Bayesian Analysis of Structural VAR Models

The Identification Problem

- Classic questions in empirical macroeconomics:
  - ➢ What is the effect of a policy intervention (interest rate increase, fiscal stimulus) on macroeconomic aggregates (output, inflation, employment,…)?
  - ➢ Which structural shocks drive aggregate fluctuations?
- What we are interested in is the dynamic causal effect of structural shock u_t on a vector of macro time series y_t:

$$\frac{\partial y_{t+s}}{\partial u_t}, s = 0, 1, 2, 3, \ldots$$

Dynamic structural model:

$$\underset{(n\times n)}{\mathbf{A}}\underset{(n\times 1)}{\mathbf{y}_t} = \underset{(n\times 1)}{\boldsymbol{\lambda}} + \underset{(n\times n)}{\mathbf{B}_1}\underset{(n\times 1)}{\mathbf{y}_{t-1}} + \cdots + \underset{(n\times n)}{\mathbf{B}_m}\underset{(n\times 1)}{\mathbf{y}_{t-m}} + \underset{(n\times n)}{\mathbf{D}^{1/2}}\underset{(n\times 1)}{\mathbf{v}_t}$$

$\mathbf{v}_t \sim$ i.i.d. $N(\mathbf{0}, \mathbf{I}_n)$

$$\mathbf{x}'_{t-1} = (1, \mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \ldots, \mathbf{y}'_{t-m})'$$

$$\mathbf{D}^{1/2} = \begin{bmatrix} \sqrt{d_{11}} & 0 & \ldots & 0 \\ 0 & \sqrt{d_{22}} & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \sqrt{d_{nn}} \end{bmatrix}$$

Example: supply and demand

$$q_t = \lambda^s + \alpha^s p_t + b_{11}^s p_{t-1} + b_{12}^s q_{t-1} + b_{21}^s p_{t-2} + b_{22}^s q_{t-2} + \cdots + b_{m1}^s p_{t-m} + b_{m2}^s q_{t-m} + \sqrt{d_s}\, v_t^s$$

$$q_t = \lambda^d + \beta^d p_t + b_{11}^d p_{t-1} + b_{12}^d q_{t-1} + b_{21}^d p_{t-2} + b_{22}^d q_{t-2} + \cdots + b_{m1}^d p_{t-m} + b_{m2}^d q_{t-m} + \sqrt{d_d}\, v_t^d$$

$$\mathbf{y}_t = \begin{bmatrix} q_t \\ p_t \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} 1 & -\alpha^s \\ 1 & -\beta^d \end{bmatrix}$$

Reduced-form (forecasting equations):

$$\mathbf{y}_t = \mathbf{\Phi}\, \mathbf{x}_{t-1} + \boldsymbol{\varepsilon}_t$$

$$\boldsymbol{\varepsilon}_t \sim \text{i.i.d. } N(\mathbf{0}, \mathbf{\Omega})$$

$$\hat{\mathbf{\Phi}}_T = \left(\sum_{t=1}^{T} \mathbf{y}_t \mathbf{x}'_{t-1}\right)\left(\sum_{t=1}^{T} \mathbf{x}_{t-1} \mathbf{x}'_{t-1}\right)^{-1}$$

$$\hat{\boldsymbol{\varepsilon}}_t = \mathbf{y}_t - \hat{\mathbf{\Phi}}_T \mathbf{x}_{t-1}$$

$$\hat{\mathbf{\Omega}}_T = T^{-1} \sum_{t=1}^{T} \hat{\boldsymbol{\varepsilon}}_t \hat{\boldsymbol{\varepsilon}}'_t$$

Nonorthogonalized impulse responses:

$$\frac{\partial \mathbf{y}_{t+s}}{\partial \boldsymbol{\varepsilon}'_t} = \mathbf{\Psi}_s$$

Reminder of Bayesian Principles

- Bayesian idea: before observing a data sample $\boldsymbol{y}_T$, the researcher has some beliefs about how likely different parameter values $\mathbf{\Phi}$ are which can be expressed in the form of a distribution

  $\Rightarrow$ prior density $p(\mathbf{\Phi})$

- Combine prior information with information in the data through the likelihood function to obtain the posterior distribution:

$$p(\mathbf{\Phi} \mid \mathbf{y}_T) \propto p(\mathbf{y}_T \mid \mathbf{\Phi})\, p(\mathbf{\Phi})$$

$\Rightarrow$ "The posterior is proportional to the likelihood times the prior."

Why Bayesian Estimation of VARs?

- Given the large number of parameters in VARs, estimates of objects of interest (e.g. impulse responses, forecasts) can become *imprecise* in large models.
  - ➢ need restrictions (e.g., lag length, which variables to include)
  - ➢ Bayesian philosophy: use "soft" restrictions
    - ⇒ guide estimates toward prior restrictions, but don't insist
    - ⇒ incorporating prior information generally yields more precise estimates
- Bayesian simulation methods provide an easy way to characterize estimation uncertainty.

Prior for VAR Coefficients

Stacking the T observations, the system can be written as:
$$\underset{(T\times n)}{\mathbf{Y}} = \underset{(T\times k)}{\mathbf{X}}\underset{(k\times n)}{\mathbf{\Phi}} + \underset{(T\times n)}{\boldsymbol{\varepsilon}}$$

Define $\boldsymbol{\phi} = vec(\mathbf{\Phi})$ and assume that the prior for the VAR coefficients is normal:

$$p(\boldsymbol{\phi}) \sim N(\boldsymbol{\phi}_0, \mathbf{V}_0)$$

Conditional on $\mathbf{\Omega}$, the posterior for the VAR coefficients is normal:

$$p(\boldsymbol{\phi} \mid \mathbf{\Omega}, \mathbf{Y}) \sim N(\boldsymbol{\phi}^*, \mathbf{V}^*)$$

where $\mathbf{V}^* = (\mathbf{V}_0^{-1} + \mathbf{\Omega}^{-1} \otimes \mathbf{X}'\mathbf{X})^{-1}$

$$\boldsymbol{\psi}^* = (\mathbf{V}^*)^{-1}(\mathbf{V}_0^{-1}\boldsymbol{\phi}_0 + \mathbf{\Omega}^{-1} \otimes \mathbf{X}'\mathbf{X}\,\widehat{\boldsymbol{\phi}}_{OLS})$$

Prior for Variance Matrix

- Univariate regression:

$$\sigma^2 = E(\varepsilon_t^2)$$

Let $z_i \sim N(0, \lambda^{-1})$ for $i = 1, 2, \ldots \upsilon$

then, $W = (z_1^2 + z_2^2 + \ldots + z_\upsilon^2)$

$$\sim \Gamma(\upsilon, \lambda)$$

Prior for Variance Matrix

- Univariate regression:

$$\sigma^2 = E(\varepsilon_t^2)$$

Let $z_i \sim N(0, \lambda^{-1})$ for $i = 1, 2, \ldots \upsilon$

then, $W = (z_1^2 + z_2^2 + \ldots + z_\upsilon^2)$

$$\sim \Gamma(\upsilon, \lambda)$$

- Multivariate regression:

$$\mathbf{\Omega} = E(\boldsymbol{\varepsilon}' \boldsymbol{\varepsilon})$$

Now $\mathbf{z}_i \sim N(\mathbf{0}, \mathbf{\Lambda}^{-1})$ where $\mathbf{z}_i$ is $1 \times n$

then, $\mathbf{W} = (\mathbf{z}_1^2 + \mathbf{z}_2^2 + \ldots + \mathbf{z}_\upsilon^2)$

$$\sim W(\upsilon, \mathbf{\Lambda})$$

Prior for Variance Matrix

- The conjugate prior for the VAR covariance matrix $\mathbf{\Omega}$ is an inverse Wishart distribution:

$$p(\mathbf{\Omega}) \sim IW(\upsilon_0, \mathbf{S}_0)$$

where υ_0 are the prior degrees of freedom

$\mathbf{S}_0$ is the prior scale matrix

- Conditional on $\boldsymbol{\phi}$, the posterior for $\mathbf{\Omega}$ is also inverse Wishart:

$$p(\mathbf{\Omega} \mid \boldsymbol{\phi}, \mathbf{Y}) \sim IW(\upsilon_0 + T, \mathbf{S}^*)$$

where $\mathbf{S}^* = \mathbf{S}_0 + (\mathbf{Y} - \mathbf{X}\,\mathbf{\Phi})'(\mathbf{Y} - \mathbf{X}\,\mathbf{\Phi})$

Key Prior Distributions for VARs

1. Minnesota prior (Litterman, 1980; Doan, Litterman, and Sims, 1984)
2. Normal-inverse Wishart prior (see Uhlig, *JME* 2005)
3. Independent Normal-inverse Wishart prior
4. Dummy observations (see Banbura, Giannone and Reichlin, *JAE* 2010)
5. Steady state priors (see Villani, *JAE* 2009; Giannone, Lenza, and Primiceri, *RESTAT* 2015)

Minnesota Prior

Structured prior beliefs:

1) The endogenous variables in the VAR follow a random walk or an AR(1) process.

<u>Example</u>: bivariate VAR(2) model

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix} \begin{pmatrix} y_{t-2} \\ x_{t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

$$\Longrightarrow \begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} b_{11} & 0 \\ 0 & b_{22} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_{t-2} \\ x_{t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

Prior mean: $\widetilde{\boldsymbol{b}}_{\mathbf{0}} = (0 \;\; b_{11}^0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; b_{22}^0 \;\; 0 \;\; 0)'$

Under RW assumption: $b_{11}^0 = b_{22}^0 = 1$

Minnesota Prior

2) The variance $\tilde{\mathbf{V}}_0$ of the prior for the VAR coefficients b_{ij} is set based on the following observations:

a. Greater confidence that coefficients on *higher* lags are zero.

b. Greater confidence that coefficients *other than own* lags are zero.

To make this operational define a set of *hyperparameters* that control the **tightness** of the prior:

- $\boldsymbol{\lambda_1}$: controls the standard deviation of the prior on **own** lags (overall confidence in the prior)

  ⇨ as $\lambda_1 \to 0$, greater weight is given to RW or AR(1)

- $\boldsymbol{\lambda_2}$: controls the standard deviation of the prior on **lags** of variables **other than** the dependent variable

  ⇨ with $\lambda_2 = 1$, no distinction between own and other lags

- $\boldsymbol{\lambda_3}$: controls the degree to which coefficients on **lags higher than 1** are likely to be zero

  ⇨ for $\lambda_3 = 0$, all lags are given equal weight

- $\boldsymbol{\lambda_4}$: controls the prior variance of the constant

  ⇨ as $\lambda_4 \to 0$, constant terms are shrunk to zero

Formulas for Minnesota Prior

- $\left(\frac{\lambda_1}{m^{\lambda_3}}\right)^2 \quad if \quad i = j$
- $\left(\frac{\sigma_i \lambda_1 \lambda_2}{\sigma_j m^{\lambda_3}}\right)^2 \quad if \quad i \neq j$
- $(\sigma_i \lambda_4)^2$ for the constant

$\Rightarrow$ σ_i and σ_j are the standard deviations of error terms from AR regressions estimated via OLS

$\Rightarrow$ the ratio of σ_i and σ_j accounts for the possibility that variables i and j may have different *scales*

What does $\widetilde{\mathbf{V}}_0$ look like?

$$\widetilde{\mathbf{V}}_0 = \begin{pmatrix} (\sigma_1\lambda_4)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & (\lambda_1)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \left(\frac{\sigma_1\lambda_1\lambda_2}{\sigma_2}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \left(\frac{\sigma_1\lambda_1\lambda_2}{\sigma_2 2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & (\sigma_2\lambda_4)^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_2\lambda_1\lambda_2}{\sigma_1}\right)^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & (\lambda_1)^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_2\lambda_1\lambda_2}{\sigma_1 2^{\lambda_3}}\right)^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 \end{pmatrix}$$

Typical values for hyperparameters used in the literature (see Doan 2013):

$$\lambda_1 = 0.2 \qquad \lambda_2 = 0.7 \qquad \lambda_3 = 1 \; or \; 2 \qquad \lambda_4 = 100$$

Prior Using Dummy Observations

- The computation of the moments of the conditional posterior distribution for the VAR coefficients requires the computation of the inverse of an $(nk \times nk)$ matrix:

$$(\mathbf{V}_0^{-1} + \mathbf{\Omega}^{-1} \otimes \mathbf{X}'\mathbf{X})^{-1}$$

- Alternative approach: use dummy observations to represent prior densities

- Dummy observations can be obtained as follows:
  1. Use actual observations from other countries or pre-sample
  2. Use observations generated by simulating macro model
  3. Use observations generated from "introspection"

  $\Rightarrow$ represent by hyperparameters

Prior Using Dummy Observations

- Augment actual data $\mathbf{Y}$ and $\mathbf{X}$ with "artificial" data $\mathbf{Y_D}$ and $\mathbf{X_D}$: $\mathbf{Y}^* = [\mathbf{Y}; \mathbf{Y_D}]$ and $\mathbf{X}^* = [\mathbf{X}; \mathbf{X_D}]$

- For example:

  - Prior mean: $\psi_0 = (\mathbf{X_D'}\mathbf{X_D})^{-1}\mathbf{X_D'}\mathbf{Y_D}$
  - Posterior mean: $\psi^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^*$

$\Rightarrow$ the weight placed on the artificial data determines how much confidence researcher has in the prior

System Dynamics

Vector autoregression:

$$\underset{(n\times 1)}{\mathbf{y}_t} = \underset{(n\times 1)}{\mathbf{c}} + \underset{(n\times n)}{\mathbf{\Phi}_1}\,\mathbf{y}_{t-1} + \mathbf{\Phi}_2\mathbf{y}_{t-2} + \cdots + \mathbf{\Phi}_m\mathbf{y}_{t-m} + \boldsymbol{\varepsilon}_t$$

$\boldsymbol{\varepsilon}_t$ ~ white noise

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \vdots \\ \mathbf{y}_{t-m+1} \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi}_1 & \mathbf{\Phi}_2 & \mathbf{\Phi}_3 & \cdots & \mathbf{\Phi}_{m-1} & \mathbf{\Phi}_m \\ \mathbf{I}_n & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_n & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \mathbf{y}_{t-3} \\ \vdots \\ \mathbf{y}_{t-m} \end{bmatrix} + \begin{bmatrix} \mathbf{c} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_t \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}$$

$$\underset{(nm\times nm)}{\mathbf{F}} = \begin{bmatrix} \mathbf{\Phi}_1 & \mathbf{\Phi}_2 & \mathbf{\Phi}_3 & \cdots & \mathbf{\Phi}_{m-1} & \mathbf{\Phi}_m \\ \mathbf{I}_n & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_n & \mathbf{0} \end{bmatrix}$$

Rewriting a VAR(m) as a VAR(1)

$$\boldsymbol{\xi}_t = \mathbf{F}\boldsymbol{\xi}_{t-1} + \tilde{\mathbf{v}}_t$$

$$\underset{(nm\times 1)}{\boldsymbol{\xi}_t} = (\mathbf{y}_t', \mathbf{y}_{t-1}', \ldots, \mathbf{y}_{t-m+1}')'$$

$$\tilde{\mathbf{v}}_t = \mathbf{k} + \mathbf{v}_t$$

$$\underset{(nm\times 1)}{\mathbf{k}} = (\mathbf{c}', \mathbf{0}', \ldots, \mathbf{0}')'$$

$$\underset{(nm\times 1)}{\mathbf{v}_t} = (\boldsymbol{\varepsilon}_t', \mathbf{0}', \ldots, \mathbf{0}')'$$

$$\boldsymbol{\xi}_t = \mathbf{F}\boldsymbol{\xi}_{t-1} + \tilde{\mathbf{v}}_t$$

This is called the *companion form* of the VAR(*m*).

And ***F*** is called the *companion matrix*.

$$\boldsymbol{\xi}_t = \mathbf{F}\boldsymbol{\xi}_{t-1} + \tilde{\mathbf{v}}_t$$

$$\boldsymbol{\xi}_{t+1} = \mathbf{F}\boldsymbol{\xi}_t + \tilde{\mathbf{v}}_{t+1}$$

$$\boldsymbol{\xi}_t = \mathbf{F}\boldsymbol{\xi}_{t-1} + \tilde{\mathbf{v}}_t$$

$$\boldsymbol{\xi}_{t+1} = \mathbf{F}\boldsymbol{\xi}_t + \tilde{\mathbf{v}}_{t+1}$$

$$\boldsymbol{\xi}_t = \mathbf{F}\boldsymbol{\xi}_{t-1} + \tilde{\mathbf{v}}_t$$

$$\boldsymbol{\xi}_{t+1} = \mathbf{F}\boldsymbol{\xi}_t + \tilde{\mathbf{v}}_{t+1}$$

$$\boldsymbol{\xi}_{t+1} = \mathbf{F}^2\boldsymbol{\xi}_{t-1} + \mathbf{F}\tilde{\mathbf{v}}_t + \tilde{\mathbf{v}}_{t+1}$$

$$
\begin{aligned}
\boldsymbol{\xi}_t &= \mathbf{F}\boldsymbol{\xi}_{t-1} + \tilde{\mathbf{v}}_t \\
\boldsymbol{\xi}_{t+1} &= \mathbf{F}\boldsymbol{\xi}_t + \tilde{\mathbf{v}}_{t+1} \\
\boldsymbol{\xi}_{t+1} &= \mathbf{F}^2\boldsymbol{\xi}_{t-1} + \mathbf{F}\tilde{\mathbf{v}}_t + \tilde{\mathbf{v}}_{t+1} \\
&\vdots \\
\boldsymbol{\xi}_{t+h} &= \mathbf{F}^{h+1}\boldsymbol{\xi}_{t-1} + \mathbf{F}^h\tilde{\mathbf{v}}_t + \mathbf{F}^{h-1}\tilde{\mathbf{v}}_{t+1} \\
&\quad + \cdots + \tilde{\mathbf{v}}_{t+h}
\end{aligned}
$$

$$\boldsymbol{\xi}_{t+h} = \mathbf{F}^{h+1}\boldsymbol{\xi}_{t-1} + \mathbf{F}^{h}\tilde{\mathbf{v}}_t + \mathbf{F}^{h-1}\tilde{\mathbf{v}}_{t+1}$$

$$+ \cdots + \tilde{\mathbf{v}}_{t+h}$$

$$\tilde{\mathbf{v}}_t = \mathbf{k} + \mathbf{v}_t$$

$$\boldsymbol{\xi}_{t+h} = \mathbf{v}_{t+h} + \mathbf{F}\mathbf{v}_{t+h-1} + \mathbf{F}^2\mathbf{v}_{t+h-2} + \cdots + \mathbf{F}^h\mathbf{v}_t$$

$$+ \mathbf{k}_h + \mathbf{F}^{h+1}\boldsymbol{\xi}_{t-1}$$

$$\mathbf{k}_h = (\mathbf{I}_{mn} + \mathbf{F} + \mathbf{F}^2 + \cdots + \mathbf{F}^h)\mathbf{k}$$

$$\boldsymbol{\xi}_{t+h} = \mathbf{v}_{t+h} + \mathbf{F}\mathbf{v}_{t+h-1} + \mathbf{F}^2\mathbf{v}_{t+h-2} + \cdots + \mathbf{F}^h\mathbf{v}_t + \mathbf{k}_h + \mathbf{F}^{h+1}\boldsymbol{\xi}_{t-1}$$

First n rows give value of $\mathbf{y}_{t+h}$.

First n rows of $\mathbf{F}^h\mathbf{v}_t$ are given by $\boldsymbol{\Psi}_h\boldsymbol{\varepsilon}_t$ where $\boldsymbol{\Psi}_h$ denotes first n rows and columns of $\mathbf{F}^h$.

$$\mathbf{y}_{t+h} = \boldsymbol{\varepsilon}_{t+h} + \boldsymbol{\Psi}_1\boldsymbol{\varepsilon}_{t+h-1} + \boldsymbol{\Psi}_2\boldsymbol{\varepsilon}_{t+h-2} + \cdots + \boldsymbol{\Psi}_h\boldsymbol{\varepsilon}_t + \boldsymbol{\Pi}_{h+1}\mathbf{x}_{t-1}$$

Vector MA(∞) Representation

$$\mathbf{y}_{t+h} = \boldsymbol{\varepsilon}_{t+h} + \boldsymbol{\Psi}_1 \boldsymbol{\varepsilon}_{t+h-1} + \boldsymbol{\Psi}_2 \boldsymbol{\varepsilon}_{t+h-2} + \cdots$$
$$+ \boldsymbol{\Psi}_h \boldsymbol{\varepsilon}_t + \boldsymbol{\Pi}_{h+1} \mathbf{x}_{t-1}$$

$$\boldsymbol{\Pi}_{h+1} \mathbf{x}_{t-1} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} (\mathbf{k}_h + \mathbf{F}^{h+1} \boldsymbol{\xi}_{t-1})$$

$$\boldsymbol{\xi}_{t-1} = (\mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \ldots, \mathbf{y}'_{t-m})'$$

Coefficient on $\mathbf{y}_{t-1}$ is $\boldsymbol{\Psi}_{h+1}$.

Stability Condition

$$\underset{(nm\times nm)}{\mathbf{F}} = \begin{bmatrix} \mathbf{\Phi}_1 & \mathbf{\Phi}_2 & \mathbf{\Phi}_3 & \cdots & \mathbf{\Phi}_{m-1} & \mathbf{\Phi}_m \\ \mathbf{I}_n & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_n & \mathbf{0} \end{bmatrix}$$

$\Rightarrow$ if the eigenvalues of $\mathbf{F}$ all lie inside the unit circle, the VAR model is stable

$\Rightarrow$ implies: any shock must eventually die out

Nonorthogonalized Impulse Responses

$$\begin{aligned}\mathbf{y}_{t+h} &= \boldsymbol{\varepsilon}_{t+h} + \boldsymbol{\Psi}_1\boldsymbol{\varepsilon}_{t+h-1} + \boldsymbol{\Psi}_2\boldsymbol{\varepsilon}_{t+h-2} + \cdots \\ &\quad + \boldsymbol{\Psi}_h\boldsymbol{\varepsilon}_t + \boldsymbol{\Pi}_{h+1}\mathbf{x}_{t-1}\end{aligned}$$

What is effect on $\mathbf{y}_{t+h}$ if ε_{jt} goes up with all other ε's fixed?

Answer: column j of $\boldsymbol{\Psi}_h$

$$\boldsymbol{\Psi}_h = \frac{\partial \mathbf{y}_{t+h}}{\partial \boldsymbol{\varepsilon}_t'}$$

Nonorthogonalized Impulse Responses

$$\boldsymbol{\Psi}_h = \frac{\partial \mathbf{y}_{t+h}}{\partial \boldsymbol{\varepsilon}_t'}$$

- The moving average coefficients are called dynamic multipliers

  $\Rightarrow$ they allow us to see how a one-time shock propagates through time i.e. how a shock affects the endogenous variables in the system at a specific horizon h

Structural model:

$$\mathbf{A}\mathbf{y}_t = \mathbf{B}\,\mathbf{x}_{t-1} + \mathbf{D}^{1/2}\,\mathbf{v}_t \qquad \mathbf{v}_t \sim \text{i.i.d. } N(\mathbf{0}, \mathbf{I}_n)$$

Structural impulse responses:

$$\frac{\partial \mathbf{y}_{t+s}}{\partial \mathbf{v}_t'} = \frac{\partial \mathbf{y}_{t+s}}{\partial \boldsymbol{\varepsilon}_t'} \frac{\partial \boldsymbol{\varepsilon}_t}{\partial \mathbf{v}_t'} = \boldsymbol{\Psi}_s \mathbf{H} \qquad \boldsymbol{\Psi}_0 = \mathbf{I}_n$$

$$\mathbf{H} = \frac{\partial \boldsymbol{\varepsilon}_t}{\partial \mathbf{v}_t'} = \mathbf{A}^{-1}\mathbf{D}^{1/2}$$

Reduced form:

$$\mathbf{y}_t = \boldsymbol{\Phi}\,\mathbf{x}_{t-1} + \boldsymbol{\varepsilon}_t \qquad \boldsymbol{\varepsilon}_t \sim \text{i.i.d. } N(\mathbf{0}, \boldsymbol{\Omega})$$

$$\boldsymbol{\Phi} = \mathbf{A}^{-1}\mathbf{B}$$

$$\boldsymbol{\varepsilon}_t = \mathbf{A}^{-1}\mathbf{D}^{1/2}\mathbf{v}_t = \mathbf{H}\mathbf{v}_t$$

$$E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = \boldsymbol{\Omega} = \mathbf{A}^{-1}\mathbf{D}(\mathbf{A}^{-1})'$$

The Identification Problem

$$\mathbf{\Omega} = (\mathbf{A}^{-1})\mathbf{D}(\mathbf{A}^{-1})'$$

Supply and demand example:

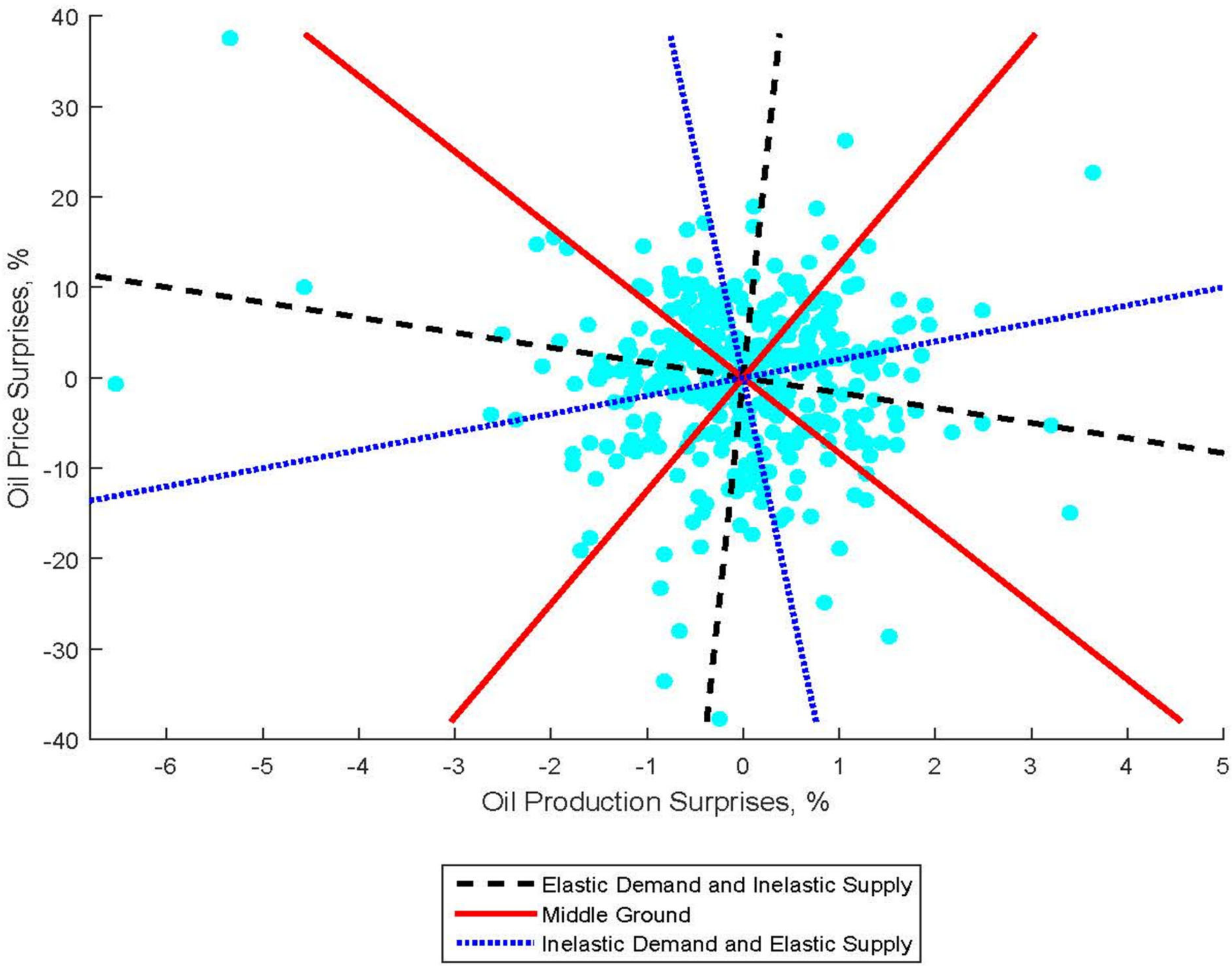
4 structural parameters in $\mathbf{A}$ and $\mathbf{D}$

$$(\alpha^s, \beta^d, d_s, d_d)$$

BUT can only estimate 3 parameters in $\mathbf{\Omega}$ by OLS

$$(\omega_{11}, \omega_{12}, \omega_{22})$$

The Identification Problem



Structural model:

$$\mathbf{A}\mathbf{y}_t = \boldsymbol{\lambda} + \mathbf{B}_1\mathbf{y}_{t-1} + \cdots + \mathbf{B}_m\mathbf{y}_{t-m} + \mathbf{u}_t \qquad \mathbf{u}_t = \mathbf{D}^{1/2}\mathbf{v}_t$$

$\mathbf{u}_t \sim$ i.i.d. $N(\mathbf{0}, \mathbf{D})$ $\quad$ $\mathbf{D}$ diagonal

Intuition:

If we knew row i of $\mathbf{A}$ (denoted $\mathbf{a}_i'$),

then we could estimate coefficients for

i^{th} structural equation ($\mathbf{b}_i'$) by OLS

regression of $\mathbf{a}_i'\mathbf{y}_t$ on $\mathbf{x}_{t-1}$:

$$\hat{\mathbf{b}}_i = \left(\sum_{t=1}^{T} \mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right)^{-1}\left(\sum_{t=1}^{T} \mathbf{x}_{t-1}\mathbf{y}_t'\mathbf{a}_i\right) = \hat{\boldsymbol{\Phi}}_T'\mathbf{a}_i$$

$$\hat{d}_{ii} = \mathbf{a}_i'\hat{\boldsymbol{\Omega}}_T\mathbf{a}_i \qquad \hat{\mathbf{D}} = \mathbf{A}\hat{\boldsymbol{\Omega}}_T\mathbf{A}' \text{ (diagonal)}$$

Approaches to Identification: Exact versus Inexact Prior Information

- Exact prior information:

  Researcher is certain of some features of $\mathbf{A}$ before seeing the data.

- Inexact prior information:

  Researcher has some ideas about $\mathbf{A}$ but is not certain of them.

Exact Identification

Put enough restrictions on $\mathbf{A}$ and $\mathbf{D}$ so that for any $\mathbf{\Omega}$ there is a unique $\mathbf{A}$ and $\mathbf{D}$ for which $\mathbf{\Omega} = \mathbf{A}^{-1}\mathbf{D}(\mathbf{A}^{-1})'$

$\Longrightarrow$ Point identification

Example

- Assume that short-run price elasticity of supply $\alpha^s = 0$

$$\mathbf{A} = \begin{bmatrix} 1 & -\alpha^s \\ 1 & -\beta^d \end{bmatrix}$$

$\Longrightarrow$ means $\mathbf{A}$ and $\mathbf{A}^{-1}$ are lower triangular

- If $\mathbf{D}$ is diagonal, then $\mathbf{H}$ is also lower triangular:

$$\mathbf{H} = \frac{\partial \mathbf{y}_t}{\partial \mathbf{v}_t'} = \mathbf{A}^{-1}\mathbf{D}^{1/2}$$

Algorithm to Estimate Structural IRFs

(1) Estimate reduced-form VAR parameters by OLS:

$$\widehat{\Phi} = \left[\sum_{t=1}^{T} \mathbf{y}_t \mathbf{x}_{t-1}'\right]\left[\sum_{t=1}^{T} \mathbf{x}_{t-1} \mathbf{x}_{t-1}'\right]^{-1}$$

$$\hat{\boldsymbol{\Omega}} = T^{-1} \sum_{t=1}^{T} (\mathbf{y}_t - \widehat{\boldsymbol{\Phi}} \mathbf{x}_{t-1})(\mathbf{y}_t - \widehat{\boldsymbol{\Phi}} \mathbf{x}_{t-1})'$$

(2) Find Cholesky factorization $\hat{\boldsymbol{\Omega}} = \hat{\mathbf{P}}\hat{\mathbf{P}}'$

Then we infer dynamic structural responses:

$$\frac{\partial \mathbf{y}_{t+s}}{\partial \mathbf{v}_t'} = \frac{\partial \mathbf{y}_{t+s}}{\partial \boldsymbol{\varepsilon}_t'} \frac{\partial \boldsymbol{\varepsilon}_t}{\partial \mathbf{v}_t'} = \widehat{\boldsymbol{\Psi}}_s \widehat{\mathbf{P}}$$

Algorithm to Estimate Structural IRFs

Example: Cholesky identification

$$\hat{\mathbf{A}}^{-1}\hat{\mathbf{D}}^{1/2} = \hat{\mathbf{P}}$$

$$\frac{\partial \mathbf{y}_{t+s}}{\partial \mathbf{u}_t'} = \mathbf{\Psi}_s \mathbf{A}^{-1} = \mathbf{\Psi}_s \hat{\mathbf{P}}\hat{\mathbf{D}}^{-1/2}$$

Units: what happens to $\mathbf{y}_{t+s}$ if Fed sets interest rate 1% higher than its usual rule?

Many researchers report $\mathbf{\Psi}_s \hat{\mathbf{P}}$

Units: what happens to $\mathbf{y}_{t+s}$ if Fed sets interest rate 1 standard deviation higher than its usual rule?

Algorithm to Estimate Structural Parameters

(1) Estimate reduced-form VAR parameters by OLS:

$$\widehat{\Phi} = \left[\sum_{t=1}^{T} \mathbf{y}_t \mathbf{x}_{t-1}'\right]\left[\sum_{t=1}^{T} \mathbf{x}_{t-1} \mathbf{x}_{t-1}'\right]^{-1}$$

$$\hat{\mathbf{\Omega}} = T^{-1} \sum_{t=1}^{T} (\mathbf{y}_t - \widehat{\mathbf{\Phi}}\mathbf{x}_{t-1})(\mathbf{y}_t - \widehat{\mathbf{\Phi}}\mathbf{x}_{t-1})'$$

(2) Find Cholesky factorization $\hat{\mathbf{\Omega}} = \hat{\mathbf{P}}\hat{\mathbf{P}}'$

(3) Set diagonal elements of $\hat{\mathbf{D}}$ to squares of diagonal elements of $\hat{\mathbf{P}}$.

(4) Set $\hat{\mathbf{A}} = \hat{\mathbf{D}}^{1/2}\hat{\mathbf{P}}^{-1}$.

Link to Instrumental Variables

- If we know the parameter of the supply equation, we could use the supply shock as instrument to estimate the parameter in the demand equation (see Baumeister and Hamilton, ET 2024).

$\Longrightarrow$ MLE of β^d conditional on knowing α^s can be found from IV regression of $\widehat{\varepsilon}_{qt}$ on $\widehat{\varepsilon}_{pt}$ using $(\widehat{\varepsilon}_{qt} - \alpha^s \widehat{\varepsilon}_{pt})$ as instrument:

$$\widehat{\beta}^d_{MLE} = \frac{\sum_{t=1}^{T} (\widehat{\varepsilon}_{qt} - \alpha^s \widehat{\varepsilon}_{pt}) \widehat{\varepsilon}_{qt}}{\sum_{t=1}^{T} (\widehat{\varepsilon}_{qt} - \alpha^s \widehat{\varepsilon}_{pt}) \widehat{\varepsilon}_{pt}} = \frac{(\widehat{\omega}_{qq} - \alpha^s \widehat{\omega}_{qp})}{(\widehat{\omega}_{qp} - \alpha^s \widehat{\omega}_{pp})}$$

Point Identification: Example

- Application 2a: Simple bivariate supply and demand model of the global oil market

$$\mathbf{y}_t = (\Delta q_t, p_t)'$$

  - Δq_t = oil production growth
  - p_t = real price of oil
  - monthly VAR(24) for 1975M2 to 2007M12

⇨ Compute IRFs for oil supply and oil demand shocks assuming that supply elasticity $\alpha^s = 0$

Point Identification: Example

- Oil demand elasticity estimate

$$\hat{\boldsymbol{\Omega}} = \begin{bmatrix} 2.2825 & -0.4720 \\ -0.4720 & 32.2594 \end{bmatrix}$$

$$\widehat{\beta}^{d}_{MLE} = \frac{(\widehat{\omega}_{qq} - \alpha^{s}\widehat{\omega}_{qp})}{(\widehat{\omega}_{qp} - \alpha^{s}\widehat{\omega}_{pp})}$$

$$\Rightarrow (\widehat{\beta}^{d}_{MLE} \mid \alpha^{s} = 0) = \frac{\widehat{\omega}_{qq}}{\widehat{\omega}_{qp}} = \frac{2.2825}{-0.472} = -4.8358$$

Inexact Identification

- Assumption that short-run supply elasticity $\alpha^s = 0$ is very strong

$\Longrightarrow$ Can we make inference using weaker assumptions?

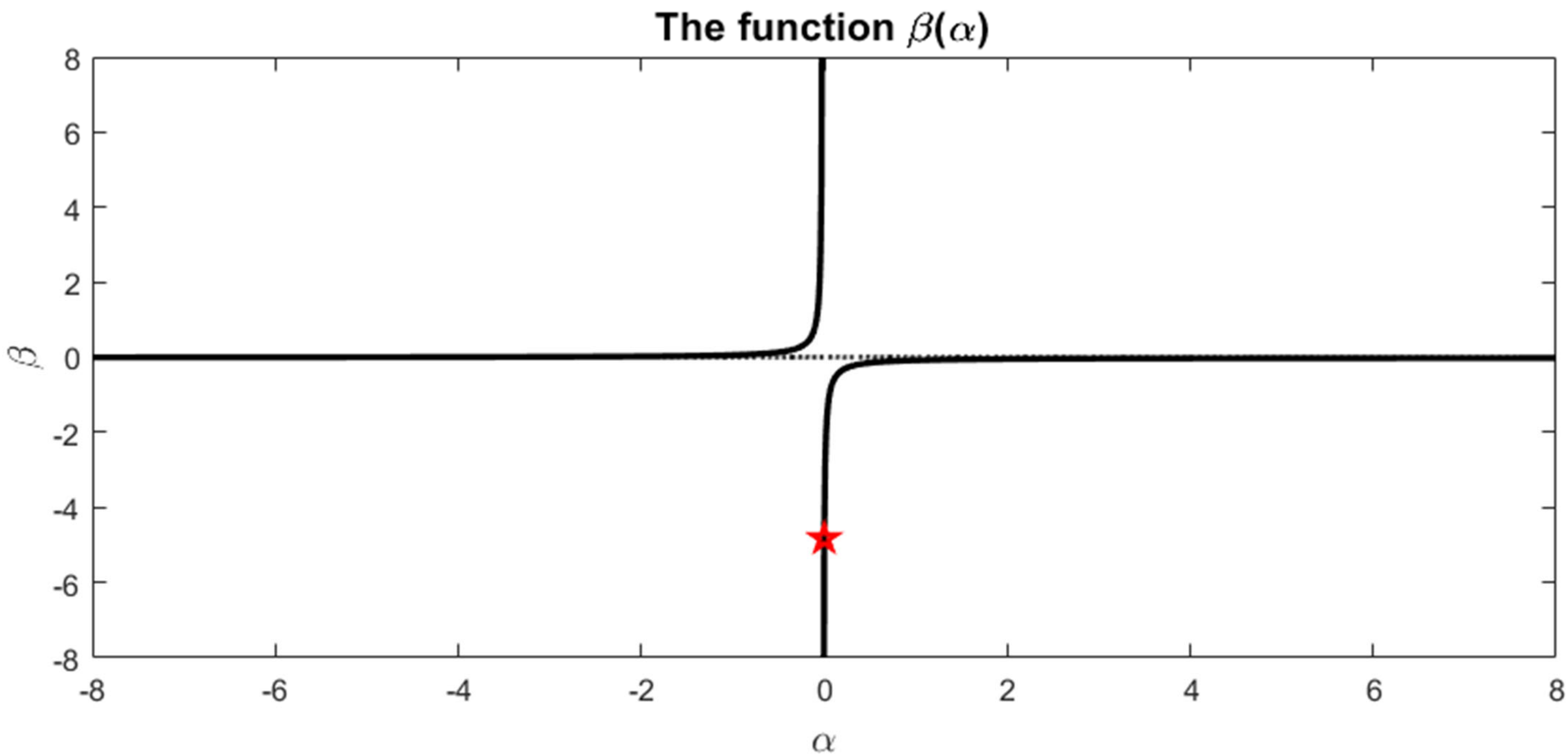
Identified Set

If we know nothing at all about α^s or β^d before seeing the data, MLE is a *set* of values for α^s and β^d.

Every element in this set maximizes the likelihood function.

The MLE $\{\hat{\alpha}^s, \hat{\beta}^d\}_{MLE}$ is the set of all values of (α^s, β^d) for which $\widehat{\beta}^d = \dfrac{(\widehat{\omega}_{qq} - \alpha^s \widehat{\omega}_{qp})}{(\widehat{\omega}_{qp} - \alpha^s \widehat{\omega}_{pp})}$.

Every value of (α^s, β^d) on this curve maximizes the likelihood function.



Red star: Cholesky-identified oil demand elasticity

Additional Information

- We may have additional exact prior information that $\alpha^s \geq 0$ and $\beta^d \leq 0$.

$\Longrightarrow$ MLE of (α^s, β^d) is still a set, but a subset of the original.

Inexact Identification: Identification Using Inequality Constraints

- We may have confidence in signs:

$$\mathbf{H} = \begin{bmatrix} \partial q_t/\partial v_t^d & \partial q_t/\partial v_t^s \\ \partial p_t/\partial v_t^d & \partial p_t/\partial v_t^s \end{bmatrix} = \begin{bmatrix} + & - \\ + & + \end{bmatrix}$$

⇨ Set identification: want set of all $(n \times n)$ $\mathbf{H}$ such that

(1) $\mathbf{H}\mathbf{H}' = \mathbf{\Omega}$

(2) $\mathbf{H}$ satisfies certain sign restrictions

(3) columns of $\mathbf{H}$ are orthogonal to each other

How Do We Obtain Such an **H**?

- Claim: the set of all $\mathbf{H}$ such that $\mathbf{HH}' = \mathbf{\Omega}$ is the set defined by $\mathbf{H} = \mathbf{PQ}$ where $\mathbf{Q}$ is the set of all orthogonal matrices (all $(n \times n)$ matrices satisfying $\mathbf{QQ}' = \mathbf{I}_n$)

$$\Rightarrow \quad \mathbf{HH}' = \mathbf{PQ}(\mathbf{PQ})' = \mathbf{PQQ}'\mathbf{P}' = \mathbf{PI}_n\mathbf{P}' = \mathbf{\Omega}$$

- $\mathbf{Q}$ is called a *rotation matrix* because it allows to "rotate" the initial Cholesky factor (recursive matrix) while maintaining the property that shocks are uncorrelated

One strategy:

(1) Generate a million matrices $\mathbf{Q}^{(j)}$ $j = 1, \ldots, 10^6$ drawn "uniformly" from the set of all orthogonal matrices.

(2) For each $\mathbf{Q}^{(j)}$ calculate $\mathbf{H}^{(j)} = \hat{\mathbf{P}}\mathbf{Q}^{(j)}$ for $\hat{\mathbf{P}}\hat{\mathbf{P}}' = \hat{\mathbf{\Omega}}$.

(3) Keep $\mathbf{H}^{(j)}$ if it satisfies restrictions.

How to Generate a Draw for $\mathbf{Q}$?

Rubio-Ramírez, Waggoner, and Zha (2010) algorithm to generate $\mathbf{Q}^{(j)}$:

(1) Generate an $(n \times n)$ matrix $\mathbf{X}^{(j)}$ of independent $N(0,1)$ variables

(2) Calculate the QR decomposition $\mathbf{X}^{(j)} = \mathbf{Q}^{(j)}\mathbf{R}^{(j)}$ where $\mathbf{Q}^{(j)}$ is orthogonal and $\mathbf{R}^{(j)}$ is upper triangular

(e.g. use qr command in Matlab)

$\Rightarrow$ Caution: need to undo the sign normalization conventions in some software implementations of *QR*

Problem:

If we do this for fixed $\hat{\mathbf{\Omega}}$, the resulting set just reports uncertainty resulting from incomplete identification.

But we also don't know the true $\mathbf{\Omega}$.

Solution:

(i) Generate $(\mathbf{\Omega}^{(j)})^{-1}$ from Wishart with

$T-p$ degrees of freedom and scale matrix $T\hat{\mathbf{\Omega}}$.

(ii) Calculate $\mathbf{P}^{(j)}\mathbf{P}^{(j)\prime} = \mathbf{\Omega}^{(j)}$ and $\mathbf{H}^{(j)} = \mathbf{P}^{(j)}\mathbf{Q}^{(j)}$.

Traditional Sign Restriction Algorithm

Step 1 Take a draw $(\mathbf{\Phi}, \mathbf{\Omega})$ from the posterior of the reduced-form coefficients

Step 2 Compute the Cholesky factor $\mathbf{P}$ of $\mathbf{\Omega}$

Step 3 Generate an $(n \times n)$ matrix $\mathbf{X} = [x_{ij}]$ from $N(0,1)$

Step 4 Take the QR decomposition of $\mathbf{X} = \mathbf{QR}$ for $\mathbf{Q}$ orthogonal and $\mathbf{R}$ upper triangular. Normalize the elements in $\mathbf{Q}$ such that the diagonal entries of $\mathbf{R}$ are positive.

Step 5 Compute IRFs using $\mathbf{H} = \mathbf{PQ}$

Step 6 Keep $\mathbf{H}$ if it satisfies sign restrictions; otherwise, discard it.

Let's Look at the Algorithm

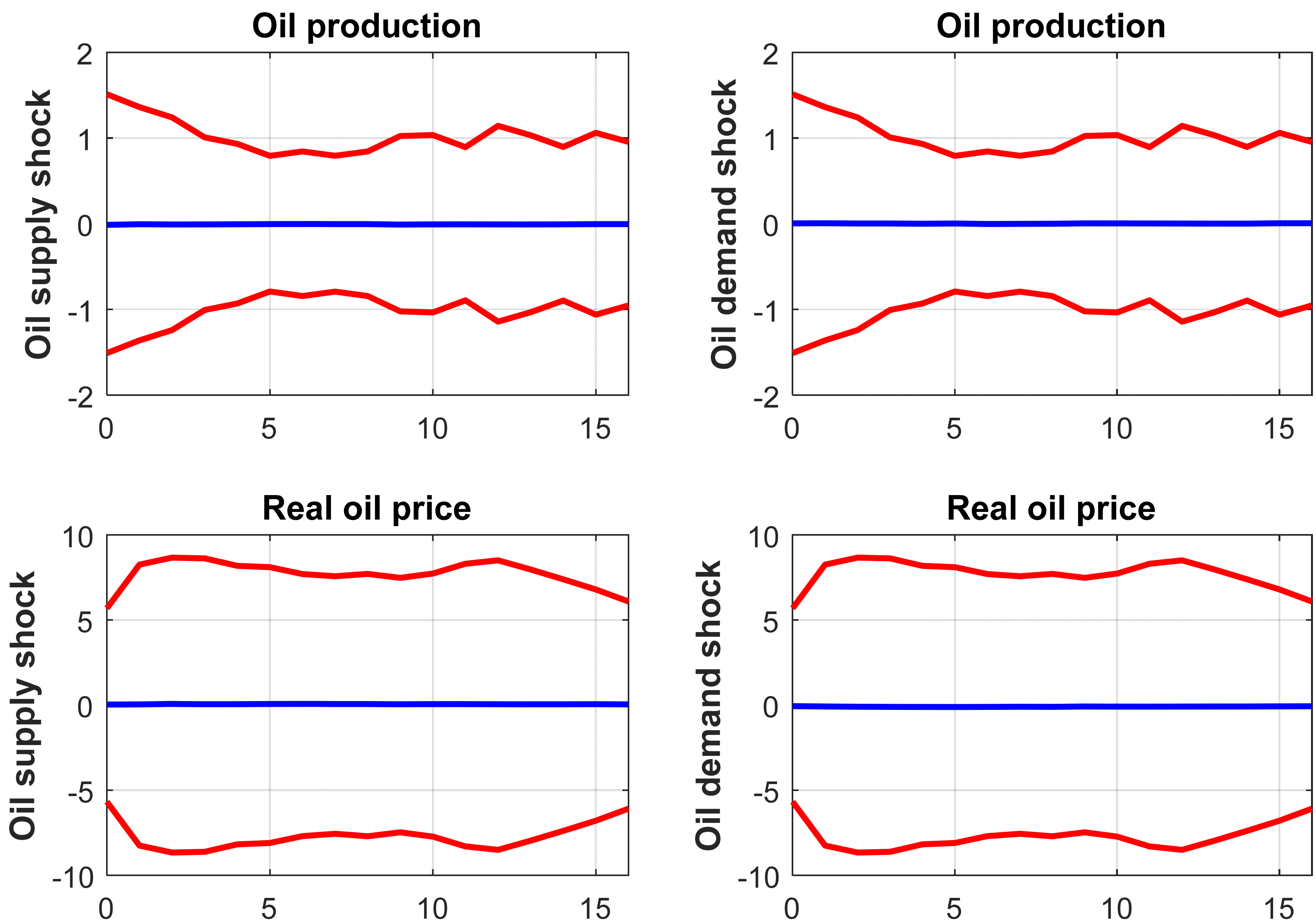
- Application 2b: Simple bivariate supply and demand model of the global oil market

$$\mathbf{y}_t = (\Delta q_t, p_t)'$$

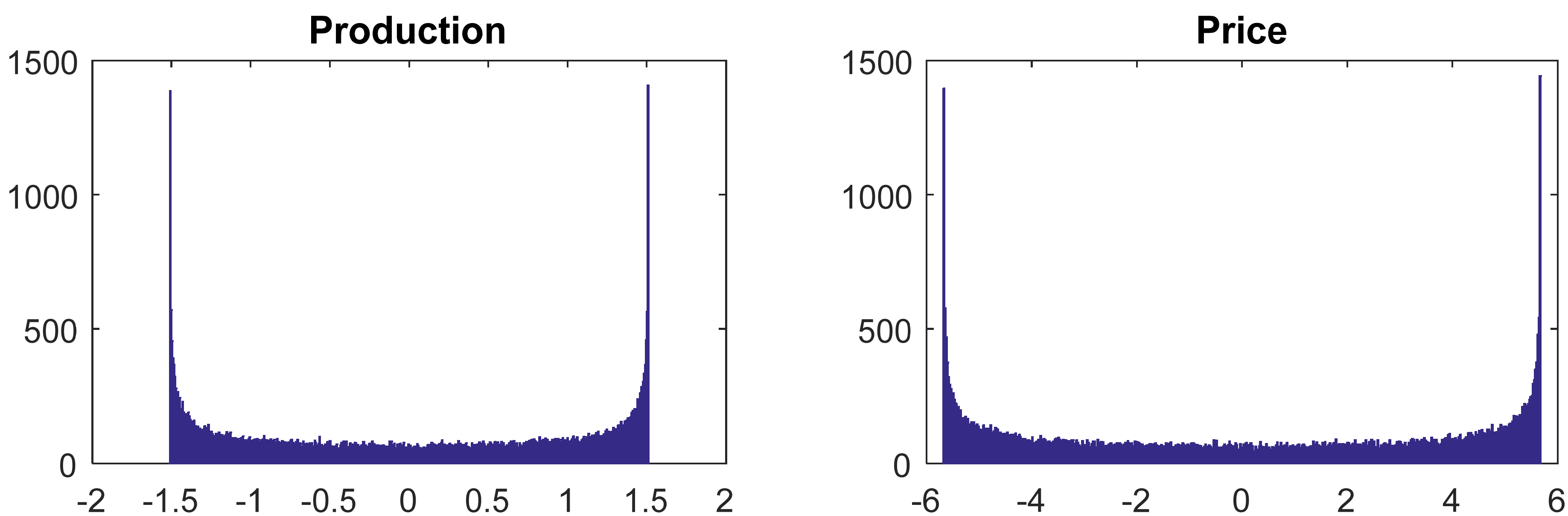
  - Δq_t = oil production growth
  - p_t = real price of oil
  - monthly VAR(24) for 1975M2 to 2007M12

- Question: What happens if we don't impose ANY restrictions at all?

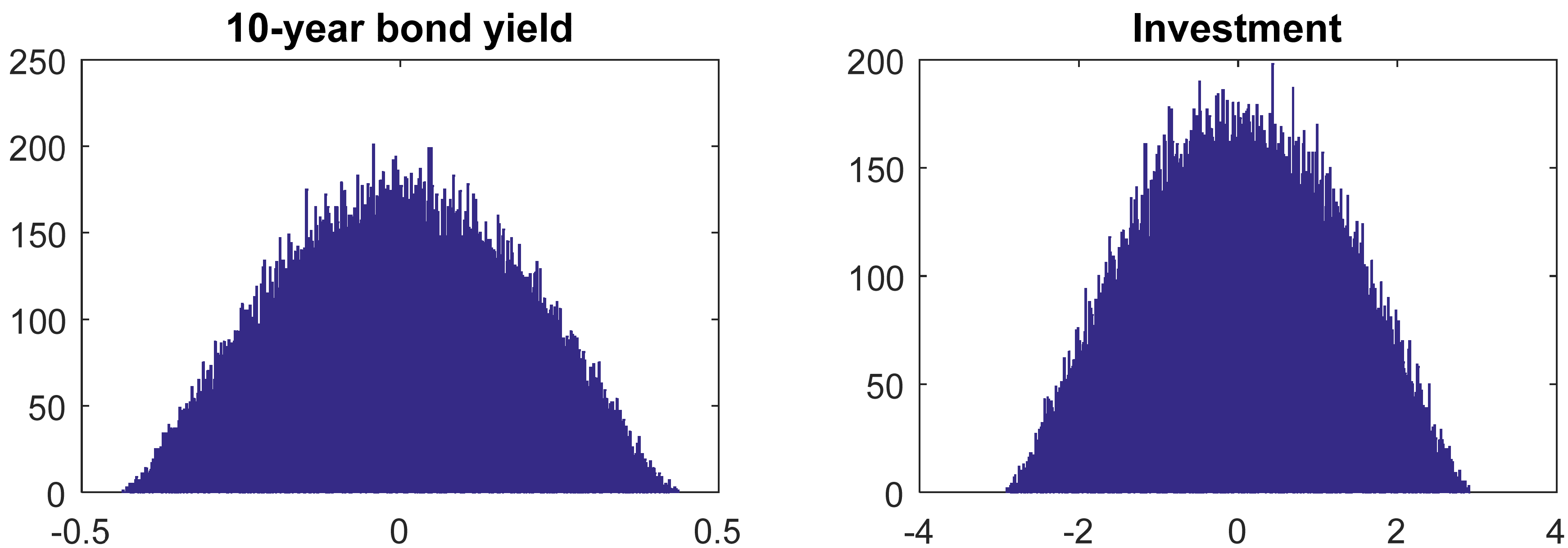
Impulse responses for one-standard-deviation supply and demand shocks



Histogram of impact effect of one-standard-deviation shocks: bivariate model



Histogram of impact effect of one-standard-deviation shocks: 6-variable VAR



⇨ What is going on here?

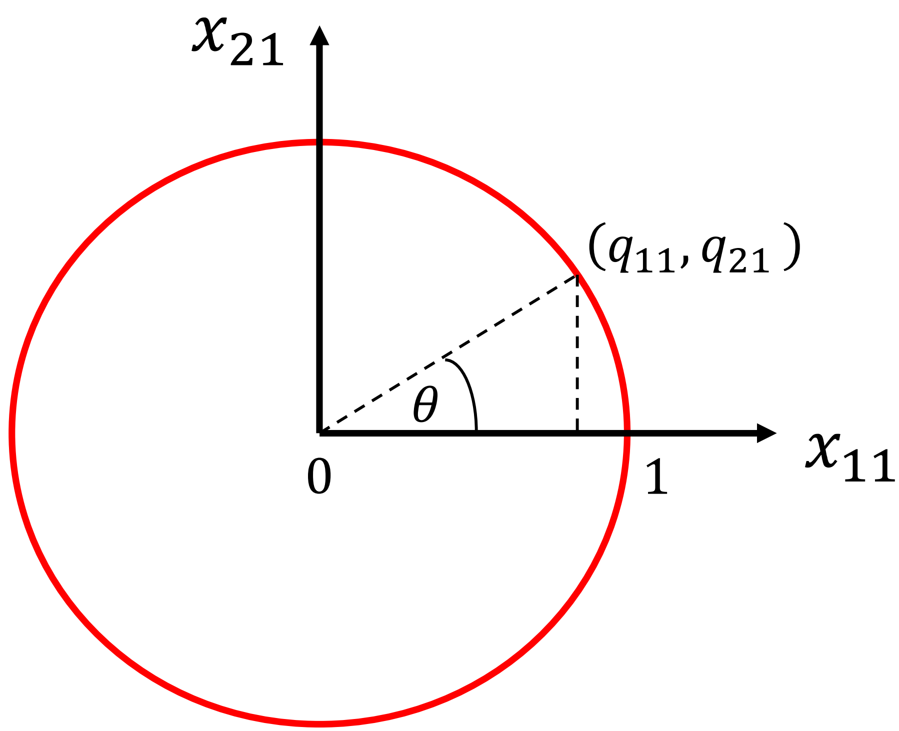
A Closer Look at $\mathbf{Q}$

- $\mathbf{Q}$ is an orthogonal matrix which means that its columns and rows are orthogonal unit vectors (orthonormal vectors)

- First column of $\mathbf{Q}$ $=$ first column of $\mathbf{X}$ normalized to have unit length:

$$\begin{bmatrix} q_{11} \\ \vdots \\ q_{n1} \end{bmatrix} = \begin{bmatrix} x_{11}/\sqrt{x_{11}^2 + \cdots + x_{n1}^2} \\ \vdots \\ x_{n1}/\sqrt{x_{11}^2 + \cdots + x_{n1}^2} \end{bmatrix}$$

Bivariate Case

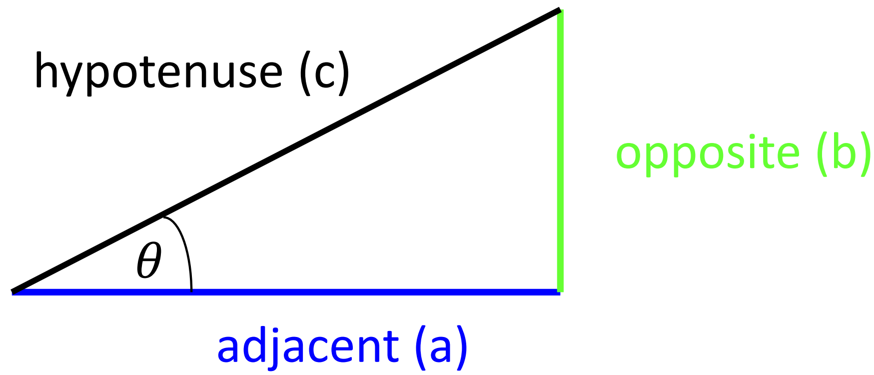


$$q_{11} = x_{11}/\sqrt{x_{11}^2 + x_{21}^2}$$

$$q_{21} = x_{21}/\sqrt{x_{11}^2 + x_{21}^2}$$

θ is the angle between $(1,0)$ and (x_{11}, x_{21})

Some Trigonometry



cosine of the angle θ gives the length of the x-component:

$$\cos(\theta) = \frac{adjacent}{hypotenus} = \frac{a}{c}$$

sine of the angle θ gives the length of the y-component:

$$\sin(\theta) = \frac{opposite}{hypotenus} = \frac{b}{c}$$

Pythagoras: $a^2 + b^2 = c^2 \Rightarrow c = \sqrt{a^2 + b^2}$

Rotation Matrix

$$q_{11} = x_{11} / \sqrt{x_{11}^2 + x_{21}^2} = \cos(\theta)$$

$$q_{21} = x_{21} / \sqrt{x_{11}^2 + x_{21}^2} = \sin(\theta)$$

$$\mathbf{Q} = \begin{cases} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} & \text{with prob } 1/2 \\ \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix} & \text{with prob } 1/2 \end{cases}$$

$$\theta \sim U(-\pi, \pi)$$

$$q_{i1} = x_{i1}/\sqrt{x_{11}^2 + \cdots + x_{n1}^2}$$

$$q_{i1}^2 = x_{i1}^2/(x_{11}^2 + \cdots + x_{n1}^2)$$

Recall: $x_{11} \sim N(0,1) \Rightarrow x_{11}^2 \sim \chi^2(1)$

$$\Rightarrow (x_{11}^2 + \cdots + x_{n1}^2) \sim \chi^2(n)$$

In general: if $X \sim \chi^2(\alpha)$ and $Y \sim \chi^2(\beta)$ are independent,

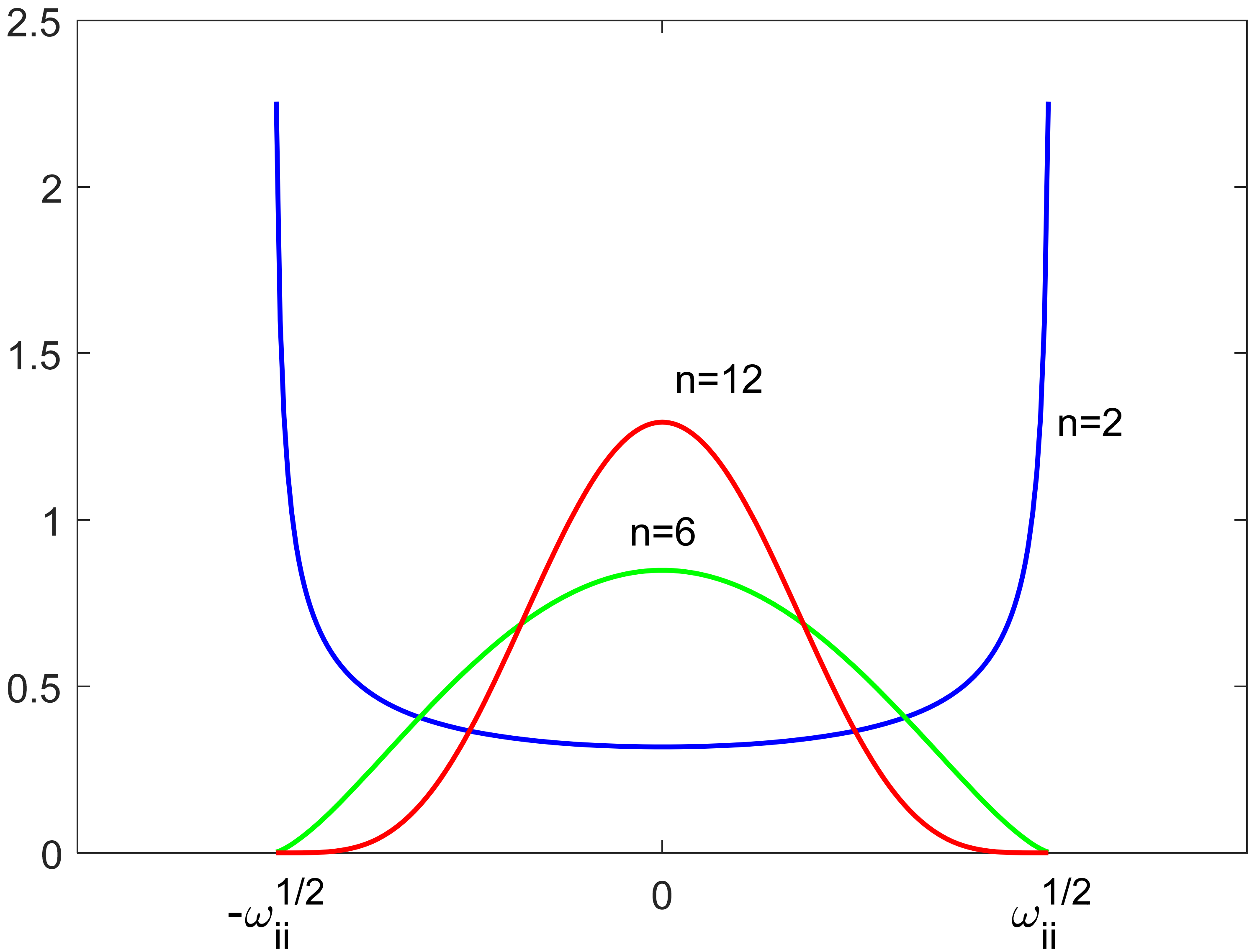
then $\frac{X}{X+Y} \sim$ Beta$(\frac{\alpha}{2}, \frac{\beta}{2})$

$\Rightarrow q_{i1}^2 \sim$ Beta$(1/2, (n-1)/2)$

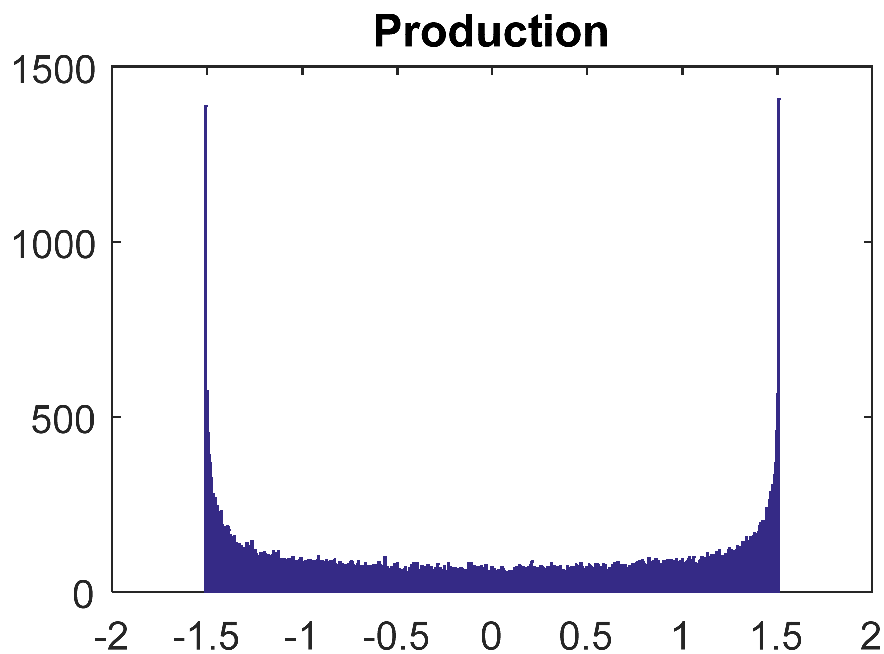
$$p(q_{i1}) = \begin{cases} \frac{\Gamma(n/2)}{\Gamma(1/2)\Gamma((n-1)/2)}(1-q_{i1}^2)^{(n-3)/2} & \text{if } q_{i1} \in [-1,1] \\ 0 & \text{otherwise} \end{cases}$$

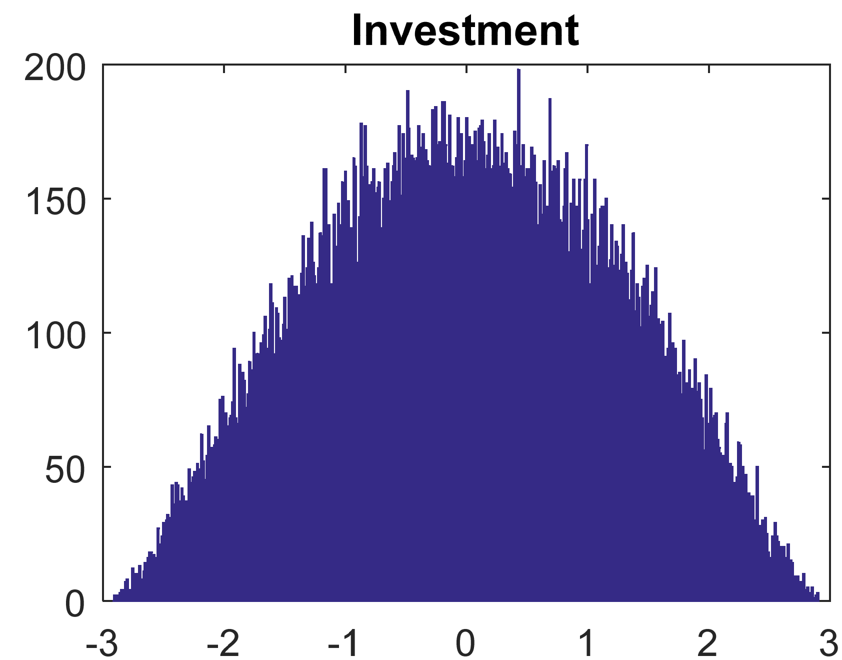
$$h_{11} = p_{11}q_{11} = \sqrt{\omega_{11}}\,q_{11}$$

Impact effect of one-standard-deviation shock on variable *i*: analytic distribution



Impact effect of one-standard-deviation shock on variable i: evidence





$$\hat{\mathbf{\Omega}}_T = \begin{bmatrix} 2.28 & -0.47 \\ -0.47 & 32.26 \end{bmatrix}$$

$\sqrt{\omega_{66}} = 2.94$ (6-variable VAR)

$\sqrt{\omega_{11}} = 1.51$ (bivariate)

Take-Away #1

- A prior that is UNINFORMATIVE about a parameter (here: the angle of rotation θ) is in general ***informative*** about nonlinear transformations of θ

⇒ there is an *informative prior* for IRFs implicit in the traditional sign restriction algorithm

Why is this Problematic?

- This prior implies that before seeing the data, we would have the same prior knowledge regardless of the dataset, economic content, sample period …

  $\Rightarrow$ reasonable?

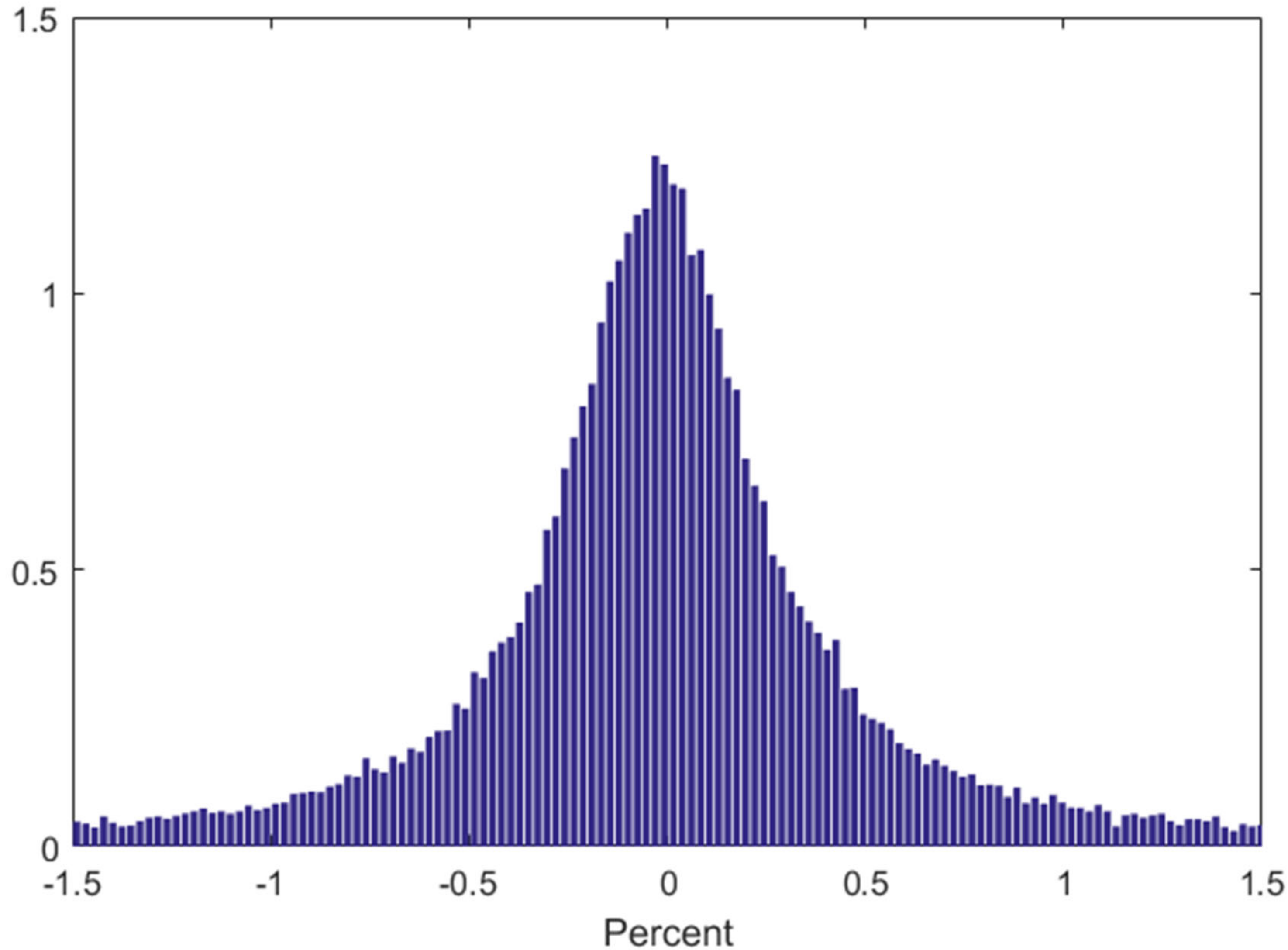
Other Objects of Interest

- Suppose we are interested in the effect of a shock that raises the price by 1% on quantity
- In the bivariate case with $\boldsymbol{y}_t = [p_t \ \ q_t]'$, we normalize the impact matrix **H** by dividing the first column by its first element

$$\mathbf{H} = \begin{bmatrix} 1 & \dots \\ \frac{h_{21}}{h_{11}} & \dots \end{bmatrix}$$

$\Longrightarrow$ this ratio is called a price elasticity (only valid in bivariate case)

Impact effect of 1% increase in price on quantity (without sign restrictions)



What's the implicit prior distribution here?

$$\mathbf{H} = \mathbf{PQ}$$

$$\Rightarrow \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} = \begin{bmatrix} p_{11} & 0 \\ p_{21} & p_{22} \end{bmatrix} \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix}$$

If we normalize shock 1 as something

that raises variable 1 by 1 unit:

$$h_{21}^{*} = \frac{h_{21}}{h_{11}} = \frac{p_{21}q_{11}+p_{22}q_{21}}{p_{11}q_{11}} = \frac{p_{21}}{p_{11}} + \frac{p_{22}}{p_{11}}\frac{x_{21}}{x_{11}}$$

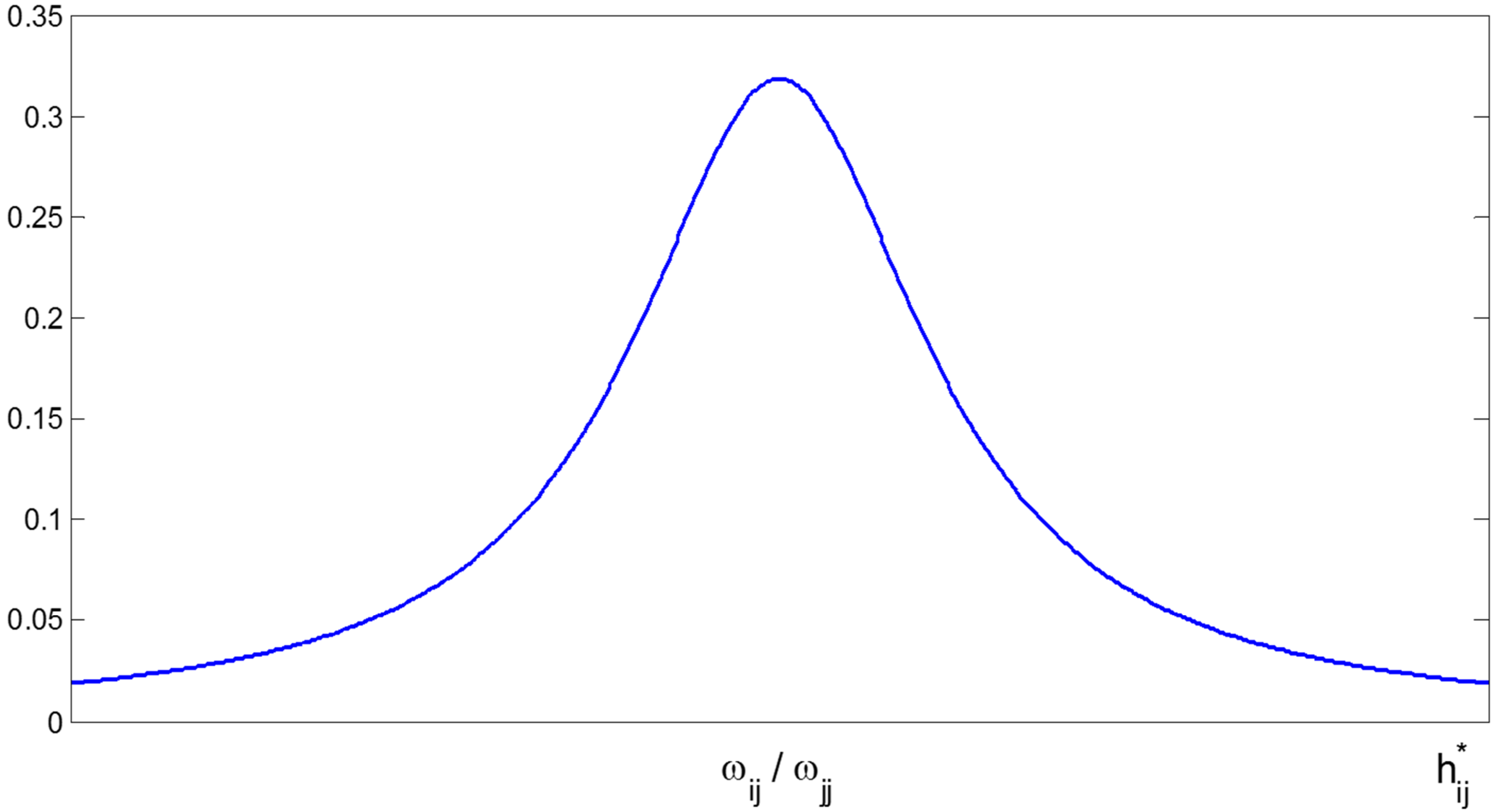
x_{21}/x_{11} ~ Cauchy$(0,1)$

$\Rightarrow h_{ij}^{*}|\mathbf{\Omega}$ ~ Cauchy$(c_{ij}^{*}, \sigma_{ij}^{*})$

location parameter: $c_{ij}^{*} = \omega_{ij}/\omega_{jj}$

scale parameter: $\sigma_{ij}^{*} = \sqrt{\frac{\omega_{ii}-\omega_{ij}^{2}/\omega_{jj}}{\omega_{jj}}}$

Impact effect on variable i of shock that increases j by one unit



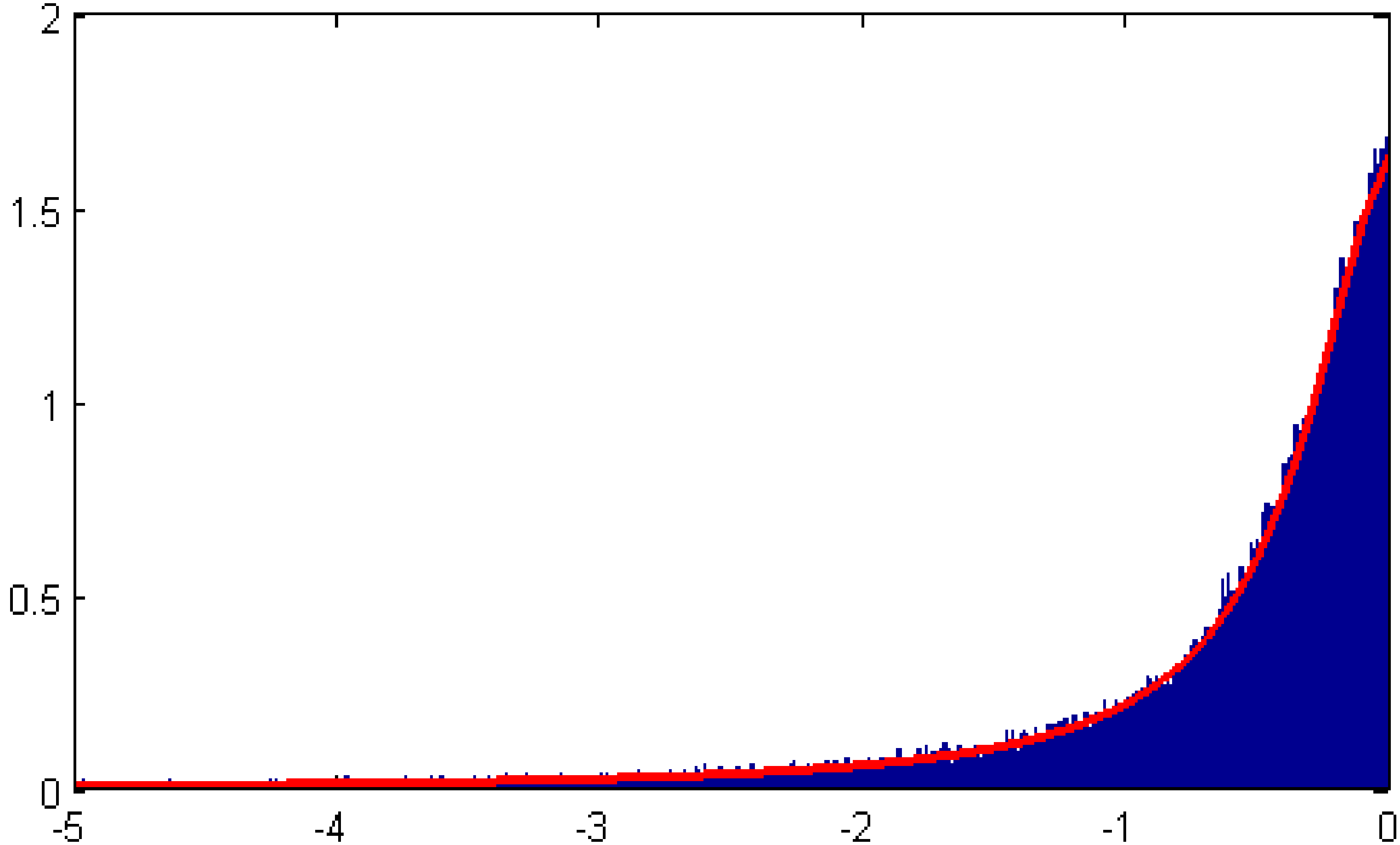
What Happens If We Impose Sign Restrictions?

- Sign restrictions confine these distributions to particular regions but do not change their basic features.
- Apply traditional algorithm to 8-lag VAR fit to growth rates of U.S. real compensation per worker Δw_t and of U.S. employment Δn_t for the period 1970:Q1-2014:Q2

  ⇒ Application 3: sign-identified labor market model
- Identify supply and demand shock by sign restrictions:

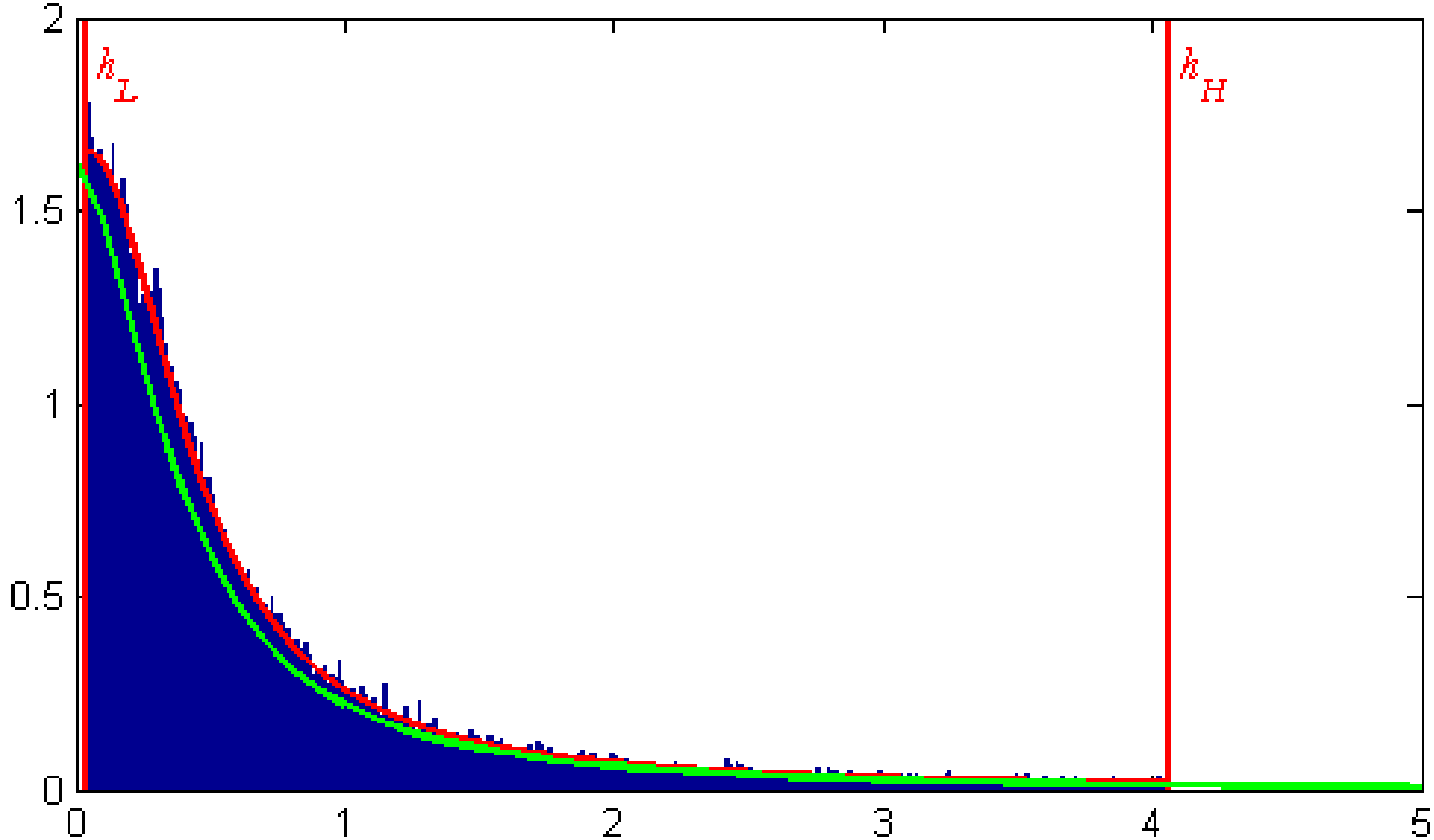
$$\begin{bmatrix} \Delta w_t \\ \Delta n_t \end{bmatrix} = \begin{bmatrix} + & + \\ + & - \end{bmatrix}$$

Implied elasticity of labor demand ($= h_{22}^*$)



Red = truncated Cauchy
Blue = output of traditional algorithm

Implied elasticity of labor supply ($= h_{21}^{*}$)



Red = truncated Cauchy
Blue = output of traditional algorithm

What's the Nature of this Truncation?

$$\mathbf{H} = \mathbf{PQ}$$

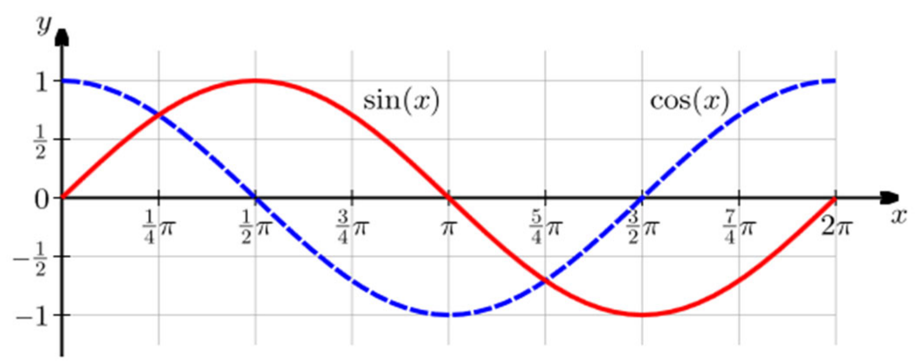
$$\Rightarrow \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} = \begin{bmatrix} p_{11} & 0 \\ p_{21} & p_{22} \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}$$

$$= \begin{bmatrix} p_{11}\cos\theta & p_{11}\sin\theta \\ (p_{21}\cos\theta + p_{22}\sin\theta) & (p_{21}\sin\theta - p_{22}\cos\theta) \end{bmatrix}$$

variable 1 = wage, variable 2 = employment

shock 1 = demand, shock 2 = supply

$$\begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} = \begin{bmatrix} + & + \\ + & - \end{bmatrix}$$



$$h_{11}, h_{12} \geq 0 \Rightarrow \theta \in [0, \pi/2]$$

What's the Nature of this Truncation?

$$\begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} = \begin{bmatrix} p_{11}\cos\theta & p_{11}\sin\theta \\ (p_{21}\cos\theta + p_{22}\sin\theta) & (p_{21}\sin\theta - p_{22}\cos\theta) \end{bmatrix}$$

If $p_{21} > 0$, then $h_{21} \geq 0$ for all $\theta \in [0, \pi/2]$

But $h_{22} \leq 0 \Rightarrow \theta \in [0, \tilde{\theta}]$ for $\cot\tilde{\theta} = p_{21}/p_{22}$

$$\hat{\mathbf{\Omega}} = \begin{bmatrix} 0.5920 & 0.0250 \\ 0.0250 & 0.1014 \end{bmatrix}$$

What Does This Imply for the Elasticities?

$$\begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} = \begin{bmatrix} p_{11}\cos\theta & p_{11}\sin\theta \\ (p_{21}\cos\theta + p_{22}\sin\theta) & (p_{21}\sin\theta - p_{22}\cos\theta) \end{bmatrix}$$

$$h_{21}^* = \frac{h_{21}}{h_{11}} = \frac{(p_{21}\cos\theta + p_{22}\sin\theta)}{p_{11}\cos\theta} = \frac{p_{21}}{p_{11}} + \frac{p_{22}}{p_{11}}\tan\theta$$

$$h_{22}^* = \frac{h_{22}}{h_{12}} = \frac{(p_{21}\sin\theta - p_{22}\cos\theta)}{p_{11}\sin\theta} = \frac{p_{21}}{p_{11}} - \frac{p_{22}}{p_{11}}\cot\theta$$

for $\theta \in [0, \widetilde{\theta}] \Rightarrow h_{21}^* \in [\omega_{21}/\omega_{11}, \omega_{22}/\omega_{21}]$

$$h_{22}^* \in (-\infty, 0]$$

for $\theta = 0$: $h_{21}^{*} = \frac{p_{21}}{p_{11}}$ since $\tan 0 = 0$

$h_{22}^{*} = -\infty$ since $\cot 0 = \infty$

for $\theta = \widetilde{\theta}$: $h_{21}^{*} = \frac{p_{21}}{p_{11}} + \frac{p_{22}}{p_{11}} \frac{p_{22}}{p_{21}} = \frac{p_{21}^{2} + p_{22}^{2}}{p_{11} p_{21}}$

since $\tan \widetilde{\theta} = 1/\cot \widetilde{\theta}$ and $\cot \widetilde{\theta} = \frac{p_{21}}{p_{22}}$

$h_{22}^{*} = \frac{p_{21}}{p_{11}} - \frac{p_{22}}{p_{11}} \frac{p_{21}}{p_{22}} = 0$

since $\cot \widetilde{\theta} = \frac{p_{21}}{p_{22}}$

$$\mathbf{\Omega} = \mathbf{P}\mathbf{P}' = \begin{bmatrix} p_{11} & 0 \\ p_{21} & p_{22} \end{bmatrix} \begin{bmatrix} p_{11} & p_{21} \\ 0 & p_{22} \end{bmatrix}$$

$$= \begin{bmatrix} p_{11}^{2} & p_{21} p_{11} \\ p_{21} p_{11} & p_{21}^{2} + p_{22}^{2} \end{bmatrix} = \begin{bmatrix} \omega_{11} & \omega_{21} \\ \omega_{21} & \omega_{22} \end{bmatrix}$$

Take-Away #2

- The sign restrictions may end up implying no or trivial restrictions on the feasible set

$$\beta^d \in (-\infty, 0]$$

$$\alpha^s \in [0.04, 4.06]$$

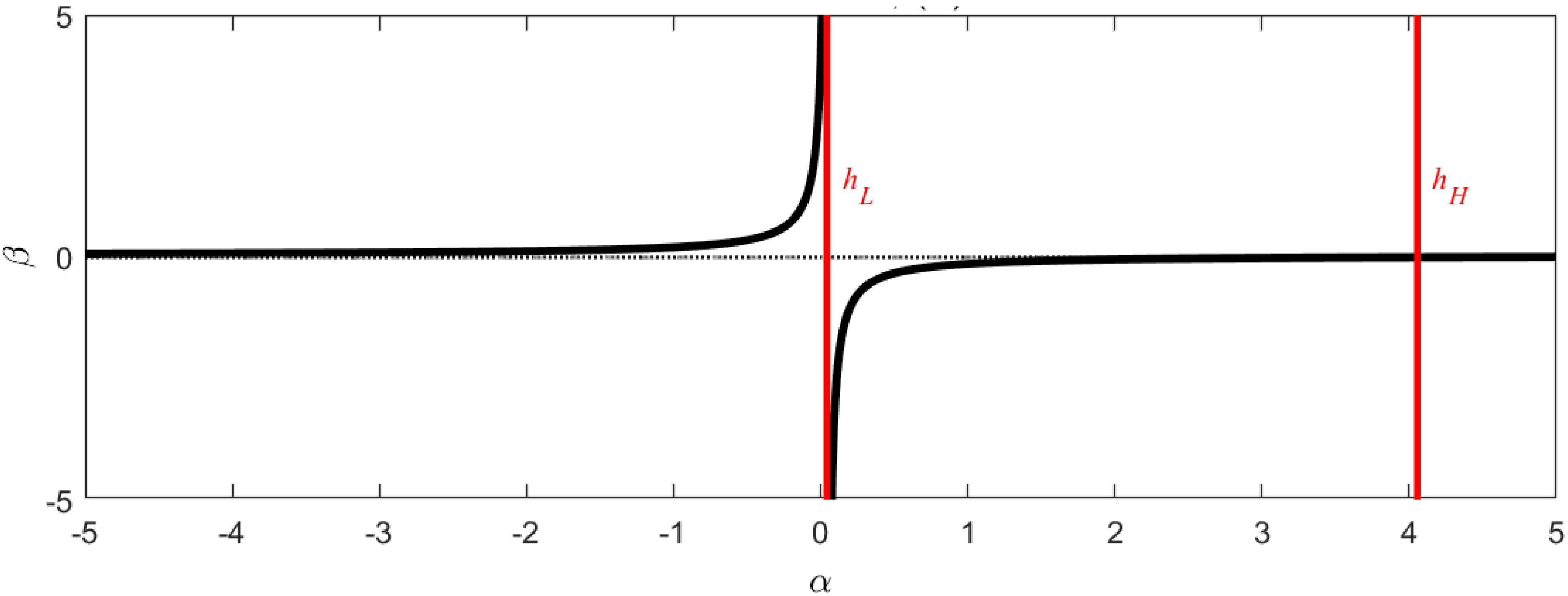
⇨ the allowable set is uselessly large

- Remedy? Add information in the form of more and tighter constraints

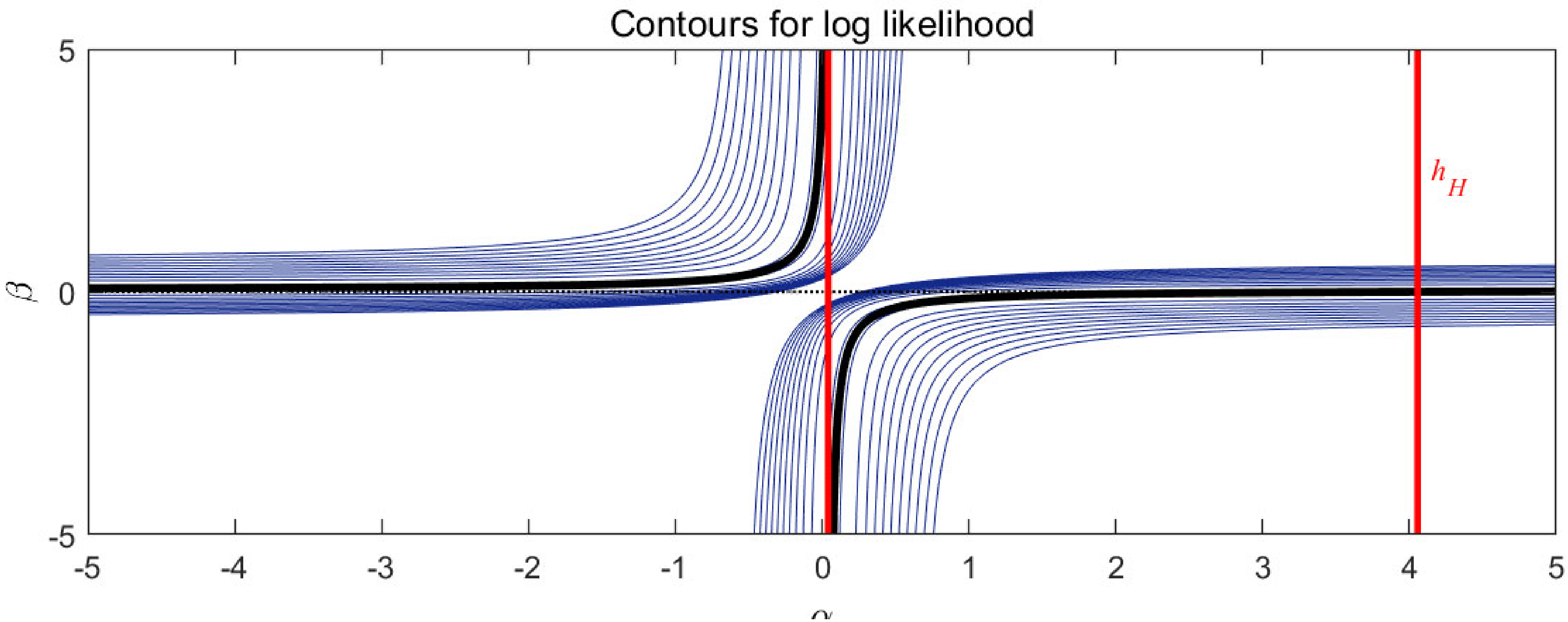
⇨ Example: Kilian and Murphy (2012, 2014)

BUT raises inference and computational issues

Every value of (α^s, β^d) on this curve between h_L and h_U maximizes the likelihood function and satisfies the sign restrictions.



But true Ω could be in a range of values around $\hat{\Omega}$. So true value of (α^s, β^d) could be in a range around the MLE $\{\hat{\alpha}^s, \hat{\beta}^d\}_{MLE}$.



Meaning: True h_L could be less than $\hat{h}_L$,

true h_H could be bigger than $\hat{h}_H$.

Reason: Sampling uncertainty

As $T \to \infty$, $\hat{\mathbf{\Omega}} \xrightarrow{p} \mathbf{\Omega}_0$ and the MLE $\{\hat{\alpha}^s, \hat{\beta}^d\}_{MLE}$ would converge to a fixed curve in $\mathbb{R}^2$.

This curve is called the "identified set."

For this example, the identified set is

$$\left\{(\alpha^s, \beta^d) \,:\, \beta^d = \frac{(\omega_{0,nn} - \alpha^s \omega_{0,wn})}{(\omega_{0,wn} - \alpha^s \omega_{0,ww})}\right\}$$

With an infinite sample, we would know that (α^s, β^d) must be in the identified set, but that is all we could know

- Researchers typically report median and 68% of retained set of values.

- BUT if our prior information comes only in the form of sign restrictions, the model is only set identified.

- This means that there is a set of values of the response of variable *i* to shock *j* after *s* periods that are all associated with the highest possible value for the likelihood function and that are consistent with *all* of the information available to the researcher: report upper and lower bound

Take-Away #3

- Every researcher who uses RWZ and reports a median as if it were a point estimate or 68% or 95% confidence bands is implicitly relying on some prior information *in addition to* sign restrictions.
- Rubio-Ramirez and co-authors argue that using this prior information is justified because the prior implies a uniform distribution with respect to a certain measure.
- BUT a frequentist who is unpersuaded by the validity of the implicit Bayesian prior would find the reported confidence bands to be much too narrow.
- Practitioners make no effort to persuade anyone that some elements of the MLE of the identified set are a priori unlikely.

Why is This Such a Big Deal?

- In set-identified structural VARs, the credibility of the prior is key because the influence of the prior does not vanish asymptotically (Baumeister and Hamilton, 2015; Giacomini and Kitagawa, 2021)

- For some parameters, data are completely uninformative, and you get back the prior.

  ⇨ Bayesian with uniform prior would claim to rule out some values for those parameters even though there is no basis in the data for doing so

  ⇨ Thus, prior is used in an informative way!

Quantitative Relevance of the Prior

- Inoue and Kilian (2024) argue that prior for $\mathbf{Q}$ is negligible in **tightly** identified VAR models given that they imply a narrow identified set.

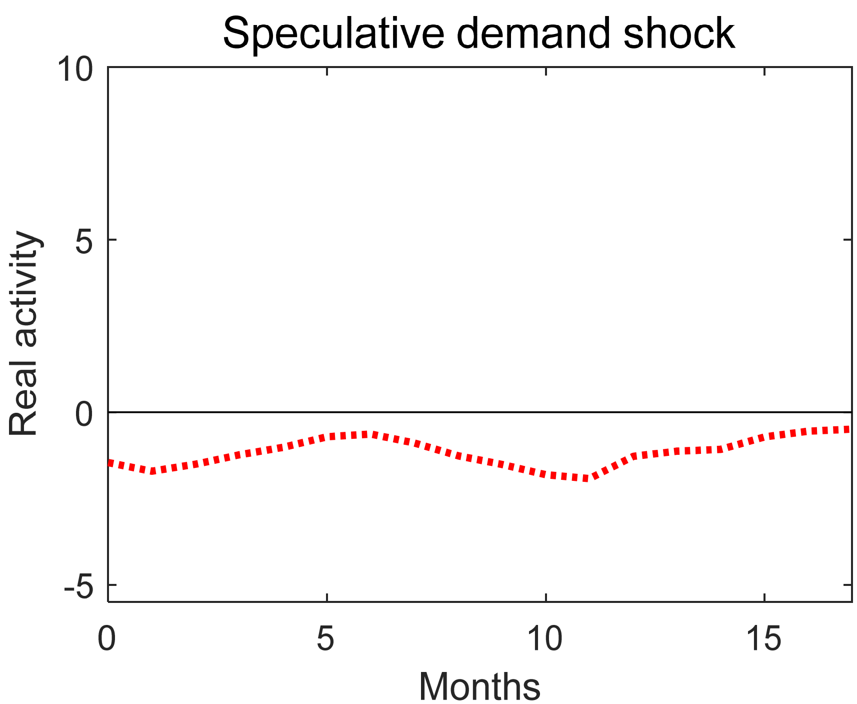
- As we add information in the form of more and tighter constraints, we will end up discarding most of the RWZ draws.

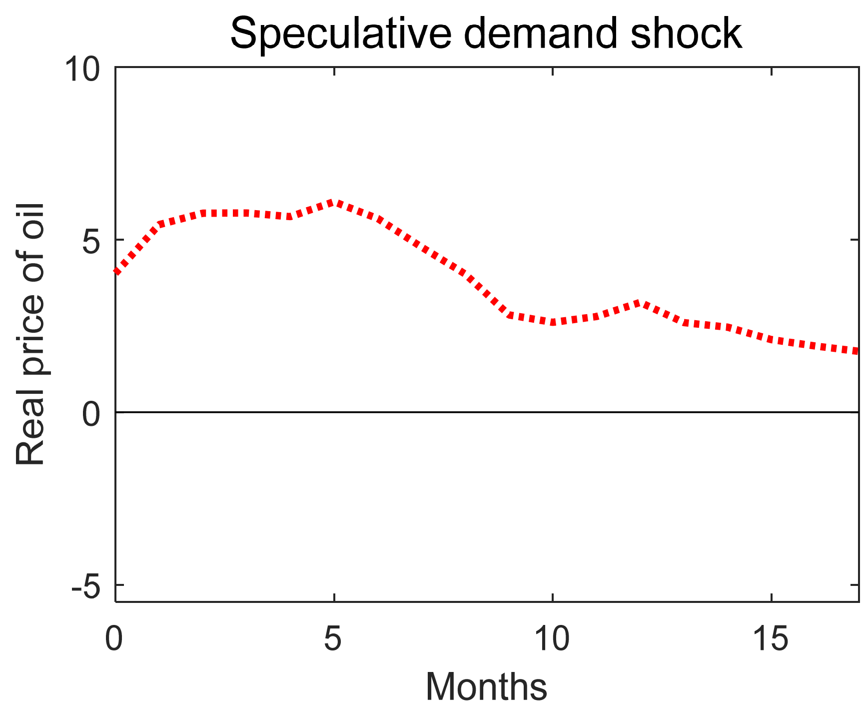
- BUT using only a handful of numbers may give a very **inaccurate** estimate of the identified set.

Illustration

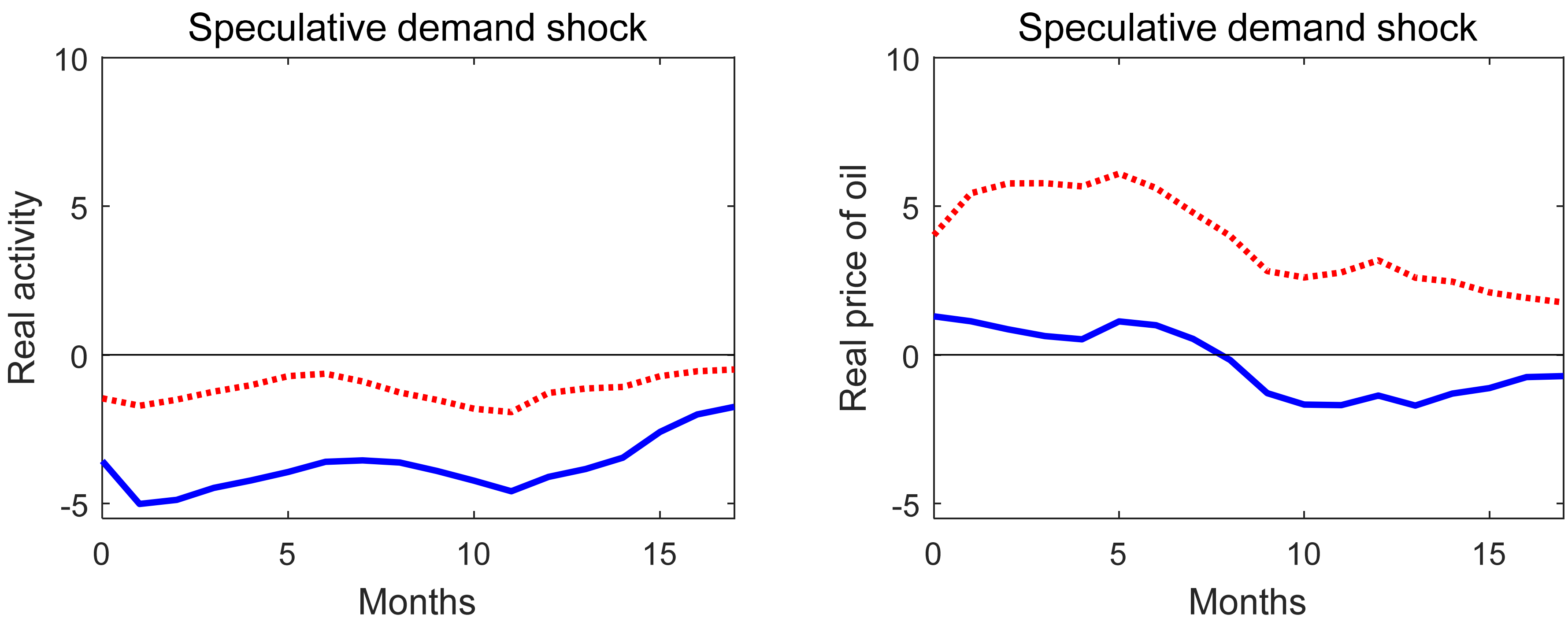
- Kilian and Murphy (2014) generated 5 million draws for the vector of possible parameters.
- Rule out various draws based on a long list of criteria (e.g., supply elasticity < 0.0258)
- The end result of running the code is that only 16 of the original 5 million draws remain.
- Select the draw with “impact price elasticity of oil demand in use closest to the posterior median”

Effect of speculative demand shock on real activity and price as originally reported in KM14





Effect of speculative demand shock calculated using KM14 code with two different random number seeds



Dashed red: seed = 316 (used by KM14). Blue: seed = 613.

Take-Away #4

- Causal interpretation of correlations requires information about the economic structure.

- Instead of having an implicit prior built into some mechanical algorithm, make prior beliefs about the economic structure explicit and defend them.

- Use **all** the (exact and inexact) information that you may have.

Bayesian Inference in Set-Identified Structural VAR Models

Structural model:

$$\mathbf{A}\mathbf{y}_t = \mathbf{B}\mathbf{x}_{t-1} + \mathbf{u}_t$$

$$\mathbf{u}_t \sim N(\mathbf{0}, \mathbf{D})$$

$\mathbf{D}$ diagonal

Bayesian approach:

Summarize whatever information we have that helps identify $\mathbf{A}$ in the form of a density $p(\mathbf{A})$.

$p(\mathbf{A})$ is highest for values of $\mathbf{A}$ we think are most plausible.

$p(\mathbf{A}) = 0$ for values of $\mathbf{A}$ we rule out altogether.

$p(\mathbf{A})$ can also impose sign restrictions and zeros

Bayesian begins with prior beliefs

before seeing data:

$$p(\mathbf{A}, \mathbf{D}, \mathbf{B}) = p(\mathbf{A})p(\mathbf{D}|\mathbf{A})p(\mathbf{B}|\mathbf{D}, \mathbf{A})$$

$p(\mathbf{A})$ could be any density

Use natural conjugate priors for $\mathbf{D}$ and $\mathbf{B}$:

diagonals of $\mathbf{D}^{-1}|\mathbf{A}$ are independent gamma

rows of $\mathbf{B}|\mathbf{D}, \mathbf{A}$ are independent normal

Prior for $p(\mathbf{D}|\mathbf{A})$

$$d_{ii}^{-1}|\mathbf{A} \sim \Gamma(\kappa_i, \tau_i)$$

where

$$E(d_{ii}^{-1}|\mathbf{A}) = \kappa_i/\tau_i$$

$$Var(d_{ii}^{-1}|\mathbf{A}) = \kappa_i/\tau_i^2$$

uninformative priors: $\kappa_i, \tau_i \to 0$

Prior for $p(\mathbf{B}|\mathbf{D}, \mathbf{A})$

$$\mathbf{B} = \begin{bmatrix} \boldsymbol{\lambda} & \mathbf{B}_1 & \mathbf{B}_2 & \cdots & \mathbf{B}_m \end{bmatrix}$$

$$\mathbf{b}_i|\mathbf{A}, \mathbf{D} \sim N(\mathbf{m}_i, d_{ii}\mathbf{M}_i)$$

uninformative priors: $\mathbf{M}_i^{-1} \to \mathbf{0}$

Likelihood:

$$p(\mathbf{Y}_T|\mathbf{A},\mathbf{D},\mathbf{B}) = (2\pi)^{-Tn/2}|\det(\mathbf{A})|^T|\mathbf{D}|^{-T/2} \times$$
$$\exp\left[-(1/2)\sum_{t=1}^{T}(\mathbf{A}\mathbf{y}_t - \mathbf{B}\mathbf{x}_{t-1})'\mathbf{D}^{-1}(\mathbf{A}\mathbf{y}_t - \mathbf{B}\mathbf{x}_{t-1})\right]$$

prior:

$$p(\mathbf{A},\mathbf{D},\mathbf{B}) = p(\mathbf{A})p(\mathbf{D}|\mathbf{A})p(\mathbf{B}|\mathbf{A},\mathbf{D})$$

posterior:

$$p(\mathbf{A},\mathbf{D},\mathbf{B}|\mathbf{Y}_T) = \frac{p(\mathbf{Y}_T|\mathbf{A},\mathbf{D},\mathbf{B})p(\mathbf{A},\mathbf{D},\mathbf{B})}{\int p(\mathbf{Y}_T|\mathbf{A},\mathbf{D},\mathbf{B})p(\mathbf{A},\mathbf{D},\mathbf{B})d\mathbf{A}d\mathbf{D}d\mathbf{B}}$$

$$= p(\mathbf{A}|\mathbf{Y}_T)p(\mathbf{D}|\mathbf{A},\mathbf{Y}_T)p(\mathbf{B}|\mathbf{A},\mathbf{D},\mathbf{Y}_T)$$

prior:

$$\mathbf{b}_i | \mathbf{A}, \mathbf{D} \sim N(\mathbf{m}_i, d_{ii}\mathbf{M}_i)$$

posterior:

$$\mathbf{b}_i | \mathbf{A}, \mathbf{D}, \mathbf{Y}_T \sim N(\mathbf{m}_i^*, d_{ii}\mathbf{M}_i^*)$$

As $T \to \infty$, $\mathbf{m}_i^* \to$ true value and $\mathbf{M}_i^* \to \mathbf{0}$

prior:

$$d_{ii}^{-1}|\mathbf{A} \sim \Gamma(\kappa_i, \tau_i)$$

posterior:

$$d_{ii}^{-1}|\mathbf{A}, \mathbf{Y}_T \sim \Gamma(\kappa_i^*, \tau_i^*)$$

As $T \to \infty$, $d_{ii} \xrightarrow{p}$ true value

Posterior distribution for **A**

prior: $p(\mathbf{A})$

If $\mathbf{M}_i^{-1} = \mathbf{0}$, and $\tau_i = \kappa_i = 0$,

posterior: $p(\mathbf{A}|\mathbf{Y}_T) = \dfrac{k_T p(\mathbf{A})|\det(\mathbf{A}\hat{\mathbf{\Omega}}_T\mathbf{A}')|^{T/2}}{\left\{\det\left[\text{diag}(\mathbf{A}\hat{\mathbf{\Omega}}_T\mathbf{A}')\right]\right\}^{T/2}}$

$k_T =$ constant that makes this integrate to 1

Posterior distribution for **A**

$$p(\mathbf{A}|\mathbf{Y}_T) = \frac{k_T p(\mathbf{A}) |\det(\mathbf{A}\hat{\mathbf{\Omega}}_T\mathbf{A}')|^{T/2}}{\left\{\det\left[\text{diag}(\mathbf{A}\hat{\mathbf{\Omega}}_T\mathbf{A}')\right]\right\}^{T/2}}$$

If we evaluate the posterior at an $\mathbf{A}$ that diagonalizes $\hat{\mathbf{\Omega}}_T$, i.e. $\mathbf{A}\mathbf{\Omega}_T\mathbf{A}' = \text{diag}(\mathbf{A}\mathbf{\Omega}_T\mathbf{A}')$, then $p(\mathbf{A}|\mathbf{Y}_T) = k_T p(\mathbf{A})$

Posterior distribution for **A**

$$p(\mathbf{A}|\mathbf{Y}_T) = \frac{k_T p(\mathbf{A})|\det(\mathbf{A}\hat{\mathbf{\Omega}}_T\mathbf{A}')|^{T/2}}{\left\{\det\left[\text{diag}(\mathbf{A}\hat{\mathbf{\Omega}}_T\mathbf{A}')\right]\right\}^{T/2}}$$

If we evaluate the posterior at an $\mathbf{A}$ for which $\mathbf{A}\mathbf{\Omega}_T\mathbf{A}' \neq \text{diag}(\mathbf{A}\mathbf{\Omega}_T\mathbf{A}')$, from Hadamard's inequality it follows

$$\det\left[\text{diag}(\mathbf{A}\hat{\mathbf{\Omega}}_T\mathbf{A}')\right] > \det(\mathbf{A}\hat{\mathbf{\Omega}}_T\mathbf{A}')$$

and as the sample size grows

$$p(\mathbf{A}|\mathbf{Y}_T) \to 0$$

What Does this Mean?

- If there is more than one matrix $\mathbf{A}$ that diagonalizes $\hat{\boldsymbol{\Omega}}_T$, then the model is **under-identified**.
- If the model is under-identified, the influence of the **prior will not vanish** even if sample size T goes to ∞.
- Posterior is a **re-scaled** version of the prior:
$p(\mathbf{A}|\mathbf{Y}_T) = k_T p(\mathbf{A})$

Is This a Bad Thing?

- If enough restrictions are imposed so that the model is exactly identified, then there exists only one allowable value for **A** that diagonalizes **Ω** and the posterior converges to that value as $T \to \infty$.

  BUT need to be absolutely certain about identifying assumptions

- If you are *not* certain about some of the identifying assumptions, you want to be able to take that uncertainty into account. You do that in the form of $p(\mathbf{A})$.

  ⇨ Doubts about the identification itself will be reflected in the posterior distribution.

Application 4: Labor Market Dynamics

demand:

$$\Delta n_t = k^d + \beta^d \Delta w_t + b_{11}^d \Delta w_{t-1} + b_{12}^d \Delta n_{t-1} + b_{21}^d \Delta w_{t-2}$$
$$+ b_{22}^d \Delta n_{t-2} + \cdots + b_{m1}^d \Delta w_{t-m} + b_{m2}^d \Delta n_{t-m} + u_t^d$$

supply:

$$\Delta n_t = k^s + \alpha^s \Delta w_t + b_{11}^s \Delta w_{t-1} + b_{12}^s \Delta n_{t-1} + b_{21}^s \Delta w_{t-2}$$
$$+ b_{22}^s \Delta n_{t-2} + \cdots + b_{m1}^s \Delta w_{t-m} + b_{m2}^s \Delta n_{t-m} + u_t^s$$

Prior for the Elasticities

for $\mathbf{y}_t = (\Delta w_t, \Delta n_t)' \, : \, \mathbf{A} = \begin{bmatrix} -\beta^d & 1 \\ -\alpha^s & 1 \end{bmatrix}$

Any arbitrary prior density that best summarizes our prior beliefs about labor supply and demand elasticities: $p(\mathbf{A}) = p(\alpha^s)p(\beta^d)$

What do we know about the short-run wage elasticity of labor demand?

- Hamermesh (1996) surveys microeconometric studies: 0.1 to 0.75
- Lichter et al. (2014) conduct meta-analysis of 942 estimates: lower end of Hamermesh range
- Theoretical macro models can imply value above 2.5 (Akerlof and Dickens, 2007; Galí, Smets and Wouters 2012)

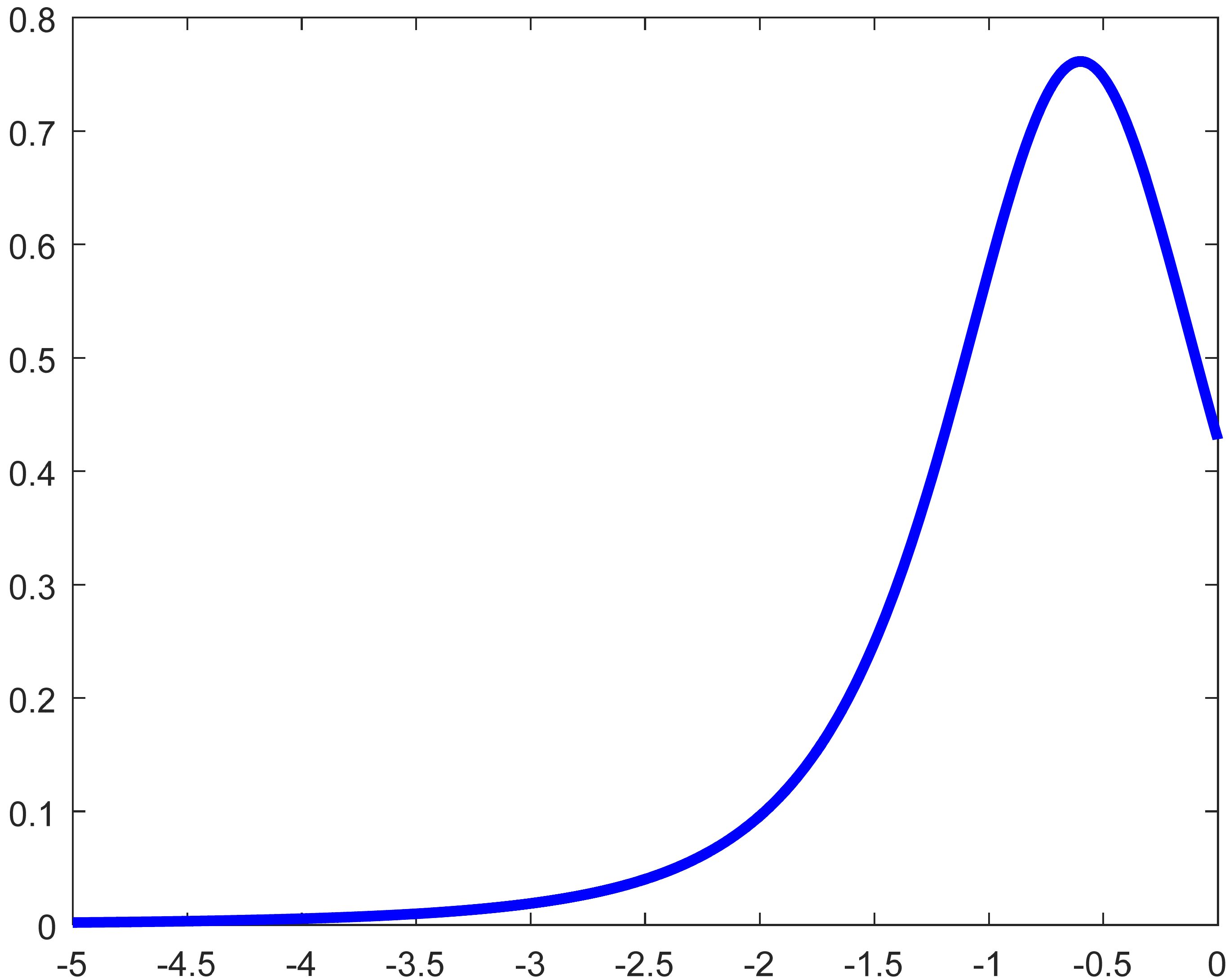
Prior for β^d: Student t with location c_β, scale σ_β, d.f. ν_β, truncated by $\beta^d \leq 0$

$c_\beta = -0.6, \sigma_\beta = 0.6, \nu_\beta = 3$

$\Rightarrow$ Prob$(\beta < -2.2) = 0.05$

Prob$(\beta > -0.1) = 0.05$

Student *t* prior for labor demand elasticity



What do we know about the wage elasticity of labor supply?

- Long run: often assumed to be zero because income and substitution effects cancel (e.g., Kydland and Prescott, 1982)
- Short run: often interpreted as Frisch elasticity
- Reichling and Whalen survey of microeconometric studies: 0.27-0.53
- Chetty et al. (2013) review 15 quasi-experimental studies: < 0.5
- Macro models often assume value greater than 2 (Kydland and Prescott, 1982, Cho and Cooley, 1994, Smets and Wouters, 2007)

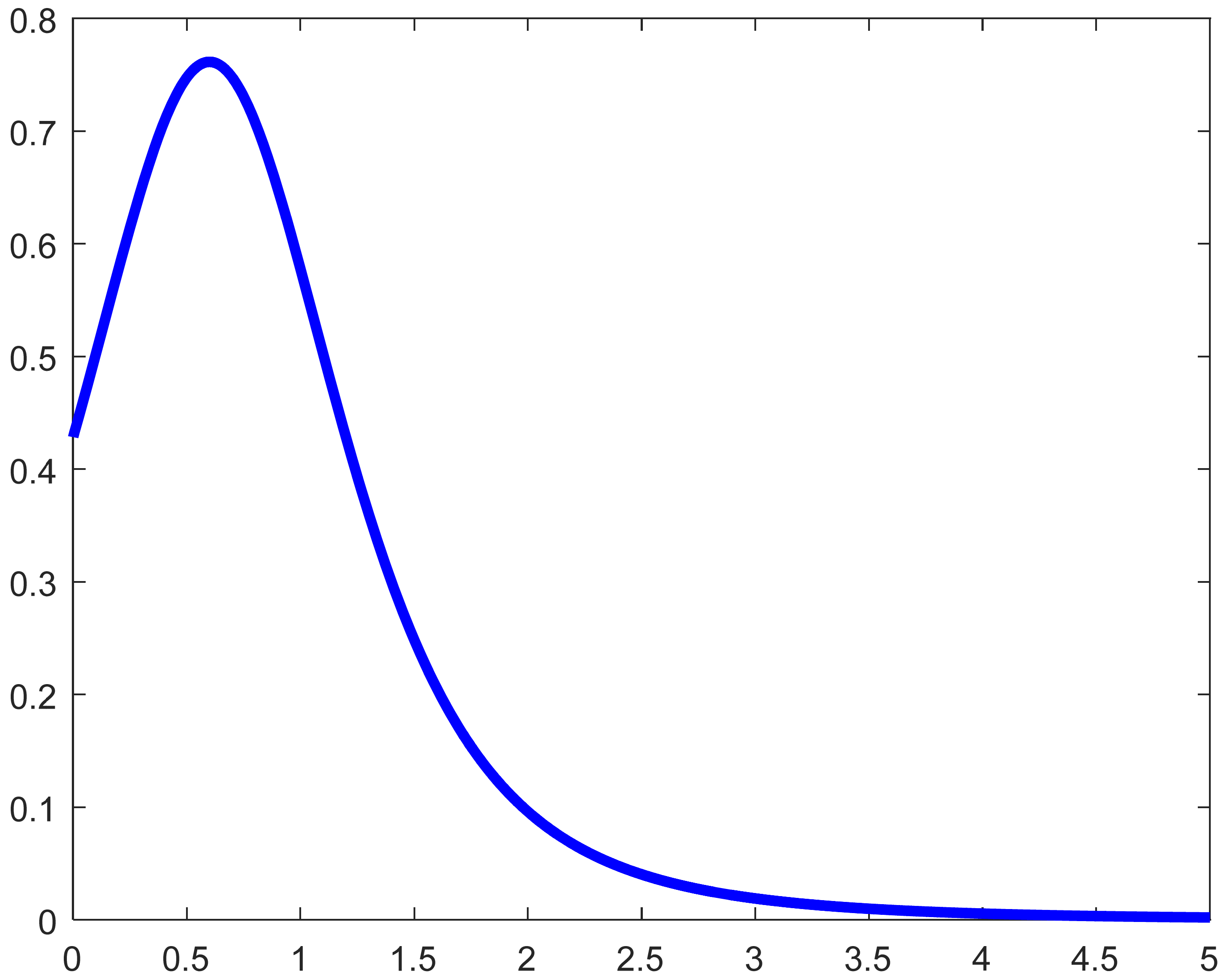
Prior for α^s: Student t with location c_α, scale σ_α, d.f. ν_a, truncated by $\alpha^s \geq 0$

$c_\alpha = 0.6, \sigma_\alpha = 0.6, \nu_\alpha = 3$

$\Rightarrow$ Prob$(\alpha < 0.1) = 0.05$

Prob$(\alpha > 2.2) = 0.05$

Student *t* prior for labor supply elasticity



Prior for the inverse of the structural variances

- Recall: $d_{ii}^{-1}|\mathbf{A} \sim \Gamma(\kappa_i, \tau_i)$
- Considerations:
  - Prior should in part reflect the *scale* of the data
  - Scales of individual innovations are obtained from residuals of univariate $AR(m)$ denoted $\hat{e}_{it}$
  - $\widehat{\boldsymbol{S}} = T^{-1} \sum_{t=1}^{T} \hat{\boldsymbol{e}}_t \hat{\boldsymbol{e}}_t'$ is the sample variance matrix of residuals with $\hat{\boldsymbol{e}}_t = (\hat{e}_{1t}, \hat{e}_{2t})'$
  - Prior mean $E\left(d_{ii}^{-1} \middle| \boldsymbol{A}\right) = \kappa_i / \tau_i$ is set equal to the reciprocal of the i^{th} diagonal element of $\boldsymbol{A}\widehat{\boldsymbol{S}}\boldsymbol{A}'$

    $\rightarrow \tau_i(\boldsymbol{A}) = \kappa_i \, a_i' \widehat{\boldsymbol{S}} a_i$
- We set $\kappa_i = 2$ which puts modest weight on our prior beliefs (equivalent to 4 observations of data).

Prior for the lagged structural coefficients

- Recall: $\mathbf{b}_i|\mathbf{A},\mathbf{D} \sim N(\mathbf{m}_i, d_{ii}\mathbf{M}_i)$
- Basic idea: Minnesota prior (Doan, Litterman, and Sims 1984, Sims and Zha 1998)
  - Prior mean $\mathbf{m}_i$
    - The most useful variable for predicting any variable is its own lagged value.
    - Prior belief that macro variables behave like random walks.
  - Prior variance $\mathbf{M}_i$
    - Coefficients on higher lags are more likely to be zero
      $\rightarrow$ implies: smaller values for the diagonal elements of $\mathbf{M}_i$ for higher lags
    - Prior variance is governed by a few hyperparameters λ_i

Prior mean $\mathbf{m}_i$

Reduced form might look like a random walk:

$$\boldsymbol{\Phi} = \mathbf{A}^{-1}\mathbf{B}$$

$$E(\boldsymbol{\Phi}) = \begin{bmatrix} \underset{(n\times n)}{\mathbf{I}_n} & \underset{(n\times(k-n))}{\mathbf{0}} \end{bmatrix} = \underset{(n\times k)}{\boldsymbol{\eta}}$$

$$\Rightarrow E(\mathbf{B}|\mathbf{A}) = \mathbf{A}\boldsymbol{\eta}$$

$$\mathbf{m}_i(\mathbf{A}) = E(\mathbf{b}_i|\mathbf{A}) = \boldsymbol{\eta}'\mathbf{a}_i$$

Prior variance $\mathbf{M}_i$

Variance reflects increasing confidence in prior expectation as lag order increases:

- λ_1 : confidence in higher-order lags = 0

$$\underset{(1\times m)}{\mathbf{v}_1'} = (1/(1^{2\lambda_1}), 1/(2^{2\lambda_1}), \ldots, 1/(m^{2\lambda_1}))$$

- s_{ii} are the diagonal elements of $\hat{\boldsymbol{S}}$

$$\underset{(1\times n)}{\mathbf{v}_2'} = (s_{11}^{-1}, s_{22}^{-1}, \ldots, s_{nn}^{-1})'$$

- λ_0 : overall confidence in prior (λ_3 for constant term)

$$\mathbf{v}_3 = \lambda_0^2 \begin{bmatrix} \mathbf{v}_1 \otimes \mathbf{v}_2 \\ \lambda_3^2 \end{bmatrix}$$

Prior variance $\mathbf{M}_i$

$$\mathbf{v}_3 = \lambda_0^2 \begin{bmatrix} \mathbf{v}_1 \otimes \mathbf{v}_2 \\ \lambda_3^2 \end{bmatrix}$$

$\mathbf{M}_i$ is a diagonal matrix whose (r, r) element is the r^{th} element of $\mathbf{v}_3$:

$$M_{i,rr} = v_{3r}$$

Hyperparameters for labor market example:

$\lambda_0 = 0.2,\ \lambda_1 = 1,\ \lambda_3 = 100$

Posterior Distribution for $\mathbf{B}$

Posterior for $\mathbf{B}$ given $\mathbf{A}, \mathbf{D}$:

$$\mathbf{b}_i | \mathbf{Y}, \mathbf{A}, \mathbf{D} \sim N(\mathbf{m}_i^*(\mathbf{A}), d_{ii}\mathbf{M}_i^*)$$

$$\mathbf{m}_i^*(\mathbf{A}) = (s_i^{\mathbf{xx}})^{-1} s_i^{\mathbf{yx}}(\mathbf{A})'$$

$$s_i^{\mathbf{xx}} = \sum_{t=1}^{T} \mathbf{x}_{t-1}\mathbf{x}_{t-1}' + \mathbf{M}_i^{-1}$$

$$s_i^{\mathbf{yx}}(\mathbf{A}) = \mathbf{a}_i' \sum_{t=1}^{T} \mathbf{y}_t \mathbf{x}_{t-1}' + \mathbf{m}_i(\mathbf{A})'\mathbf{M}_i^{-1}$$

$$\mathbf{M}_i^*(\mathbf{A}) = (s_i^{\mathbf{xx}})^{-1}$$

Could implement with augmented regression

$$\mathbf{m}_i^*(\mathbf{A}) = (\tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i)^{-1}\tilde{\mathbf{X}}_i'\tilde{\mathbf{y}}_i$$

$\tilde{\mathbf{X}}_i$ is $(T+k)\times k$

first T rows contain $\mathbf{x}_{t-1}'$

last k rows are $\mathbf{P}_i$ (the Cholesky factor of $\mathbf{M}_i^{-1}$)

$$s_i^{\mathbf{xx}} = \tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i = \sum_{t=1}^{T} \mathbf{x}_{t-1}\mathbf{x}_{t-1}' + \mathbf{M}_i^{-1}$$

$$\underset{[k\times(T+k)]}{\tilde{\mathbf{X}}_i'} = \begin{bmatrix} \mathbf{x}_0 & \cdots & \mathbf{x}_{T-1} & \mathbf{P}_i \end{bmatrix}$$

$$s_i^{\mathbf{yx}}(\mathbf{A}) = \mathbf{a}_i' \sum_{t=1}^{T} \mathbf{y}_t \mathbf{x}_{t-1}' + \mathbf{m}_i(\mathbf{A})' \mathbf{M}_i^{-1}$$

$\tilde{\mathbf{y}}_i$ is $(T+k) \times 1$

first T elements are $\mathbf{a}_i'\mathbf{y}_t$

last k elements are $\mathbf{m}_i'(\mathbf{A})\mathbf{P}_i$

$$s_i^{\mathbf{yx}}(\mathbf{A})' = \tilde{\mathbf{X}}_i'\tilde{\mathbf{y}}_i$$

$$\underset{[1\times(T+k)]}{\tilde{\mathbf{y}}_i} = (\mathbf{a}_i'\mathbf{y}_1, \ldots, \mathbf{a}_i'\mathbf{y}_T, \mathbf{m}_i'\mathbf{P}_i) \text{ where } \mathbf{P}_i\mathbf{P}_i' = \mathbf{M}_i^{-1}$$

Posterior Distribution for $\mathbf{D}$

$$d_{ii}^{-1} | \mathbf{A}, \mathbf{Y} \sim \Gamma(\kappa_i^*, \tau_i^*(\mathbf{A}))$$

$$\kappa_i^* = \kappa_i + T/2$$

$$\tau_i^*(\mathbf{A}) = \tau_i(\mathbf{A}) + (1/2)\zeta_i^*(\mathbf{A})$$

$$\zeta_i^*(\mathbf{A}) = s_i^{\mathbf{yy}}(\mathbf{A}) - s_i^{\mathbf{yx}}(\mathbf{A})(s_i^{\mathbf{xx}})^{-1} s_i^{\mathbf{yx}}(\mathbf{A})'$$

$$s_i^{\mathbf{yy}}(\mathbf{A}) = \mathbf{a}_i' \sum_{t=1}^{T} \mathbf{y}_t \mathbf{y}_t' \mathbf{a}_i + \mathbf{m}_i(\mathbf{A})' \mathbf{M}_i^{-1} \mathbf{m}_i(\mathbf{A})$$

$$= \tilde{\mathbf{y}}_i' \tilde{\mathbf{y}}_i$$

Posterior Distribution for $\mathbf{A}$

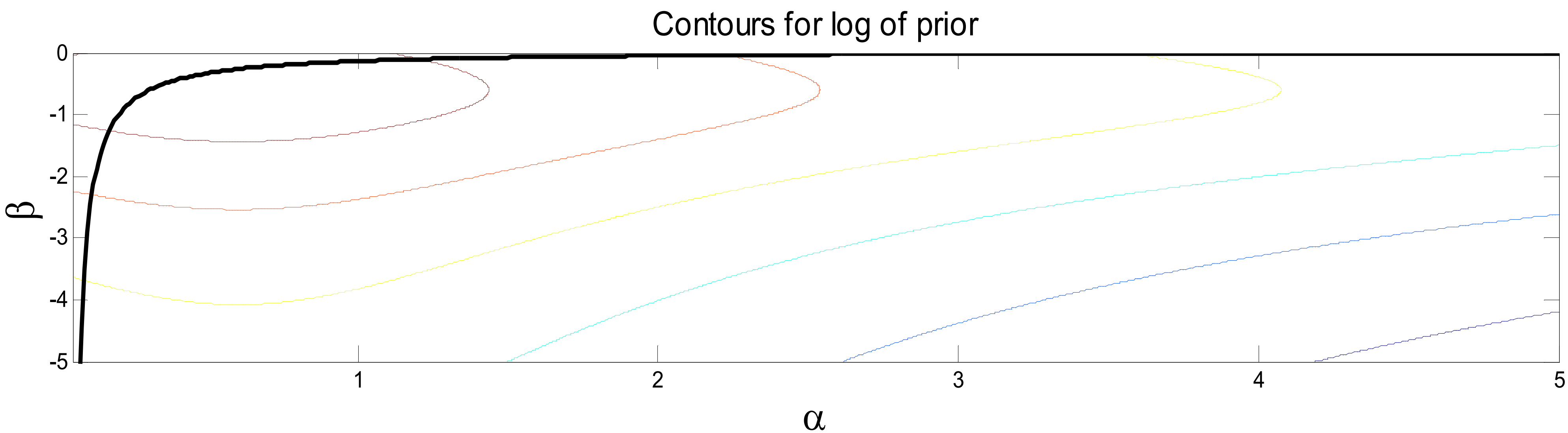
$$p(\mathbf{A}|\mathbf{Y}_T) = \frac{k_T p(\mathbf{A})[\det(\mathbf{A}\hat{\mathbf{\Omega}}_T\mathbf{A}')]^{T/2}}{\prod_{i=1}^n [(2/T)\tau_i^*(\mathbf{A})]^{\kappa_i^*}} \prod_{i=1}^n \tau_i(\mathbf{A})^{\kappa_i}$$

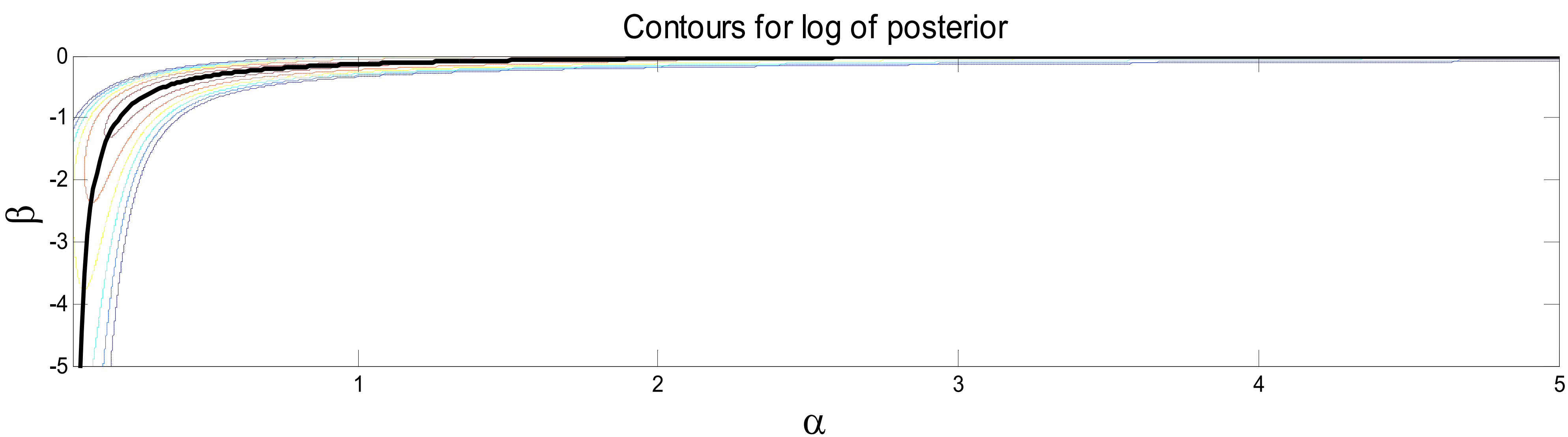
where $\hat{\mathbf{\Omega}}_T$ is the sample variance matrix for the reduced-form VAR residuals:

$$\hat{\mathbf{\Omega}}_T = T^{-1}\left\{\sum_{t=1}^T \mathbf{y}_t\mathbf{y}_t' - \left(\sum_{t=1}^T \mathbf{y}_t\mathbf{x}_{t-1}'\right)\left(\sum_{t=1}^T \mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right)^{-1}\left(\sum_{t=1}^T \mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right)\right\}$$

Set Identification with Informative Priors

- Prior induces ranking in posterior





Baumeister-Hamilton Algorithm

- Goal:

  Generate draws from the joint posterior distribution

$$p(\mathbf{A}, \mathbf{D}, \mathbf{B} | \mathbf{Y}_T)$$

- Procedure:

  Draw $\mathbf{A}^{(\ell)}$ from $p(\mathbf{A} | \mathbf{Y}_T)$

  Draw $\mathbf{D}^{(\ell)}$ from $p(\mathbf{D} | \mathbf{A}, \mathbf{Y}_T)$

  Draw $\mathbf{B}^{(\ell)}$ from $p(\mathbf{B} | \mathbf{A}, \mathbf{D}, \mathbf{Y}_T)$

  Repeat for $\ell = 1, \ldots, 10^6$

- $\{\mathbf{A}^{(\ell)}, \mathbf{D}^{(\ell)}, \mathbf{B}^{(\ell)}\}_{\ell=1}^{N}$ is a draw from the joint posterior

How to Generate Draws for $\mathbf{A}$?

- Problem:

  Posterior distribution for **A** is of unknown form

  $\Longrightarrow$ cannot directly sample from it

- Solution:

  Use random-walk Metropolis-Hastings algorithm to approximate the posterior distribution $p(\mathbf{A}|\mathbf{Y}_T)$

Metropolis-Hastings Algorithm

- Goal:

  Draw samples from a distribution with unusual form (referred to as the target density)

  $$\pi(\mathbf{\Phi})$$

  where $\mathbf{\Phi}$ is a $(K \times 1)$ vector of parameters

- How can we do that?
  - Specify a candidate-generating (proposal) density $q(\mathbf{\Phi}^{G+1}|\mathbf{\Phi}^{G})$ or $q(\mathbf{\Phi}^{G+1})$ from which we can generate draws easily.
  - Evaluate $\dfrac{\pi(\mathbf{\Phi}^{G+1})/q(\mathbf{\Phi}^{G+1})}{\pi(\mathbf{\Phi}^{G})/q(\mathbf{\Phi}^{G})}$ to decide whether to keep the candidate draw or to discard it.

Random-Walk Metropolis-Hastings Algorithm

- As the name suggests, the proposal density is a random walk:

$$\mathbf{\Phi}^{G+1} = \mathbf{\Phi}^{G} + e$$

where $e \sim i.i.d. N(0, \mathbf{\Sigma})$ is a $(K \times 1)$ vector

- Given that $e = \mathbf{\Phi}^{G+1} - \mathbf{\Phi}^{G}$ is normally distributed, the density $p(\mathbf{\Phi}^{G+1} - \mathbf{\Phi}^{G}) = p(\mathbf{\Phi}^{G} - \mathbf{\Phi}^{G+1})$ due to symmetry.

- Thus, $q(\mathbf{\Phi}^{G+1} | \mathbf{\Phi}^{G}) = q(\mathbf{\Phi}^{G} | \mathbf{\Phi}^{G+1})$ and the acceptance probability simplifies to: $\alpha = \frac{\pi(\mathbf{\Phi}^{G+1})}{\pi(\mathbf{\Phi}^{G})}$

Random-Walk Metropolis-Hastings Algorithm

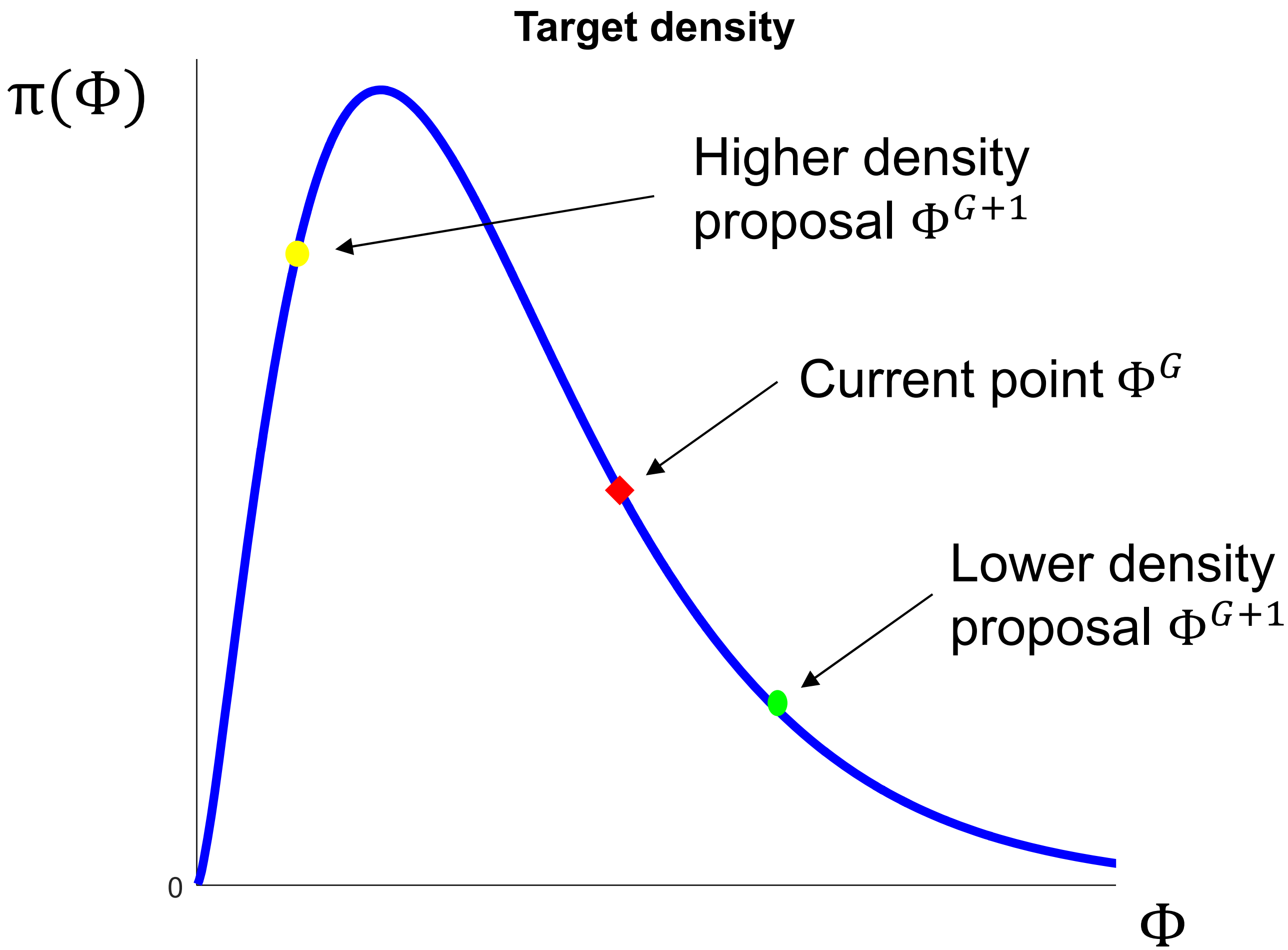
$$\alpha = \frac{\pi(\mathbf{\Phi}^{G+1})}{\pi(\mathbf{\Phi}^{G})}$$

- if $\pi(\mathbf{\Phi}^{G+1}) > \pi(\mathbf{\Phi}^{G})$, the chain moves to $\mathbf{\Phi}^{G+1}$
- otherwise, it moves with probability α

In other words:

- If the jump goes “uphill”, it is always accepted.
- If the jump goes “downhill”, it is accepted with a nonzero probability.

Graphical Illustration



RW-MH Algorithm Step by Step

Step 1 Specify a starting value for the parameters $\mathbf{\Phi}$ denoted $\mathbf{\Phi}^0$ and set the variance Σ of the shocks to the random walk.

Step 2 Draw a new value for the parameters $\mathbf{\Phi}^{new}$ using

$$\mathbf{\Phi}^{new} = \mathbf{\Phi}^{old} + e$$

where for the first draw $\mathbf{\Phi}^{old} = \mathbf{\Phi}^0$.

Step 3 Compute the acceptance probability

$$\alpha = \min\left(\frac{\pi(\mathbf{\Phi}^{new})}{\pi(\mathbf{\Phi}^{old})}, 1\right)$$

If $\alpha > u \sim U(0,1)$, retain $\mathbf{\Phi}^{new}$; otherwise, retain $\mathbf{\Phi}^{old}$.

Step 4 Repeat steps (2) and (3) M times and use the last L draws for inference.

Generating draws for **A**

- Generate a candidate $\tilde{\boldsymbol{\alpha}}^{(\ell+1)} = \boldsymbol{\alpha}^{(\ell)} + \xi(\hat{\mathbf{Q}}_{\Lambda}^{-1})'\mathbf{v}_{\ell+1}$

  where

  $\mathbf{v}_{\ell+1}$ is a (2×1) vector of independent standard Student t variables with 2 degrees of freedom,

  $\hat{\mathbf{Q}}$ is the Cholesky factor of $\hat{\boldsymbol{\Lambda}}$ (see below), and

  ξ is a scalar tuning parameter to get a 30% acceptance rate

- Check whether the sign restrictions for the elements in $\tilde{\boldsymbol{\alpha}}$ are satisfied:

  $\beta^d < 0$ and $\alpha^s > 0$

Generating draws for **A**

- For those $\tilde{\alpha}$ that have the correct signs, calculate the log of the target function:

$$q(\mathbf{A}) = \log(p(\mathbf{A})) + (T/2)\log\left[\det\left(\mathbf{A}\hat{\mathbf{\Omega}}_T\mathbf{A}'\right)\right] - \sum_{i=1}^{n} \kappa_i^* \log[(2/T)\tau_i^*(\mathbf{A})] + \sum_{i=1}^{n} \kappa_i \tau_i(\mathbf{A})$$

where

$$\mathbf{A} = \begin{bmatrix} -\beta^d & 1 \\ -\alpha^s & 1 \end{bmatrix}$$

Generating draws for **A**

- To enhance the efficiency of the algorithm, find the value $\hat{\boldsymbol{\alpha}}$ that maximizes the target function numerically
  - ➢ Use $\hat{\boldsymbol{\alpha}}$ as a starting value for RW-MH step
  - ➢ Use matrix of second derivatives (Hessian) as the variance matrix

$$\hat{\boldsymbol{\Lambda}} = -\frac{\partial^2 q(\mathbf{A}(\boldsymbol{\alpha}))}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} \Bigg|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}$$

Summary of BH Algorithm

(1) Use random-walk Metropolis-Hastings to generate a draw from $p(\mathbf{A}|\mathbf{Y})$.

(2) Given this value for $\mathbf{A}$, generate diagonal elements of **D** from inverse-gamma $(\kappa_i^*, \tau_i^*(\mathbf{A}))$

(3) Given this value for **D**, generate ith row of $\mathbf{B}$ from $N(\mathbf{m}_i^*(\mathbf{A}), d_{ii}\mathbf{M}_i^*)$ distribution

(4) The generated triple $(\mathbf{A}, \mathbf{D}, \mathbf{B})$ is a draw from the posterior distribution $p(\mathbf{A}, \mathbf{D}, \mathbf{B}|\mathbf{Y})$ which allows us to calculate the posterior distribution of any object of interest.

Reduced-form VAR coefficients:

$$\mathbf{\Pi} = \mathbf{A}^{-1}\mathbf{B}$$

Reduced-form IRF:

$\mathbf{\Psi}_h$ is first $(n \times n)$ block of $\mathbf{F}^h$

Structural IRF:

$$\frac{\partial \mathbf{y}_{t+h}}{\partial \mathbf{u}_t'} = \mathbf{\Psi}_h \mathbf{A}^{-1}$$