and 1% Value-at-Risk forecasts using direct simulation of returns (black lines) and simulation of only 'high loss' price paths



A.2.2. Other empirical examples on non-elliptical shapes

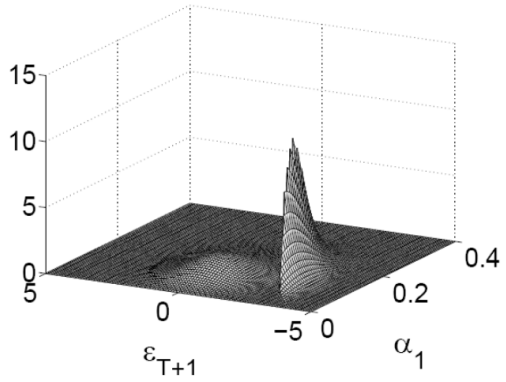
2-regime mixture model for US GNP growth: As a different empirical example (that does not involve the IV model) of a 'non-standard posterior' under flat priors, we consider a 2-regime mixture model for US GNP growth. On purpose we show here a case where the model is different than the class of reduce rank models. This is done to indicate that non-elliptical densities occur in many empirical econometric models. This model was also analyzed in Hoogerheide (2006) (Ch. 3), where percentage growth in US GNP has two mean levels:for $t = 1 \ldots, T$, where $p \in (0, 1)$, $\varepsilon_t \sim N(0, \sigma^2)$ and $\beta_1 < \beta_2$ for identification.

$$y_t = \begin{cases} \beta_1 + \varepsilon_t & \text{with probability } p, \\ \beta_2 + \varepsilon_t & \text{with probability } 1 - p. \end{cases} \tag{A.10}$$

Figure A.7 presents the highest posterior density credible sets for the model parameters in (33) together with the marginal posterior densities (illustrated by histograms of draws from these densities) of $(\beta_1, \beta_2, \sigma, p)$. The effect of the strongly non-elliptical shape on forecasting and policy advice constitutes an important topic for future research.

Value-at-Risk forecasts of stock returns: As a third empirical example we consider a model for forecasting 10-day ahead value-at-risk (VaR) for a 'given' distribution for stock returns where the parameters of the distribution are fixed. In this case, the issue of a boundary or near-boundary refers to rare events, i.e. 'thick tails' or extreme scenarios, which lead to non-elliptical predictive distributions.
For this model, simulated return paths using a 'direct simulation' are given by the black lines in Figure A.5. We refer to Hoogerheide and Van Dijk (2010) for the details of this return model and simulation algorithm. Using these simulated paths, the 99% VaR is calculated as the 1% quantile of simulated prices after

Fig A.6: Optimal importance density for Bayesian estimation of 1% quantile of return distribution in an ARCH(1) model (with variance targeting) with ARCH(1) parameter α_1 and standardized error ε_{t+1}, where half the probability mass of the importance density lies in the 'high loss region'.

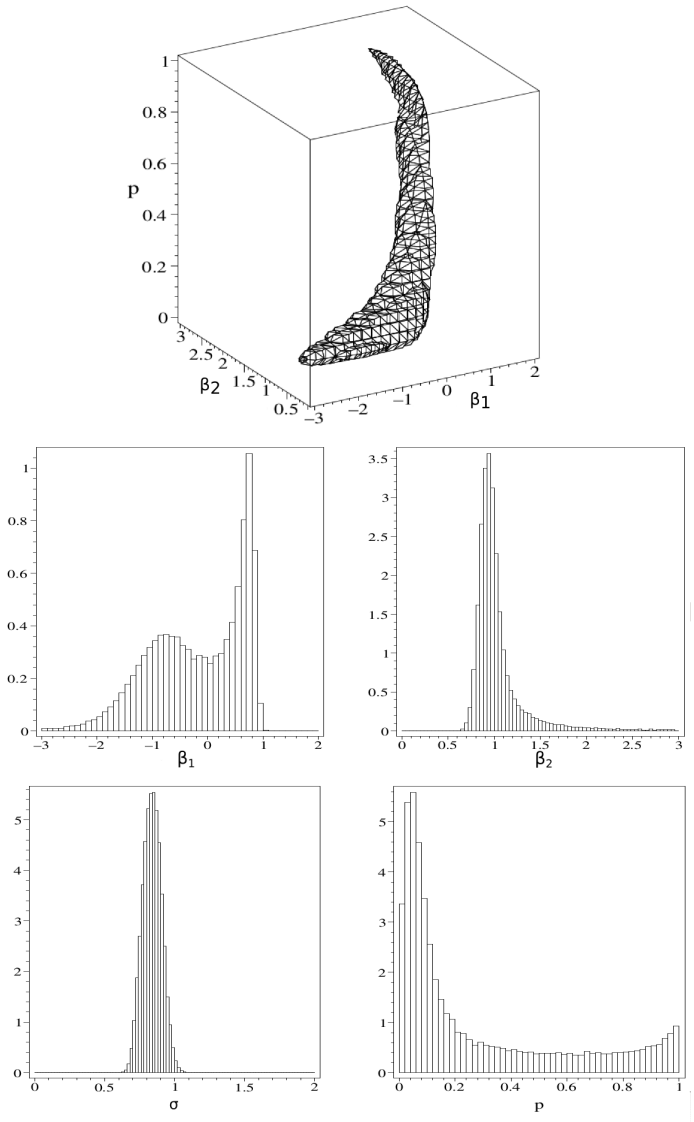


10 days. The disadvantage of this direct simulation is the computing time: estimated VaR may be inaccurate since it is based on only 1% of the simulation paths. An alternative, more efficient, simulation example is given by the red lines in Figure A.5, where the red lines correspond to 'high loss scenarios' which should constitute 50% of the draws from the importance density for optimal Bayesian estimation of the 1% .
This example is closely related to the near-boundary issue due to the bimodality of the optimal importance density for the model parameters (e.g., α_1 in an ARCH(1) model) and simulated future return paths, where we observe a 'middle mode' and a 'high loss mode' in Figure A.6.

A.3. Appendix for section 3: Basic model structures, nonstandard likelihood shapes and posterior existence

We start this appendix by listing some matrix and determinant properties and by giving definitions of the distributions and corresponding density functions that are needed to obtain the properties of the likelihood and posterior in the three basic model structures. We then provide detailed derivations of the posterior distributions of the cointegration model, including the requirements for the existence of posterior distributions. We refer to Baştürk et al. (2016b) and Kleijn (2016) for some more background details on the definitions and derivations.

Fig A.7: Highest posterior density credible set for parameters $(\beta_1, \beta_2, \sigma, p)$ for the 2-regime mixture model for US GNP growth.



A.3.1. Results on matrices and determinants and definitions of matric-variate distributions

Decomposition of sum of squares in linear regression

$$(Y - X\beta)'(Y - X\beta) = Y'M_X Y + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}), \tag{A.11}$$

where $\hat{\beta} = (X'X)^{-1}X'Y$ and $M_X = I - X(X'X)^{-1}X$.
Frisch-Waugh:

$$\begin{aligned}(Y - X_1\beta_1 - X_2\beta_2)'(Y - X_1\beta_1 - X_2\beta_2)& \\ =(Y - X_2\beta_2)'M_{X_1}(Y - X_2\beta_2) + (\beta_1 - \hat{\beta}_1)'X_1'X_1(\beta_1 - \hat{\beta}_1)& \\ =Y'M_X Y + (\beta_2 - \hat{\beta}_2)'X_2'M_{X_1}X_2(\beta_2 - \hat{\beta}_2)& \\ + (\beta_1 - \hat{\beta}_1)'X_1'X_1(\beta_1 - \hat{\beta}_1)&\end{aligned} \tag{A.12}$$

where $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y$ and $\hat{\beta}_2 = (X_2'M_{X_1}X_2)^{-1}X_2'M_{X_1}Y$.
From Anderson (2003, ch.14):

$$(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} = A_{22}^{-1}A_{21}(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}A_{12}A_{22}^{-1} + A_{22}^{-1} \tag{A.13}$$

Orthogonal complements From Johansen (1988)

$$\beta_\perp(\alpha_\perp'\beta_\perp)^{-1}\alpha_\perp' + \alpha(\beta'\alpha)^{-1}\beta' = I, \tag{A.14}$$

where $\beta'\alpha$ has full rank. If we choose $\alpha = \beta = A^{-1/2}\tilde{\alpha}$ with $\alpha_\perp = \beta_\perp = A^{1/2}\tilde{\alpha}_\perp$ (so that $\alpha'\alpha_\perp = \tilde{\alpha}'\tilde{\alpha}_\perp = 0$), where $\tilde{\alpha}$ has full rank, we have:

$$A^{1/2}\tilde{\alpha}_\perp(\tilde{\alpha}_\perp'A\tilde{\alpha}_\perp)^{-1}\tilde{\alpha}_\perp'A^{1/2} + A^{-1/2}\tilde{\alpha}(\tilde{\alpha}'A^{-1}\tilde{\alpha})^{-1}\tilde{\alpha}'A^{-1/2} = I, \tag{A.15}$$

Pre- and post-multiplying with $A^{-1/2}$ yields:

$$\tilde{\alpha}_\perp(\tilde{\alpha}_\perp'A\tilde{\alpha}_\perp)^{-1}\tilde{\alpha}_\perp' + A^{-1}\tilde{\alpha}(\tilde{\alpha}'A^{-1}\tilde{\alpha})^{-1}\tilde{\alpha}'A^{-1} = A^{-1}. \tag{A.16}$$

Derivation of a matrix equation **Theorem:** Consider two matrices A ($T \times m_1$) and B ($T \times m_2$), where $m_1 \leq T$ and $m_2 \leq T$. Suppose that A has full rank, so that $(A'A)^{-1}$ exists. Then we can decompose the determinant $|(A\ B)'(A\ B)|$ as follows:

$$|(A\ B)'(A\ B)| = |A'A||B'M_A B| \tag{A.17}$$

where M_A is the $T \times T$ projection matrix defined as $M_A = I - A(A'A)^{-1}A'$.
Proof: First, note that

$$|(A\ B)'(A\ B)| = \left|\begin{pmatrix} A'A & A'B \\ B'A & B'B \end{pmatrix}\right|, \tag{A.18}$$

and

$$|A'A|^{-1} = |(A'A)^{-1}| = \left|\begin{pmatrix} (A'A)^{-1} & 0 \\ 0' & I_{m_2} \end{pmatrix}\right|, \tag{A.19}$$

where 0 is the $m_1 \times m_2$ zero matrix, and I_{m_2} is the $m_2 \times m_2$ identity matrix. From (41) and (42), we have:

$$\begin{aligned} |(A\ B)'(A\ B)||A'A|^{-1} &= \left|\begin{pmatrix} (A'A)^{-1} & 0 \\ 0' & I_{m_2} \end{pmatrix}\right| \left|\begin{pmatrix} A'A & A'B \\ B'A & B'B \end{pmatrix}\right| \\ &= \left|\begin{pmatrix} (A'A)^{-1} & 0 \\ 0' & I_{m_2} \end{pmatrix}\begin{pmatrix} A'A & A'B \\ B'A & B'B \end{pmatrix}\right| \\ &= \left|\begin{pmatrix} I_{m_1} & (A'A)^{-1}A'B \\ B'A & B'B \end{pmatrix}\right|, \end{aligned} \tag{A.20}$$

where I_{m_1} is the $m_1 \times m_1$ identity matrix. Multiplying (43) by

$$1 = \left|\begin{pmatrix} I_{m_1} & 0 \\ -B'A & I_{m_2} \end{pmatrix}\right| \tag{A.21}$$

we have:

$$\begin{aligned} |(A\ B)'(A\ B)||A'A|^{-1} &= \left|\begin{pmatrix} I_{m_1} & 0 \\ -B'A & I_{m_2} \end{pmatrix}\right| \left|\begin{pmatrix} I_{m_1} & (A'A)^{-1}A'B \\ B'A & B'B \end{pmatrix}\right| \\ &= \left|\begin{pmatrix} I_{m_1} & (A'A)^{-1}A'B \\ 0' & B'B - B'A(A'A)^{-1}A'B \end{pmatrix}\right| \\ &= |B'B - B'A(A'A)^{-1}A'B|. \end{aligned} \tag{A.22}$$

Finally, multiplying (45) by $|A'A|$ yields:

$$|(A\ B)'(A\ B)| = |A'A||B'B - B'A(A'A)^{-1}A'B||A'A||B'M_AB|, \tag{A.23}$$

where $M_A = I - A(A'A)^{-1}A'$. □

Corollary: If additionally B has full rank, then

$$|B'M_AB| = \frac{|A'M_BA||B'B|}{|A'A|}, \tag{A.24}$$

where M_B is the $T \times T$ projection matrix defined as $M_B = I - B(B'B)^{-1}B'$.

Proof: Note that if B has full rank, we can switch the matrices A and B. So, in that case we have:

$$|A'A||B'M_AB| = |(A\ B)'(A\ B)| = |(B\ A)'(B\ A)| = |B'B||A'M_BA|. \tag{A.25}$$

and the result follows immediately. □

Inverted Wishart distribution Let Σ be an $n \times n$ random symmetric positive definite matrix. Σ has an inverted Wishart distribution if its density function is

$$p(\Sigma|Q,\nu) = c|\Sigma|^{-\frac{1}{2}(\nu+n+1)}|Q|^{\nu/2}\exp\left[-\frac{1}{2}\operatorname{tr}(Q\Sigma^{-1})\right], \text{ for } |\Sigma| > 0 \tag{A.26}$$

where Σ is a symmetric positive definite $n \times n$ matrix and $\nu \geq n-1$. The constant c is given by $c^{-1} = 2^{\frac{1}{2}\nu n}\Gamma_n(\nu/2)$ where

$$\Gamma_n(x) = \pi^{n(n-1)/4} \prod_{i=1}^{n} \Gamma[x + (1-i)/2] \tag{A.27}$$

is the multivariate gamma function.
If Σ has the above inverted Wishart density, $\Psi = \Sigma^{-1}$ has a Wishart distribution with scale Q^{-1} and degrees of freedom ν. An algorithm to generate random draws from an inverted Wishart distribution is derived in Zellner, Bauwens, and Van Dijk (1988, pp.67-71).
Using (49) it follows directly that

$$\int |\Sigma|^{-\frac{1}{2}(\nu+n+1)} \exp\left[-\frac{1}{2}\operatorname{tr} Q\Sigma^{-1}\right] d\Sigma \propto |Q|^{-\frac{1}{2}\nu}, \tag{A.28}$$

which is often denoted as the inverted Wishart step in integrating out the covariance matrix in the posterior distribution.

Matric-variate t distribution The $p \times q$ random matrix T has a matric-variate t (Mt) distribution (see Zellner (1971) and Kleibergen and Van Dijk (1998)) with parameters P, Q, n if, and only if, its probability density function is:

$$p_{Mt}(T|P,Q,n) = k\, \frac{|P|^{q/2}\,|Q|^{(n-p)/2}}{|Q + T'\,P\,T|^{n/2}} = k\, \frac{|P|^{-(n-q)/2}\,|Q|^{-p/2}}{|P^{-1} + T\,Q^{-1}\,T'|^{n/2}}, \tag{A.29}$$

where the equality follows from the following equality:

$$\frac{|Q + T'PT|}{|P|\,|Q|} = |P^{-1} + TQ^{-1}T'| \tag{A.30}$$

and k is given by:

$$k = \frac{1}{\pi^{pq/2}}\, \frac{\prod_{i=1}^{q}\Gamma[(n-i+1)/2]}{\prod_{i=1}^{q}\Gamma[(n-p-i+1)/2]}, \tag{A.31}$$

and we have $n > p+q-1$ and P and Q are positive definite symmetric (PDS) matrices of size $p \times p$ and $q \times q$, respectively. Alternatively, the matric-variate t-distribution is often parameterized in terms of the degrees of freedom parameter $\nu = n - p$ instead of n.

Marginal and conditional matric-variate t distributions: For a random matrix T which has a matric-variate t distribution, marginal and conditional distributions of partitions of T are also matric-variate t distribution.
First note that (52) does not contain a location parameter. A location parameter M can be introduced using

$$p_{Mt}(T|M,P,Q,n) = p_{Mt}(T - M|P,Q,n). \tag{A.32}$$

Next, consider the following partitioning of T, location parameter M and the scale parameters P:

$$T = \begin{pmatrix} T_1^{(p_1 \times q)} \\ T_2^{(p_2 \times q)} \end{pmatrix}, \quad M = \begin{pmatrix} M_1^{(p_1 \times q)} \\ M_2^{(p_2 \times q)} \end{pmatrix}, \quad P = \begin{pmatrix} P_{11}^{(p_1 \times p_1)} & P_{12}^{(p_1 \times p_2)} \\ P_{21}^{(p_2 \times p_1)} & P_{22}^{(p_2 \times p_2)} \end{pmatrix} \tag{A.33}$$

where $p_1 + p_2 = p$ and $A^{(a,b)}$ denotes that matrix A has dimensions $a \times b$. Then the following conditional and marginal densities hold:

$$p(T_1 \mid T_2, M, P, Q) = p_{Mt}\left(T_1 | M_{T_1|T_2}, P_{11}, Q_{T_1|T_2}, n\right), \tag{A.34}$$

$$p(T_2 \mid M, P, Q) = p_{Mt}\left(T_2 | M_2, P_{22.1}, Q, n - p_1\right), \tag{A.35}$$

where $M_{T_1|T_2} = M_1 - P_{11}^{-1}P_{12}(T_2 - M_2)$, $Q_{T_1|T_2} = Q + (T_2 - M_2)' P_{22.1}(T_2 - M_2)$ and $P_{22.1} = P_{22} - P_{21}P_{11}^{-1}P_{12}$. See Zellner (1971), appendix B.5 and Bauwens et al. (1999), appendix A.2.

A.3.2. Derivations of conditional and marginal densities of α, β_2 and Σ of the cointegration model under linear normalization and non-existence of the marginal posterior of α

Full conditional posterior densities of α, β_2, Σ: As mentioned in the main text, a basic feature of the joint posterior density function of α, β_2, Σ, conditional upon the data, is that it is *a proper density even though the marginal density of α is unbounded around $\alpha = 0$ and it has heavy tails.*

The conditional posterior of Σ given α, β_2 is given as an inverse-Wishart density with T degrees of freedom and a finite scale matrix, $(\Delta Y - Y_{-1}\beta\alpha')' (\Delta Y - Y_{-1}\beta\alpha')$, which is PDS for all values of α and β_2 and $T > k - 1$. These conditions are satisfied when the data have full rank. We note that the sample size requirement is usually not binding for time series in the field of econometrics. We refer for more details to the definition of the inverse-Wishart distribution and the inverse-Wishart integration step.

The conditional posterior of α given β_2 and Σ is a matrix normal density with well-defined parameters.

In order to save on notation, we note that one can apply similar steps as presented below for the conditional matrix t distribution for α given β_2 and marginal with respect to Σ and derive similar location and scale matrices for the normal as for the matrix t distribution.

The conditional posterior of β_2 given α and Σ is a matrix normal density with well-defined parameters This conditional posterior is obtained in the same way as listed above. Details are left to interested readers.

Conditional and marginal posterior densities of α, β_2 after integrating out Σ.

The conditional posterior density of $\alpha|\beta_2, Y$ is proportional to the joint posterior

density $p(\alpha, \beta_2 \mid Y)$ [15]:

$$p(\alpha \mid \beta_2, Y) \propto \left|(\Delta Y - Y_{-1}\beta\alpha')'(\Delta Y - Y_{-1}\beta\alpha')\right|^{-T/2}. \tag{A.36}$$

Completing the squares on α in (59) yields:

$$p(\alpha \mid \beta_2, Y) \propto \left|\left(\Delta Y' M_{Y_{-1}\beta}\Delta Y + (\alpha - \hat{\alpha})\left(\beta' Y'_{-1}Y_{-1}\beta\right)(\alpha - \hat{\alpha})'\right)\right|^{-T/2}, \tag{A.37}$$

where $\hat{\alpha}' = \left(\beta' Y'_{-1}Y_{-1}\beta\right)^{-1}\beta' Y'_{-1}\Delta Y$ and (60) holds under the conditions that β has full column rank, which is due to the normalization condition and that $\text{rank}(Y_{-1}) \geq r$, hence the $r \times r$ matrix $\left(\beta' Y'_{-1}Y_{-1}\beta\right)$ has rank r and is invertible:

$$\text{rank}\left(\beta' Y'_{-1}Y_{-1}\beta\right) = \min(\text{rank}(Y_{-1}), r) = r. \tag{A.38}$$

From (60) and using the first definition of a matric-variate t density, see also Dickey (1967), it follows that the conditional density of α given β_2 is a matric-variate t density:

$$p(\alpha \mid \beta_2, Y) \propto p_{Mt}\left(\alpha|\hat{\alpha}, (\Delta Y' M_{Y_{-1}\beta}\Delta Y)^{-1}, (\beta' Y'_{-1}Y_{-1}\beta)^{-1}, T\right), \tag{A.39}$$

where the matrix $\hat{\alpha}$ contains location parameter, $\beta' Y'_{-1}Y_{-1}\beta$ and $\Delta Y' M_{Y_{-1}\beta}\Delta Y$ are matrices that contain scale parameters with $T-k$ degrees of freedom, where $T > k+r-1$ is a sample size requirement. For sample sizes that are usually given in econometrics, the latter condition is fulfilled. The matric-variate t density property holds under the condition that $\beta' Y'_{-1}Y_{-1}\beta$ and $\Delta Y' M_{Y_{-1}\beta}\Delta Y$ are positive definite for all values of β_2 which holds under linear normalization, see also below.

Conditional posterior of $(\beta_2|\alpha)$ The conditional posterior density of $\beta_2|\alpha, Y$ is proportional to the joint posterior density $p(\alpha, \beta_2 \mid Y)$. This conditional is obtained in three steps. First, by completing the squares on $\Pi = \beta\alpha'$. Next, by completing the squares on β and thirdly by completing the squares on β_2 and using the decomposition of the joint matric-variate t density of β into a conditional matric-variate t density of β_2 and a marginal density of β_1 evaluated at $\beta_1 = I$, see Dickey (1967), Zellner (1971) or Bauwens et al. (1999) for background on the matric-variate t density.
The first step, completing the squares on $\beta\alpha'$, which is the restricted value of Π, proceeds as follows:

$$\begin{aligned} p(\alpha, \beta_2 \mid Y) &\propto \left|(\Delta Y - Y_{-1}\beta\alpha')'(\Delta Y - Y_{-1}\beta\alpha')\right|^{-T/2}, && \text{(A.40)} \\ &= \left|\Delta Y' M_{Y_{-1}}\Delta Y + (\beta\alpha' - \hat{\Pi})' Y'_{-1}Y_{-1}(\beta\alpha' - \hat{\Pi})\right|^{-T/2} && \text{(A.41)} \\ &\propto \left|(Y'_{-1}Y_{-1})^{-1} + (\beta\alpha' - \hat{\Pi})D^{-1}(\beta\alpha' - \hat{\Pi})'\right|^{-T/2} && \text{(A.42)} \end{aligned}$$

[15] This relation holds since $p(\alpha \mid \beta_2, Y) = \frac{p(\alpha, \beta_2|Y)}{p(\beta_2|Y)} \propto p(\alpha, \beta_2 \mid Y)$.

where $\hat{\Pi} = (Y'_{-1}Y_{-1})^{-1}Y'_{-1}\Delta Y$ and $D = \Delta Y' M_{Y_{-1}} \Delta Y$, which only depends on given data. In the last line we made use of the determinant equality.
The second step is completing the squares on β in (65). Here we use:

$$\hat{\beta} = \hat{\Pi} D^{-1}\alpha \left(\alpha' D^{-1}\alpha\right)^{-1}. \tag{A.43}$$

Hence (65) can be written as:

$$\begin{aligned} p\left(\alpha, \beta_2 \mid Y\right) \quad \propto \quad & \Big| (Y'_{-1}Y_{-1})^{-1} + (\beta - \hat{\beta})\alpha' D^{-1}\alpha(\beta - \hat{\beta})' \\ & + \hat{\Pi}\left(D^{-1} - D^{-1}\alpha(\alpha' D^{-1}\alpha)^{-1}\alpha' D^{-1}\right)\hat{\Pi}' \Big|^{-T/2}, \end{aligned} \tag{A.44}$$

where the term $\hat{\Pi}\left(D^{-1} - D^{-1}\alpha(\alpha' D^{-1}\alpha)^{-1}\alpha' D^{-1}\right)\hat{\Pi}'$ results from the difference between $(\beta\alpha' - \hat{\Pi})D^{-1}(\beta\alpha' - \hat{\Pi})'$ from (65) and $(\beta - \hat{\beta})\alpha' D^{-1}\alpha(\beta - \hat{\beta})'$ from (67).
Equation (67) can be simplified as follows:

$$\begin{aligned} p\left(\alpha, \beta_2 \mid Y\right) \propto & \Big| (\beta - \hat{\beta})\alpha' D^{-1}\alpha(\beta - \hat{\beta})' \qquad\qquad (A.45) \\ & + \left(Y'_{-1}Y_{-1} - Y'_{-1}\Delta Y \alpha_\perp (\alpha'_\perp \Delta Y' \Delta Y \alpha_\perp)^{-1}\alpha'_\perp \Delta Y' Y_{-1}\right)^{-1} \Big|^{-T/2} \\ = & \Big| \left(Y'_{-1} M_{\Delta Y \alpha_\perp} Y_{-1}\right)^{-1} + (\beta - \hat{\beta})\alpha' D^{-1}\alpha(\beta - \hat{\beta})' \Big|^{-T/2}, \qquad (A.46) \end{aligned}$$

where $r \times r$ values of β are fixed due to the normalization restriction and the orthogonal complement $\alpha_\perp$ satisfies $\alpha'\alpha_\perp = 0$ and $\alpha'_\perp \alpha_\perp = I$. Equation (69) takes the form of a matric-variate t density for the unrestricted parameter matrix $\beta = (\beta'_1, \beta'_2)'$, when the linear normalization is not used. Note that the matrices $Y'_{-1}M_{\Delta Y\alpha_\perp}Y_{-1}$ and $\alpha' D^{-1}\alpha$ are required to be positive definite for all values of the random variable matrix α. If one or more columns of α go to zero, then the matrix $\alpha' D^{-1}\alpha$ becomes singular. We can ignore that here since this event has probability measure zero within the space of α. When columns are very close but not equal to zero, then the matrix is nearly singular. We investigate the limiting behavior when the columns become arbitrarily close to zero in the next section. This is a a situation of near-unit roots which is an empirical relevant issue that received a lot of attention in the literature. We comment on this also in the conclusions.
We next make use of the decomposition of a matric-variate t density into a

conditional and marginal one, as mentioned before:

$$\begin{aligned} p(\alpha,\beta \mid Y) &\propto \left|\left(Y_{-1}'M_{\Delta Y\alpha_\perp}Y_{-1}\right)^{-1}+(\beta-\hat\beta)\alpha' D^{-1}\alpha(\beta-\hat\beta)'\right|^{-T/2} \\ &= \left|P^{-1}+(\beta-\hat\beta)Q^{-1}(\beta-\hat\beta)'\right|^{-T/2} \qquad \text{(A.47)} \\ &\overset{(52)}{=} \frac{|P|^{(T-r)/2}|Q|^{k/2}}{c(k,r,T)}p_{Mt}(\beta \mid \hat\beta,P,Q,T) \qquad \text{(A.48)} \\ &\overset{(57,58)}{=} \frac{|P|^{(T-r)/2}|Q|^{k/2}}{c(k,r,T)}p_{Mt}(\beta_2 \mid \hat\beta_{2|1},P_{22},Q_{2|1},T) \\ &\quad \times p_{Mt}(\beta_1 \mid \hat\beta_1,P_{11.2},Q,T-k+r) \qquad \text{(A.49)} \end{aligned}$$

$$\begin{aligned} P &= Y_{-1}'M_{\Delta Y\alpha_\perp}Y_{-1} \\ Q &= (\alpha' D^{-1}\alpha)^{-1} \\ c(k,r,T) &= \frac{1}{\pi^{kr/2}}\frac{\prod_{i=1}^r \Gamma[(T-i+1)/2]}{\prod_{i=1}^r \Gamma[(T-k-i+1)/2]} \\ P_{11.2} &= P_{11}-P_{12}P_{22}^{-1}P_{21} \\ Q_{2|1} &= Q+(\beta_1-\hat\beta_1)'P_{11.2}(\beta_1-\hat\beta_1) \\ \hat\beta_{2|1} &= \hat\beta_2-P_{22}^{-1}P_{21}(\beta_1-\hat\beta_1) \end{aligned}$$

Since $p(\beta_2 \mid \alpha,Y)$ is proportional to $p(\alpha,\beta_2 \mid Y)$ and $p_{Mt}(\beta_2 \mid \hat\beta_{2|1},P_{22},Q_{2|1},T)$ is the only factor that depends on β_2 it follows that

$$\begin{aligned} p(\beta_2|\alpha,Y) &\propto p_{Mt}(\beta_2 \mid \hat\beta_{2|1},P_{22},Q_{2|1},T) \qquad \text{(A.50)} \\ &\propto |Q_{2|1}|^{(T-k+r)/2}|P_{22}|^{r/2} \\ &\quad \times \left|Q_{2|1}+(\beta_2-\hat\beta_{2|1})'P_{22}(\beta_2-\hat\beta_{2|1})\right|^{-T/2} \qquad \text{(A.51)} \end{aligned}$$

Marginal posterior of β_2 From (59) and (62) we obtain:

$$\begin{aligned} p\left(\beta_2 \mid Y\right) &= \int p(\alpha,\beta_2 \mid Y)d\alpha \\ &\propto \left|\beta' Y_{-1}'Y_{-1}\beta\right|^{-k/2}\left|\Delta Y' M_{Y_{-1}\beta}\Delta Y\right|^{-(T-r)/2}. \qquad \text{(A.52)} \end{aligned}$$

The second factor in (75) can be written as:

$$\left|\Delta Y' M_{Y_{-1}\beta}\Delta Y\right| = \frac{\left|\beta' Y_{-1}'M_{\Delta Y}Y_{-1}\beta\right|\left|\Delta Y'\Delta Y\right|}{\left|\beta' Y_{-1}'Y_{-1}\beta\right|} \propto \frac{\left|\beta' Y_{-1}'M_{\Delta Y}Y_{-1}\beta\right|}{\left|\beta' Y_{-1}'Y_{-1}\beta\right|}. \qquad \text{(A.53)}$$

Inserting (76) in (75), we obtain:

$$p\left(\beta_2 \mid Y\right) \propto \frac{\left|\beta' Y_{-1}'M_{\Delta Y}Y_{-1}\beta\right|^{-(T-r)/2}}{\left|\beta' Y_{-1}'Y_{-1}\beta\right|^{-(T-k-r)/2}}. \qquad \text{(A.54)}$$

We next analyze the right hand side of (77) as function of β_2 using the identification restrictions: $\beta = (I\ \beta_2')'$, hence $Y_{-1}\beta = Y_{-1,1} + Y_{-1,2}\beta_2$ and thus the denominator becomes:

$$\beta' Y_{-1}' Y_{-1} \beta = (Y_{-1,1} + Y_{-1,2}\beta_2)' (Y_{-1,1} + Y_{-1,2}\beta_2) . \tag{A.55}$$

Using similar results for completing the squares on β_2 in the denominator of (77) yields:

$$\beta' Y_{-1}' Y_{-1} \beta = Y_{-1,1}' M_{Y_{-1,2}} Y_{-1,1} + (\beta_2 - \bar{\beta}_2)' Y_{-1,2}' Y_{-1,2} (\beta_2 - \bar{\beta}_2) \tag{A.56}$$

where

$$\bar{\beta}_2 = -(Y_{-1,2}' Y_{-1,2})^{-1} Y_{-1,2}' Y_{-1,1}. \tag{A.57}$$

Analogously, completing the squares on β_2 in the numerator of (77) yields

$$\begin{aligned} \beta' Y_{-1}' M_{\Delta Y} Y_{-1} \beta &= Y_{-1,1}' M_{\Delta Y\, Y_{-1,2}} Y_{-1,1} \\ &\quad + (\beta_2 - \tilde{\beta}_2)' Y_{-1,2}' M_{\Delta Y} Y_{-1,2} (\beta_2 - \tilde{\beta}_2), \end{aligned} \tag{A.58}$$

where

$$\tilde{\beta}_2 = -(Y_{-1,2}' M_{\Delta Y} Y_{-1,2})^{-1} Y_{-1,2}' M_{\Delta Y} Y_{-1,1}. \tag{A.59}$$

Using these two decompositions, the marginal posterior density of β_2 in (77) is:

$$p(\beta_2 \mid Y) \propto \frac{\left| Y_{-1,1}' M_{\Delta Y\, Y_{-1,2}} Y_{-1,1} + (\beta_2 - \tilde{\beta}_2)' Y_{-1,2}' M_{\Delta Y} Y_{-1,2} (\beta_2 - \tilde{\beta}_2) \right|^{-\frac{T-r}{2}}}{\left| Y_{-1,1}' M_{Y_{-1,2}} Y_{-1,1} + (\beta_2 - \bar{\beta}_2)' Y_{-1,2}' Y_{-1,2} (\beta_2 - \bar{\beta}_2) \right|^{-\frac{T-k-r}{2}}}, \tag{A.60}$$

where $\bar{\beta}_2$ and $\tilde{\beta}_2$ are defined in (80) and (82), respectively.

Proposition *Given the standard form of a cointegration model under linear normalization and using a diffuse class of priors, the marginal posterior of the cointegration parameters β_2 is proportional to a matric-variate t density times a polynomial in β_2:*

$$\begin{aligned} p(\beta_2 \mid Y) \propto & p_{Mt}(\beta_2 | \tilde{\beta}_2, Y_{-1,2}' M_{\Delta Y} Y_{-1,2}, Y_{-1,1}' M_{\Delta Y\, Y_{-1,2}} Y_{-1,1}, T-r) \\ & \times \left| Y_{-1,1}' M_{Y_{-1,2}} Y_{-1,1} + (\beta_2 - \bar{\beta}_2)' Y_{-1,2}' Y_{-1,2} (\beta_2 - \bar{\beta}_2) \right|^{(T-k-r)/2} . \end{aligned} \tag{A.61}$$

Conditions that guarantee that this is a proper density are discussed next.

Existence of the marginal posterior of $\beta_2|Y$ We first rewrite (77) as follows

$$p(\beta_2 \mid Y) \propto \left| \beta' Y_{-1}' Y_{-1} \beta \right|^{-k/2} \left(\frac{\left| \beta' Y_{-1}' M_{\Delta Y} Y_{-1} \beta \right|}{\left| \beta' Y_{-1}' Y_{-1} \beta \right|} \right)^{-(T-r)/2}, \tag{A.62}$$

where the second factor is a matrix generalization of a Rayleigh quotient. Similarly to the vector case, also in this matrix case the quotient is bounded from above and below by functions of the given data which we can show by defining

$$B = (Y'_{-1}Y_{-1})^{1/2}\beta(\beta'Y'_{-1}Y_{-1}\beta)^{-1/2} \tag{A.63}$$

$$W = (Y'_{-1}Y_{-1})^{-1/2}Y'_{-1}M_{\Delta Y}Y_{-1}(Y'_{-1}Y_{-1})^{-1/2} \tag{A.64}$$

and rewriting the quotient as:

$$\begin{aligned}\frac{\left|\beta'Y'_{-1}M_{\Delta Y}Y_{-1}\beta\right|}{\left|\beta'Y'_{-1}Y_{-1}\beta\right|} &= \left|\beta'Y'_{-1}Y_{-1}\beta\right|^{-1/2}\left|\beta'Y'_{-1}M_{\Delta Y}Y_{-1}\beta\right|\left|\beta'Y'_{-1}Y_{-1}\beta\right|^{-1/2} \\ &= \left|(\beta'Y'_{-1}Y_{-1}\beta)^{-1/2}\beta'Y'_{-1}M_{\Delta Y}Y_{-1}\beta(\beta'Y'_{-1}Y_{-1}\beta)^{-1/2}\right| \\ &= |B'WB|\,. \end{aligned} \tag{A.65}$$

Since it holds that $B'B = I_r$ we can apply Theorem 11.15 from Magnus and Neudecker (1995) which states that $|B'WB|$ is bounded having the product of the r smallest eigenvalues of the matrix W as its lower bound and the product of the r largest eigenvalues of W as its upper bound.
Therefore, integrability of the function (85) depends on the integrability of the factor $\left|\beta'Y'_{-1}Y_{-1}\beta\right|^{-k/2}$. Using (79) we rewrite the integral as

$$\begin{aligned}&\int \left|\beta'Y'_{-1}Y_{-1}\beta\right|^{-k/2} d\beta_2 = \\ &\qquad \int \left|Y'_{-1,1}M_{Y_{-1,2}}Y_{-1,1} + (\beta_2 - \bar{\beta}_2)'Y'_{-1,2}Y_{-1,2}(\beta_2 - \bar{\beta}_2)\right|^{-k/2} d\beta_2. \end{aligned} \tag{A.66}$$

The integrand is proportional to a matric-variate t density with r degrees of freedom which exists under the condition that $k > (k-r) + r - 1 = k - 1$.

Proposition *Given the standard form of a cointegration model under linear normalization and using a diffuse class of priors, the marginal posterior distribution of the cointegration parameters β_2, with density (89), exists independent of the cointegrating rank r, but no first or higher order moments exist.*

It is noteworthy that this result is also independent of the difference $k - r$. We come back to this point below. This result extends the analysis and results of Kleibergen and Van Dijk (1994). Note that the choice of the prior parameter h does not play a role in the existence condition for the function (85).

Marginal posterior of β_2 for $k = 2$, $r = 1$ For the special case $k = 2$, $r = 1$, positive definiteness of the left hand side of (89) is trivial if for convenience the data matrices are scaled and rotated such that $Y'_{-1}Y_{-1} = I_k$:

$$\int \left|\beta'Y'_{-1}Y_{-1}\beta\right|^{-k/2} d\beta_2 \quad = \quad \int \left(1 + \beta_2^2\right)^{-1} d\beta_2 \tag{A.67}$$

The integrand is proportional to a Cauchy density. Hence, the integral is finite and the marginal posterior of β_2 exists but no finite first or higher order moments.

Marginal posterior of α Using equations (36) and (51), one can derive the marginal posterior of α as:

$$\begin{aligned}
p(\alpha \mid Y) &= \int p(\alpha, \beta \mid Y)\, d\beta_2 \\
&\propto |P|^{(T-r)/2} |Q|^{k/2} p_{Mt}\left(\beta_1 \mid \hat{\beta}_1, P_{11.2}, Q, T-k+r\right) && \text{(A.68)} \\
&= |P|^{(T-r)/2} |Q|^{k/2} |P_{11.2}|^{r/2} |Q|^{(T-k+r-r)/2} \\
&\quad \times \left|Q + (\beta_1 - \hat{\beta}_1)' P_{11.2} (\beta_1 - \hat{\beta}_1)\right|^{-(T-k+r)/2} && \text{(A.69)} \\
&\overset{\beta_1 \equiv I_r}{=} |P|^{(T-r)/2} |Q|^{T/2} |P_{11} - P_{12} P_{22}^{-1} P_{21}|^{r/2} \\
&\quad \times \left|Q + (I_r - \hat{\beta}_1)' (P_{11} - P_{12} P_{22}^{-1} P_{21}) (I_r - \hat{\beta}_1)\right|^{-(T-k+r)/2} && \text{(A.70)} \\
&= |P|^{T/2} |P_{22}|^{-r/2} |Q|^{T/2} \\
&\quad \times \left|Q + (I_r - \hat{\beta}_1)' (P_{11} - P_{12} P_{22}^{-1} P_{21}) (I_r - \hat{\beta}_1)\right|^{-(T-k+r)/2} && \text{(A.71)}
\end{aligned}$$

where we have used $|P_{11.2}| = |P_{11} - P_{12} P_{22}^{-1} P_{21}| = |P|/|P_{22}|$.

Proposition *Given the standard form of a cointegration model under linear normalization and using a diffuse class of priors, the marginal posterior density of the adjustment parameters α is a rational function of α, given as*:

$$p(\alpha \mid Y) \propto |P|^{T/2} |P_{22}|^{-r/2} |Q|^{T/2} \left|Q + (I_r - \hat{\beta}_1)' P_{11.2} (I_r - \hat{\beta}_1)\right|^{-(T-k+r)/2}, \tag{A.72}$$

and this density is not proportional to a known form of densities.

Existence of the marginal posterior of α given Y For the existence of the distribution with density (95), we first show that the first two factors in the right hand side of (95) are bounded. Consider:

$$|P_{22}|^{-r/2} |P|^{T/2} = |(Y_{-1}' M_{\Delta Y \alpha_\perp} Y_{-1})_{\{l,k-r\}}|^{-r/2} |Y_{-1}' M_{\Delta Y \alpha_\perp} Y_{-1}|^{T/2}$$

where $A_{\{l,b\}}$ denotes the $b \times b$ lower-right minor of matrix A. From Theorem 11.16 in Magnus and Neudecker (1995) we have that $|(Y_{-1}' M_{\Delta Y \alpha_\perp} Y_{-1})_{\{l,k-r\}}|$ has its lower bound equal to the product of the $k-r$ smallest eigenvalues of $Y_{-1}' M_{\Delta Y \alpha_\perp} Y_{-1}$ and its upper bound is equal to the product of the $k-r$ largest eigenvalues. Note that the matrix $Y_{-1}' M_{\Delta Y \alpha_\perp} Y_{-1}$ is positive definite in the typical set up of econometrics, *e.g.* $\operatorname{rank}(M_{\Delta Y \alpha_\perp}) = T - \operatorname{rank}(\Delta Y \alpha_\perp) \geq T - r$ and hence these products of eigenvalues are bounded.

Using (47) we have

$$|Y'_{-1}M_{\Delta Y\alpha_\perp}Y_{-1}| = \frac{|\alpha'_\perp \Delta Y' M_{Y_{-1}} \Delta Y \alpha_\perp|}{|\alpha'_\perp \Delta Y' \Delta Y \alpha_\perp|} |Y'_{-1}Y_{-1}| \tag{A.73}$$

of which the last factor is constant given the data and the first factor is bounded by products of the r smallest and largest eigenvalues of

$$(\Delta Y'\Delta Y)^{-1/2}\Delta Y' M_{Y_{-1}} \Delta Y (\Delta Y'\Delta Y)^{-1/2} \tag{A.74}$$

by similar arguments as in subsection A.3.2.
The density in (95) integrates to a finite value if the product of the last two factors $|Q|^{T/2}\left|Q + (I_r - \hat{\beta}_1)' P_{11.2}(I_r - \hat{\beta}_1)\right|^{-(T-k+r)/2}$ has a finite integral. Note again that Q here is a function of α with $Q = (\alpha' D^{-1}\alpha)^{-1}$.
Since $(I_r - \hat{\beta}_1)' P_{11.2}(I_r - \hat{\beta}_1)$ is positive semidefinite and therefore

$$|Q| \leq \left|Q + (I_r - \hat{\beta}_1)' P_{11.2}(I_r - \hat{\beta}_1)\right| \tag{A.75}$$

we have that

$$\begin{aligned} |Q|^{T/2}\left|Q + (I_r - \hat{\beta}_1)' P_{11.2}(I_r - \hat{\beta}_1)\right|^{-(T-k+r)/2} &\leq |Q|^{T/2}\,|Q|^{-(T-k+r)/2} \\ &= |Q|^{(k-r)/2} \end{aligned} \tag{A.76}$$

So the integral of the product of these factors is bounded by $\int \left|\alpha' D^{-1}\alpha\right|^{-(k-r)/2} d\alpha$. Hence, a sufficient condition for the existence of the posterior of α is:

$$\int \left|\alpha' D^{-1}\alpha\right|^{-(k-r)/2} d\alpha < \infty. \tag{A.77}$$

The integrand has an asymptote at $\alpha = 0_{(k\times r)}$. We analyze two shape features of the posterior density: an asymptote in the interior and tail behavior when α tends to infinity. We show that the determinant in (100) is integrable around $\alpha = 0$ despite the asymptote at $\alpha = 0_{(k\times r)}$ and that the tails of the posterior are integrable. This is analyzed in the main text for the 2-dimensional case. For completeness we discuss here the marginal posterior density of $\alpha|Y$ for the general vector and the matrix case.

General vector case $r = 1$ We consider the case of $r = 1$ but we relax the restriction $k = 2$. First we focus on the parameter space around the origin (where the asymptote is located). Regard $h(\alpha'\alpha) = |\alpha'\alpha|^{-(k-r)/2} = (\alpha'\alpha)^{-(k-1)/2}$ as the kernel of a spherical density. Following e.g. theorems 1.5.5 and 2.1.3 from

Muirhead (1982) this can be transformed to polar coordinates as

$$\begin{aligned}
\alpha_1 &= \lambda \cos\theta_1 \\
\alpha_2 &= \lambda \sin\theta_1 \cos\theta_2 \\
\alpha_3 &= \lambda \sin\theta_1 \sin\theta_2 \cos\theta_3 \\
&\vdots \\
\alpha_{k-1} &= \lambda \sin\theta_1 \sin\theta_2 \ldots \sin\theta_{k-2} \cos\theta_{k-1} \\
\alpha_k &= \lambda \sin\theta_1 \sin\theta_2 \ldots \sin\theta_{k-2} \sin\theta_{k-1}
\end{aligned} \tag{A.78}$$

with $\theta \in \Theta \subset \mathbb{R}^{k-1}$ with $\Theta = \{\theta : \theta_{k-1} \in (0, 2\pi], \theta_i \in (0, \pi] \text{ for } i \neq k-1\}$ and $\lambda > 0$, such that $\lambda^2 = \alpha'\alpha$ and the Jacobian is given by $|J| = \lambda^{k-1} \prod_{i=1}^{k-2} \sin^{k-1-i}\theta_i$. The $\lambda, \theta_1, \ldots, \theta_{k-1}$ are independent. All θ_i have bounded density functions on a bounded support and can therefore be integrated out of the joint density resulting in a factor $2\pi^{k/2}/\Gamma(k/2)$. Hence the integral $h(\alpha'\alpha)$ over a ball with radius R around the origin (where the asymptote is located) can be expressed as

$$\begin{aligned}
\int_{\alpha'\alpha \leq R^2} h(\alpha'\alpha) d\alpha &= \int_{\lambda=0}^{R} \int_{\theta \in \Theta} h(\lambda^2) |J| d\theta d\lambda \\
&= \int_{\lambda=0}^{R} (\lambda^2)^{-(k-1)/2} \lambda^{k-1} \int_{\theta \in \Theta} \prod_{i=1}^{k-2} \sin^{k-1-i}\theta_i d\theta d\lambda \\
&= \frac{2\pi^{k/2}}{\Gamma(k/2)} \int_0^R 1 d\lambda = \frac{2\pi^{k/2} R}{\Gamma(k/2)}.
\end{aligned} \tag{A.79}$$

Note that the existence of this expression does not depend on k and that it is equal to R times the surface area of a unit sphere in $\mathbb{R}^k$.
So also in the general vector case the asymptote poses no problems and for any finite R the integral is bounded.
If however R tends to ∞ the integral in (102) again goes to ∞ at a rate R so that the *sufficient* condition is not satisfied then. However, the tails are integrable and the marginal posterior of α is proper. The easiest way to see this is as follows. We have shown in equation (89) that the marginal posterior of β_2 is proper but it has no first or higher order moments. Further, the conditional posterior of α given β_2 is proper for each value of β_2, see (62) and (81). Therefore, the joint posterior of (α, β_2) is proper. We could simulate α from its (marginal) posterior by simulating β_2 from its marginal posterior and simulating α given the draw of β_2. We emphasize that the line of reasoning to show that the tails are integrable is a general one. That is, it holds for the bivariate case, the general vector case and the matrix case.

Matrix case For the analysis of the asymptote in the matrix case we can use the transformation between α and its singular value decomposition $\alpha = USV'$ where U is a $k \times r$ semi-orthogonal matrix with $U'U = I_r$, V is a orthogonal

$r \times r$ matrix with $V'V = I_r$ and S is a $r \times r$ diagonal matrix with λ_i, $i = 1...r$, as diagonal elements. The λ_i's denote the singular values in descending order, that is $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$. Rennie (2006) shows using wedge product notation that the Jacobian of the transformation is proportional to

$$|J| \propto \prod_{i<j\leq r} (\lambda_i^2 - \lambda_j^2) \prod_{i\leq r} \lambda_i^{k-r}. \tag{A.80}$$

up to volume elements[16] of both the Stiefel manifold $\mathbb{V}_{k,r} = \{U \in \mathbb{R}^{k\times r} : U'U = I_r\}$ related to U and the orthogonal group $\mathbb{O}_r = \mathbb{V}_{r,r} = \{V \in \mathbb{R}^{r\times r} : V'V = I_r\}$ related to V. Note that the singular values of α by definition are equal to the square roots of the eigenvalues of $\alpha'\alpha$ and hence its determinant occurring in the integrand is equal to the product of the squared singular values λ_i, that is $|\alpha'\alpha| = \prod_{i=1}^{r} \lambda_i^2$.

Finally, we consider the area of integration $\|\alpha'\alpha\|_2 \leq R$ around the asymptote where $\|\cdot\|_2$ denotes the spectral norm which by definition equals the largest singular value (which is λ_1 in our case), that is $\lambda \in \Lambda_R = \{\lambda \in \mathbb{R}^r : \lambda : 0 \leq \lambda_r \leq \lambda_{r-1} \leq \lambda_1 \leq R\}$. Note that this is a consistent generalization of the restriction $\alpha'\alpha \leq R^2$ in terms of the Euclidean dot product in the vector case. Note also that the Frobenius norm (square root of the sum of the squared elements which equals the sum of the singular values) would be an equally valid generalization. The integral can thus be expressed as

$$\begin{aligned}
\int_{\|\alpha\|_2\leq R} |\alpha'\alpha|^{-(k-r)/2} d\alpha =& 2^{-r} \int_{U'U=I} \int_{V'V=I} \int_{\Lambda_R} \prod_{i=1}^{r} (\lambda_i^2)^{-(k-r)/2} |J| d\lambda dV dU \\
=& 2^{-r} \operatorname{Vol}(\mathbb{O}_r) \operatorname{Vol}(\mathbb{V}_{k,r}) \int_{\Lambda_R} \prod_{i<j\leq r} (\lambda_i^2 - \lambda_j^2) d\lambda \\
=& \frac{2^r \pi^{r(k+r)/2}}{\Gamma_r(k/2)\Gamma_r(r/2)} \int_{\Lambda_R} \prod_{i<j\leq r} (\lambda_i^2 - \lambda_j^2) d\lambda
\end{aligned} \tag{A.81}$$

using the fact that the volume of the Stiefel manifolds is given by $\operatorname{Vol}(\mathbb{V}_{k,r}) = \frac{2^r \pi^{kr/2}}{\Gamma_r(k/2)}$ and $\operatorname{Vol}(\mathbb{O}_r) = \operatorname{Vol}(\mathbb{V}_{r,r}) = \frac{2^r \pi^{r^2/2}}{\Gamma_r(r/2)}$ and the factor 2^{-r} arises because of the uniqueness of the singular value decomposition up to simultaneous sign changes of corresponding columns of U and V which could be enforced for instance by imposing a positive sign for the first nonzero entry in each column of U. In the special case $r = 1$ the product in the integrand is empty and the integrand becomes equal to 1, and therefore (104) simplifies further.

The integrand is a polynomial in the λ_i's and the area of integration is bounded. Hence, we conclude the integral over this bounded region is finite despite the fact that it contains an asymptote at $|\alpha'\alpha| = 0$.

For the analysis of the tail behaviour of the posterior density of the matrix α

[16] For ease of exposition we use this slightly less formal notation omitting wedge products in the intermediate step at the right hand side at the top of (104). See e.g. Muirhead (1982) or Rennie (2006) for formal wedge product notation.

we refer to the text presented after the vector case. Also in the matrix case, the sufficient condition is not satisfied. However, the marginal posterior of α is proper as explained before. We note that the marginal posterior density of α is a matrix polynomial and can have all kinds of shapes. To analyze these shapes is more a computational topic and is outside the scope of this appendix.

A.3.3. Likelihood shape and posterior existence in an instrumental variable model

In this section we summarize the derivation of the posterior results of an IV model. The scheme of derivations is given in Figure A.8. A well-known way to specify the instrumental variable model (also known as the Incomplete Simultaneous Equation model, see Zellner et al. (1988) is to write a standard regression equation where the right hand side variables are possibly endogenous and to add a second equation where these right hand side, so-called instrumental, variables are linked to exogenous variables. Specifically, the matrix representation of this model reads

$$y = X\beta + u, \tag{A.82}$$
$$X = Z\Pi + V, \tag{A.83}$$

where y is a $T \times 1$ dimensional vector of observations on economic variables at time t; β is a $r \times 1$ vector of parameters belonging to the T observations on the r possibly endogenous variables arranged in the matrix X ; the disturbances ε_t for $t = 1, \ldots, T$ have independent Gaussian distributions with Σ as a positive definite symmetric (PDS) parameter matrix. The observations on X are connected to T observations on k exogenous (or predetermined) variables arranged in the matrix Z through the matrix Π which is $k \times r$ with usually k the number of instrumental variables greater or equal to the number r of endogenous variables. This condition plays a central role in the analysis. We assume that the data matrix $(y \;\; X \;\; Z)$ has full column rank $m + k + 1$.
The marginal posterior density of $p(\beta, \Pi|Y)$ under a flat prior is given in Figure A.8 in the middle. Highly non-elliptical posterior shapes may result from the local non-identification of β if Π does not have full column rank, which is easily seen from the restricted reduced form given in Figure A.8 and from the motivating examples in Section 2.
The marginal posterior densities of β and Π were derived by Bauwens and Van Dijk (1990) and Kleibergen and Van Dijk (1998), respectively. In Zellner et al. (2014) the following result is shown.

Proposition *Given the standard form of an instrumental variable model under linear normalization and using a diffuse class of priors, the marginal posterior of the parameters β is proportional to a multivariate t density times a polynomial in β, while the marginal posterior of the parameters Π is proportional to a matrix t density times a rational function in Π. Both densities are improper for $k \leq r$ (exact or under-identification) and proper densities for $k > r$*

Fig A.8: Derivation scheme for posterior densities of an IV Model with r endogenous variables and k instruments, under a diffuse prior

Model and posterior

$$(Y \;\; X) = Z(0 \;\; \Pi)\begin{pmatrix}1 & 0' \\ \beta & I_r\end{pmatrix} + (u \;\; V)\begin{pmatrix}1 & 0' \\ \beta & I_r\end{pmatrix} \quad (u \;\; V) \sim N(0, \Sigma \otimes I_T)$$

.

Parameter matrix has a triangular structure

$$(Y \;\; X) = Z\Pi(\beta \;\; I_r) + (u + \beta V \;\; V)$$

Identification restriction is linear normalisation

β is a r vector, Π is $k \times r$

posterior has a ridge at $\Pi = 0$*, joint density is improper for* $r \geq k$ *and proper for* $k > r$

Conditional posteriors

$p(\beta, \Pi, \Sigma \mid \text{data}), \quad \text{data} = \{X, Y, Z\}$

↓	↓	↓
complete sum of squares in β	complete sum of squares in Π	use Inverse-Wishart distribution
↓	↓	↓
$p(\beta \mid \Pi, \Sigma, \text{data}) \propto$ Normal density	$p(\Pi \mid \beta, \Sigma, \text{data}) \propto$ multivariate Normal density	$p(\Sigma \mid \beta, \Pi, \text{data}) \propto$ inverse-Wishart density

Conditional moments of $p(\beta|\Pi, \Sigma, data)$*,* $p(\Pi|\beta, \Sigma, data)$ *and* $p(\Sigma|\beta, \Pi, data)$ *exist for all values of* Π *in their domain and for any number of instruments,* $k = 1, 2, \ldots, K$*.*

Marginal posteriors of β and Π

$p(\beta, \Pi, \Sigma \mid \text{data})$

↓

Inverse-Wishart step on Σ

↓

$p(\beta, \Pi \mid \text{data}) \propto \left|(u \;\; V)'(u \;\; V)\right|^{-T/2}$ for $u = Y - X\beta$, $V = X - Z\Pi$

↓	↓
apply determinant decomposition $\left\|(u \;\; V)'(u \;\; V)\right\| = (u'u)(V'M_uV)$	apply determinant decomposition $\left\|(u \;\; V)'(u \;\; V)\right\| = (V'V)(u'M_Vu)$
↓	↓
complete sum of squares on Π	complete sum of squares on β
↓	↓
$p(\Pi \mid \beta, \text{data}) \propto$ a matrix t-density *Conditional moments exist for all values of* β *in its domain.*	$p(\beta \mid \Pi, \text{data}) \propto$ a multivariate t-density *The conditional posterior of* β *given* Π *does not exist for* $\Pi = 0$*.*
↓	↓
matrix t-density step on Π	multivariate t-density step on β
↓	↓
use matrix decomposition and properties of the projection matrix:	use matrix decomposition and properties of the projection matrix:
↓	↓
$p(\beta \mid data)$ *is proportional to a multivariate* t*-density times a polynomial in* β*. It is an improper density for an exactly identified model* $(r = k)$*; and a proper density for an overidentified model* $(k > r)$*.*	$p(\Pi \mid data)$ *is proportional to matrix* t*-density times a rational function in* Π*. It is an improper density for an exactly identified model* $(r = k)$*; and a proper density for an overidentified model* $(k > r)$*.*

(over-identification).

We conclude that the use of the triangular structure on the parameters and the zero restrictions specified in the first box shown in Figure A.8 make that the posteriors in an instrumental variable regression model are proper densities given enough instruments.

A.3.4. Likelihood shape and posterior existence in a static factor model

In this section we summarize the derivation of the results. A basic static factor model can be specified in matrix notation as follows:

$$Y = F\Lambda + E, \tag{A.84}$$
$$F = 0 + U, \tag{A.85}$$

where Y is the $T \times p$ matrix of observations, F is the $T \times r$ matrix of factors, Λ is the $r \times p$ matrix of factor loadings, E is the $T \times p$ matrix of disturbances and U is the $T \times r$ matrix of disturbances. In addition, $\text{cor}(U, E) = 0$, $E \sim MN(0, \Sigma, I_T)$ and $U \sim MN(0, I_p, I_T)$. In this notation, $MN(X, \Omega, \Phi)$ denotes the matric-variate normal distribution with mean M and scale parameters Ω, Φ, and I_k is the $k \times k$ identity matrix.
In Figure A.9 we show this model and summarize the derivation steps to obtain the marginal posteriors of the parameters. It can be easily seen that this static factor without the normal prior on F and the restriction on the matrix Σ will lead to identical results for the shape of the posterior densities and existence of moments as for the case of the parameters of the cointegration model. Thus, we state the following well-known result:

Proposition *Given the standard form of an static factor model, a normal prior on the factors F and a diagonal matrix $\Sigma = D$ of the disturbances, and given a linear normalization with a diffuse class of priors, the marginal posterior of the parameters F is proportional to a polynomial in F multiplied by an exponential function in F, while the marginal posterior of the loading parameters Λ_2 is proportional to a polynomial function in Λ_2. Both densities are proper, since the exponential tails of the normal density of F dominate the polynomial tails of t type densities. The restriction of the diagonal covariance matrix Σ leads to proper posterior density of the matrix of factor loadings Λ_2.*

We emphasize that a static factor model analysis using linear normalization is subject to the criticism, already listed Section 3.1, in that estimation results depend upon the ordering of the factors. Using the Lasso type prior with orthogonal normalization, the empirical results are now independent of the factor ordering. Details are given in Baştürk et al. (2017). The topic of finding a factor normalization that leads to results that are independent of the factor ordering is extensively studied nowadays. We refer to only a few papers, such as Kaufmann and Schumacher (2013) and Chan et al. (2017) and the references cited there.

Fig A.9: Derivation scheme for posterior densities of a static factor model with k variables and $r << k$ factors under a diffuse prior.

Model and posterior

$$Y = I_T F\Lambda + U,\ U \sim MN(0, \Sigma, I_T)$$
$$F = 0 + V,\ V \sim MN(0, I_r, I_T)$$

Identification restriction is linear normalization on Λ and diagonal matrix on Σ

$$Y = I_T F(I_r\ \Lambda_2) + U,\ \Lambda_2 \text{ is } r \times (k-r),\ F \text{ is } T \times r$$

posterior has ridge at $F = 0$*, joint density is proper*

Conditional posteriors

$p(F, \Lambda_2, \Sigma = D \mid \text{data}),\ \text{data} = \{Y\}$

↓	↓	↓
complete sum of squares in Λ_2	complete sum of squares in F	use Inverse-Gamma dist.
↓	↓	↓
$p(\Lambda_2 \mid F, \Sigma = D, \text{data}) \propto$ matrix Normal density	$p(F \mid \Lambda_2, \Sigma = D, \text{data}) \propto$ matrix Normal density	$p(\Sigma = D \mid F, \Lambda_2, \text{data}) \propto$ inverse-Gamma densities

Conditional moments of $p(F \mid \Lambda_2, \Sigma = D, data)$*,* $p(\Lambda_2 \mid F, \Sigma = D, data)$ *and* $p(\Sigma = D \mid F, \Lambda_2, data)$ *exist for all values of the conditioning parameters in their domains and for all finite* k *and* r*.*

Marginal posteriors of Λ_2 and F

$p(F, \Lambda_2, \Sigma = D \mid \text{data})$

↓

Inverse-Gamma steps on $\Sigma = D$

↓

$p(F, \Lambda_2 \mid \text{data}) \propto \exp\left(-\frac{1}{2}\text{tr}(F'F)\Pi_{k=1}^{K}\left[(y_k - F\lambda_k)'(y_k - F\lambda_k)\right]^{-T/2}\right)$

↓	↓
complete sum of squares on F in two steps	complete sum of squares on the columns of Λ
↓	↓
$p(F \mid \Lambda_2, \text{data}) \propto$ matrix t density times an exponential function in F *conditional moments exist for all values of* Λ_2 *in its domain*	$p(\lambda_k \mid F, \text{data}) \propto$ a multivariate t density *conditional moments exist for all values of* F *in its domain*
↓	↓
no further analytical integration step on F	product of t-density steps on the columns of Λ_2
↓	↓
$p(\Lambda_2 \mid \text{data})$ *is not member of a known class of densities, but it is proper due to the diagonal matrix* $\Sigma = D$	$p(F \mid \text{data})$ *is a polynomial in* F *times an exponential function in* F *and it is proper due to the normal prior on* F*.*

A.4. Appendix for Section 4: Regularization priors

Prior choice and existence of posterior moments In the specification $\Pi = \beta\Lambda\alpha'$ uniform priors can be specified for α en β on their respective Stiefel manifolds. For Σ we again specify a diffuse prior and we assume all marginal prior to be independent, that is

$$p(\beta, \alpha, \lambda, \Sigma) = p(\beta)p(\alpha)p(\lambda)p(\Sigma) \tag{A.86}$$

with

$$p(\alpha) \propto \quad 1 \quad \text{if } \alpha'\alpha = I_r,\ 0 \text{ otherwise} \tag{A.87}$$
$$p(\beta) \propto \quad 1 \quad \text{if } \beta'\beta = I_r,\ 0 \text{ otherwise} \tag{A.88}$$
$$p(\Sigma) = \quad |\Sigma|^{-h/2} \quad \text{if } \Sigma \text{ is symmetric and positive definite} \tag{A.89}$$

and we use again the specific case of $h = k+1$. We now discuss the choice of the prior on the singular values λ in more detail when we explore the integrability of the posterior in relation to this prior $p(\lambda)$.
Due to the similarity in the prior and the likelihood, with the distinction (i) that all elements of β, not only elements of β_2, are now random variables, (ii) that we now have $\alpha^* = \alpha\Lambda$ instead of α and (iii) that we now include the prior $p(\lambda)$ (which is independent of the priors on the other parameters), we can write

$$\begin{aligned} p(\alpha, \beta, \lambda) &\propto \left|(\Delta Y - Y_{-1}\beta\Lambda\alpha')'(\Delta Y - Y_{-1}\beta\Lambda\alpha')\right|^{-T/2} p(\lambda) \\ &= \left|\Delta Y' M_{Y_{-1}\beta\Lambda}\Delta Y + (\alpha - \hat{\alpha})\left(\Lambda\beta' Y'_{-1}Y_{-1}\beta\Lambda\right)(\alpha - \hat{\alpha})'\right|^{-T/2} p(\lambda). \end{aligned} \tag{A.90}$$

When we integrate this posterior with respect to α over the manifold $\alpha'\alpha = I_r$ we can derive the following bound:

$$\begin{aligned} &\int_{\alpha'\alpha=I} p(\alpha, \beta, \lambda|Y)d\alpha \\ &\quad = \int_{\alpha'\alpha=I} \left|\Delta Y' M_{Y_{-1}\beta\Lambda}\Delta Y + (\alpha - \hat{\alpha})\left(\Lambda\beta' Y'_{-1}Y_{-1}\beta\Lambda\right)(\alpha - \hat{\alpha})'\right|^{-T/2} p(\lambda)d\alpha \\ &\quad \leq \int_{\alpha'\alpha=I} \left|\Delta Y' M_{Y_{-1}\beta\Lambda}\Delta Y\right|^{-T/2} p(\lambda)d\alpha \\ &\quad = \left|\Delta Y' M_{Y_{-1}\beta\Lambda}\Delta Y\right|^{-T/2} p(\lambda) \int_{\alpha'\alpha=I} d\alpha \\ &\quad \propto \left(\frac{\left|\Lambda\beta' Y'_{-1}M_{\Delta Y}Y_{-1}\beta\Lambda\right|}{\left|\Lambda\beta' Y'_{-1}Y_{-1}\beta\Lambda\right|}\right)^{-T/2} p(\lambda)\,\mathrm{Vol}(\mathbb{V}_{k,r}) \\ &\quad = \left(\frac{\left|\beta' Y'_{-1}M_{\Delta Y}Y_{-1}\beta\right|}{\left|\beta' Y'_{-1}Y_{-1}\beta\right|}\right)^{-T/2} p(\lambda)\,\mathrm{Vol}(\mathbb{V}_{k,r}), \end{aligned} \tag{A.91}$$

where the second to last step follows from (47) and the last step follows since both numerator and denominator are of the form $|\Lambda X \Lambda|$ for some matrix X

and can be written as $|\Lambda||X||\Lambda|$ such that the factors $|\Lambda|$ in numerator and denominator cancel against each other.

The first factor in (114) is bounded by products of eigenvalues similar to the previous section. The volume $\mathrm{Vol}(\mathbb{V}_{k,r})$ of the Stiefel manifold of $k \times r$ orthogonal matrices α is a finite constant. So integrability of (113) depends on $p(\lambda)$.

The alternative route via integrating over β proceeds as follows:

$$\begin{aligned}\int_{\beta'\beta=I} p(\alpha,\beta,\lambda|Y)d\beta &= \int_{\beta'\beta=I} \Big|\big(Y'_{-1}M_{\Delta Y\alpha_\perp}Y_{-1}\big)^{-1} \\ &\quad +(\beta-\hat{\beta}\Lambda^{-1})\Lambda\alpha' D^{-1}\alpha\Lambda(\beta-\hat{\beta}\Lambda^{-1})'\Big|^{-T/2} p(\lambda)d\beta \\ &\leq \int_{\beta'\beta=I} \Big|\big(Y'_{-1}M_{\Delta Y\alpha_\perp}Y_{-1}\big)^{-1}\Big|^{-T/2} p(\lambda)d\beta \\ &= \Big|\big(Y'_{-1}M_{\Delta Y\alpha_\perp}Y_{-1}\big)^{-1}\Big|^{-T/2} p(\lambda)\int_{\beta'\beta=I} d\beta \\ &\propto \left(\frac{|\alpha'_\perp\Delta Y' M_{Y_{-1}}\Delta Y\alpha_\perp|}{|\alpha'_\perp\Delta Y'\Delta Y\alpha_\perp|}\right)^{-T/2} p(\lambda)\,\mathrm{Vol}(\mathbb{V}_{k,r}), \quad \text{(A.92)}\end{aligned}$$

where the last step follows again from (47). The first factor, which is a function of α, is bounded by products of eigenvalues, and again integrability of (115) depends on $p(\lambda)$.

As a starting point for the specification of an uninformative prior, suppose that we specify a diffuse prior on Π on the manifold of $k \times k$ matrices with rank r, that is

$$p(\Pi) \propto 1 \;\text{ if }\; \mathrm{rank}(\Pi) = r\,, 0 \text{ otherwise} \tag{A.93}$$

then using the Jacobian of the transformation $\Pi = \beta\Lambda(\lambda)\alpha'$, we obtain that the implied prior for (α,β,λ) equals

$$\begin{aligned} p(\alpha,\beta,\lambda) &\propto p(\Pi(\alpha,\beta,\lambda))\left|\frac{\partial\,\mathrm{vec}(\Pi(\alpha,\beta,\lambda))}{\partial\,\mathrm{vec}(\alpha,\beta,\lambda)'}\right| &\text{(A.94)}\\ &\propto \prod_{i<j\leq r}(\lambda_i^2-\lambda_j^2)\prod_{i\leq r}\lambda_i^{k-r}. &\text{(A.95)}\end{aligned}$$

This implies independent priors on α, β and λ with α and β uniform similar to (110) and (111). However, the implied prior on the singular values

$$p(\lambda) = \prod_{i<j\leq r}(\lambda_i^2-\lambda_j^2)\prod_{i\leq r}\lambda_i^{k-r} \tag{A.96}$$

is not integrable as $\lambda_i \to \infty$. The factor $\prod_{i<j\leq r}(\lambda_i^2-\lambda_j^2)$ in (119) results from the ordering of the singular values that we assume in the singular value decomposition and regularizes the posterior by letting the prior go to 0 whenever two (or more) singular values λ_i and λ_j for $i \neq j$ are equal, because in that case the factor $\lambda_i^2-\lambda_j^2$ will equal 0.

We note that the other factor $\prod_{i\leq r}\lambda_i^{k-r}$ in (119) can be shown to correspond

to the embedding prior of Kleibergen and Paap (2002) on (α^*, β) conditional on the rank reduction under a flat prior on Π. Their prior is in that case given by

$$p(\beta, \alpha^*) \propto |\beta'\beta|^{(k-r)/2} |\alpha^{*\prime}\alpha^*|^{(k-r)/2}. \tag{A.97}$$

We can rewrite their prior adapted for our normalization using $\alpha^* = \alpha\Lambda$, $\alpha'\alpha = I_r$ and $\beta'\beta = I_r$ as

$$\begin{aligned} |\beta'\beta|^{(k-r)/2} |\alpha^{*\prime}\alpha^*|^{(k-r)/2} &= |\beta'\beta|^{(k-r)/2} |\Lambda\alpha'\alpha\Lambda|^{(k-r)/2} && \text{(A.98)} \\ &= |I_r|^{(k-r)/2} |\Lambda I_r \Lambda|^{(k-r)/2} = |\Lambda\Lambda|^{(k-r)/2} \\ &= |\Lambda|^{k-r} = \prod_{i=1}^{r} \lambda_i^{k-r}. && \text{(A.99)} \end{aligned}$$

The connection with the previous section is that this prior regularizes the vertical asymptote at $|\alpha^{*\prime}\alpha^*| = 0$ with $\alpha^* = \alpha\,\mathrm{diag}(\lambda)$.

We now try to specify an uninformative or weakly informative prior on λ using a more direct approach. Initially we disregard the ordering of singular values and we could then use the following approach. Since $\lambda_i > 0$ specifying a diffuse prior on $\log \lambda_i$ would correspond to $p(\lambda_i) \propto \lambda_i^{-1}$ which is analogous to a diffuse prior $p(\sigma^2) \propto \sigma^{-2}$ for a variance parameter σ^2. In this case the prior for the vector λ equals $p(\lambda) \propto \prod_{i=1}^{r} \lambda_i^{-1}$. Note that λ_i equals $(\alpha_i^{*\prime}\alpha_i^*)^{1/2}$ such that both correspond to the singular values of Π. The implied prior in the specification $\Pi = \beta\alpha^*$ is thus given by $p(\alpha^*) = \prod_{i=1}^{r} (\alpha_i^{*\prime}\alpha_i^*)^{-1/2} = |\alpha^{*\prime}\alpha^*|^{-1/2}$.

If we also include the ordering of the singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$ in the prior specification we first note that the singular values λ_i for all $i > 1$ are bounded by λ_{i-1}. So given λ_1 the other λ_i are jointly restricted to a bounded (hyper-)triangular region $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$. Conditional on λ_1 we specify a joint uniform prior on $(\lambda_2, \ldots, \lambda_r)$ on this support. Only the largest singular value λ_1 has infinite (conditionally upon the other λ_i) support and requires a prior of which the tails go to zero fast enough.

We now use the connection to a Dirichlet distribution to find a prior for λ_1 that is consistent with the other λ_i. In order to do this we transform the λ_i into the increments δ_i as follows:

$$\delta_i = \lambda_i - \lambda_{i+1}, \text{ for } i = 1, \ldots r \tag{A.100}$$

with inverse transformation $\lambda_i = \sum_{j=i}^{r} \delta_j$. Its Jacobian is given by

$$\left| \frac{\partial \lambda}{\partial \delta'} \right| = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ 0 & 1 & \ldots & 1 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{vmatrix} = 1, \tag{A.101}$$

which means that we can easily transform from λ to δ and vice versa. In particular a joint uniform distribution on λ_i also implies a joint uniform distribution on δ_i over the simplex on which it is defined, that is $\delta_i > 0$ and $\sum_{i=1}^{r} \delta_i = \lambda_1$.

A Dirichlet distribution $D(1,\ldots,1)$ also corresponds to a uniform distribution on a simplex. This Dirichlet distribution can be constructed from r i.i.d. random variables from an exponential distribution with any rate $\theta > 0$. Let $p(\delta_i) \sim \exp(-\delta_i\theta)$. Then

$$\frac{\delta_i}{\sum_{i=1}^{r}\delta_i} = \frac{\delta_i}{\lambda_1} \sim D(1,\ldots,1), \tag{A.102}$$

as required. We now also can find a prior for λ_1 that is fully consistent with the joint uniform prior on the λ_i for $i > 1$ on its support since all λ_i are derived from the same i.i.d. joint distribution of the increments δ_i:

$$\lambda_1 = \sum_{j=i}^{r}\delta_i \sim \text{Gamma}(r,\theta) \tag{A.103}$$

and its density is thus given by

$$p(\lambda_1) \propto \lambda_1^{r-1}\exp(-\lambda_i\theta). \tag{A.104}$$

The density of the uniform prior $p(\lambda_2,\ldots\lambda_r|\lambda_1)$ equals the inverse of the volume of the simplex on which it is defined. The volume equals $\lambda_1^{r-1}/(r-1)!$ where the factor $1/(r-1)!$ results from the Jacobian of the transformation from (unit) box to (unit) simplex. So the joint prior of λ on its support $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$ in this approach is thus given by

$$p(\lambda) \quad = \quad p(\lambda_2,\ldots\lambda_r|\lambda_1)p(\lambda_1) \propto \exp(-\lambda_i\theta). \tag{A.105}$$

We can summarize the results from this section as follows. Using the parametrization $\Pi = \beta\Lambda\alpha'$ and the normalizing restrictions $\alpha'\alpha = I_r$, $\beta'\beta = I_r$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$ all parameters except λ_1 are defined on bounded sets. A natural choice for an uninformative prior is the uniform prior over these sets. Only λ_1 is defined on an infinite interval (conditionally upon the other λ_i). A natural choice for λ_1 that is consistent with the uniform prior on the simplex for $\lambda_2,\ldots,\lambda_r|\lambda_1$ is the exponential distribution. Another way to look at this, is that although $\lambda \in [0,\infty)$ has infinite support, it can also be transformed to the unit interval on which a uniform prior can be specified. By doing so, all model parameters (except the covariance matrix Σ) are bounded to finite areas. Specifically, when either the transformation $\lambda^\flat = \exp(-\lambda) \in [0,1)$ or $\lambda^\sharp = 1-\exp(-\lambda) \in [0,1)$ is used and a standard uniform density is specified on $\lambda^\flat$ or $\lambda^\sharp$ then λ also has a standard exponential distribution. Using a similar argument the rate parameter θ could be included by specifying a uniform prior on e.g. $\exp(-\theta\lambda)$. A final remark concerns the rate θ of the exponential distribution. By choosing θ to a value close to 0, the exponential distribution tends towards a flat distribution over the positive real numbers.

A.5. Appendix for Section 5: Model probabilities under regularization priors and possibly irregular likelihoods

A.5.1. Model probabilities for the AR(1) model with near unit roots

We illustrate in this subsection the issues that one encounters in evaluating posterior and predictive probabilities for a basic time series model using weak and regularizing prior information. For expository purposes we make use of a univariate Auto-Regressive model of order one, AR(1), specified as:

$$y_t = \alpha y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0, \sigma^2). \tag{A.106}$$

where y_t is the observed dependent variable and the $\varepsilon_t's$ are independently and identically distributed. By substituting successively for $y_{t-1}, y_{t-2}, \ldots$ in equation (129), one obtains

$$y_t = \varepsilon_t + \alpha\varepsilon_{t-1} + \alpha^2\varepsilon_{t-2} + \alpha^3\varepsilon_{t-3} + \ldots. \tag{A.107}$$

Thus $E[y_t] = 0$ and given that the innovations or shocks $\varepsilon_t, \varepsilon_{t-1}, \ldots$ are independent, each with the constant σ^2, the variance of y_t is given by

$$V(y_t) = 1 + \alpha^2\sigma^2 + \alpha^4\sigma^2 + \alpha^6\sigma^2 + \ldots \tag{A.108}$$

One may distinguish several cases with important different dynamic characteristics, depending on the value of the dynamic adjustment parameter α.
Case I: If the stationarity condition $|\alpha| < 1$ holds so that the infinite series converges, then $V(y_t) = \frac{\sigma^2}{1-\alpha^2}$. Thus, in long term forecasting one has bounded uncertainty with respect to a forecast of y_t.
Case II: If $\alpha = 1$ and $y_0 = 0$, then the system crosses a border where the stationary state stops. There exists a discontinuity in the dynamic behavior: the unconditional variance of y_t increases linearly (and without bounds) with t:

$$V(y_t) = E\left[y_t^2\right] = E\left[\left(\sum_{i=1}^{t}\varepsilon_i\right)^2\right] = \sum_{i=1}^{t}\sum_{j=1}^{t} E\left[\varepsilon_i\varepsilon_j\right] = t\sigma^2, \tag{A.109}$$

since $E[\varepsilon_i\varepsilon_j] = \sigma^2$ if $i = j$, and 0 if $i \neq j$. This trending value tends to infinity when the length of the series becomes large. Thus, in the unit root case one has a tendency of getting large and unbounded uncertainty with respect to the very long run and asymptotic forecasts of y_t.
Case III: Roots (substantially) greater than unity are easily detected as the explosive character of the series is clear with fairly small samples. This may lead to a switching behavior in model structure. Therefore, we do not consider that case here in our basic analysis. Similarly, we do restrict our attention to series that are positively correlated.
Another way to motivate the analysis of probability model evaluation for an AR(1) model is given in the graphical representation of a few AR(1) processes

Fig A.10: Stationary AR(1) processes vs Unit Root process (black).

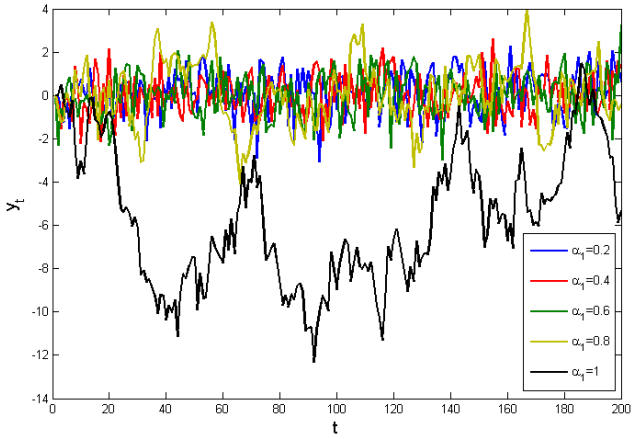
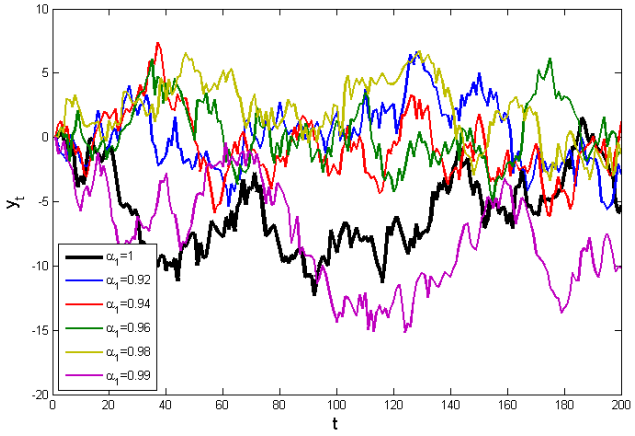


Fig A.11: Close-to-unit-root processes versus an exact unit root process (thick black).



with several choices for the α parameter. Figure A.10 presents the time series generated with AR(1) with $\sigma = 1$ and α equal to: $(0.2, 0.4, 0.6, 0.8, 1.0)$, respectively. Simply by means of visual inspection[17] one is sometimes able to distinguish stationary time series from a unit root series. The stationary processes have all an unconditional mean of zero and finite unconditional variance.

[17]This is the well-known 'Rotterdam Eyeball-test for Nonstationarity' that has been used by Van Dijk to illustrate the issue to numerous students in the econometrics program at Erasmus University.

They are tied to the zero means in the sense that deviations from them cannot accumulate indefinitely. In contrast, the process with a single root of exactly unity has an unconditional variance which increases over time. This process will tend to wander widely and is not expected to cross the origin regularly. In Figure A.11 time series patterns are presented that are generated with an α equal to 0.92, 0.94, 0.96, 0.98, 0.99, respectively. The time series are now very similar. In summary, comparing the probability of a model with a unit root with the probability of a model that is nearly non-stationary is non-trivial in spite of well-defined statistical properties of the different structures. Frequentist estimation and testing procedure tend to have poor sampling properties. In Bayesian econometrics there is also a need for accurate methods and careful specification of prior information in order to analyze near and at-the-boundary behavior of autoregressive time processes. This is what we explore next.

Model evaluation and the Bartlett/Jeffreys/Lindley paradox: The principal Bayesian tool for a probabilistic model evaluation and comparison is the posterior odds ratio specified as the product of a prior odds ratio and the so-called Bayes Factor (ratio of marginal likelihoods), given as

$$\frac{\Pr(M_1|y)}{\Pr(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)} \times \frac{\Pr(M_1)}{\Pr(M_2)}. \tag{A.110}$$

where the marginal likelihood of model $M_i, i = 1, 2$ is defined as the density of the data after marginalizing the joint posterior density with respect to the model parameters

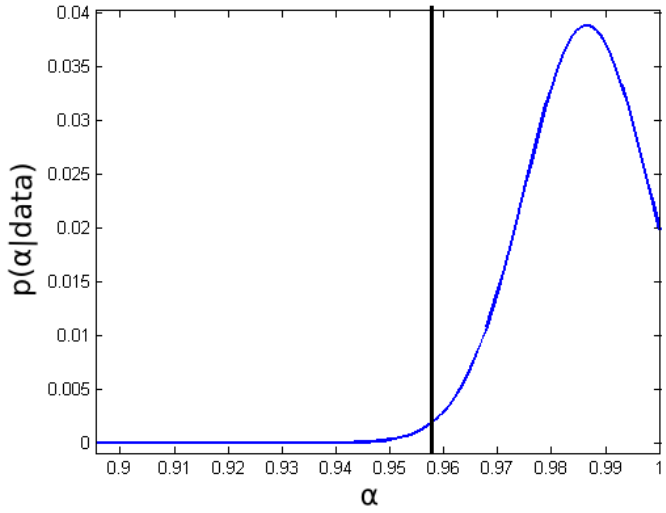
$$p(y|M_i) = \int p(y|\phi_i, M_i)p(\phi_i|M_i)d\phi_i, \tag{A.111}$$

where $p(y|\phi_i, M_i)$ is the likelihood in model M_i and $p(\phi_i|M_i)$ is the prior density of the parameters ϕ_i in model M_i. If one assigns the prior odds, $\Pr(M_1)$ divided by $\Pr(M_2)$, to be equal to one meaning that no specification is favored *a priori*, then the posterior odds is equal to the Bayes Factor. Assuming a uniform prior on the bounded interval $(c, 1)$, where $0 < c < 1$, one can derive

$$\frac{p(M_1 : \alpha = 1|y)}{p(M_2 : \alpha \neq 1|y)} = \frac{p(1|y)}{\frac{1}{1-c}\int_c^1 p(\alpha|y)d\alpha}. \tag{A.112}$$

where $p(\alpha|y)$ is a univariate t density truncated at the value $\alpha = 1$. For details on the derivation we refer to Schotman and Van Dijk (1991a). Equation (135) reveals the sensitivity of Bayes Factor with respect to the prior specification. As an illustration, consider the case of a posterior density for α when the DGP has value of $\alpha = 0.98$. This is shown in Figure A.12. In this case 99% of the probability mass is located to the right of the value $\alpha = 0.957$. Values of α to the left of it have almost no probability weight. However, they influence strongly the average value of posterior and substantially decrease $\frac{1}{M}\sum_{i=1}^{M} p(\alpha^i|y)$. As a consequence the posterior odds favors the unit root model while the DGP

Fig A.12: Posterior density for α in the AR(1) model, when the DGP is defined by $\alpha = 0.98$.



is stationary. More specifically, the height of the posterior density at $\alpha = 1$ is compared with, loosely speaking, the average value of the posterior over the interval $(c, 1)$. Clearly, the height is a fixed number (in the sense that it does not depend on c) while the average height depends very much on the length of the interval and thus the choice of c. In Figure A.13, it is seen that allowing the lower bound of the interval c to be far to the left of $\alpha = 0.957$ and even to choose the value 0 leads to a much higher value of the posterior odds ratio than the data tell us. The unit root model is artificially often favored. When c tends to minus infinity one always favors the unit root model irrespective of the data information. Clearly, to consider a priori many irrelevant values of parameters of interest is a bad approach. This is known as the Bartlett/Jeffreys/Lindley paradox, see Jeffreys (1939), Lindley (1957) and Bartlett (1957). We emphasize that the paradox is usually listed for the extreme case of an unbounded interval and an improper flat prior, while Figure A.12 shows that the posterior odds is already too much favoring the unit root on the bounded interval $(c, 1)$, where c is situated in a region where there is almost no posterior probability mass.
In summary, we are able to construct examples using improper as well as proper priors where the Bayesian model evaluation procedure shows artificially large odds in favor of the model $\alpha = 1$. Next, for the simple illustrative case of an AR(1) process, we discuss several regularization priors for sensible model comparison. The first approach proposes a data driven prior while a second approach aims at restricting the parameter set to a sensible bounded set via orthogonal normalization.

Data driven prior We start with introducing a 'locally uniform' prior. A simple interpretation of this prior is that it is an approximation to a proper distribution which is nearly flat in the effective range of the likelihood, see Lindley (1965). This explanation justifies the use of the improper prior that is combined with the likelihood, defined on the parameter region where the data provide information.[18] The approach also constitutes a basis for application of the concept of predictive likelihood in the evaluation of model probabilities, which we cover later.

As an illustrative example consider the case of the unit root $\alpha = 1$. It was clear from Figure A.12 that the bounded interval $[0, 1]$ contains many irrelevant alternatives in case the DGP has a true value of α close to 1, usually when $\alpha \in (0.9, 1)$. A direct solution is limiting the range of α where 99% of posterior probability mass is situated. The uniform prior is imposed only on this region and the irrelevant values of α are automatically deleted, what brings a more realistic balance in the Bayes Factor. In a simple simulation we evaluate the Bayes Factors in favor of the unit root model for 10 distinct DGPs with $\alpha = 0.90 + i \times 0.01,\ i \in 1, \ldots, 10$ and $\sigma = 1$. For each DGP we simulate 100 time series of length 500. In Figure A.13 we present the average posterior probability (over 100 datasets) of the unit root derived from respective Bayes Factors. The three lines in the plot correspond, respectively, to the uniform prior on the bounded interval $[0, 1]$, and data driven priors elicited on the regions corresponding to 99% and 98% posterior probability mass.

Based on this simple simulation exercise we observe that this class of data driven priors is able to substantially limit the over-acceptance of unit root model for DGPs with α in the interval $(0, 1)$. Obviously this comes at a cost: for true value $\alpha = 1$ the posterior probability $\Pr(\alpha = 1|data)$ drops from a value around 0.87 to a value around 0.65.

Normalization Another route to improve model evaluation procedures via Bayes Factors with weakly informative priors is a more theoretical one using orthogonal normalization. By means of this normalization the parameter space is automatically restricted to a bounded set which results in a diffuse prior being proper and making model comparison possible. We illustrate this within a general specification of the AR(1) model. Consider

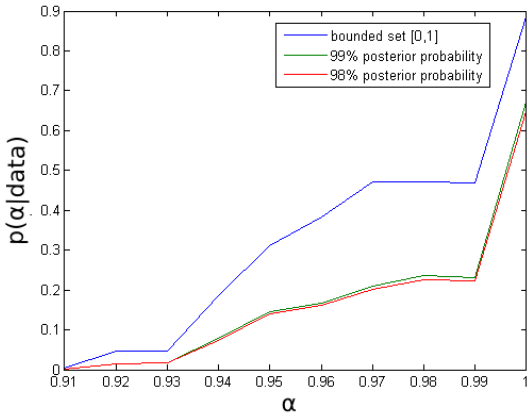
$$\alpha_0 y_t = \alpha_1 y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0, \sigma), \tag{A.113}$$

where a linear restriction $\alpha_0 = 1$ translates equation (136) into equation (129). The unbounded parameter space in equation (136) is defined in the two dimensional plane by the line $\alpha_0 = 1$. This domain is for our case bounded to the unit interval by a restriction $\alpha_1 \in [0, 1]$, in order to focus solely on positively serially correlated processes that are either stationary or a unit root process.

Alternatively, consider the orthogonal specification defined by the condition

[18] A theoretical argument to show that an improper prior distributions can be better explained as the limits of some data adaptive prior distribution rather than as the limits of some proper prior distributions is given in Akaike (1980).

Fig A.13: Posterior probability of unit root $\Pr(\alpha = 1|data)$ for DGP's ranging from $\alpha = 0.91$ up to $\alpha = 1.00$ under a uniform prior on the bounded interval $[0, 1]$ and data driven priors imposed on region corresponding to respectively 99% and 98% posterior probability mass. The reported probabilities are calculated as the average probability from 100 simulated time series of length 500.



$\alpha_0^2 + \alpha_1^2 = 1$ which restricts the parameter space to the unit circle with the center at the origin. For our consideration of positively autocorrelated series near or at the unit root, we restrict attention to the northeast quarter of the unit circle specified by the angle $\theta \in [0, \pi/2]$, see Figure A.14.
Assuming a uniform prior on α_1, and with a standard marginalization step with respect to the nuisance parameter σ one obtains under the linear normalization

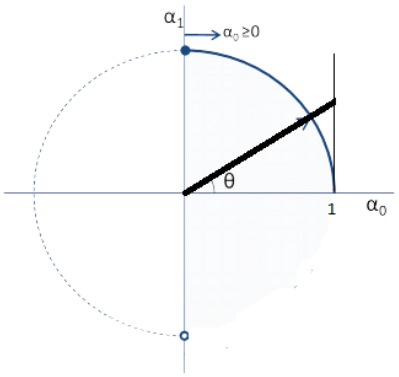
$$p(\alpha_1|\mathbf{y}) \propto \left[(\mathbf{y} - \alpha_1\mathbf{y}_{-1})'(\mathbf{y} - \alpha_1\mathbf{y}_{-1})\right]^{-\frac{1}{2}T} \tag{A.114}$$

and under the orthogonal normalization

$$p(\alpha_1|\mathbf{y}) \propto (1-\alpha_1^2)^{\frac{1}{2}T}\left[(\mathbf{y} - \frac{\alpha_1}{\sqrt{1-\alpha_1^2}}\mathbf{y}_{-1})'(\mathbf{y} - \frac{\alpha_1}{\sqrt{1-\alpha_1^2}}\mathbf{y}_{-1})\right]^{-\frac{1}{2}T}, \tag{A.115}$$

where $\mathbf{y}$ is a T-dimensional vector of observations. Given a flat prior, there exist three differences in the evaluation of the Bayes Factor for a linear and for an orthogonal normalization. First, the length of the parameter space in the linear normalization is defined by $\alpha_0 = 1$ and $\alpha_1 \in [0, 1]$. When we project this interval onto the unit circle its length changes as the parameter space is defined by $\alpha_1 \in [0, \frac{\sqrt{2}}{2}]$, $\alpha_0^2 + \alpha_1^2 = 1$. The length of this region is computed as $(\arcsin \frac{\sqrt{2}}{2})/\frac{\pi}{2} \times \frac{\pi}{2} = \frac{\pi}{4}/\frac{\pi}{2} \times \frac{\pi}{2} = \frac{\pi}{4}$, where $\frac{\pi}{2}$ corresponds to the length of a quarter of the unit circle. Thus, a uniform prior on the unit interval, given as $p(\alpha_1) \propto 1$, may be compared with a uniform prior on 1/8 of the unit circle., given as $p(\theta) \propto \frac{1}{\frac{\pi}{4}} = \frac{4}{\pi}$ and the difference in the length of the parameter space

Fig A.14: Projection with respect to $(\alpha_0, \alpha_1) = (0, 0)$.



automatically leads to the difference in Bayes Factors. Second, the likelihood also contributes to the difference in the evaluation of the Bayes Factor. In case of orthogonal normalization the likelihood is specified on an angular space and the Bayes Factor for the unit root under the linear normalization may be compared with a Bayes Factor for $\alpha_1 = \frac{\sqrt{2}}{2}$ under the orthogonal one.
Thirdly, note that mapping a uniform density defined on the unit circle onto the unit interval leads to a Cauchy type density defined on this interval. Thus, a uniform prior on the unit circle introduces a tendency towards stationary models. Conversely, imposing a uniform prior on the unit interval results in a nontrivial prior on the unit circle, which tends to favor the unit root model. We derive these results next.

Implied priors Assume a uniform prior on the respective part of the unit circle. According to our derivations above it is equal to $p(\theta) \propto \frac{1}{\pi/4} = 4/\pi$. We note that $\tan\theta = \alpha_1$ and consequently $\arctan\alpha_1 = \theta$. Then the cumulative probability mass located on the unit circle, bounded by $\theta_0 = 0$ and θ, is equal to $\frac{2\theta}{\pi}$. As the projection from the unit circle onto the unit interval (with respect to the origin of coordinate system) preserves the total probability mass, the cumulative probability mass on the unit interval up to the point $(1, \alpha_1)$ is also given by $\frac{2\theta}{\pi}$. Denote the cumulative distribution function (c.d.f) on the unit interval by F. Then we have

$$\begin{aligned} F(\alpha_1(\theta)) &= \frac{4\theta}{\pi} \\ F'(\alpha_1)\alpha_1'(\theta) &= \frac{4}{\pi} \\ f(\alpha_1) &= \frac{4}{\pi}\frac{1}{\alpha_1'(\theta)} \\ f(\alpha_1) &= \frac{4}{\pi}\frac{1}{1+\tan^2\theta} = \frac{4}{\pi}\frac{1}{1+\alpha_1^2}. \end{aligned}$$

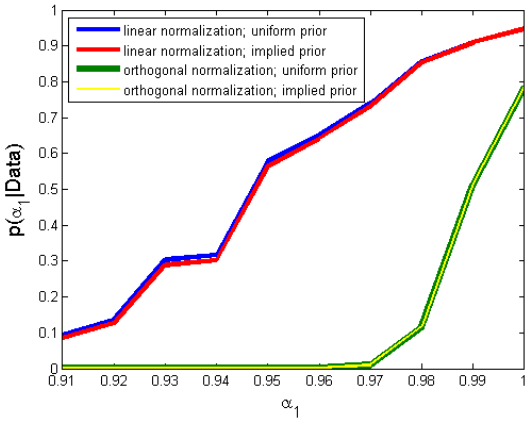
Given a uniform density on the unit circle, one obtains an implied probability density function on the unit interval that is proportional to a Cauchy density. Alternatively one can derive the prior on the unit circle, when a uniform prior is imposed on the unit interval. The uniform prior on the unit interval is defined by $p(\alpha_1) \propto 1$. Then the c.d.f. on the unit interval is given by $P(\alpha_1) = \alpha_1$. Again, as the projection with respect to $(\alpha_0, \alpha_1) = (0, 0)$ preserves the total probability mass, we observe that the cumulative probability mass projected on the unit circle is given by $F(\theta) = \alpha_1$, where F denotes now the c.d.f. on the unit circle. Then we have

$$\begin{aligned} F(\theta) &= \alpha_1 \\ F'(\theta(\alpha_1))\theta'(\alpha_1) &= 1 \\ f(\theta) &= \frac{1}{\theta'(\alpha_1)} \\ f(\theta) &= (1 + \alpha_1^2) = (1 + \tan^2 \theta). \end{aligned}$$

Thus we can define the prior on the unit circle with respect to the angle θ or α_1, again. In both cases the Uniform prior on one parameter space leads to a nontrivial implied prior on another parameter space. We refer to those priors as implied priors.

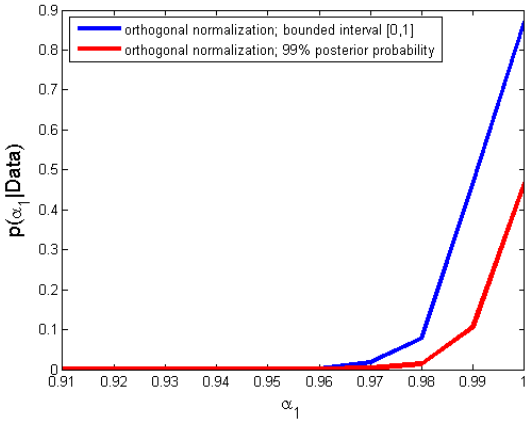
In simple simulation we investigate the implications of an orthogonal normalization for model evaluation. Again we work with DGPs ranging from $\alpha_1 = 0.91$ up to the unit root. For each DGP we simulate 100 processes of length 500. In Figure A.15 we present the average probability of the unit root model under both linear and orthogonal normalizations. We consider two different priors: a

Fig A.15: Posterior probability of unit root $\Pr(\alpha = 1|data)$ for DGP ranging from $\alpha_1 = 0.91$ up to $\alpha_1 = 1.00$ under linear and orthogonal normalization.



uniform prior on the unit interval and a uniform prior on the unit circle. We note that the implied priors did not affect our results in this case.

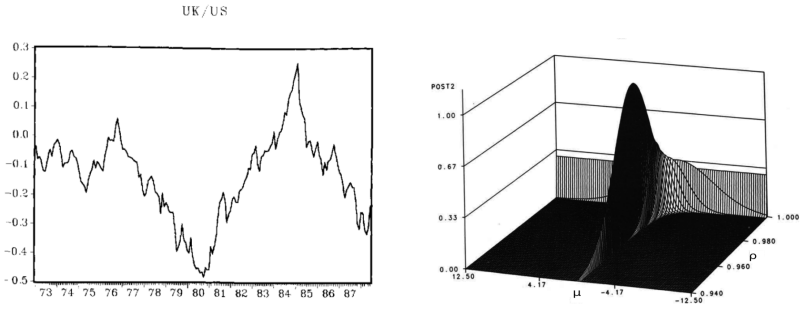
Fig A.16: Posterior probability of unit root $\Pr(\alpha = 1|data)$ for DGP ranging from $\alpha_1 = 0.91$ up to $\alpha_1 = 1.00$ under the orthogonal normalization and data driven prior.



The orthogonal normalization is evidently leading to a lower rate of over-acceptance of the unit root model. In our simulation, in the region $(0.91, 0.97)$ the model estimated under linear normalization of parameters results in relatively high posterior probability of the unit root model, compared to the orthogonal normalization, which leads to a much more reasonable probability of such behavior. Obviously this again comes at a cost: for the true value $\alpha = 1$, the posterior probability $\Pr(\alpha = 1|data)$ drops from a value around 0.95 to a value around 0.78.

Locally uniform and orthogonal normalization Given the insights from the simulation studies, as a final step we combine the data driven prior with the orthogonal normalization into one framework. We use the simulation experiment in order to investigate if the data driven prior can lead to further improvement in the model evaluation under the orthogonal normalization. In Figure A.16 we present the results. Evidently the combination of parameterization on the unit circle and a data driven prior substantially improves the results of the model evaluation procedure.
Clearly, there is a need in Bayesian inference to carefully analyze posterior odds when the priors are weakly informative. A sensitivity analysis like using data driven priors defined in a plausible bounded domain is to be recommended.
These results constitute a solid motivation for development of methods for evaluation of model probabilities for multivariate time series model.

Fig A.17: Log of US/UK real exchange rates between Dec 1972 and Jun 1988 (left panel) and posterior of (μ, ρ) with the uniform prior (right panel). See Schotman and Van Dijk (1991a).



A.5.2. AR(1) model with weakly identified mean and near unit root

We briefly mention this issue and refer for more information to Schotman and Van Dijk (1991a). The basic AR(1) model around a mean is specified as:

$$y_t - \mu = \alpha(y_{t-1} - \mu) + \varepsilon_t, \ \varepsilon_t \sim N(0, \sigma^2) \tag{A.116}$$

which can be rewritten as:

$$y_t - y_{t-1} = (\alpha - 1)(y_{t-1} - \mu) + \varepsilon_t, \ \varepsilon_t \sim N(0, \sigma^2). \tag{A.117}$$

Clearly when α tends to one, the mean becomes unidentified. This is shown in Figure A.17. We will not analyze the computation of model probabilities for this case but move on to the same issue in an IV model presented in the main text.

A.6. Appendix for Section 6: Bayesian mixtures to analyze the education effect on earned income in US states

TABLE A.1
US regions and divisions

Division	States	Number of observations
	Northeast Region	
1. New England	Connecticut (CT), Maine (ME), Massachusetts (MA), New Hampshire (NH), Rhode Island and Vermont (RI)	20120
2. Middle Atlantic	New Jersey (NJ), New York (NY) and Pennsylvania (PA)	64364
	Midwest Region	
3. East North Central	Illinois (IL), Indiana (IN), Michigan (MI), Ohio (OH) and Wisconsin (WI)	67047
4. West North Central	Iowa (IA), Kansas (KS), Minnesota (MN), Missouri (MO), Nebraska (NE), North Dakota (ND) and South Dakota (SD)	35220
	South Region	
5. South Atlantic	Delaware (DE), District of Columbia (DC), Florida (FL), Georgia (GA), Maryland (MD), North Carolina (NC), South Carolina (SC), Virginia (VA) and West Virginia (WV)	48072
6. East South Central	Alabama (AL), Kentucky (KY), Mississippi (MS) and Tennessee (TN)	31668
7. West South Central	Arkansas (AR), Louisiana (LA), Oklahoma (OK) and Texas (TX)	34651
	West Region	
8. Mountain	Arizona (AZ), Colorado (CO), Idaho (ID), Montana (MT), Nevada (NV), New Mexico (NM), Utah (UT) and Wyoming (WY)	11228
9. Pacific	Alaska (AK), California (CA), Hawaii (HI), Oregon (OR) and Washington (WA)	17139

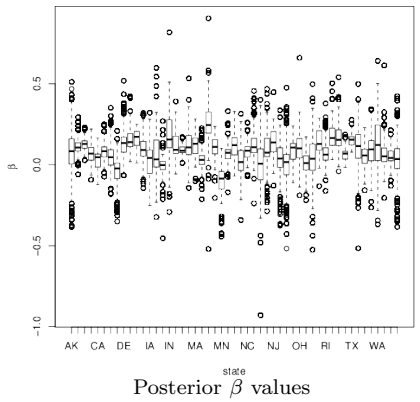
Note: The table reports US states included in the dataset, with the respective regions, divisions and the number of observations.

TABLE A.2
Income-education effects in US states: Posterior results for parameters

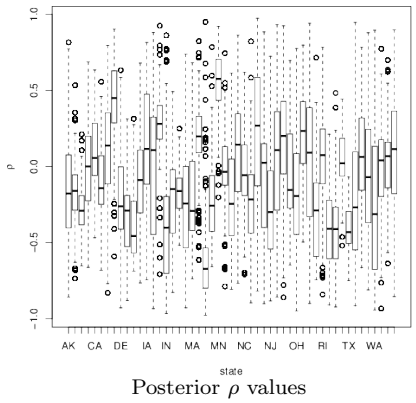
	β	Π_2	Π_3	Π_4	ρ		β	Π_2	Π_3	Π_4	ρ
AK	0.1	1.54	1.5	-0.3	-0.2	MT	0.02	0.01	0.09	0.39	0.08
	(0.12)	(0.76)	(0.77)	(0.87)	(0.33)		(0.10)	(0.13)	(0.14)	(0.14)	(0.34)
AL	0.1	0.03	0.33	0.3	-0.2	NY	0.09	-0.04	0.13	0.24	-0.06
	(0.05)	(0.08)	(0.07)	(0.08)	(0.21)		(0.05)	(0.06)	(0.06)	(0.07)	(0.25)
AR	0.1	-0.18	0.1	0.4	-0.3	ND	0.13	-0.48	-0.19	-0.08	-0.25
	(0.04)	(0.09)	(0.08)	(0.08)	(0.17)		(0.09)	(0.13)	(0.14)	(0.15)	(0.30)
AZ	0.1	0.65	-0	-0.3	-0	NE	0	-0.13	-0.15	-0.12	0.20
	(0.05)	(0.21)	(0.20)	(0.22)	(0.27)		(0.16)	(0.09)	(0.10)	(0.08)	(0.44)
CA	0	0.26	0.22	0.1	0.08	NH	0.1	-0.14	0.1	0.2	-0.08
	(0.06)	(0.05)	(0.05)	(0.05)	(0.26)		(0.10)	(0.19)	(0.18)	(0.2)	(0.39)
CO	0.1	0.31	0.44	0.4	-0.1	NJ	0.13	0.06	0.07	0.21	-0.24
	(0.06)	(0.11)	(0.10)	(0.11)	(0.25)		(0.08)	(0.06)	(0.06)	(0.08)	(0.31)
CT	0.1	0.3	0.12	0.1	0.1	NM	0.04	0.23	0.11	0.39	0.06
	(0.09)	(0.09)	(0.10)	(0.10)	(0.37)		(0.10)	(0.18)	(0.19)	(0.17)	(0.41)
DC	-0	-0.44	-0.4	-0.5	0.43	NV	0	-0.03	0.07	0.85	0.17
	(0.09)	(0.16)	(0.17)	(0.16)	(0.27)		(0.11)	(0.38)	(0.37)	(0.35)	(0.35)
DE	0.1	0.56	0.69	0.2	-0.2	NV	0.09	0.1	0.05	-0.05	-0.05
	(0.08)	(0.22)	(0.22)	(0.23)	(0.31)		(0.07)	(0.04)	(0.04)	(0.04)	(0.30)
FL	0.2	0.32	0.24	0.3	-0.3	OH	0.11	-0.04	0.05	0.05	-0.17
	(0.07)	(0.10)	(0.09)	(0.09)	(0.26)		(0.10)	(0.06)	(0.05)	(0.05)	(0.37)
GA	0.2	-0.25	0.05	0	-0.4	OK	0.01	-0.04	0.16	0.23	0.21
	(0.05)	(0.05)	(0.06)	(0.06)	(0.22)		(0.07)	(0.08)	(0.07)	(0.07)	(0.28)
HI	0.1	0.09	1.55	0.8	-0.1	OR	0.05	0.12	0.11	0.06	0.01
	(0.07)	(0.41)	(0.39)	(0.35)	(0.28)		(0.15)	(0.12)	(0.12)	(0.12)	(0.47)
IA	0	-0.04	-0	0.1	0.11	PA	0.15	0.02	0.01	0.05	-0.33
	(0.10)	(0.06)	(0.06)	(0.06)	(0.36)		(0.07)	(0.03)	(0.03)	(0.03)	(0.26)
ID	0.1	0.16	-0	0.1	0.02	RI	0.07	-0.39	0.11	0.12	0.04
	(0.13)	(0.16)	(0.17)	(0.14)	(0.45)		(0.07)	(0.15)	(0.17)	(0.18)	(0.3)
IL	0	0.07	-0.1	0.1	0.24	SC	0.17	-0.11	-0.05	0.31	-0.39
	(0.08)	(0.03)	(0.04)	(0.04)	(0.29)		(0.07)	(0.09)	(0.07)	(0.08)	(0.25)
IN	0.2	0.04	0.08	0	-0.3	SD	0.17	0.35	0.3	0.56	-0.42
	(0.15)	(0.05)	(0.05)	(0.05)	(0.44)		(0.08)	(0.14)	(0.13)	(0.15)	(0.24)
KS	0.1	0.3	0.34	0.2	-0.2	TN	0.06	-0.06	0.19	0.47	0.07
	(0.07)	(0.07)	(0.08)	(0.08)	(0.26)		(0.03)	(0.08)	(0.07)	(0.07)	(0.17)
KY	0.1	0.08	0.35	0.6	-0.2	TX	0.16	-0.04	0.23	0.26	-0.43
	(0.03)	(0.07)	(0.07)	(0.07)	(0.17)		(0.06)	(0.05)	(0.05)	(0.07)	(0.19)
LA	0.1	0.1	0.26	0.3	-0.2	UT	0.11	-0.02	-0.15	-0.25	-0.20
	(0.10)	(0.09)	(0.09)	(0.10)	(0.36)		(0.13)	(0.14)	(0.15)	(0.16)	(0.45)
MA	0.1	0.13	0.17	0.3	-0.2	VA	0.06	0.08	0.3	0.32	0.05
	(0.07)	(0.06)	(0.06)	(0.07)	(0.32)		(0.07)	(0.08)	(0.09)	(0.08)	(0.31)
MD	0	0.38	0.43	0.3	0.17	VT	0.09	0.22	0.47	0.33	-0.06
	(0.06)	(0.10)	(0.10)	(0.09)	(0.24)		(0.10)	(0.19)	(0.22)	(0.21)	(0.39)
ME	0.3	0.01	0.28	0	-0.6	WA	0.13	0.14	0.12	0	-0.22
	(0.13)	(0.09)	(0.10)	(0.11)	(0.26)		(0.15)	(0.09)	(0.08)	(0.11)	(0.47)
MI	0.1	0.15	0.11	0.1	-0.2	WI	0.07	0.21	0.01	0.1	-0.01
	(0.07)	(0.03)	(0.04)	(0.04)	(0.28)		(0.08)	(0.06)	(0.08)	(0.07)	(0.28)
MN	-0.1	-0.2	-0.2	-0.1	0.55	WV	0.05	-0.04	0.09	0.27	0.04
	(0.10)	(0.06)	(0.06)	(0.05)	(0.21)		(0.06)	(0.09)	(0.07)	(0.07)	(0.26)
MO	0.1	-0.08	0.09	0	-0	WY	0.03	0.14	0.38	-0.13	0.09
	(0.08)	(0.06)	(0.05)	(0.05)	(0.30)		(0.11)	(0.23)	(0.22)	(0.25)	(0.38)
MS	0.1	0.07	0.22	0.3	-0.2						
	(0.08)	(0.08)	(0.09)	(0.08)	(0.36)						

Note: The table reports posterior means for the parameters and posterior standard deviations (in parentheses) for each state. Π_2, Π_3 and Π_4 are the coefficients for the 2nd, 3rd and 4th quarter of birth dummies, respectively. Posterior results are achieved by 30000 draws (3000 burn-in).

Fig A.18: Income-education effects in US states: Boxplots for income effects and degree of endogeneity

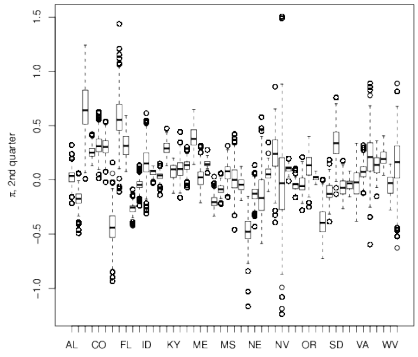


Posterior β values

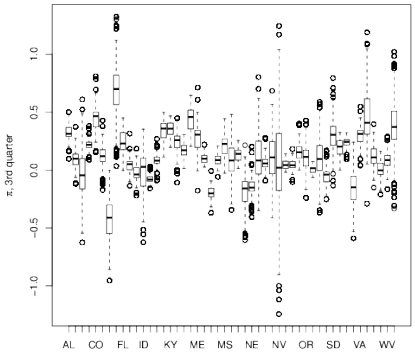


Posterior ρ values

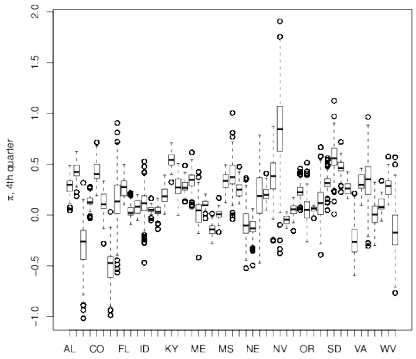
Fig A.19: Income-education effects in US states: Boxplots for the quarter of birth effects on education



Posterior Π_2 values



Posterior Π_3 values



Posterior Π_4 values

References

Akaike, H. (1980). "The interpretation of improper prior distributions as limits of data dependent proper prior distributions." *Journal of the Royal Statistical Society. Series B (Methodological)*, 42: 46–52. 73

Anderson, T. (2003). *An introduction to multivariate statistical analysis*. Wiley, New York, 3rd edition. 48

Anderson, T. W. and Rubin, H. (1950). "The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations." *The Annals of Mathematical Statistics*, 570–582. 1

— (1956). "Statistical inference in factor analysis." In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 5, 1. 1

Angrist, J. D. and Krueger, A. B. (1991). "Does compulsory school attendance affect schooling and earnings?" *The Quarterly Journal of Economics*, 106(4): 979–1014. 6, 7, 27, 31, 35, 36, 40

Bartlett, M. S. (1957). "A Comment on D. V. Lindley's Statistical Paradox." *Biometrika*, 44(3/4): 533–534. 3, 72

Baştürk, N., Çakmaklı, C., Ceyhan, S., and Van Dijk, H. K. (2014a). "On the Rise of Bayesian Econometrics after Cowles Foundation Monographs 10, 14." *Œconomia*, 4: 381–447. 2, 4

Baştürk, N., Çakmaklı, C., Ceyhan, S. P., and Van Dijk, H. K. (2014b). "Posterior-Predictive Evidence on US Inflation using Extended Phillips Curve Models with non-filtered Data." *Journal of Applied Econometrics*, 29(7): 1164–1182. 23

Baştürk, N., Grassi, S., Hoogerheide, L., and Van Dijk, H. K. (2017). "Bayesian static and dynamic factor modeling with industry momentum strategies." Unpublished manuscript, Erasmus University Rotterdam. 63

Baştürk, N., Grassi, S., Hoogerheide, L. F., and Van Dijk, H. K. (2016a). "Time-varying combinations of Bayesian dynamic models and equity momentum strategies." Technical Report 16-099/III, Tinbergen Institute. 34

Baştürk, N., Hoogerheide, L., Kleijn, R., and Van Dijk, H. K. (2016b). "Prior Ignorance, Likelihood Shape and Posterior Existence in a Cointegration Model." Unpublished manuscript, Erasmus University Rotterdam. 46

Baumeister, C. and Hamilton, J. D. (2015). "Sign restrictions, structural vector autoregressions, and useful prior information." *Econometrica*, 83(5): 1963–1999. 22

Bauwens, L., Lubrano, M., and Richard, J.-F. (1999). *Bayesian inference in dynamic econometric models*. Advanced texts in Econometrics. Oxford: Oxford University Press. 51, 52

Bauwens, L. and Van Dijk, H. K. (1990). "Bayesian limited information analysis revisited." In Gabszewicz, J., Richard, J., and Wolsey, L. (eds.), *Economic Decision-Making: Games, Econometrics and Optimisation: Contributions in Honour of Jacques H. Drèze*, chapter 18, 385–424. Amsterdam: North-Holland. 15, 61

Berger, J. O., Pericchi, L. R., et al. (2004). "Training samples in objective

Bayesian model selection." *The Annals of Statistics*, 32(3): 841–869. 23

Billio, M., Casarin, R., Ravazzolo, F., and Van Dijk, H. K. (2013). "Time-varying combinations of predictive densities using nonlinear filtering." *Journal of Econometrics*, 177(2): 213–232. 34

Casarin, R., Grassi, S., Ravazzolo, F., and Van Dijk, H. K. (2015). "Dynamic predictive density combinations for large data sets in economics and finance." Technical Report 2015-084/III, Tinbergen Institute. 34

Chan, J., Leon-Gonzalez, R., and Strachan, R. W. (2017). "Invariant inference and efficient computation in the static factor model." *Journal of the American Statistical Association*, (just-accepted). 34, 63

Chan, J. C., Koop, G., Leon-Gonzalez, R., and Strachan, R. W. (2012). "Time Varying Dimension Models." *Journal of Business & Economic Statistics*, 30(3): 358–367. 34

Del Negro, M. and Schorfheide, F. (2004). "Priors from general equilibrium models for VARs." *International Economic Review*, 45(2): 643–673. 23

— (2011). "Bayesian macroeconometrics." In Geweke, J., Koop, G., and Van Dijk, H. K. (eds.), *The Oxford handbook of Bayesian econometrics*, 293–389. Oxford University Press: Oxford. 34

Dickey, J. (1967). "Matricvariate generalizations of the multivariate t distribution and the inverted multivariate t distribution." *The Annals of Mathematical Statistics*, 38(2): 511–518. 52

— (1971). "The Weighted Likelihood Ratio, Linear Hypothesis on Normal Location Parameters." *The Annals of Mathematical Statistics*, 42: 240–223. 29

Doan, T., Litterman, R., and Sims, C. (1992). "Forecasting and Conditional Projections using Realistic Prior Distributions." *Econometric Reviews*, 3: 1–100. 23

Drèze, J. H. (1976). "Bayesian limited information analysis of the simultaneous equations model." *Econometrica: Journal of the Econometric Society*, 1045–1075. 15, 41

— (1977). "Bayesian regression analysis using poly-t densities." *Journal of Econometrics*, 6(3): 329–354. 38

Eklund, J. and Karlsson, S. (2007). "Forecast combination and model averaging using predictive measures." *Econometric Reviews*, 26(2-4): 329–363. 24

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media. 3, 4, 34

Gelfand, A. E. and Dey, D. K. (1994). "Bayesian model choice: asymptotics and exact calculations." *Journal of the Royal Statistical Society. Series B (Methodological)*, 501–514. 24

Gelman, A., Meng, X., and Stern, H. (1996). "Posterior predictive assessment of model fitness via realized discrepancies." *Statistica Sinica*, 733–760. 23

Geweke, J. (1996). "Bayesian reduced rank regression in econometrics." *Journal of Econometrics*, 75(1): 121–146. 15

Geweke, J., Koop, G., and Van Dijk, H. K. (2011). *The Oxford handbook of Bayesian econometrics*. Oxford University Press. 4, 34

Giannone, D., Lenza, M., and Primiceri, G. (2015). "Prior Selection for Vector Autoregressions." *The Review of Economics and Statistics*, 97(2): 436–451.

23

Hamilton, J., Waggoner, D., and Zha, T. (2007). "Normalization in econometrics." *Econometric Reviews*, 26(2–4): 221–252. 20

Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press. 23

Herbst, E. P. and Schorfheide, F. (2015). *Bayesian Estimation of DSGE Models*. Princeton University Press. 34

Hoogerheide, L. (2006). "Essays on Neural Network Sampling Methods and Instrumental Variables." Ph.D. thesis, Erasmus University Rotterdam. 35, 45

Hoogerheide, L., Kleibergen, F., and Van Dijk, H. K. (2007a). "Natural conjugate priors for the instrumental variables regression model applied to the Angrist–Krueger data." *Journal of Econometrics*, 138(1): 63–103. 41

Hoogerheide, L. and Van Dijk, H. K. (2010). "Bayesian forecasting of value at risk and expected shortfall using adaptive importance sampling." *International Journal of Forecasting*, 26(2): 231–247. 45

— (2013). "Bridging two key issues in Bayesian inference: The relationship between the Lindley paradox and non-elliptical credible sets." In Singpurwalla, N., Dawid, P., and O'Hagan, A. (eds.), *Festschrift for Dennis Lindleys ninetienth birthday*, volume 2, 511–530. Blurb Publishers. 24

Hoogerheide, L. F., Kaashoek, J. F., and Van Dijk, H. K. (2007b). "On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: An application of flexible sampling methods using neural networks." *Journal of Econometrics*, 139(1): 154–180. 4

Hoogerheide, L. F. and Van Dijk, H. K. (2001). "Comparison of the Anderson-Rubin test for overidentification and the Johansen test for cointegration." Technical Report 2001-04, Econometric Institute Research Papers. 9

— (2006). "A reconsideration of the Angrist-Krueger analysis on returns to education." *Econometric Institute Report*. 28, 29, 31

— (2008). "Possibly ill-behaved posteriors in econometric models." Technical Report 08-036/4, Tinbergen Institute. 6, 27, 40

Jeffreys, H. (1939). *The theory of probability*. Oxford: The Clarendon Press, 1st edition. 3, 72

Jochmann, M. and Koop, G. (2015). "Regime-switching cointegration." *Studies in Nonlinear Dynamics & Econometrics*, 19(1): 35–48. 34

Jochmann, M., Koop, G., Leon-Gonzalez, R., and Strachan, R. W. (2013). "Stochastic search variable selection in vector error correction models with an application to a model of the UK macroeconomy." *Journal of Applied Econometrics*, 28(1): 62–81. 34

Johansen, S. (1988). "Statistical analysis of cointegration vectors." *Journal of Economic Dynamics and Control*, 12(2): 231–254. 48

— (1991). "Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models." *Econometrica*, 59(6): 1551–1580. 1

— (1995). *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford University Press on Demand. 9

Kaufmann, S. and Schumacher, C. (2013). "Bayesian estimation of sparse dy-

namic factor models with order-independent identification." Working Paper 13.04, Study Center Gerzensee. 63

Kleibergen, F. and Paap, R. (2002). "Priors, posteriors and Bayes factors for a Bayesian analysis of cointegration." *Journal of Econometrics*, 111(2): 223–249. 10, 17, 19, 67

Kleibergen, F. and Van Dijk, H. K. (1994). "On the shape of the likelihood/posterior in cointegration models." *Econometric Theory*, 10(3-4): 514–551. 10, 11, 16, 56

— (1998). "Bayesian simultaneous equations analysis using reduced rank structures." *Econometric Theory*, 14(06): 701–743. 16, 17, 19, 50, 61

Kleibergen, F. and Zivot, E. (2003). "Bayesian and classical approaches to instrumental variable regression." *Journal of Econometrics*, 114(1): 29–72. 17

Kleijn, R. (2016). "Essays on Bayesian model averaging using economic time series." Ph.D. thesis, Erasmus University Rotterdam. 46

Koop, G., Leon-Gonzalez, R., and Strachan, R. W. (2011). "Bayesian inference in a time varying cointegration model." *Journal of Econometrics*, 165(2): 210–220. 34

Koop, G., Strachan, R., Van Dijk, H. K., and Villani, M. (2006). "Bayesian approaches to cointegration." In *The Palgrave Handbook of Theoretical Econometrics*, 871–898. Palgrave Macmillan. 17

Lindley, D. V. (1957). "A statistical paradox." *Biometrika*, 44(1/2): 187–192. 3, 72

— (1965). *Introduction to probability and statistics, part 2: Inference*, volume 2. London: Cambridge University Press. 73

Maddala, G. (1976). "Weak priors and sharp posteriors in simultaneous equation models." *Econometrica: Journal of the Econometric Society*, 345–351. 41

Magnus, J. and Neudecker, H. (1995). *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons. 56, 57

Martin, G. M. (2001). "Bayesian analysis of a fractional cointegration model." *Econometric Reviews*, 20(2): 217–234. 16

Martin, G. M. and Martin, V. L. (2000). "Bayesian inference in the triangular cointegration model using a Jeffreys prior." *Communications in Statistics-Theory and Methods*, 29(8): 1759–1785. 8, 16

Mengersen, K., Robert, C., and Titterington, M. (2011). *Mixtures: estimation and applications*, volume 896. John Wiley & Sons. 4

Muirhead, R. (1982). *Aspects of Multivariate Statistical Theory*. New York: John Wiley. 59, 60

Pole, A., West, M., and Harrison, J. (1994). *Applied Bayesian forecasting and time series analysis*. CRC press. 23

Primiceri, G. E. (2005). "Time varying structural vector autoregressions and monetary policy." *The Review of Economic Studies*, 72(3): 821–852. 34

Rennie, J. D. (2006). "Jacobian of the singular value decomposition with application to the trace norm distribution."
URL http://qwone.com/~jason/writing/svdJacobian.pdf 60

Schotman, P. and Van Dijk, H. K. (1991a). "A Bayesian analysis of the unit

root in real exchange rates." *Journal of Econometrics*, **49**: 195–238. 23, 71, 78

Schotman, P. C. and Van Dijk, H. K. (1991b). "On Bayesian routes to unit roots." *Journal of Applied Econometrics*, 6(4): 387–401. 22, 23

Siegel, P. M. and Hodge, R. W. (1968). "A causal approach to the study of measurement error." *Methodology in Social Research. New York: McGraw-Hill*, 28: 59. 35

Sims, C. (1980). "Macroeconomics and Reality." *Econometrica*, 48(1): 1–48. 9, 34

Sims, C. A. (2004). "Econometrics for Policy Analysis: Progress and Regress." *De Economist*, 152(2): 167–175. 23

— (2005). "Dummy Observation Priors Revisited." Manuscript, Princeton University. 23

— (2007). "Thinking about instrumental variables." Manuscript, Princeton University. 17

— (2008). "Making macro models behave reasonably." Manuscript, Princeton University. 3, 16

Sims, C. A. and Zha, T. (1998). "Bayesian methods for dynamic multivariate models." *International Economic Review*, 949–968. 34

Strachan, R. and Van Dijk, H. K. (2004). "Valuing structure, model uncertainty and model averaging in vector autoregressive processes." Technical Report EI 2004-23, Econometric Institute. 17

— (2013). "Evidence on features of a DSGE business cycle model from Bayesian model averaging." *International Economic Review*, 54(1): 385–402. 4, 23

Strachan, R. W. (2003). "Valid Bayesian estimation of the cointegrating error correction model." *Journal of Business & Economic Statistics*, 21(1): 185–195. 34

Strachan, R. W. and Inder, B. (2004). "Bayesian analysis of the error correction model." *Journal of Econometrics*, 123(2): 307–325. 17

Sugita, K. et al. (2016). "Bayesian inference in Markov switching vector error correction model." *Economics Bulletin*, 36(3): 1534–1546. 34

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288. 20

Tuyl, F., Gerlach, R., and Mengersen, K. (2008). "A comparison of Bayes–Laplace, Jeffreys, and other priors: the case of zero events." *The American Statistician*, 62(1): 40–44. 23

Uhlig, H. (1994). "On Jeffreys prior when using the exact likelihood function." *Econometric Theory*, 633–644. 16

Van Dijk, H. K. and Kloek, T. (1980). "Further experience in Bayesian analysis using Monte Carlo integration." *Journal of Econometrics*, 14(3): 307–328. 23

Verdinelli, I. and Wasserman, L. (1995). "Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ratio." *Journal of the American Statistical Association*, 90: 614–618. 29

Villani, M. (1998). "On the representation of ignorance regarding a matrix of reduced rank." Technical report, Stockholm University, Department of Statistics. 17

— (2000). "Aspects of Bayesian cointegration." Ph.D. thesis, Stockholm University, Sweeden. 17

— (2005). "Bayesian reference analysis of cointegration." *Econometric Theory*, 21(02): 326–357. 17

— (2009). "Steady-state priors for vector autoregressions." *Journal of Applied Econometrics*, 24(4): 630–650. 23

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. John Wiley and Sons, New York. Reprinted in 1996. 50, 51, 52

Zellner, A., Ando, T., Baştürk, N., Hoogerheide, L., and Van Dijk, H. K. (2014). "Bayesian analysis of instrumental variable models: Acceptance-rejection within direct Monte Carlo." *Econometric Reviews*, 33: 3–35. 4, 38, 40, 41, 61

Zellner, A., Bauwens, L., and Van Dijk, H. K. (1988). "Bayesian specification analysis and estimation of simultaneous equation models using Monte Carlo methods." *Journal of Econometrics*, 38(1): 39–72. 15, 50, 61