Topics in Bayesian Vector Autoregressions

Giovanni Ricco

10th November 2023

Hyperpriors

Hierarchical Modelling

Giannone, Lenza and Primiceri (2014)

Hyperparameters $\{\lambda_i\}$ are additional unknown coefficients

- Model parameters: $\theta \equiv A, \Sigma$
- Hyperparameters: $\gamma \equiv \{\lambda_1, \lambda_2, \ldots, \lambda_n, \ldots\}$

1. Specify prior distribution for γ

$$\textit{Hyperprior} : p(\gamma)$$

2. Compute:

$$p(\gamma|y) \propto \underbrace{\int p(y|\theta,\gamma)p(\theta|\gamma)d\theta}_{p(y|\gamma)} \times p(\gamma)$$

and $\quad \gamma^* = \textit{argmax}\, p(\gamma|y)$

Hierarchical Modelling

Giannone, Lenza and Primiceri (2014)

$$p(y|\gamma) \propto \underbrace{\left| \left(V_\varepsilon^{\text{posterior}} \right)^{-1} V_\varepsilon^{prior} \right|^{\frac{T-p+d}{2}}}_{\textit{Fit}} \underbrace{\prod_{t=p+1}^{T} \left| V_{t+1|t} \right|^{-\frac{1}{2}}}_{\textit{Penalty}}$$

- $V_\varepsilon^{\text{posterior}}$ and V_ε^{prior} are the posterior and prior mean of Σ
- $V_{t+1|t}$ is the variance (conditional on Σ) of the 1-step-ahead forecast of y_t, averaged across all possible a priori realisations of Σ

$$V_{t+1|t} \equiv \mathbb{E}_\Sigma \left[\mathbb{V}ar(y_t|y^{t-1}, \Sigma) \right]$$

Hierarchical Modelling

Giannone, Lenza and Primiceri (2014)

Intuition for the mechanism:

- **First term** increases when $V_{\varepsilon}^{posterior}$ falls relative to $V_{\varepsilon}^{\text{prior}}$
  $\Longrightarrow$ ML favours hyperparameter values that generate smaller residuals

- **Second term** increases with the a priori residual variances and the uncertainty of the parameter estimates
  $\Longrightarrow$ ML penalises models potentially overfitting data

- Standard **trade-off between model fit and complexity**

Hyperpriors

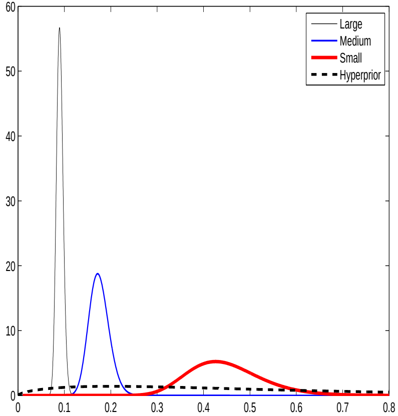

Figure: Posterior distribution of the hyperparameter λ, the parameter governing the standard deviation of the Minnesota prior in a small, medium, and large BVARs and its prior distribution (Giannone et al, 2015)

Hyperpriors

TABLE 2.—MSFE OF POINT FORECASTS

		Small (S)		Medium (M)		Large (L)			
Horizons	Variables	VAR	BVAR	VAR	BVAR	VAR	BVAR	Factor M	RW
One quarter	Real GDP	13.49	9.61	19.15	7.97		8.18	7.29	10.23
	GDP deflator	1.53	1.32	2.26	1.35		1.10	1.14	5.19
	Federal funds rates	1.61	1.04	1.82	1.03		1.00	1.25	1.06
One year	Real GDP	5.40	3.85	12.10	3.42		3.97	3.52	3.98
	GDP deflator	1.61	1.45	2.25	1.58		0.96	1.01	4.65
	Federal funds rates	0.58	0.32	0.56	0.31		0.36	0.32	0.31

The table reports the mean squared forecast errors of the BVARs and the competing models (VAR: flat-prior VAR, RW: random walk in levels with drift: factor M: factor augmented regression), for each variable and horizon. The evaluation sample is 1975Q1–2008Q4 for the one-quarter-ahead forecasts and 1975Q4–2008Q4 for the one-year-ahead forecasts.

(Priors for) VARs with Trending Variables

Trends in Variables

- Applied statisticians and macroeconomists often treat low frequency (trends) and high frequency (seasonal) variation as a nuisance
- **Usual approach:** get rid of it in a way that leaves inference about the other frequencies minimally affected
  - Linear or log-linear **deterministic trend**
  - First or second **differences**
  - **Hodrick-Prescott filter**...

Trends in Variables

Frequency domain

- With seasonality there often is a clean separation of seasonal and non-seasonal variation
- Separating the trend from the business cycle variation is much less clear
- Granger's 'typical spectral shape'

Trends

Remark:

- **Sample information** about variation at **frequencies with wavelength** $\sim T$ in a sample of size T is inherently **weak**
- Only one observation of a cycle of wavelength T!
- (Explicit or implicit) **prior beliefs dominate** sample information

VARs with Trending Variables

- Let us consider an AR(1) model

$$y_t = \mu + \phi y_{t-1} + u_t \qquad u_t \sim \mathcal{N}(0, \sigma^2)$$

- For $|\phi| < 1$, the model is stationary and stable.
- The unconditional distribution for y_t is

$$y_t \sim \mathcal{N}\left(\frac{\mu}{1-\phi}, \frac{\sigma^2}{1-\phi^2}\right)$$

VARs with Trending Variables

- Iterating back to time 0

$$
\begin{aligned}
y_t &= \underbrace{\phi^t y_0 + \sum_{j=0}^{t-1} \phi^j \mu}_{\textit{Deterministic Component}} + \underbrace{\sum_{j=0}^{t-1} \phi^j u_{t-j}}_{\textit{Stochastic Comp.}} \quad (1) \\
&= \underbrace{\left(y_0 - \frac{\mu}{1-\phi} \right) \phi^t}_{\textit{Det. Return to Trend}} + \underbrace{\frac{\mu}{1-\phi}}_{\textit{S.S.}} + \underbrace{\sum_{j=0}^{t-1} \phi^j u_{t-j}}_{\textit{Stochastic}} \quad (2)
\end{aligned}
$$

- In principle the unconditional mean can be far away from the initial observations

VARs with Trending Variables

Observation: Unit roots convert constants into polynomial trends!

$$\frac{\mu}{1-\phi} \xrightarrow{\phi \to 1} \pm\infty$$

$$DC_t = \begin{cases} y_0 + t\mu & \textit{if } \phi = 1 \\ \frac{\mu}{1-\phi} + \left(y_0 - \frac{\mu}{1-\phi}\right)\phi^t & \textit{if } \phi \neq 1 \end{cases}$$

Intuition: (V)AR model conditional on initial observations y_0 (e.g. OLS or conditional ML)

- the estimator will try to 'fit' the low frequency components of the data by using the deterministic component
- 'reversion to the mean' from the initial conditions!

VARs with Trending Variables

- VARs estimated conditional on initial observations – OLS or conditional ML – tend to imply that initial conditions are implausibly accurate predictor of the trend or long-run swings in the sample

- The criterion of fit applies no penalty to parameter values that make the initial conditions highly implausible as draws from the model's implied unconditional distribution for y_t

- The model attributes the low-frequency behaviour of the data to a process of return to the steady state from these 'unlikely initial conditions'

VARs with Trending Variables

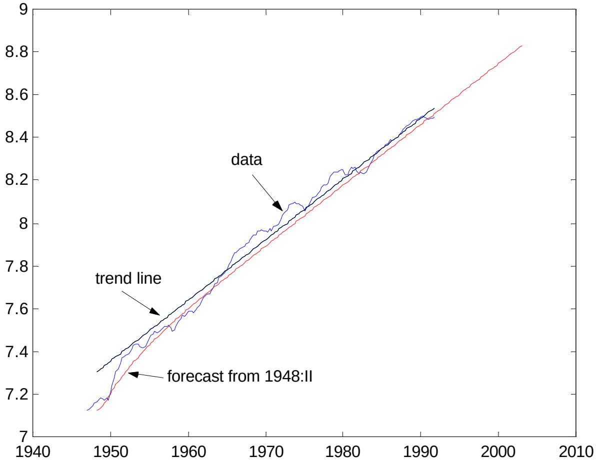

Figure: Log GDP: actual, estimated linear trend, deterministic forecast (Sims 1996)

VARs with Trending Variables

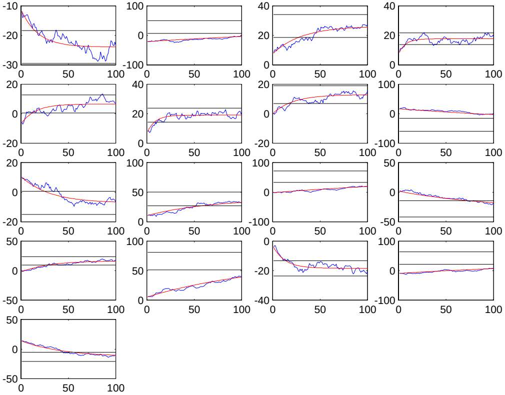

Figure: Sims (1998)'s **Initial Conditions Rogues Gallery** – Rougher lines are RW Monte Carlo data. Smoother curved lines are deterministic components. Horizontal lines are 95% probability bands around the unconditional mean.

VARs with Trending Variables

Sims (1996)

- In a univariate, one-lag model, return-to-trend dynamics can only take the exponential form
$$(y_0 - Ey)\phi^t$$
- With k lags, a univariate model can produce return-to-trend dynamics that are linear combinations of k exponentials
- If all the observations behave like a k-th order polynomial, the AR can predict them perfectly
- A VAR with k lags on n variables has kn roots and can fit perfectly an arbitrary collection of kn-th order polynomials

Sums-of-coefficients Priors

Litterman (1986), Sims and Zha (1998)

- The Sums-of-coefficients priors can be implemented using the dummy observations

$$y_d = \mathrm{diag}(\delta_1\mu_1, \ldots, \delta_n\mu_n)/\lambda_4$$

$$x_d = ((1_{1\times p}) \otimes \mathrm{diag}(\delta_1\mu_1, \ldots, \delta_n\mu_n)/\lambda_4 \quad 0_{n\times 1})$$

- Expresses a belief that when the average of lagged values of a variable is at some level μ_i, that value is likely to be a good forecast of $y_{i,t}$
- ... and that knowing the average of lagged values of variable j does not help in predicting a variable $i \neq j$
- Introduce correlation among coefficients on a given variable in a given equation

Sums-of-coefficients Priors

Litterman (1986), Sims and Zha (1998)

Example (n=2, p=2):

$$\begin{pmatrix} \frac{\delta_1\mu_1}{\lambda_4} & 0 \\ 0 & \frac{\delta_2\mu_2}{\lambda_4} \end{pmatrix} = \begin{pmatrix} \frac{\delta_1\mu_1}{\lambda_4} & 0 & \frac{\delta_1\mu_1}{\lambda_4} & 0 & 0 \\ 0 & \frac{\delta_2\mu_2}{\lambda_4} & 0 & \frac{\delta_2\mu_2}{\lambda_4} & 0 \end{pmatrix} A + U^d$$

- When $\lambda_4 \to 0$ the model can be expressed in terms of differenced data, with as many unit roots as variables. For each variable i we have

$$(1 - A_{1,ii} - A_{2,ii} - \cdots - A_{p,ii}) \frac{\mu_i}{\lambda_4} = u_t$$

or

$$(1 - A_{ii}(1)) \frac{\mu_i}{\lambda_4} = u_t$$

Co-persistence Prior

Sims (1993), Sims and Zha (1998)

- The Co-persistence prior (or dummy initial observation prior) can be implemented using the dummy observations

$$y_d = [\delta_1\mu_1/\lambda_5, \ldots, \delta_n\mu_n/\lambda_5]$$

$$x_d = [\delta_1\mu_1/\lambda_5, \ldots, \delta_n\mu_n/\lambda_5 \;\; 1/\lambda_5]$$

Co-persistence Prior

Sims (1993), Sims and Zha (1998)

- Write the VAR using the lag operator

$$(I - A(L))y_t = C + u_t$$

then this prior can be written as

$$\left((I - A(1))\,\underline{\mu} - C\right)\frac{1}{\lambda_5} = u_t$$

- There can be a single common unit root in the system if C is small, otherwise the system is stationary and stable around μ
- What is this prior for? Avoiding overfitting using deterministic trends/components
- SoC and Co-persistence taken together, favour unit roots and cointegration

Co-persistence Prior

Sims (1993), Sims and Zha (1998)

- **Intuition – How to fix the issue?** Do not condition on initial observation $\Longrightarrow$ unconditional ML: add $p(y_0|\theta)$ to the likelihood
- Hence for an AR(1)

$$y_0 \sim \mathcal{N}\left(\frac{\mu}{1-\phi}, \frac{\sigma^2}{1-\phi}\right)$$

- This implies

$$y_0 \sim \frac{\mu}{1-\phi} \qquad \Longrightarrow \qquad (1-\phi)y_0 - \mu \sim 0$$

- This is what is enforced with $\lambda_5 \to 0$

$$(1 - A(1))\underline{\mu} - C = 0$$

Priors for the Long Run

Giannone, Lenza, Primiceri (2019)

VAR(1)

$$y_t = C + Ay_{t-1} + \varepsilon_t$$

Rewrite the VAR in terms of levels and differences:

$$\Delta y_t = C + \Pi y_{t-1} + \varepsilon_t \qquad \text{where} \qquad \Pi = A - \mathbb{I}_n$$

- Usual prior for the long run $\Longrightarrow$ prior on Π centred at 0
- Standard approach: push coefficients towards all variables being independent random walks

Priors for the Long Run

Giannone, Lenza, Primiceri (2019)

$$\Delta y_t = C + \Pi y_{t-1} + \varepsilon_t \qquad \text{where} \qquad \Pi = A - \mathbb{I}_n$$

Rewrite as

$$\Delta y_t = C + \underbrace{\Pi H^{-1}}_{\Lambda} \underbrace{H y_{t-1}}_{\tilde{y}_{t-1}} + \varepsilon_t$$

- Choose H and put prior on Λ conditional on H
- Economic theory suggests that some linear combinations of y are less (more) likely to exhibit long-run trends
- Loadings associated with these combinations are less (more) likely to be 0

Priors for the Long Run

Example: 3-variable VAR

Let's consider

$$\Delta y_t = C + \underbrace{\Pi H^{-1}}_{\Lambda} \underbrace{\begin{bmatrix} 1 & 1 & 1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ c_{t-1} \\ i_{t-1} \end{bmatrix}}_{Hy_{t-1} = \tilde{y}_{t-1}} + \varepsilon_t$$

hence

$$\begin{bmatrix} \Delta x_t \\ \Delta c_t \\ \Delta i_t \end{bmatrix} = C + \begin{bmatrix} \Lambda_{11} & \Lambda_{12} & \Lambda_{13} \\ \Lambda_{21} & \Lambda_{22} & \Lambda_{23} \\ \Lambda_{31} & \Lambda_{32} & \Lambda_{33} \end{bmatrix} \begin{bmatrix} x_{t-1} + c_{t-1} + i_{t-1} \\ c_{t-1} - x_{t-1} \\ i_{t-1} - c_{t-1} \end{bmatrix} + \varepsilon_t$$

- Red: Possibly stationary linear combinations
- Blue: Common trend

Priors for the Long Run

- If the i-th row of H contains the coefficients of a linear combination of y_t that is a priori nonstationary
  $\Longrightarrow$ prior tight around zero (no error-correction)

- If the i-th row of H contains the coefficients of a linear combination of y_t that is a priori likely to be stationary
  $\Longrightarrow$ likely not zero (error-correction)

- Different priors on the loadings associated with linear combinations of y_t with different degrees of stationarity

Priors for the Long Run

$$\Delta y_t = C + \underbrace{\Pi H^{-1}}_{\Lambda} \underbrace{H y_{t-1}}_{\tilde{y}_{t-1}} + \varepsilon_t$$

- Prior on Λ

$$\Lambda_i | H, \Sigma \sim \mathcal{N}\left(0, \phi_i^2 \frac{\Sigma}{(H_i \bar{y}_0)^2}\right) \qquad i = 1, \ldots, n$$

- Hyperparameters ϕ_i
- H_i is i-th row of H
- $\bar{y}_0$ column vector containing the average of the initial p observations of each variable of the model

Priors for the Long Run

- Conjugate priors!
- Can be implemented with dummy observations in VARs in levels
- Can be easily combined with existing priors
- ML in closed form
- Hierarchical modelling and setting of ϕ_i

Priors for the Long Run

$$\tilde{y}_t = \underbrace{\begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}}_{H} \underbrace{\begin{pmatrix} Y_t \\ C_t \\ I_t \\ W_t \\ H_t \\ \pi_t \\ R_t \end{pmatrix}}_{y_t} \begin{array}{l} \rightarrow \text{real trend} \\ \rightarrow \text{log consumption-to-GDP ratio} \\ \rightarrow \text{log investment-to-GDP ratio} \\ \rightarrow \text{log labor share} \\ \rightarrow \text{log hours} \\ \rightarrow \text{nominal trend} \\ \rightarrow \text{real interest rate.} \end{array}$$

Priors for the Long Run

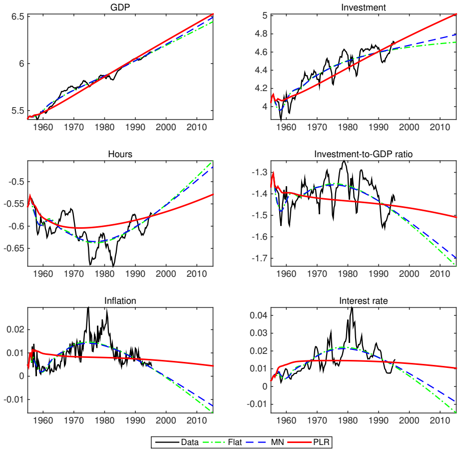

Figure: **Deterministic components** for selected variables implied by various **7-variable VARs**. Flat: BVAR with a flat prior; MN: BVAR with the Minnesota prior; PLR: BVAR with the prior for the long run (Giannone et al, 2019)

Priors for the Long Run

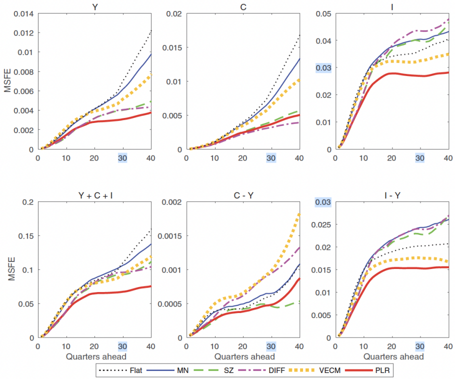

Figure: Mean squared forecast errors in models with three variables. Flat: BVAR with a flat prior; MN: BVAR with the Minnesota prior; SZ: BVAR with the Minnesota and sum-of-coefficient priors; DIFF: VAR with variables in first differences; VECM: vector error-correction model that imposes the existence of a common stochastic trend for Y, C, and I, without any additional prior information; PLR: BVAR with the Minnesota prior and the prior for the long run.

VARs with Unit Roots

Frequentist Inference vs Bayesian Inference

BVARs in levels?

- Don't we know that when variables are non stationary we should do things 'differently'?
- Bayesian inference is the same for stationary and non-stationary data!

Bayesians vs Frequentists: An Helicopter Tour

Sims and Uhlig (ECMA 91)

- Frequentist econometrics: data Y are random, parameters θ are not
  - Concerned about the properties of estimators

$$\hat{\theta} = f(y_1, y_2, \ldots, y_T)$$

    and inference procedures (tests, etc.) in repeated samples
  - Implications:
    1. There is no role for probabilistic statements about θ, such as: 'after observing the data, I believe that the probability that $\theta \leq 0$ is less than 5%.'
    2. Inference is not only based on the observed data, but also on the properties of the sampling distribution

A Simple Example

- The parameter space is $\Theta = \{0, 1\}$, and the sample space is
$$\mathcal{Y} = \{0, 1, 2, 3, 4\}$$
- Assume $P(y|\theta)$:

	0	1	2	3	4
$P_{\theta=0}$ (y)	.70	.250	.04	.005	.005
$P_{\theta=1}$ (y)	.75	.140	.04	.037	.033

Frequentist approach for testing $H_0 : \theta = 0$

- Construct a rejection region $\mathcal{C}$ – e.g. Test 0: reject H_0 if $\{y \geq 2\}$
- Size of Test 0 = Probability of rejecting H_0 if true = 5%

If instead $H_0 : \theta = 1$

- Propose Test 1A: reject H_0 if $\{y \geq 3\}$
- Size of Test 1A = Probability of rejecting H_0 if true = 7%
- Note that the size of Test 1B: reject H_0 if $\{y \geq 2\} = 11\%$

A Simple Example

- Say we observe $y = 2$
- P-value $=$ size of the test $\mathcal{C} = \{y \geq 2\}$
- Frequentist procedure seem to favour $\theta = 1$
- The p-value of $H_0 : \theta = 0$ is 5%, while that of $H_0 : \theta = 1$ is 11%

A Simple Example

- Bayesian econometrics: assume flat prior $p(\theta = 0) = p(\theta = 1) = .5$
- What is the posterior odds ratio $p(\theta = 0|y = 2)/p(\theta = 1|y = 2)$?
- Easy to compute

$$p(\theta = 0|y = 2) = \frac{p(y = 2|\theta = 0)p(\theta = 0)}{p(y = 2)}$$

$$p(\theta = 1|y = 2) = \frac{p(y = 2|\theta = 1)p(\theta = 1)}{p(y = 2)}$$

where $p(y = 2) = p(y = 2|\theta = 0)p(\theta = 0) + p(y = 2|\theta = 1)p(\theta = 1)$

- Hence:

$$\frac{p(\theta = 0|y = 2)}{p(\theta = 1|y = 2)} = \frac{p(y = 2|\theta = 0)p(\theta = 0)}{p(y = 2|\theta = 1)p(\theta = 1)} = \frac{.04}{.04} = 1$$

A Simple Example

- A Bayesian would say that the observed data are uninformative about θ
- Why difference in the conclusion?
- Driven by the properties of the sampling distribution under data that were not observed, namely $y = 3$ and $y = 4$
- The fact that

$$p(y = 3|\theta = 0) + p(y = 4|\theta = 0) = 1\%$$

while

$$p(y = 3|\theta = 1) + p(y = 4|\theta = 1) = 7\%$$

	0	1	2	3	4
$P_{\theta=0}$ (y)	.70	.250	.04	.005	.005
$P_{\theta=1}$ (y)	.75	.140	.04	.037	.033

Sims and Uhlig (1991)'s Helicopter Tour

Why Bayesian and frequentist inference differs in the case of non stationarity?

1. Generate parameter

$$\rho \sim \mathcal{U}[.8, 1.1]$$

2. Generate data from

$$y_t = \rho y_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim \mathcal{N}(0,1)$$

   with initial condition $y_0 = 0$

3. Compute the estimator

$$\hat{\rho} = \left(\sum_{t=1}^{T} y_{t-1}^2\right)^{-1} \sum_{t=1}^{T} y_t y_{t-1}$$

4. Plot $p(\rho, \hat{\rho})$

Sims and Uhlig's Helicopter Tour

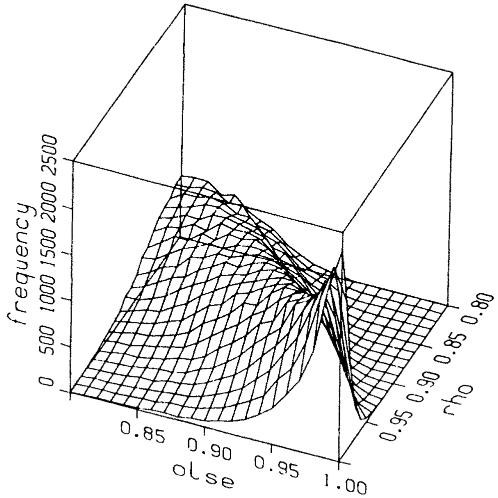

Figure: Joint frequency distribution of ρ and $\hat{\rho}$.

Sims and Uhlig's Helicopter Tour

We can understand the difference of Bayesian vs Frequentist inference by looking at the graph from different angles.

Intuition:

- **Bayesian $p(\rho|\hat{\rho})$:** Slice $p(\rho, \hat{\rho})$ for a given $\hat{\rho}$

- **Frequentist $p(\hat{\rho}|\rho)$:** Slice $p(\rho, \hat{\rho})$ for a given ρ

Sims and Uhlig's Helicopter Tour

Bayesian $p(\rho|\hat{\rho})$: The distribution of ρ for given $\hat{\rho}$ is symmetric with respect to $\hat{\rho}$

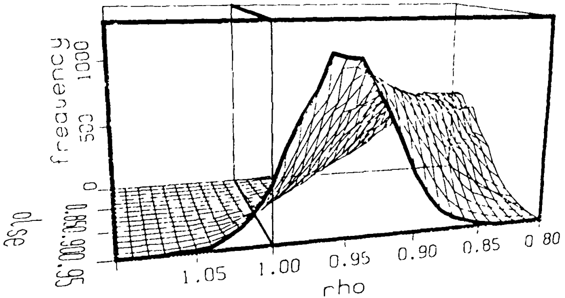

Figure: Joint frequency distribution of $\hat{\rho}$ and ρ sliced along $\hat{\rho} = .95$

Sims and Uhlig's Helicopter Tour

Bayesian $p(\rho|\hat{\rho})$: The distribution is symmetric also for $\hat{\rho} = 1$

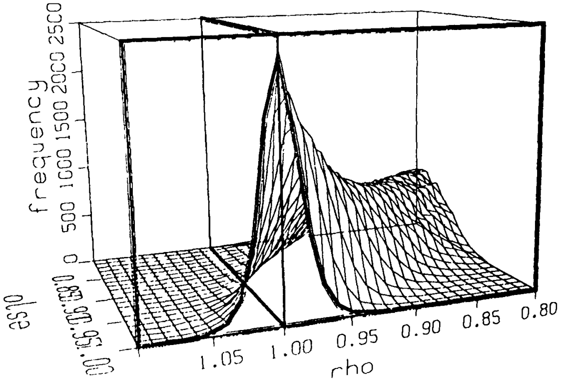

Figure: Joint frequency distribution of $\hat{\rho}$ and ρ sliced along $\hat{\rho} = 1$

Sims and Uhlig's Helicopter Tour

Frequentist $p(\hat{\rho}|\rho)$: Still symmetric for $\rho << 1$, but skewed (fat left tail) for $\rho = 1$

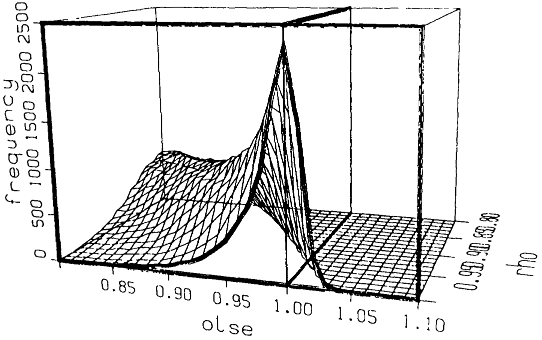

Figure: Joint frequency distribution of ρ and $\hat{\rho}$ sliced along $\rho = 1$

Sims and Uhlig (1991)'s Helicopter Tour

Assume one finds $\hat{\rho} = .95$, is there a unit root or not? (test $\rho > 1$ vs $\rho < .9$)

- **Bayesian**: The data are not informative (dotted lines in the chart)

$$E[\rho < .9 | \hat{\rho} = .95] = E[\rho > 1 | \hat{\rho} = .95]$$

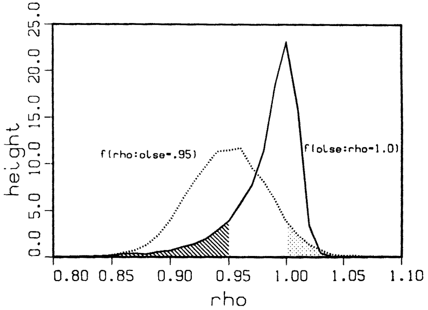

Figure: P-value vs. posterior probability

Sims and Uhlig (1991)'s Helicopter Tour

Assume one finds $\hat{\rho} = .95$, is there a unit root or not? (test $\rho > 1$ vs $\rho < .9$)

- **Frequentist**: do not reject null of unit root (solid line in the chart):
  - If $H_0 : \rho = 1$, then p-value is .12,
  - if $H_0 : \rho = .9$, then p-value is .04

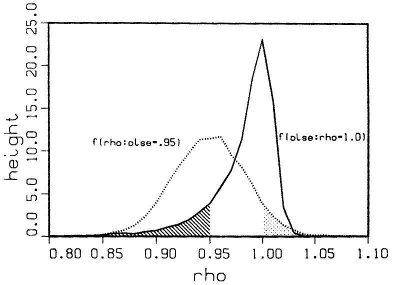

Figure: P-value vs. posterior probability

Sims and Uhlig (1991)'s Helicopter Tour

Remark: It is the fact that under $\rho = 1$ we might observe $\hat{\rho} << .95$ that makes the p-value greater. This is due to the properties of the sampling distribution!