BAYESIAN FACTOR MODEL SHRINKAGE FOR LINEAR IV REGRESSION WITH MANY INSTRUMENTS

P. RICHARD HAHN[1], JINGYU HE[2], AND HEDIBERT LOPES[3]

Abstract. A Bayesian approach for the many instruments problem in linear instrumental variable models is presented. The new approach has two components. First, a slice sampler is developed, which leverages a decomposition of the likelihood function that is a Bayesian analogue to two-stage least squares. The new sampler permits non-conjugate shrinkage priors to be implemented easily and efficiently. The new computational approach permits a Bayesian analysis of problems that were previously infeasible due to computational demands that scaled poorly in the number of regressors. Second, a new predictor-dependent shrinkage prior is developed specifically for the many instruments setting. The prior is constructed based on a factor model decomposition of the matrix of observed instruments, allowing many instruments to be incorporated into the analysis in a robust way. Features of the new method are illustrated via a simulation study and three empirical examples.

1. Introduction

Linear instrumental variable (IV) regression is a common method for calculating treatment effects for endogenous regressors. While a single valid instrument is theoretically sufficient to identify a treatment effect in this setting, one might hope to increase the precision of IV estimates by including additional instruments (both new instruments as well as interactions between, and nonlinear transformations of, existing instruments). This strategy, in conjunction with the increasing availability of high-dimensional data, has spurred recent interest in treatment effect estimation procedures which utilize a large number of instrumental variables. The need for special methods in this context is due to the fact that the usual two-stage least squares (2SLS) estimator is recognized to have non-negligible bias when the

[1] Booth School of Business, University of Chicago.
[2] Department of Statistics, University of Chicago.
[3] INSPER — Institute of Education and Research.

The first author thanks the Booth School of Business for supporting this research.

sample size is inadequate relative to the number of instruments [Bekker, 1994, Newey and Smith, 2004].

In this paper, we describe a Bayesian model that is tailored to the many-instruments setting. We first infer a factor model governing the covariation structure among the instrumental variables. Based on this structure, we develop a shrinkage prior for the first-stage regression coefficients (the treatment variable as a function of the instruments) which favors the assumption that this common factor structure also predicts the treatment variable. Our paper is similar in spirit to non-Bayesian methods which regularize the first stage regression [Chamberlain and Imbens, 2004, Okui, 2011, Carrasco, 2012]. Our approach differs from these earlier approaches in two ways. First, we use a factor-motivated shrinkage prior that encodes substantive biases. Second, we use a fully Bayesian model, in which the first and second stage estimations mutual influence one another. This joint influence can be observed concretely in the form of our proposed slice sampler.

The body of literature on Bayesian IV is extensive and long-established. Seminal references include Lindley and El-Sayed [1968] and Dreze [1976]. Other notable papers include Geweke [1996], Chao and Phillips [1998] and Lopes and Polson [2014]. An excellent textbook treatments is Rossi et al. [2006], chapter 7. Recent work has looked at variations on the typical Gaussian IV model, such as non-parametric error terms [Conley et al., 2008] and violated exclusion restrictions [Conley et al., 2012, Chan and Tobias, 2014].

Despite this abundance of literature, most of it has not specifically concerned the many instruments setting. Three notable exceptions are Chamberlain and Imbens [2004], Hahn and Hansen [2011] and Koop et al. [2012]. Chamberlain and Imbens [2004] propose a hierarchical model which serves to regularize the first-stage regression, while Hahn and Hansen [2011] revisit the problem of prior specification in the many instruments setting and develop a reparametrization that relates the use of diffuse priors to limited information maximum likelihood (LIML) methods. Our work differs from these contributions in that we focus on a class of informative, instrument-dependent, priors. Koop et al. [2012] considers instrument selection using variable selection priors, whereas our approach focuses on shrinkage rather than on selection. Conceptually, our approach presumes (with probability one) that all of the available instruments contribute to the first stage regression, whereas a variable selection approach puts non-zero probability on the possibility that some candidate instruments are spurious (with exactly zero first-stage coefficients).

Our approach also provides an *a priori* bias that directions of shared covariation among the instruments are more likely to explain the observed treatment. In this, we complement recent non-Bayesian work considering the use of factor models in the many instruments context [Groen and Kapetanios, 2009, Ng and Bai, 2009, Kapetanios and Marcellino, 2010]. Our approach differs from these earlier uses of factor models in two important respects. First, we supply our biases via a shrinkage prior, rather than via a direct dimension reduction step; as such, our approach is not complicated by a separate factor selection step. Second, the factor modeling ideas utilized in our prior are based on a factor model that permits additive idiosyncratic errors, as distinct from factor models which simply rotate the instrument matrix (e.g., singular value decompositions). See section 3.2 for further details on this distinction.

The balance of the paper is organized as follows. Section 2 reviews the Bayesian IV model and presents a slice sampler for use with general shrinkage priors. Section 3 introduces the new factor shrinkage prior. Section 4 reports on simulation studies and section 5 demonstrates our method on empirical data.

2. "Two-stage" Bayesian IV

This section describes a reparametrization of the usual Gaussian instrumental variables (IV) model. Building on this representation, we develop a slice sampler which facilitates the use of non-conjugate shrinkage priors for the instrument coefficients.

2.1. **A reparametrization of Bayesian linear IV.** The starting point of Bayesian approaches to endogenous regressors is the structural equation model

$$\begin{aligned} y_i &= \beta x_i + \epsilon_y \\ x_i &= \mathrm{z}_i^t \boldsymbol{\delta} + \epsilon_x. \end{aligned} \tag{1}$$

where (ϵ_x, ϵ_y) are jointly Gaussian with mean zero and covariance

$$\operatorname{cov}\begin{pmatrix}\epsilon_x \\ \epsilon_y\end{pmatrix} := \mathbf{S} = \begin{pmatrix}\sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2\end{pmatrix}.$$

The variable x_i is referred to as the treatment variable, y_i is the response variable and z_i is a vector of *instruments*. The unknown parameters in this model are β, δ, σ_x^2, σ_y^2 and $\sigma_{xy} = \sigma_{yx}$; the parameter of interest is β. Because of the implied covariance between x_i and ϵ_y, valid estimates of β cannot be obtained from just a regression of y_i onto x_i.

The joint distribution of the observables can be found by substitution

$$\begin{aligned} x_i &= \mathrm{z}_i^t\boldsymbol{\delta} + \epsilon_x, \\ y_i &= \mathrm{z}_i^t\boldsymbol{\delta}\beta + \beta\epsilon_x + \epsilon_y. \end{aligned} \tag{2}$$

A further reparametrization yields

$$\begin{aligned} x_i &= \mathrm{z}_i^t\boldsymbol{\delta} + \nu_x, \\ y_i &= \mathrm{z}_i^t\boldsymbol{\delta}\beta + \nu_y, \end{aligned} \tag{3}$$

with

$$\operatorname{cov}\begin{pmatrix}\nu_x \\ \nu_y\end{pmatrix} = \boldsymbol{\Omega} = \mathbf{U}\mathbf{S}\mathbf{U}^t,$$

where $\mathbf{U} = \begin{pmatrix}1 & 0 \\ \beta & 1\end{pmatrix}$. Equation (3) is referred to as the "reduced form" equations, in contrast to (1), the "structural" equations.

The focus in this paper will be on priors for $\boldsymbol{\delta}$ when the number of instruments p is large relative to the number of available observations n. Priors over the remaining parameters are determined by a factorization of the likelihood based on $\epsilon_y \mid \epsilon_x \sim \mathrm{N}(\alpha\epsilon_x, \xi^2)$, where

$$\alpha = \frac{\sigma_y}{\sigma_x}\rho; \quad \xi^2 = (1-\rho^2)\sigma_y^2, \tag{4}$$

with $\rho \equiv \frac{\sigma_{xy}}{\sigma_x\sigma_y}$. The matrix Ω can be written in terms of β, α, ξ^2 and σ_x^2,

$$\Omega = \begin{pmatrix}\sigma_x^2 & (\beta+\alpha)\sigma_x^2 \\ (\beta+\alpha)\sigma_x^2 & (\beta+\alpha)^2\sigma_x^2 + \xi^2\end{pmatrix}, \tag{5}$$

which in turn corresponds to the following factorization of the joint likelihood over observables (x, y):

$$
\begin{aligned}
f(x, y \mid \mathrm{z}) &= f(y \mid x, \mathrm{z}) f(x \mid \mathrm{z}) \\
&= \mathrm{N}_{y|x}(x\beta + \alpha(x - \mathrm{z}^t\boldsymbol{\delta}), \xi^2)\times \\
&\quad \mathrm{N}_x(\mathrm{z}^t\boldsymbol{\delta}, \sigma_x^2).
\end{aligned} \tag{6}
$$

The appearance of $\boldsymbol{\delta}$ in both factors on the right-hand side means that observations of (y_i, z_i) allow one to disentangle β and α. It is concievable, of course, that in a given applied problem one instead has

$$
f(x, y \mid \mathrm{z}) = f(y \mid x, \mathrm{z}) f(x \mid \mathrm{z}) = \mathrm{N}_{y|x}(x\beta + \alpha(x - \mathrm{z}^t\boldsymbol{\delta}), \xi^2)\mathrm{N}_x(\mathrm{z}^t\boldsymbol{\delta}^*, \sigma_x^2), \tag{7}
$$

with $\boldsymbol{\delta}^* \neq \boldsymbol{\delta}$. The assumption that $\boldsymbol{\delta}^* = \boldsymbol{\delta}$ is referred to as the instrument exclusion restriction and in general is untestable. See Conley et al. [2012] and Chan and Tobias [2014] for approaches which weaken this assumption, yielding only partial identification of β. In this paper, the exclusion restriction will be assumed.

2.2. **A slice sampler for Bayesian IV with arbitrary shrinkage priors.** In this section we adapt the elliptical slice sampler of Murray et al. [2009] to the instrumental variables setting. Murray et al. [2009] develop an algorithm for sampling from posteriors proportional to $f(\mathbf{y} \mid \boldsymbol{\delta})\pi(\boldsymbol{\delta})$ when $\pi(\boldsymbol{\delta})$ is Gaussian; that is, they advertise their algorithm as applying to the case of arbitrary likelihood and Gaussian prior. In the present case, we want to sample from a posterior proportional to $f(\mathbf{x} \mid \boldsymbol{\delta}, \sigma_x^2) f(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\delta}, \alpha, \beta, \xi^2)\pi(\boldsymbol{\delta})$ where $f(\mathbf{x} \mid \boldsymbol{\delta}, \sigma_x^2)$ is a Gaussian *likelihood* (as described in the previous section). Letting $\pi(\boldsymbol{\delta} \mid \mathbf{x}, \mathbf{Z}, \sigma_x^2)$ denote the Gaussian posterior for $\boldsymbol{\delta}$ under a flat prior, and integrating α, β and ξ^2 from the model a priori yields

$$
\pi(\boldsymbol{\delta} \mid \sigma_x^2, \mathbf{x}, \mathbf{y}, \mathbf{Z}) \propto \pi(\boldsymbol{\delta}, \sigma_x^2 \mid \mathbf{x}, \mathbf{Z}) f(\mathbf{y} \mid \mathbf{x}, \mathbf{Z}, \boldsymbol{\delta})\pi(\boldsymbol{\delta}), \tag{8}
$$

for arbitrary $\pi(\boldsymbol{\delta})$. Therefore, we can directly apply the algorithm of Murray et al. [2009], using $\pi(\boldsymbol{\delta} \mid \mathbf{x}, \mathbf{Z}, \sigma_x^2)$ in place of their Gaussian prior and using $f(\mathbf{y} \mid \mathbf{x}, \mathbf{Z}, \boldsymbol{\delta})\pi(\boldsymbol{\delta})$ in place of their likelihood.

In this paper, we will assume a normal-inverse-Gamma prior is used for (α, β, ξ^2), with prior mean $\mathrm{E}(\alpha) = \mathrm{E}(\beta) = 0$, covariance of $\mathrm{diag}((c_\beta^{-1}, c_\alpha^{-1})^t)$, and Gamma shape parameter of $s/2$ and scale parameter of $\kappa/2$. Define $\tilde{x}_i := (x_i, x_i - \mathrm{z}_i^t\boldsymbol{\delta})$. Let $\mathbf{M} = \mathrm{diag}((c_\beta, c_\alpha)^t) + \tilde{\mathbf{x}}^t\tilde{\mathbf{x}}$, $b = s + \mathbf{y}^t\mathbf{y} - \mathbf{y}^t\tilde{\mathbf{x}}\mathbf{M}^{-1}\tilde{\mathbf{x}}^t\mathbf{y}$, and $a = n + \kappa$. Note that $\tilde{\mathbf{x}}$, $\mathbf{M}$, a and b depend implicitly on $\boldsymbol{\delta}$; then $f(\mathbf{y} \mid \mathbf{x}, \mathbf{Z}, \boldsymbol{\delta}) \propto \det(\mathbf{M})^{-\frac{1}{2}} b^{-\frac{a}{2}}$, which is the kernel of a multivariate t-distribution. For the time being, we refer to a generic prior $\pi(\boldsymbol{\delta})$; section 3 describes our new factor shrinkage prior.

With these definitions, our sampler can be described as follows (for fixed σ_x^2). Let $\hat{\boldsymbol{\delta}} = (\mathbf{Z}\mathbf{Z}^t)^{-1}\mathbf{Z}\mathbf{x}$ and for an initial value of $\boldsymbol{\delta}$, define $\Delta := \boldsymbol{\delta} - \hat{\boldsymbol{\delta}}$.

Elliptical slice sampler for Bayesian IV

```
(1) Draw $\zeta \sim \mathrm{N}(0, \sigma_x^2(\mathbf{Z}\mathbf{Z}^t)^{-1})$.
(2) For $\upsilon \sim \mathrm{Uniform}(0, 1)$ define $\ell := \log(f(\mathbf{y} \mid \mathbf{x}, \mathbf{Z}, \boldsymbol{\delta})) + \log(\pi(\boldsymbol{\delta})) + \log(\upsilon)$.
(3) Draw angle $\varphi \sim \mathrm{Uniform}(0, 2\pi)$; set $lower \leftarrow \varphi - 2\pi$ and $upper \leftarrow \varphi$.
(4) Set $\Delta' \leftarrow \Delta \cos\varphi + \zeta \sin\varphi$ and $\boldsymbol{\delta}' \leftarrow \hat{\boldsymbol{\delta}} + \Delta'$ .
(5) while $\log(f(\mathbf{y} \mid \mathbf{x}, \mathbf{Z}, \boldsymbol{\delta}')) + \log(\pi(\boldsymbol{\delta}')) < \ell$
    (a) if $\varphi < 0$, set $lower \leftarrow \varphi$, else set $upper \leftarrow \varphi$.
    (b) Draw angle $\varphi \sim \mathrm{Uniform}(lower, upper)$
    (c) Update $\Delta' \leftarrow \Delta \cos\varphi + \zeta \sin\varphi$ and $\boldsymbol{\delta}' \leftarrow \hat{\boldsymbol{\delta}} + \Delta'$.
(6) Set $\Delta \leftarrow \Delta'$ and $\boldsymbol{\delta} \leftarrow \hat{\boldsymbol{\delta}} + \Delta'$.
```

See Murray et al. [2009] for a proof that this algorithm has the desired stationary distribution. Note that this sampler requires the ability to evaluate the (possibly unnormalized) prior density $\pi(\boldsymbol{\delta})$. Sampling the univariate parameter σ_x^2 can be done in between taking samples of $\boldsymbol{\delta}$, using either a conditionally conjugate inverse-gamma prior (which includes a flat prior as a limiting case) or an arbitrary prior in conjunction with a Metropolis-Hastings step.

Finally, given samples of $\boldsymbol{\delta}$, we sample (α, β, ξ^2) from $\pi(\alpha, \beta, \xi^2 \mid \mathbf{x}, \mathbf{Z}, \mathbf{y}, \boldsymbol{\delta})$, which is a conjugate Gaussian regression with predictor vector $\tilde{\mathbf{x}}$. More specifically, draw ξ^2 from an inverse-Gamma distribution with shape parameter $b/2$ and scale parameter $a/2$, then draw (α, β) as a vector with mean $\mathbf{M}^{-1}\tilde{\mathbf{x}}^t\mathbf{y}$ and covariance $\xi^2\mathbf{M}^{-1}$.

We stress that this approach is dramatically more efficient than existing algorithms (per sample) for implementing shrinkage priors, owing to the fact that individual iterations do not loop over individual coefficients nor are there observation-specific latent variables to sample. Thus, we are able to fit larger data sets than previous methods, both in terms of the number of observations, n, as well as the number of instruments, p.

3. A factor-based shrinkage prior

If it were possible to extract latent factors governing the correlation structure in a vector of instruments, one might suppose that these factors would make "strong" instruments. This is simply the usual factor regression rationale, applied to the treatment equation in an instrumental variable analysis.

It is worth distinguishing how this factor regression prior assumption differs from a variable selection prior. Instead of presuming we have only a few good instruments among a whole batch of candidate instruments and that we simply do not know which ones they are, we instead suppose that each available instrument may itself be weak, but that there exists a linear combination of them that is, taken together, much stronger. In this section, we discuss how this intuition can be incorporated into a prior distribution over $\boldsymbol{\delta}$.

To begin, suppose the covariance of the instruments decomposes as

$$\text{cov}(\mathrm{z}_i) = \mathbf{B}\mathbf{B}^t + \boldsymbol{\Psi}^2. \tag{9}$$

where $\mathbf{B}$ is a p-by-k matrix with and $\mathbf{\Psi}^2$ is diagonal with non-negative elements. Any covariance matrix admits such a decomposition, but we will be specifically interested in the case where $k \ll p$; see the following section for additional details.

Next, we consider the factor regression model

$$x_i = \boldsymbol{\theta}\hat{\mathrm{f}}_i + \varepsilon_i, \tag{10}$$

where $\hat{f}_i = \mathrm{E}(f_i \mid \mathrm{z}_i, \mathbf{B}) = \mathbf{B}^t(\mathbf{B}\mathbf{B}^t + \mathbf{\Psi}^2)^{-1}\mathrm{z}_i = \mathbf{A}\mathrm{z}_i$. This expression follows by positing a joint Gaussian distribution between k-by-1 latent factors f and instrument vector z, with cross covariance $\mathbf{B}$ and marginal covariance $\mathrm{cov}(\mathrm{f}) = \mathbf{I}_k$. This model encodes the assumption that the conditional mean of $\mathbf{x}$ lies in the same subspace as the directions of common covariation in z. By substitution, we see that it implies the $\boldsymbol{\delta}$ can be written as $\boldsymbol{\delta}^t = \boldsymbol{\theta}\mathbf{A}$.

If $\boldsymbol{\delta}$ does not actually lie in the row space of $\mathbf{A}$, the model has a misspecified support, which can dramatically degrade inference. To accommodate this possibility, we instead consider the over-parametrized factor model

$$x_i = \boldsymbol{\theta}\hat{\mathrm{f}}_i + \boldsymbol{\eta}\hat{r}_i + \varepsilon_i, \tag{11}$$

where $\hat{r}_i = \mathrm{z}_i - \mathbf{A}^\dagger\mathbf{A}\mathrm{z}_i$. Such models, referred to as "partial factor models" [Hahn et al., 2013], entail that

$$\boldsymbol{\delta}^t = \boldsymbol{\theta}\mathbf{A} + \boldsymbol{\eta}(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}). \tag{12}$$

With this formulation, a prior over $\boldsymbol{\delta}$ can be induced by placing priors over $\tilde{\boldsymbol{\delta}}^t \equiv (\boldsymbol{\theta}, \boldsymbol{\eta})$. More specifically, a strong shrinkage prior over the $p+k$ regression coefficients comprising $\tilde{\boldsymbol{\delta}}$ embeds the prior bias that the derived factors $\hat{\mathrm{f}}_i$ are likely to play a strong role in determining the conditional mean of the treatment. Meanwhile, the prior still gives $\boldsymbol{\delta}$ full support in $\mathbb{R}^p$; $\boldsymbol{\theta}$ constitutes the part of $\boldsymbol{\delta}$ that lies in the row space of $\mathbf{A}$, while $\boldsymbol{\eta}$ constitutes the part lying in the corresponding orthogonal complement. Observe also that sparsity in the over-parametrized $\tilde{\boldsymbol{\delta}}$ basis (in the sense of a few very large coefficients and many more nearly zero ones) does not in general result in a sparse $\boldsymbol{\delta}$ vector.

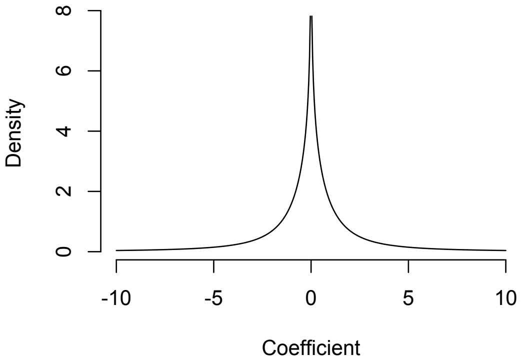

FIGURE 1. In the over-parametrized space the individual coefficients have a prior with a pole at the origin and polynomial tails.

3.1. **A latent variable specification.** One obvious approach to furnishing a bias towards factor structure would be to sample $\tilde{\boldsymbol{\delta}}$ directly. However, this approach proves to be inefficient precisely because $\tilde{\boldsymbol{\delta}}$ is unidentified; exploring the corresponding multimodal posterior is difficult. Instead, we will leverage results from the theory of pseudo-inverses to specify a prior that permits drawing posterior samples of $\boldsymbol{\delta}$ directly, while still imposing the sparsity penalty in the $\tilde{\boldsymbol{\delta}}$ partial factor representation.

First, note that $\boldsymbol{\delta} = \mathbf{H}\tilde{\boldsymbol{\delta}}$ where

$$\mathbf{H}^t = \begin{pmatrix} \mathbf{A} \\ \mathbf{I} - \mathbf{A}^\dagger\mathbf{A} \end{pmatrix}$$

Therefore, by the theory of pseudo-inverses, we have that $\tilde{\boldsymbol{\delta}} = \mathbf{H}^\dagger\boldsymbol{\delta} + (\mathbf{I} - \mathbf{H}^\dagger\mathbf{H})\omega$ for any $(p+k)$-dimensional real vector ω. Using this identity, conditional on ω, we evaluate the density $\pi(\boldsymbol{\delta} \mid \omega) \propto \prod_{j=1}^{p+k} \log(1 + 4/\tilde{\delta}_j^2)$ (up to an unknown Jacobian factor). This choice of prior density in $\tilde{\boldsymbol{\delta}}$ space is motivated by an analytic approximation of the so-called horseshoe prior; see Carvalho et al. [2010] for details and operating characteristics of this prior. Figure 1 depicts this prior for a single coefficient. We complete the specification with independent standard normal priors over the elements of ω, which is convenient, maintains the full support, and is demonstrated to work well in simulation studies.

However, observing that $(\mathbf{I} - \mathbf{H}^{\dagger}\mathbf{H})$ has rank k and is idempotent, it follows that for $\mathbf{U}\mathbf{U}^t = (\mathbf{I} - \mathbf{H}^{\dagger}\mathbf{H})$ with $\mathbf{U}$ a $(p+k)$-by-k matrix, our prior may be specified via $\tilde{\boldsymbol{\delta}} = \mathbf{H}^{\dagger}\boldsymbol{\delta} + \mathbf{U}\mathrm{w}$, where w is a k dimensional vector, also with a standard normal prior. Putting these pieces together, our prior can be expressed as

$$
\begin{aligned}
\pi(\boldsymbol{\delta}) &= \int \pi(\boldsymbol{\delta} \mid \mathrm{w})\phi(\mathrm{w})d\mathrm{w}, \\
&= h \int \prod_j \log\left(1 + \frac{4}{\tilde{\delta}_j^2}\right)\phi(\mathrm{w})d\mathrm{w}, \qquad (13) \\
\tilde{\boldsymbol{\delta}} &= \mathbf{H}^{\dagger}\boldsymbol{\delta} + \mathbf{U}\mathrm{w},
\end{aligned}
$$

where h is a normalizing constant and $\phi(\cdot)$ denotes the (multivariate, independent) standard normal density function. To implement this prior, we alternately sample from w and $\boldsymbol{\delta}$, using the slice sampler previously described. Note that w appears in $\pi(\boldsymbol{\delta} \mid \mathrm{w})$ as a location parameter, and so does not appear in its normalizing constant (and vice-versa, interchanging the roles of $\boldsymbol{\delta}$ and w), which allows the algorithm to proceed using only evaluations of $\prod_j \log\left(1 + \frac{4}{\tilde{\delta}_j^2}\right)$. Finally, our implementation incorporates a global scale parameter v, so that our prior is evaluated as $\pi(\boldsymbol{\delta} \mid \mathrm{w}, v) \propto v^{-p-k}\prod_j \log\left(1 + \frac{4}{(\tilde{\delta}_j/v)^2}\right)$, where v is sampled via a Metropolis-Hastings step.

In practice, of course, $\mathbf{B}$ and $\boldsymbol{\Psi}^2$ are unknown and must be estimated in some fashion. Although a natural option would be to infer these parameters hierarchically, we will instead use point estimates. The reason to avoid a Bayesian model for the instrument vector is mainly computational; current factor model sampling algorithms scale poorly in n. Additionally, it is often the case that instruments are not continuous, which substantially complicates the factor modeling [Murray et al., 2013]. The concrete form of our empirical plug-in estimate of $\mathbf{B}$ and Ψ is the topic of the next subsection.

3.2. **The Frisch decomposition.** The notion of "shared factors" among vectors of measurements can be characterized in terms of an optimization problem motivated by the early work of Ragnar Frisch on "confluence analysis" [Frisch, 1934]. Specifically, given a covariance matrix $\mathbf{\Sigma}$, consider the following *rank minimization problem*:

$$\begin{aligned} \min_{\mathbf{D}} \quad & \operatorname{rank}(\mathbf{\Sigma} - \mathbf{D}) \\ \text{s.t.} \quad & \mathbf{D} \text{ diagonal}, \\ & \mathbf{\Sigma} - \mathbf{D} \geq 0. \end{aligned} \tag{14}$$

If $\mathbf{D}^*$ is a solution to (14), denote a matrix pair $(\mathbf{\Psi}^2, \mathbf{B})$ a *Frisch decomposition* of $\mathbf{\Sigma}$, if

$$\mathbf{\Psi}^2 = \mathbf{D}^*; \quad \mathbf{B}\mathbf{B}^t = \mathbf{\Sigma} - \mathbf{D}^*. \tag{15}$$

By assuming $\mathbf{\Sigma}$ known, this problem is non-statistical in nature, yet it readily captures an intuition about what makes factor models appealing as descriptions of data. Factor models are popular not merely because they decompose covariance structure into a common component and an independent (diagonal) component, but because it is anticipated that this decomposition can be done parsimoniously. Indeed, any p-by-p covariance matrix has a $p-1$ dimensional factor representation (let $\mathbf{\Psi}^2 = \iota_p \mathbf{I}$ for ι_p the smallest eigenvalue of the singular-value decomposition), whereas the Frisch decomposition demands that we have the most concise of all such descriptions.

Although solving (14) exactly is quite difficult, high quality approximations are available using a surrogate objective function based on the matrix trace [Fazel, 2002]:

$$\begin{aligned} \min_{\mathbf{D}} \quad & \operatorname{trace}(\mathbf{\Sigma} - \mathbf{D}) \\ \text{s.t.} \quad & \mathbf{D} \text{ diagonal}, \\ & \mathbf{\Sigma} - \mathbf{D} \geq 0. \end{aligned} \tag{16}$$

The trace approximation is convex and can be routinely solved by readily available software [Grant and Boyd, 2013, 2008]. The specifics of this approximation are beyond the scope of this paper; see Ning et al. [2015] for an excellent overview with many references. The trace approximation serves to extract a “sharper” set of eigenvectors, in the sense of having a more rapidly decaying set of eigenvalues, as seen in Figure 2, which overlays the eigenvalues of an example covariance matrix $\mathbf{\Sigma}$ and $\mathbf{\Sigma} - \mathbf{D}^*$, where $\mathbf{D}^*$ solves (16). In this sense, the trace heuristic still isolates “communalities”. Henceforth, whenever a Frisch decomposition is referred to, it is to be understood that it is computed approximately using the feasible trace formulation in (16).

With $\mathbf{D}^*$ in hand, it remains to factor $\mathbf{\Sigma} - \mathbf{D}^* = \mathbf{B}\mathbf{B}^t$. Any factorization will do, in our empirical specification we use a routine singular value decomposition. Note that many empirical factor models implicitly take $\mathbf{D}$ to be the zero matrix.

Finally, we note two additional details concerning the implemented Frisch decomposition. First, the solution to (14) is invariant to row and column scaling operations, while (16) is not. This observation has motivated weighted minimum trace approximations that attempt to define and compute an optimal weight matrix [Shapiro, 1982, Ning et al., 2015]. As a crude heuristic, the approach taken here is to solve (16) applied to the sample correlation matrix as opposed to the sample covariance matrix.

Similarly, because $\mathbf{\Sigma}$ is only known up to an empirical estimate, the actual rank of $\mathbf{B}$ will tend not to be reduced, regardless of if there is true underlying factor structure or not. Accordingly, $\mathbf{H}$ may be defined by approximating $\mathbf{B}$ by its first few dominate eigenvectors. Various heuristics can be used based on inspection of the eigenvalues of $\mathbf{B}$. Our preferred heuristic, which we observe to work well empirically, is to truncate at the point of largest ratio between consecutive eigenvalues up to a pre-specified maximum $k_{max} \ll p$.

Our perspective is that the choice of how to determine k can be thought of as a *definition* of the prior being used. The factor shrinkage prior is meant to encode the idea that directions of strong covariation among the instruments are likely to be good predictors of the treatment variable. Specifying a heuristic method for selecting k serves as a formalization of what it means to be a “strong direction of covariation”. At the same time, it is important to bear in mind that the implied prior for $\boldsymbol{\delta}$ still has full support, so “misspecification” of the prior in the sense of sub-optimally selecting k is not as damaging to posterior inferences as in the usual factor regression context.

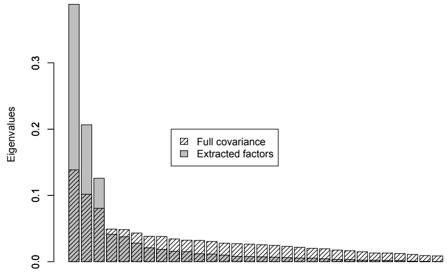

Figure 2. An illustration of how the eigenvalues of a full covariance matrix $\mathbf{\Sigma}$ can be flatter than the eigenvalues of the trace-heuristic derived loadings matrix $\mathbf{\Sigma} - \mathbf{D}^*$. This occurs when $\mathbf{\Sigma}$ has an underlying factor structure with relatively large idiosyncratic variances.

4. Simulation study

Simulation studies reveal that, when exploitable factor structure is evident among the instruments, our factor shrinkage prior (FSP) — based on an approximate Frisch decomposition of an estimated covariance matrix — delivers more precise inferences than a model using a shrinkage prior on $\boldsymbol{\delta}$ that is not dependent on $\mathbf{Z}$. Despite using empirically estimated information about the factor structure, injecting this information into our shrinkage prior yields better inferences concerning $\boldsymbol{\delta}$, which translates to better inferences concerning the treatment effect coefficent, β. By better, more precise inference, we specifically mean shorter posterior credible intervals (with comparable coverage rates) and smaller mean squared error.

Naturally, in any Bayesian simulation the performance is ultimately dictated by how closely the prior used to simulate the data matches the prior used in the model, which, at a high level, is an unilluminating finding. However, in our model, the data dependence adds an additional degree of freedom. The goal of the simulation study reported here is to describe the relationship between the instrument matrix and the underlying parameters and to understand in what situations our model is anticipated to improve over a data-independent shrinkage prior.

First, in extensive simulation studies not reported here, we determined that the qualitative behavior of the factor shrinkage prior mimics the behavior of the partial factor model described in Hahn et al. [2013]. That is, when there is strong factor structure, our model performs on par with a factor model that explicitly models the instrument matrix (in terms of mean squared estimation error). When there is no strong factor structure, our model performs on par with a direct shrinkage prior (canonically, the horseshoe prior of Carvalho et al. [2010]). We do see modest deterioration due to our empirical approximation, meaning that generally speaking our prior performed somewhere in between the full factor model and the pure shrinkage prior in all regimes, usually closer to the better of the two. The purpose of the present investigation is to determine if this improved estimation translates to improved inferences regarding β.

There are two aspects to consider when it comes simulating factor structure. The first aspect is whether or not the instrument matrix exhibits strong patterns of correlation and whether or not that pattern can be expressed in Frisch form with $k \ll p$. The second aspect is whether or not the treatment variable depends on these same patterns of dependence exhibited by the instrument matrix. Here we only consider the second issue. We consider it a given that the matrix of instruments is assumed to have some degree of factor structure, hence our simulation study reports how our model performs as we vary the strength of dependence between these factors and the treatment variable. The basic model we generate from is the following. For $n = 200$, $p = 20$, and $k = 3$,

$$\begin{aligned}
&\mathrm{z}_i \sim \mathrm{N}(0, \mathbf{B}\mathbf{B}' + \mathbf{\Psi}^2),\\
(17) \qquad &x_i = \mathrm{z}_i^t \boldsymbol{\delta} + \sigma_x \epsilon_{\mathbf{x}|\mathrm{z};i},\\
&y_i = \beta \mathbf{x}_i + \alpha(\mathbf{x}_i - \mathrm{z}_i^t \boldsymbol{\delta}) + \xi \epsilon_{\mathbf{y}|\mathrm{z},\mathbf{x};i},
\end{aligned}$$

where all error terms (the ϵ's) are independent standard normal. We draw β and α independently from normal inverse-gamma distributions with precision parameters 1 and 1/2, respectively, and common shape parameters $\kappa/2 = 16$ and scale parameter $s/2 = 4$.

In preliminary simulations, we determined two conditions that confer an advantage to our method. First, our method will *not* outperform a straight shrinkage prior if there is ample data to estimate $\boldsymbol{\delta}$ without the prior having any noticeable influence. In terms of model parameters, this means that σ_x and ξ cannot be too small, otherwise the problem is "easy" and the two priors (indeed, any sensible prior) will perform comparably. Second, we must have a sufficient sample size to estimate $\mathbf{\Sigma}$, and hence $\mathbf{B}$, reasonably well. Although $\mathbf{\Sigma}$ is a p-by-p matrix, the assumption of factor structure suggests that estimation is possible with fewer observations than if there were no factor structure because the "intrinsic dimension" is smaller than p, although some amount of regularization may be beneficial [Aswani et al., 2011]. Intuitively, our method is useful when 1) the prior matters and 2) the prior information used is good. Practically, this means that we want to trust our ability to estimate $\mathbf{\Sigma}$ reasonably well before proceeding with this prior. One case where this is sensible is when there are observations of z, without the associated x and y observations, that can improve estimation of $\mathbf{\Sigma}$ for the purpose of prior specification (a setting sometimes referred to as semi-supervised learning [Belkin et al., 2006]). In our simulation study we assure that this is the case by generating $\mathbf{B}$, $\mathbf{\Psi}^2$, $\boldsymbol{\delta}$, σ_x and ξ as follows.

- Each of the k column of $\mathbf{B}$ is drawn uniformly on the unit hypersphere. The columns of $\mathbf{B}$ are not forced to be orthogonal.
- Each ψ_j, $j = 1, \ldots, p$, is drawn uniformly between $[2a_j, 4a_j]$ where $a_j = \sqrt{\sum_l b_{jl}^2}$. In words, the standard deviation of the idiosyncratic errors in each dimension of z_i are between two and four times as big as the "signal" due to the factor loadings in the corresponding row. This is important in our simulation because if the factors associates too strongly with any single instrument, then that instrument alone will be sufficient to accurately estimate the conditional expectation of the treatment and no benefit is gained from the factor prior.
- Define $\mathbf{A} := \mathbf{B}^t(\mathbf{B}\mathbf{B}^t + \mathbf{\Psi}^2)^{-1}$, $\boldsymbol{\delta}_f := \boldsymbol{\theta}^t\mathbf{A}$, and $\boldsymbol{\delta}_{\bar{f}} := \boldsymbol{\eta}(\mathbf{I} - \mathbf{A}\mathbf{A}^\dagger)$. Draw $\boldsymbol{\theta}$ (k-by-1) and $\boldsymbol{\eta}$ (p-by-1) independently from unit hyperspheres. Then

$$\boldsymbol{\delta} = \sqrt{a}\boldsymbol{\delta}_f + \sqrt{1-a}\boldsymbol{\delta}_{\bar{f}}.$$

This construction decomposes $\boldsymbol{\delta}$ into a part in the factor space and a part in the orthogonal complement, with a defining the proportion of the total variance attributable to the factor component.

- Finally, to determine sensible noise levels, define $d := \boldsymbol{\delta}^t(\mathbf{B}\mathbf{B}^t + \boldsymbol{\Psi}^2)\boldsymbol{\delta}$ and $q = (\beta^2(1+s_x)^2 + sx^2\alpha^2)d^2$. These terms represent the variation in x attributable to z and the variation in y attributable to x and z, respectively. Then, define $\sigma_x = s_x\sqrt{d}$ and $\xi = s_y\sqrt{q}$, so that s_x and s_y control the signal-to-noise ration in the two regression equations. In our simulations we use $s_x = s_y = 2$.

Table 1. Simulation results based on 500 simulated data sets. The parameter a denotes what proportion of $||\boldsymbol{\delta}||_2^2$ is due to the part of $\boldsymbol{\delta}$ that lies in the factor regression subspace of z. A factor shrinkage prior (FSP) is compared to a straight shrinkage ("horseshoe", HS) prior in terms of the average ratio of interval lengths (ARIL) for a nominal 90% credible interval, the corresponding coverage, and the root mean squared estimation error (RMSE).

		HS	FSP
a	ARIL	(RMSE, coverage)	(RMSE, coverage)
100%	87.0%	0.37, 91.6%	0.32, 91.0%
90%	89.6%	0.38, 91.4%	0.36, 91.0%
80%	91.5%	0.38, 90.8%	0.38, 88.2%
70%	94.3%	0.34, 93.4%	0.32, 91.4%
60%	95.8%	0.40, 88.6%	0.39, 88.4%
50%	97.3%	0.37, 87.4%	0.35, 88.6%

We compare our factor shrinkage prior to a straight horseshoe prior. Priors over (β, α) are the same for both models and are set to the true normal-inverse-gamma data generating prior mentioned above. Our simulation results, based on 500 simulated data sets at each value of a, are reported in Table 1. As expected, the model performs best (in the sense of having the smallest interval length ratio) when the data generating process is a pure factor regression model ($a = 100\%$). However, the factor shrinkage prior performs well even when $a = 50\%$, indicating that a full half of the treatment regression "signal" comes from outside the factor regression model.

5. Empirical applications

This section demonstrates the factor shrinkage prior (FSP) method on three empirical applications, which have all been studied in the previous literature. Our goal in these demonstrations is not a thorough stand-alone economic analysis, but rather a direct comparison with previously reported results on a variety of data sets. We also discuss computational hurdles encountered with each data set.

Each of the three data sets has features which illustrate different aspects of the many instruments problem as it manifests in applied work. Our first example is a relatively tractable application, with 23 exogenous controls, 48 instruments and 2,217 observations. The elasticity of inter-temporal substitution example, from Yogo [2004], is rather more difficult, with 59 instruments for a mere 114 observations (and no additional controls). This example uses instruments that are numerical (as opposed to categorical or dummy instruments). Finally, the returns to schooling example has 180 instruments as well as many controls (509), and over 300 thousand observations. This data set has many dummy instruments, and we see that our method works as advertised despite the fact that fitting a factor model to the instruments themselves is infeasible with such a large sample size (using currently available factor model sampling algorithms).

To aid in direct comparison with earlier analysis, we incorporate exogenous covariates by "partialing them out" via an initial application of ordinary least squares to both the treatment and response equations, and then apply our Bayesian model to the resulting residuals. We describe in an appendix how we could instead include the variables directly into our model. In general, we feel that the latter approach is sounder, because regularization on these parameters is presumably beneficial to estimation for the usual reasons.

In all cases, priors are specified in terms of standardized response and treatment variables and then transformed to the original scale by post-processing posterior samples.

5.1. **Automobile data.** Berry et al. [1995] perform a detailed regression analysis to infer the demand for automobiles. Here we use the same data, although we follow the re-analysis of Chernozhukov et al. [2015] for comparison.

For reference, the model being fit is as in (18): $f(x, y \mid \mathrm{z}) = \mathrm{N}_{y|x}(x\beta + \alpha(x - \mathrm{z}^t\boldsymbol{\delta}_\mathrm{z} - \mathrm{w}^t\boldsymbol{\delta}_\mathrm{w}), \xi^2)\mathrm{N}_x(\mathrm{z}^t\boldsymbol{\delta} + \mathrm{w}^t\boldsymbol{\gamma}, \sigma_x^2)$, for $y_{ht} = \log(s_{ht}) - \log(s_{0t})$, with s_{it} denoting the market share of product h in market t and product 0 denoting the "outside option". (Note that the double subscripts, product h and market t, are immaterial once the response variable and instruments have been formed; we end up with the usual panel regression model by reindexing with single subscript i.)

The treatment variable x_{ht} is the corresponding product price and w_{ht} are additional product characteristics: air conditioning (AC), horsepower to weight ratio (HWR), miles per dollar (MPD), and vehicle size (VS). Following Berry et al. [1995], the instrument vector, z_{ht} is constructed by calculating the following:

- the sum of each characteristic in w, taken across models made by product h's firm,
- the sum of each characteristic in w, taken across competitor firms' products,
- the total number of models produced by product h's firm, and
- the total number of models produced by the firm's competitors.

Thus, with four attributes included in w, there are a total of 10 basic instruments. Following, Chernozhukov et al. [2015], we augment the covariate vector w to include a time trend and quadratic and cubic terms in all continuous characteristics (everything except AC, including the time trend), as well as all first-order interaction terms. This yields a total of 23 additional covariates. Using this expanded covariate vector, the instrument vector z is constructed according to the procedure described above, yielding 48 total instruments.

The most notable challenge of the automobile data is that the covariance matrix is nearly singular, so the inversion needed by our factor shrinkage prior is ill-defined. However, there is nothing illicit in our approach from pre-regularizing this covariance estimate, which we do using the CondReg package in R [Oh et al., 2015] and deciding the amount of regularization by cross-validation.

Using the largest ratio between consecutive eigenvalues less than $k_{max} = 10$ choses 2 factors. Our preferred analysis ($\kappa = 8$, $s = 2$, $c_\beta = 4$, $c_\alpha = 1$) gives a point estimate of $\bar{\beta} = -0.275$ (with standard deviation of 0.018), which is somewhat larger in magnitude than the estimate of -0.22 reported by Chernozhukov et al. [2015]. Results are not sensitive to moderate variation in prior parameters.

5.2. **The elasticity of inter-temporal substitution.** Yogo [2004] considers estimating the elasticity of inter-temporal substitution (EIS) via a linearization of the Euler equation, using macroeconomic data and an instrumental variable analysis. Ng and Bai [2009] extend this analysis by incorporating many additional macro variables (detailed in Ludvigson and Ng [2007]) as instruments and consolidating them into factors using a boosting approach. This section mimics that analysis for comparative purposes, focusing on the 1970:3 to 1998:4 quarterly data for the United States ($n = 114$). Of the 209 macro-variables used as instruments in Ng and Bai [2009], a subset of 78 are used here[1]. This number of instruments is then further reduced to 55 by dropping one of each pair of variables with correlation greater than 0.9, to lessen numerical instability due to extreme multicolinearity (many of these series are essentially identical). A representative subset of these macro-variables includes, for example, gross domestic purchases, fixed investment in durable equipment, assets abroad, and net exports[2]. Including Yogo's original four instruments (twice lagged nominal interest

[1] A complete data set was not readily available. For the purposes of our illustration, the distinction between 78 instruments and 209 instruments is minor; having strictly fewer instruments than earlier analyses is not essential to the model interpretation and comparison in the way that having the same control variables is.

[2] It is a practically relevant question as to whether or not (lagged) macroeconomic indicators serve as valid instruments in the sense of satisfying the exclusion restriction. On the one hand, under a causal interpretation it seems reasonable to assert that past indicators should only relate to the present economy via the more recent indicators—a sort of Markov property. On the other hand, this narrative falls apart when one considers latent common causes that serve to induce dependence between today's indicators, yesterdays indicators, and today's response variable. Such shared common causes clearly violate the desired exclusion restriction. That said, this possibility will not be discussed further here; rather, a narrow comparison is drawn with the results of Ng and Bai [2009], who assume the validity of the macro indicators as instruments. As such, a detailed justification of these instruments is beyond the scope of our illustration.

rate, inflation, consumption growth, and log dividend-price ratio), our model uses 59 instruments. Unlike the automobile data or the returns to schooling data in the next example, the analyses of Yogo [2004] and Ng and Bai [2009] do not include exogenous controls.

For reference, the model being fit is as in (6): $f(x, y \mid \mathrm{z}) = \mathrm{N}_{y|x}(x\beta + \alpha(x - \mathrm{z}^t\boldsymbol{\delta}), \xi^2)\mathrm{N}_x(\mathrm{z}^t\boldsymbol{\delta}, \sigma_x^2)$, where y_i is the quarterly consumption growth (i.e., the change in consumption) in the United States, x_i is the real interest rate and β denotes the elasticity of inter-temporal substitution (EIS). The instrument vector z_i consists of aforementioned macroeconomic indicators (twice lagged), in addition to the original instruments used in Yogo [2004]: twice lagged nominal interest rate, inflation, consumption growth, and log dividend-price ratio. See Yogo [2004] section II for a theoretical justification of this model.

Table 2 compares the estimates and standard errors/posterior uncertainty for various estimation methods. Straight shrinkage regularization with the larger set of instruments gives an answer similar to the corresponding OLS estimate with comparable standard deviation as well. The factor shrinkage prior with prior $\kappa = 8$, $s = 2$, $c_\beta = 4$ and $c_\alpha = 1$ gives a posterior mean of 0.12 with posterior standard deviation of 0.08. It is not surprising in this case that the result is similar to the unmodified horseshoe regression, as the Frisch decomposition returns only very small values for the elements of $\mathbf{D}$ (not strong evidence of additive measurement error in the factors). Our same heuristic choses $k = 2$ factors.

The similarity of both horseshoe and factor shrinkage to the 2SLS estimate is somewhat surprising, but turns out to hinge on the choice of prior. The signal to noise ratio in the Yogo data is very poor — the least squares regression has an R^2 of only about 0.05 (with $p = 59$ and $n = 114$) — and we observe a corresponding sensitivity to choice of prior. Through extensive investigation we were able to ascertain that the dominant prior impact is determined by the relative size of c_β and c_α; these parameters govern the *a priori* decomposition of the reduced form coefficient $(\alpha + \beta)$ into it's constituent (economically interpretable) pieces. Table 3 records how the posterior mean estimate changes as a function of c_β and c_β (the prior precision parameters). Only for implausibly high levels of prior precision are we able to match the estimates of Ng and Bai [2009].

Table 2. Estimates of the elasticity of inter-temporal substitution using the direct regression ($\psi := \beta$), by various methods: ordinary least squares (OLS), two-stage least squares (2SLS), Bayesian IV with factor shrinkage prior (FSP) or unmodified horseshoe prior (HS IV), and the boosted factor IV of Ng and Bai [2009], Table 7b (FIV$_b$). Yogo's original four instruments appear in all models. For all FSP and HS models, the hyper-parameters for (β, α) are $\kappa = 8$, $s = 2$, $c_\beta = 4$ and $c_\alpha = 1$. All figures have been rounded to two decimal places for comparison.

Method	$\hat{\beta}$	s.d.
OLS	0.12	0.05
2SLS (all)	0.12	0.07
HS IV (all)	0.12	0.07
2SLS (Yogo)	0.06	0.09
HS IV (Yogo)	0.05	0.09
FSP (2 factors)	0.12	0.08
FIV$_b$ (Ng and Bai)	0.09	0.06

Table 3. Posterior mean estimates of β across a range of hyper-parameter values.

	c_α				
c_β	0.1	1	10	100	200
0.1	0.126	0.125	0.122	0.117	0.116
1	0.125	0.124	0.121	0.116	0.116
10	0.117	0.116	0.114	0.112	0.112
100	0.072	0.072	0.075	0.085	0.087
200	0.051	0.051	0.054	0.066	0.070

Finally, we remark that the Yogo data proved to be extremely computationally demanding, owing to (we believe) multimodality in the posterior for $\boldsymbol{\delta}$, which is a well-known result of fat-tailed priors in conjunction with information-poor data. It took a staggering 50 million iterations before posterior mean estimates stabilized. The other two data sets did not require such heroic efforts, due (we believe) to larger sample sizes. So, while our sampler was fast enough to permit this amount of computation (about three hours), additional innovations in sampling such models are an open area of research.

5.3. **Returns to schooling data.**

In this section we revisit the well-known analysis of Angrist and Krueger [1991], where the causal impact of schooling on wages was studied using data from the 1980 U.S. Census on 329,509 men born between 1930 and 1939. We closely follow the analysis of Hansen and Kozbur [2014], who control for 509 variables, consisting of 9 year-of-birth indicators, 50 state-of-birth indicators, as well as the 450 interactions between them. For instruments, three quarter-of-birth indicators are used, as well as interactions with the 9 main effects for year-of-birth and 50 main effects for state-of-birth, for a total of 180 instruments. Further, one can use three quarter-of-birth dummies and their interactions with the full set of state-of-birth and year-of-birth controls to obtain a total of 1527 candidate instruments. For a detailed argument (based on compulsory schooling laws) regarding why quarter of birth serves as a valid instrument in this setting, please see the original thorough analysis of Angrist and Krueger [1991]. Our response variable $y_i = \log(\text{wage}_i)$ is the reported log wage of individual i and our treatment variable x_i is reported years of completed schooling for individual i.

In the 1527 instrument case, regularization was required to obtain a stable inverse for our prior. Due to memory considerations we used the `glasso` package in `R` [Friedman et al., 2008]. We used the same max ratio of eigenvalue strategy to select the number of factors, which choses $k = 2$ (with $k_{max} = 100$).

Table 4 compares estimates for a range of different methods, reproducing a table from Hansen and Kozbur [2014] with a column added for the results of our factor shrinkage approach. We observe that the standard error reduction, relative to the basic three-instrument case is modest to the point of being immaterial for every method except the factor shrinkage prior in the 1527 instrument analysis. However, this distinction is difficult to interpret given the conceptual differences between frequentist standard errors and Bayesian posterior credible intervals. Nonetheless, it means that a Bayesian would report with higher subjective precision by including the extra instruments, with a posterior standard deviation on the same order of magnitude as 2SLS (but a substantially higher posterior mean).

An interesting methodological question is how estimates from regularized methods behave as the number of instruments is increased. Specifically, it is observed that as more instruments are added, estimates from many estimators (including 2SLS) tend towards the OLS estimate. Therefore it is interesting to see how regularization methods mitigate this recognized bias. Again, the results are equivocal in the sense that estimates lying between the three-instrument 2SLS estimate (0.1079) and the OLS estimate (0.0673) need not be incorrect. Hansen and Kozbur [2014] suggest informally that the stability of their RJIVE estimate as the number of instruments increases (at the three-instrument 2SLS estimate) is reassuring behavior, though of course this does not follow logically [3]. In particular, we see no reason a priori to trust the RJIVE estimate (0.1067) more than the Post-LASSO estimate (0.0862) for example. We make this comparison because our Bayesian estimate happens to line up with the Post-LASSO estimate in this instance and, from a Bayesian perspective, we find no special reason to find it suspect. Furthermore, if the three-instrument 2SLS is going to serve as a gold-standard, we may as well simply use that method exclusively, especially when more elaborate methods do not appear to yield much increase in precision.

Table 4. A factor shrinkage analysis ($k = 2$ factors) of the returns to schooling data produce estimates similar to those from other regularization methods for treatment effects. Table reproduced from Hansen and Kozbur [2014].

	2SLS	Post-LASSO	JIVE	RJIVE	FSP
	A. 3 instruments				
Schooling coefficient	0.1079	0.115	0.1091	0.1091	0.1098
Estimated standard error	0.0196	0.0205	0.0202	0.0202	0.0207
	B. 180 instruments				
Schooling coefficient	0.0928	0.1125	0.1096	0.1062	0.1107
Estimated standard error	0.0097	0.0173	0.0161	0.0157	0.0183
	C. 1527 instruments				
Schooling coefficient	0.0712	0.0862	0.0816	0.1067	0.0862
Estimated standard error	0.0049	0.0254	0.5168	0.0171	0.0066

6. Conclusion

When many candidate instruments are available — including polynomial expansions and interactions of existing instruments — judicious regularization of the first-stage regression is

[3] To quote Hansen and Kozbur [2014]: "The results reported in Panel C of Table 3 are based on using the full set of 1527 instruments and are the most interesting from the standpoint of the present paper. In this case, we see that both the Post-LASSO and JIVE point estimates have shifted substantively toward the OLS estimate. In contrast, the RJIVE is very stable, remaining around the value estimated by all of the procedures using only three instruments. More interesting is the fact that standard errors from both JIVE and Post-LASSO are now pronouncedly larger than the standard error from the RJIVE".

a crucial component of a well-behaved IV estimator. This paper proposes a Bayesian model built on the assumption that the treatment variable is more likely to depend on the communalities of the instrument matrix than on the idiosyncrasies. Analysis on synthetic data reveals that the new prior performs according to intuition: when factor structure predictive of the treatment is apparent in the matrix of instruments, this concordance with the prior yields tighter inference concerning the treatment effect of interest. The new prior can be used even when the instruments are not jointly Gaussian, such as many binary instruments. Finally, the slice sampling approached described here allows us to try various shrinkage priors easily and allows fitting large data sets, both in terms of the sample size n and the number of instruments p.

References

J. D. Angrist and A. B. Krueger. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014, 1991.

A. Aswani, P. Bickel, and C. Tomlin. Regression on manifolds: Estimation of the exterior derivative. *The Annals of Statistics*, pages 48–81, 2011.

P. A. Bekker. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society*, pages 657–681, 1994.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research*, 7:2399–2434, 2006.

S. Berry, J. Levinsohn, and A. Pakes. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995.

M. Carrasco. A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–398, 2012.

C. Carvalho, N. Polson, and J. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97:465–480, 2010.

G. Chamberlain and G. Imbens. Random effects estimators with many instrumental variables. *Econometrica*, 72(1):295–306, 2004.

J. Chan and J. Tobias. Priors and posterior computation in linear endogenous variable models with imperfect instruments. *Journal of Applied Econometrics*, 2014.

J. C. Chao and P. C. Phillips. Posterior distributions in limited information analysis of the simultaneous equations model using the Jeffreys prior. *Journal of Econometrics*, 87(1): 49–86, 1998.

V. Chernozhukov, C. Hansen, and M. Spindler. Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, 105(5):486–90, May 2015.

T. G. Conley, C. B. Hansen, R. E. McCulloch, and P. E. Rossi. A semi-parametric Bayesian approach to the instrumental variable problem. *Journal of Econometrics*, 144(1):276–305, 2008.

T. G. Conley, C. B. Hansen, and P. E. Rossi. Plausibly exogenous. *Review of Economics and Statistics*, 94(1):260–272, 2012.

J. H. Dreze. Bayesian limited information analysis of the simultaneous equations model. *Econometrica: Journal of the Econometric Society*, pages 1045–1075, 1976.

M. Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

R. Frisch. Statistical confluence analysis by means of complete regression systems. Technical Report 5, University of Oslo, Economic Institute, 1934.

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

M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008.

M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.0 beta. `http://cvxr.com/cvx`, Sept. 2013.

J. J. Groen and G. Kapetanios. Parsimonious estimation with many instruments. *Federal Reserve Bank of New York, Staff Report*, (386), 2009.

J. Hahn and K. Hansen. Parameter orthogonalization and Bayesian inference with many instruments. *Economics Letters*, 112(2):207–209, 2011.

P. Hahn, C. M. Carvalho, and S. Mukherjee. Partial factor modeling: predictor-dependent shrinkage for linear regression. *Journal of the American Statistical Association*, 108(503): 999–1008, 2013.

C. Hansen and D. Kozbur. Instrumental variables estimation with many weak instruments using regularized jive. *Journal of Econometrics*, 182(2):290–308, 2014.

G. Kapetanios and M. Marcellino. Factor-GMM estimation with large sets of possibly weak instruments. *Computational Statistics & Data Analysis*, 54(11):2655–2675, 2010.

G. Koop, R. Leon-Gonzalez, and R. Strachan. Bayesian model averaging in the instrumental variable regression model. *Journal of Econometrics*, 171(2):237–250, 2012.

D. Lindley and G. El-Sayed. The Bayesian estimation of a linear functional relationship. *Journal of the Royal Statistical Society. Series B*, 30:190–202, 1968.

H. F. Lopes and N. G. Polson. Bayesian instrumental variables: priors and likelihoods. *Econometric Reviews*, 33(1-4):100–121, 2014.

S. C. Ludvigson and S. Ng. The empirical risk–return relation: A factor analysis approach. *Journal of Financial Economics*, 83(1):171–222, 2007.

I. Murray, R. P. Adams, and D. J. MacKay. Elliptical slice sampling. *arXiv preprint arXiv:1001.0175*, 2009.

J. S. Murray, D. B. Dunson, L. Carin, and J. E. Lucas. Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108(502):656–665, 2013.

W. K. Newey and R. J. Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

S. Ng and J. Bai. Selecting instrumental variables in a data rich environment. *Journal of Time Series Econometrics*, 1(1), 2009.

L. Ning, T. T. Georgiou, A. Tannenbaum, and S. P. Boyd. Linear models based on noisy data and the frisch scheme. *SIAM Review*, 57(2):167–197, 2015.

S.-Y. Oh, B. Rajaratnam, and J.-H. Won. On the solution path of regularized covariance estimators. *arXiv preprint arXiv:1502.00471*, 2015.

R. Okui. Instrumental variable estimation in the presence of many moment conditions. *Journal of Econometrics*, 165(1):70–86, 2011.

P. E. Rossi, G. M. Allenby, and R. McCulloch. *Bayesian statistics and marketing*. Series in Probability and Statistics. Wiley, 2006.

A. Shapiro. Weighted minimum trace factor analysis. *Psychometrika*, 47(3):243–264, 1982.

M. Yogo. Estimating the elasticity of intertemporal substitution when instruments are weak. *Review of Economics and Statistics*, 86(3):797–810, 2004.

Appendix A. Incorporating exogenous covariates

In many applied problems, in addition to the instrumental variables, one has available exogenous control covariates. In fact, the validity of the instruments often depends on incorporating a sufficiently rich set of control variables. Fortunately, our importance sampling approach is easily modified to accommodate this possibility. Letting w_i denote the vector of control variables we have

$$\begin{aligned} f(x,y \mid \mathrm{z}) &= f(y \mid x,\mathrm{z})f(x \mid \mathrm{z}) \\ &= \mathrm{N}_{y|x}(x\beta + \alpha(x - \mathrm{z}^t\boldsymbol{\delta}_{\mathrm{z}} - \mathrm{w}^t\boldsymbol{\delta}_{\mathrm{w}}) + \boldsymbol{\gamma}\mathrm{w}_i, \xi^2)\times \\ &\quad\ \mathrm{N}_x(\mathrm{z}^t\boldsymbol{\delta}_{\mathrm{z}} + \mathrm{w}^t\boldsymbol{\delta}_{\mathrm{w}}, \sigma_x^2). \end{aligned} \tag{18}$$

Using this model, our sampler works as before, by first drawing samples of $\boldsymbol{\delta}^t = (\boldsymbol{\delta}_{\mathrm{z}}^t, \boldsymbol{\delta}_{\mathrm{w}}^t)$ from $\pi(\boldsymbol{\delta}, \sigma_x^2 \mid \mathbf{x}, \mathbf{Z}, \mathbf{W})$ and redefining $\tilde{\mathbf{x}}_i := (x_i, x_i - \mathrm{z}_i^t\boldsymbol{\delta}_{\mathrm{z}} - \mathrm{w}_i^t\boldsymbol{\delta}_{\mathrm{w}}, \mathrm{w}_i)$.

Note that by incorporating w_i, we enlarge $\mathbf{M}$ to have size $p_{\mathrm{w}} + 2$ square, where p_{w} is the number of exogenous covariates. While this requires additional computation, evaluating $\mathbf{M}$, $\mathbf{M}^{-1}$, and $\det(\mathbf{M})$ can be done relatively efficiently using block matrix algebra, taking advantage of the fact that for each first-stage sample $\boldsymbol{\delta}$, $\tilde{\mathbf{x}}_i$ only differs in a single entry. Also, in step 3, note that it is unnecessary to draw $\boldsymbol{\gamma}$; one need only draw from the upper 2-by-2 block of the multivariate normal distribution corresponding to (α, β).