(Reduced-Form) Vector Autoregressions

Frank Schorfheide
University of Pennsylvania
Gerzensee Ph.D. Course on Bayesian Macroeconometrics

May 18, 2019
A VAR(p) is a multivariate generalization of the AR(p) model:

\[ y_t = \Phi_1 y_{t-1} + \ldots + \Phi_p y_{t-p} + u_t \]

- \( y_t \) is a \( n \times 1 \) random vector that takes values in \( \mathbb{R}^n \).
- \( u_t \sim iid(0, \Sigma) \) is a vector of reduced-form innovations.
• Derivation of likelihood function.

• Derivation of posterior for multivariate regression with unknown variance.

• Use prior distribution to cope with high-dimensional parameter space (regularization).
Likelihood Function

- Define the \((np + 1) \times 1\) vector \(x_t\) as

\[
x_t = [y_{t-1}', \ldots, y_{t-p}', 1]'
\]

- Moreover, define the matrixes

\[
Y = \begin{bmatrix}
y_1' \\
\vdots \\
y_T'
\end{bmatrix}, \quad X = \begin{bmatrix}
x_1' \\
\vdots \\
x_T'
\end{bmatrix}, \quad \Phi = [\Phi_1, \ldots, \Phi_p, \Phi_c]'
\]

- The conditional density of \(y_t\):

\[
p(y_t | Y_{1:t-1}, Y_{1-p:0}, \Phi, \Sigma) \propto |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (y_t' - x_t' \Phi) \Sigma^{-1} (y_t' - x_t' \Phi)' \right\},
\]

where \(Y_{t_0:t_1} = [y_{t_0}, \ldots, y_{t_1}]\).
Likelihood Function

- Take the product of the conditional densities of \( y_1, \ldots, y_T \) to obtain the joint density.

- Let \( Y_{1-p:0} \) be a vector with initial observations

\[
p(Y_{1:T} | Y_{1-p:0}, \Phi, \Sigma) = \prod_{t=1}^{T} p(y_t | Y_{1:t-1}, Y_{1-p:0}, \Phi, \Sigma)
\]

\[
\propto |\Sigma|^{-T/2} \exp \left\{ -\frac{1}{2} \sum_{t=1}^{T} (y'_t - x'_t \Phi) \Sigma^{-1} (y'_t - x'_t \Phi)' \right\}
\]
Some Matrix Algebra

- **Lemma:** Let $A$ and $B$ be two $n \times n$ matrices, then
  \[ \text{tr}[A + B] = \text{tr}[A] + \text{tr}[B] \]

- **Lemma:** Let $a$ be a $n \times 1$ vector, $B$ be a symmetric positive definite $n \times n$ matrix, and $tr$ the trace operator that sums the diagonal elements of a matrix. Then
  \[ a' Ba = \text{tr}[Baa'] \]

- **Deduce:**
  \[
  \sum_{t=1}^{T} (y_t' - x_t' \Phi) \Sigma^{-1} (y_t' - x_t' \Phi)' = \text{tr} \left[ \Sigma^{-1} \sum_{t=1}^{T} (y_t' - x_t' \Phi)' (y_t' - x_t' \Phi) \right].
  \]
Likelihood Function

\[ p(Y_{1:T} \mid Y_{1-p:0}, \Phi, \Sigma) \propto |\Sigma|^{-T/2} \exp \left\{ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1} \sum_{t=1}^{T} (y_{t}' - x_{t}'\Phi)'(y_{t}' - x_{t}'\Phi) \right] \right\} \]

\[ \propto |\Sigma|^{-T/2} \exp \left\{ -\frac{1}{2} \text{tr}[\Sigma^{-1}(Y - X\Phi)'(Y - X\Phi)] \right\} \]
Likelihood Function

- Define the “OLS” estimator

\[ \hat{\Phi} = (X'X)^{-1}X'Y. \]

- Define the sum of squared OLS residual matrix

\[ \hat{S} = (Y - X\hat{\Phi})'(Y - X\hat{\Phi}) = Y'Y - Y'X(X'X)^{-1}X'Y. \]

- It can be verified that

\[ (Y - X\Phi)'(Y - X\Phi) = (\Phi - \hat{\Phi})'X'X(\Phi - \hat{\Phi}) + \hat{S}. \]
This leads to the following representation of the likelihood function
\[
p(Y|\Phi, \Sigma) \propto |\Sigma|^{-T/2} \exp \left\{ -\frac{1}{2} tr[\Sigma^{-1}\hat{S}] \right\} \\
\times \exp \left\{ -\frac{1}{2} tr[\Sigma^{-1}(\Phi - \hat{\Phi})'X'X(\Phi - \hat{\Phi})] \right\}.
\]

Let \( \beta = vec(\Phi) \) and \( \hat{\beta} = vec(\hat{\Phi}) \). It can be verified that
\[
tr[\Sigma^{-1}(\Phi - \hat{\Phi})'X'X(\Phi - \hat{\Phi})] = (\beta - \hat{\beta})'[\Sigma \otimes (X'X)^{-1}]^{-1}(\beta - \hat{\beta}).
\]

The likelihood function has the alternative representation
\[
p(Y|\Phi, \Sigma) \propto |\Sigma|^{-T/2} \exp \left\{ -\frac{1}{2} tr[\Sigma^{-1}\hat{S}] \right\} \\
\times \exp \left\{ -\frac{1}{2} (\beta - \hat{\beta})'[\Sigma \otimes (X'X)^{-1}]^{-1}(\beta - \hat{\beta}) \right\}.
\]
Suppose that
\[ p(\Phi, \Sigma) \propto c. \]
Then
\[ p(\Phi, \Sigma|Y) \propto p(Y|\Phi, \Sigma). \]
• Suppose that the random matrix $\Phi$ has density

\[
p(\Phi|\Sigma, X'X) \propto |\Sigma \otimes (X'X)^{-1}|^{-1/2} \exp \left\{-\frac{1}{2} tr[\Sigma^{-1}(\Phi - \hat{\Phi})'X'X(\Phi - \hat{\Phi})]\right\}
\]

then $\Phi|(\Sigma, X'X)$ is matricvariate normal

$MN(\hat{\Phi}, \Sigma \otimes (X'X)^{-1})$.

• Let $\beta = \text{vec}(\Phi)$ and $\hat{\beta} = \text{vec}(\hat{\Phi})$. Then

$\beta|\Sigma, X'X \sim N \left(\hat{\beta}, \Sigma \otimes (X'X)^{-1}\right)$.

• To generate a draw $Z$ from a multivariate $N(\mu, \Sigma)$, decompose $\Sigma = CC'$. E.g., $C$ could be the lower triangular Cholesky factor. Then let $Z = \mu + C \cdot N(0, I)$. 
• Let \( \Sigma \) be a \( n \times n \) positive definite random matrix. \( \Sigma \) has the Inverted Wishart \( IW(S, \nu) \) distribution if its density is of the form

\[
p(\Sigma | S, \nu) \propto |S|^{\nu/2}|\Sigma|^{-(\nu+n+1)/2} \exp\left\{-\frac{1}{2} \text{tr}[\Sigma^{-1}S]\right\}
\]

• To sample a \( \Sigma \) from an Inverted Wishart \( IW(S, \nu) \) distribution, draw \( n \times 1 \) vectors \( Z_1, \ldots, Z_\nu \) from a multivariate normal \( N(0, S^{-1}) \) and let

\[
\Sigma = \left[ \sum_{i=1}^{\nu} Z_i Z_i' \right]^{-1}
\]
Likelihood Function Interpreted as PDF for \((\Phi, \Sigma)\) – Step 1: \(p(\Phi|\Sigma, Y)\)

- Interpret the likelihood as density for \((\Phi, \Sigma)\):
  \[
  p(\Phi, \Sigma|Y) \propto |\Sigma|^{-T/2} \exp\left\{-\frac{1}{2} tr[\Sigma^{-1} \hat{S}]\right\}
  \times \exp\left\{-\frac{1}{2} tr[\Sigma^{-1}(\Phi - \hat{\Phi})' X' X (\Phi - \hat{\Phi})]\right\}
  \]

- Thus,
  \[
  \Phi|(\Sigma, Y) \sim MN(\hat{\phi}, \Sigma \otimes (X' X)^{-1}).
  \]
Likelihood Function Interpreted as PDF for $(\Phi, \Sigma) – \text{Step 2: } p(\Sigma|Y)$

- Compute $p(\Sigma|Y) \propto \int p(Y|\Phi, \Sigma)d\Phi$
- Note that:
  
  $$|\Sigma \otimes (X'X)^{-1}|^{1/2} = |\Sigma|^{k/2}|X'X|^{-n/2}.$$

- Therefore,
  
  $$p(\Sigma|Y) \propto |\Sigma|^{-(T-k)/2}|X'X|^{-n/2} \exp \left\{ -\frac{1}{2} \text{tr}[\Sigma^{-1}\hat{S}] \right\}.$$

- Deduce
  
  $$\Sigma|Y \sim IW(\hat{S}, T - k - n - 1), \quad \Phi|(\Sigma, Y) \sim MN(\hat{\phi}, \Sigma \otimes (X'X)^{-1}).$$

- Write as
  
  $$(\Phi, \Sigma)|Y \sim MNIW\left(\hat{\Phi}, (X'X)^{-1}, \hat{S}, T - k - n - 1\right).$$
Bayesian Analysis with Improper Prior

- Replace (improper) prior $p(\Phi, \Sigma) \propto c$ by (improper) prior:
  \[ p(\Phi, \Sigma) \propto |\Sigma|^{-(n+1)/2}. \]

- Then, the posterior is obtained from
  \[ p(\Phi, \Sigma|Y) \propto p(Y|\Phi, \Sigma)p(\Phi, \Sigma). \]

- Our previous analysis implies that
  \[ (\Phi, \Sigma)|Y \sim MNIW\left(\hat{\Phi}, (X'X)^{-1}, \hat{S}, T - k\right). \]
Algorithm: Direct Sampling of VAR Parameters

For $s = 1, \ldots, n_{\text{sim}}$:

1. Draw $\Sigma^{(s)}$ from an $IW(\hat{S}, T - k)$ distribution.
2. Draw $\Phi^{(s)}$ from the conditional distribution $MN(\hat{\Phi}, \Sigma^{(s)} \otimes (X'X)^{-1})$. □
The Use of Priors for VARs

• Priors are used to “regularize” the VAR likelihood and cope with the dimensionality problem: the number of free parameters is often large relative to the number of observations.

• Priors add information to the estimation problem.
Consider the prior:

$$\Sigma \sim IW(\nu, S), \quad \Phi | \Sigma \sim MN(\mu_\Phi, \Sigma \otimes P_\Phi^{-1}),$$

Prior density:

$$p(\Phi, \Sigma) = (2\pi)^{-nk/2} |\Sigma|^{-k/2} |P_\Phi|^{n/2} \exp \left\{ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1} (\Phi - \mu_\Phi)' P_\Phi (\Phi - \mu_\Phi) \right] \right\}$$

$$\times C_{IW} |\Sigma|^{-(\nu+n+1)/2} \exp \left\{ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1} S \right] \right\}.$$
\[ p(Y|\Phi, \Sigma)p(\phi, \Sigma) \propto (2\pi)^{-nk/2}|\Sigma|^{-k/2}|P_{\phi}|^{n/2}C_{IW}|\Sigma|^{-(\nu+n+1)/2}(2\pi)^{-nT/2}|\Sigma|^{-T/2} \]
\[ \times \exp\left\{-\frac{1}{2}\text{tr}\left[\Sigma^{-1}(\Phi - \mu_\Phi)'P_{\phi}(\Phi - \mu_\Phi)\right]\right\} \exp\left\{-\frac{1}{2}\text{tr}\left[\Sigma^{-1}(\Phi - \hat{\Phi})X'X(\Phi - \hat{\Phi})\right]\right\} \]
\[ \times \exp\left\{-\frac{1}{2}\text{tr}\left[\Sigma^{-1}S\right]\right\} \exp\left\{-\frac{1}{2}\text{tr}\left[\Sigma^{-1}\hat{S}\right]\right\} \]
**Posterior**

- Define

\[ \bar{P}_\Phi = P_\Phi + X'X, \quad \bar{\mu}_\Phi = \bar{P}_\Phi^{-1} [P_\Phi \mu_\Phi + X'\hat{X}]. \]

- Write

\[
p(Y|\Phi, \Sigma)p(\phi, \Sigma) = (2\pi)^{-nk/2} |\Sigma|^{-k/2} |P_\Phi|^{n/2} C_{IW} |\Sigma|^{-(\nu+n+1)/2} (2\pi)^{-nT/2} |\Sigma|^{-T/2} |\bar{P}_\phi|^{n/2} |\bar{P}_\phi|^{-n/2} \times \exp \left\{ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1} (\Phi - \bar{\mu}_\Phi) \bar{P}_\phi (\Phi - \bar{\mu}_\Phi) \right] \right\} \times \exp \left\{ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1} (S + \mu_\phi' P_\phi \mu_\phi + Y'Y - \bar{\mu}_\phi' \bar{P}_\phi \bar{\mu}_\phi) \right] \right\}
\]  

Frank Schorfheide  
(Reduced-Form) Vector Autoregressions
Define

\[ \bar{S} = S + \mu' \Phi \bar{P} \mu \Phi + \bar{Y}' \bar{Y} - \bar{\mu}' \bar{P} \bar{\mu} \Phi, \quad \bar{\nu} = \nu + T \]

Deduce that

\[ \Sigma | Y \sim IW(\bar{S}, \bar{\nu}), \quad \Phi | (\Sigma, Y) \sim MN(\bar{\mu} \Phi, \Sigma \otimes \bar{P}^{-1}). \]

Draws from the posterior can be generated by direct sampling.
Marginal Data Density

Let $\bar{C}_{IW}$ be the normalization constant of the posterior IW distribution. Then,

$$p(Y) = \int \int p(Y|\Phi, \Sigma)p(\phi, \Sigma)d\Phi d\Sigma$$

$$= (2\pi)^{-nT/2} \int |P_\phi|^{k/2} |P_\phi|^{-k/2} \bar{C}_{IW} \Sigma^{-(\bar{\nu}+n+1)/2} \exp \left\{ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1} S \right] \right\} d\Sigma$$

$$= (2\pi)^{-nT/2} \frac{|P_\phi|^{n/2} \bar{C}_{IW}}{|\bar{P}_\phi|^{n/2} \bar{C}_{IW}},$$

where

$$\frac{C_{IW}}{\bar{C}_{IW}} = \frac{|S|^{\nu/2} 2^{n_y \bar{\nu}/2} \prod_{i=1}^{n_y} \Gamma((\bar{\nu} + 1 - i)/2)}{|S|^{\bar{\nu}/2} 2^{n_y \nu/2} \prod_{i=1}^{n_y} \Gamma((\nu + 1 - i)/2)}.$$
• "Performance" of Bayesian VAR is sensitive to prior variance.

• Choose prior variance in a data-driven way.

• Hierarchical model

\[ p(Y|\Phi, \Sigma)p(\Phi, \Sigma|\lambda)p(\lambda), \]

where \( \lambda \) controls features of the prior.
Selection vs. Averaging

• **Selection:**
  • Compute
    \[
    p(Y|\lambda) = \int p(Y|\Phi, \Sigma)p(\Phi, \Sigma|\lambda)d(\Phi, \Sigma).
    \]
  • Define: \( \hat{\lambda} = \text{argmax} \ p(Y|\lambda) \).
  • Work with \( p(\Phi, \Sigma|Y, \hat{\lambda}) \).

• **Averaging:**
  • Use prior \( p(\lambda) \)
  • Factorize posterior as
    \[
    p(\Phi, \Sigma, \lambda|Y) = p(\Phi, \Sigma|Y, \lambda)p(\lambda|Y),
    \]
    where \( p(\lambda|Y) \propto p(Y|\lambda)p(\lambda) \).
Illustration: Marginal Likelihood of $\lambda$
Illustration: Marginal Likelihood of $\lambda$

Prior $\lambda = \infty$

Likelihood $\lambda \to 0$

Frank Schorfheide
(Reduced-Form) Vector Autoregressions
Example: Marginal Likelihood of $\lambda$

- Suppose the VAR takes the special form of an AR(1) model:
  \[ y_t = \phi y_{t-1} + u_t, \quad u_t \sim iidN(0, 1) \]

- Suppose the prior takes the form
  \[ \phi \sim N \left( \phi^*, \frac{1}{\lambda T \gamma_0} \right). \]
  where $\gamma_0 = 1/(1 - \phi^2)$

- Define $\gamma_1 = \phi^* \gamma_0$ and denote the sample autocovariances by
  \[ \hat{\gamma}_0 = \frac{1}{T} \sum y_t^2, \quad \hat{\gamma}_1 = \frac{1}{T} \sum y_t y_{t-1}. \]

- For convenience, we standardized the prior variance by $T$. 
Example: Marginal Likelihood of $\lambda$

- After some algebra it can be shown that marginal likelihood of $\lambda$ takes the following form

$$\ln p(Y|\lambda, \phi^*) = -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \tilde{\sigma}^2(\lambda, \phi^*) - \frac{1}{2} c(\lambda, \phi^*).$$

- The term $\tilde{\sigma}^2(\lambda, \phi^*)$ measures the in-sample one-step-ahead forecast error:

$$\lim_{\lambda \to 0} \tilde{\sigma}^2(\lambda, \phi^*) = \frac{1}{T} \sum(y_t - \hat{\phi}y_{t-1})^2$$

$$\lim_{\lambda \to \infty} \tilde{\sigma}^2(\lambda, \phi^*) = \frac{1}{T} \sum(y_t - \phi^* y_{t-1})^2.$$ 

- The third term above can be interpreted as a penalty for model complexity and is of the form

$$c(\lambda, \phi^*) = \ln \left(1 + \frac{\hat{\gamma}_0}{\lambda \gamma_0}\right).$$

- As $\lambda$ approaches zero, the marginal log likelihood function tends to minus infinity.
• Recall that marginal likelihood of $\lambda$ takes the following form

$$
\ln p(Y|\lambda, \phi^*) = -T/2 \ln(2\pi) - \frac{T}{2} \hat{\sigma}^2(\lambda, \phi^*) - \frac{1}{2} c(\lambda, \phi^*).
$$

• If an interior maximum of marginal likelihood exists, it is given by

$$
\hat{\lambda} = \frac{\gamma_0^2}{T(\gamma_0 \gamma_1 - \gamma_0 \hat{\gamma}_1)^2 - (\gamma_0)^2 \hat{\gamma}_0}.
$$

Consider the following Gaussian bivariate VAR(2).

\[
\begin{bmatrix}
  y_{1,t} \\
  y_{2,t}
\end{bmatrix} = \begin{bmatrix}
  \alpha_1 \\
  \alpha_2
\end{bmatrix} + \begin{bmatrix}
  \beta_{11} & \beta_{12} \\
  \beta_{21} & \beta_{22}
\end{bmatrix} \begin{bmatrix}
  y_{1,t-1} \\
  y_{2,t-1}
\end{bmatrix} \\
+ \begin{bmatrix}
  \gamma_{11} & \gamma_{12} \\
  \gamma_{21} & \gamma_{22}
\end{bmatrix} \begin{bmatrix}
  y_{1,t-2} \\
  y_{2,t-2}
\end{bmatrix} + \begin{bmatrix}
  u_{1,t} \\
  u_{2,t}
\end{bmatrix}
\]

Minnesota Prior is centered at:

\[
\begin{bmatrix}
  y_{1,t} \\
  y_{2,t}
\end{bmatrix} = \begin{bmatrix}
  ? \\
  ?
\end{bmatrix} + \begin{bmatrix}
  1 & 0 \\
  0 & 1
\end{bmatrix} \begin{bmatrix}
  y_{1,t-1} \\
  y_{2,t-1}
\end{bmatrix} \\
+ \begin{bmatrix}
  0 & 0 \\
  0 & 0
\end{bmatrix} \begin{bmatrix}
  y_{1,t-2} \\
  y_{2,t-2}
\end{bmatrix} + \begin{bmatrix}
  u_{1,t} \\
  u_{2,t}
\end{bmatrix}
\]
Considerations

- It’s relatively easy to think about variances of marginal prior distributions of parameters, but hard to think of a full covariance matrix.

- In high-dimensional parameter spaces, independence of parameters may lead to a lot of probability mass in regions of the parameter space in which the model behaves unreasonable.

- VAR: want plausible long-run properties, e.g., co-trending, steady states, etc.

- Constructing priors from artificial (dummy) observations can help introducing reasonable correlations.
Constructing a Prior from “Dummy Observations”

• Suppose we have $T^*$ dummy observations $(Y^*, X^*)$, plug the dummy observations into the likelihood function and multiply the likelihood function by the (initial) improper prior:

$$p(\Phi, \Sigma) \propto |\Sigma|^{-(n+1)/2} p(Y^* | \Phi, \Sigma).$$

• Define

$$\Phi^* = (X^* X^*)^{-1} X^* Y^*, \quad S^* = (Y^* - X^* \Phi^*)' (Y^* - X^* \Phi^*).$$

• This leads to a prior

$$\Sigma \sim IW(\nu, S), \quad \Phi|\Sigma \sim MN(\mu_\Phi, \Sigma \otimes P_\Phi^{-1}),$$

with

$$\nu = T^* - k, \quad S = S^*, \quad \mu_\Phi = \Phi^*, \quad P_\Phi = X^* X^*.$$
Posterior with Dummy Observation Prior

• Posterior is proportional to

\[ p(\Phi, \Sigma, Y) \propto p(Y|\Phi, \Sigma)p(Y^*|\Phi, \Sigma)|\Sigma|^{-(n+1)/2} \]

• Define \( \bar{T} = T^* + T \) and

\[
\bar{\Phi} = (X^*X^* + X'X)^{-1}(X^*Y^* + X'Y) \\
\bar{S} = \left[ Y^*Y^* + Y'Y - (X^*Y^* + X'Y)'(X^*X^* + X'X)^{-1}(X^*Y^* + X'Y) \right].
\]

• Then, let \( \bar{X} = [X^*, X']' \) and deduce:

• Deduce that

\[ \Sigma|Y \sim IW(\bar{S}, \bar{T} - k), \quad \Phi|(\Sigma, Y) \sim MN(\bar{\Phi}, \Sigma \otimes (\bar{X}'\bar{X})^{-1}). \]
Example: Minnesota Prior

- Dummies for the $\beta$ coefficients:

$$Y^* = X^*\Phi + U$$

$$\begin{bmatrix}
\lambda_1s_1 & 0 \\
0 & \lambda_1s_2
\end{bmatrix} = \begin{bmatrix}
\lambda_1s_1 & 0 & 0 & 0 & 0 \\
0 & \lambda_1s_2 & 0 & 0 & 0
\end{bmatrix}\Phi + \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix}$$

The first observation implies, for instance, that

$$\lambda_1s_1 = \lambda_1s_1\beta_{11} + u_{11} \implies \beta_{11} = 1 - \frac{u_{11}}{\lambda_1s_1}$$

$$\implies \beta_{11} \sim \mathcal{N} \left( 1, \frac{\Sigma_{11}}{\lambda_1^2s_1^2} \right)$$

$$0 = \lambda_1s_1\beta_{21} + u_{12} \implies \beta_{21} = -\frac{u_{12}}{\lambda_1s_1}$$

$$\implies \beta_{21} \sim \mathcal{N} \left( 0, \frac{\Sigma_{22}}{\lambda_1^2s_1^2} \right)$$
Example: Minnesota Prior

- Dummies for the $\gamma$ coefficients:

$$
\begin{bmatrix}
0 & 0 \\
0 & 0 \\
\end{bmatrix} = 
\begin{bmatrix}
0 & 0 & \lambda_1 s_1 2^{\lambda_2} & 0 & 0 \\
0 & 0 & 0 & \lambda_1 s_2 2^{\lambda_2} & 0 \\
\end{bmatrix} \Phi + U
$$

- For lags of order $p$ the entry above would be $\lambda_1 s_i p_2^\lambda$.

- The prior for the covariance matrix is implemented by $\lambda_3$ replications of

$$
\begin{bmatrix}
s_1 & 0 \\
0 & s_2 \\
\end{bmatrix} = 
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix} \Phi + U
$$

Frank Schorfheide

(Reduced-Form) Vector Autoregressions
Example: Minnesota Prior

- Sums-of-coefficients dummy observations, reflecting the belief that when $y_i$ has been stable at its initial level, it will tend to persist at that level, regardless of the value of other variables:

$$\begin{bmatrix} \lambda_4 y_1 & 0 \\ 0 & \lambda_4 y_2 \end{bmatrix} = \begin{bmatrix} \lambda_4 y_1 & 0 & \lambda_4 y_1 & 0 & 0 \\ 0 & \lambda_4 y_2 & 0 & \lambda_4 y_2 & 0 \end{bmatrix} \Phi + U$$

- Co-persistence prior dummy observations, reflecting the belief that when data on all $y$’s are stable at their initial levels, thy will tend to persist at that level:

$$\begin{bmatrix} \lambda_5 y_1 & \lambda_5 y_2 \end{bmatrix} = \begin{bmatrix} \lambda_5 y_1 & \lambda_5 y_2 & \lambda_5 y_1 & \lambda_5 y_2 & \lambda_5 \end{bmatrix} \Phi + U$$
Example: Minnesota Prior - Hyperparameters

- $\lambda_1$ is the overall tightness of the prior. Large values imply a small prior covariance matrix.
- $\lambda_2$: the variance for the coefficients of lag $h$ is scaled down by the factor $(1-\lambda_2)^2$.
- $\lambda_3$: determines the weight for the prior on $\Sigma$. Suppose that $Z_i = \mathcal{N}(0, \sigma^2)$. Then an estimator for $\sigma^2$ is $\hat{\sigma}^2 = \frac{1}{\lambda_3} \sum_{i=1}^{\lambda_3} Z_i^2$. The larger $\lambda_3$, the more informative the estimator, and in the context of the VAR, the tighter the prior.
- $\lambda_4$ and $\lambda_5$: tuning parameters for sums-of-coefficients and co-persistence dummies.
- In addition: $s = \text{std}(Y_{-\tau,0})$ and $y = \text{mean}(Y_{-\tau,0})$, where $Y_{-\tau,0}$ is a short presample.
US GDP growth and VAR forecast (1-step ahead)

Flat-prior VAR
US GDP growth and BVAR forecast (1-step ahead)

BVAR (MN+SOC+DIO priors + hyperparameter selection)
### Mean Squared Forecast Errors

<table>
<thead>
<tr>
<th></th>
<th>Flat-prior</th>
<th>BVAR with MN prior ($\lambda=0.2$)</th>
<th>BVAR with MN+SOC+DIO</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 Quarter Ahead</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real GDP</td>
<td>19.18</td>
<td>9.61</td>
<td>7.97</td>
</tr>
<tr>
<td>GDP Deflator</td>
<td>2.27</td>
<td>1.53</td>
<td>1.35</td>
</tr>
<tr>
<td>Federal Funds Rate</td>
<td>1.83</td>
<td>1.08</td>
<td>1.03</td>
</tr>
<tr>
<td><strong>1 Year Ahead</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real GDP</td>
<td>11.90</td>
<td>5.48</td>
<td>3.42</td>
</tr>
<tr>
<td>GDP Deflator</td>
<td>2.22</td>
<td>1.85</td>
<td>1.58</td>
</tr>
<tr>
<td>Federal Funds Rate</td>
<td>0.56</td>
<td>0.40</td>
<td>0.31</td>
</tr>
</tbody>
</table>
The marginal likelihood can be calculated from the normalization constants of the MNIW distribution (see Zellner (1971, Appendix)):

\[
p(Y|\lambda) = (2\pi)^{-nT/2} \frac{\left| \tilde{X}' \tilde{X} \right|^{-\frac{n}{2}}}{\left| X^* ' X^* \right|^{-\frac{n}{2}}} \frac{\left| \tilde{S} \right|^{-\frac{T-k}{2}}}{\left| S^* \right|^{-\frac{T^* - k}{2}}} \frac{2^{n(\bar{T} - k)}}{\prod_{i=1}^{n} \Gamma[(\bar{T} - k + 1 - i)/2]} \frac{2^{n(T^* - k)}}{\prod_{i=1}^{n} \Gamma[(T^* - k + 1 - i)/2]}.
\]

The hyperparameters \((\bar{y}, \bar{s}, \lambda)\) enter through the dummy observations \(X^*\) and \(Y^*\).
Extension 1: Alternative Priors

- Estimation is typically based on conditional likelihood functions that ignore the likelihood of the initial observations.
- Example:

\[
y_t = c + \phi y_{t-1} + u_t = \phi^{t-1} y_1 + c \sum_{s=0}^{t-2} \phi^s + \sum_{s=0}^{t-2} \phi^s u_{t-s}
\]

\[
DC_t = \begin{cases} 
  y_1 + (t - 1)c & \text{if } \phi = 1 \\
  \frac{c}{1-\phi} + \phi^{t-1}(y_1 - \frac{c}{1-\phi}) & \text{if } |\phi| < 1 
\end{cases}
\]

- Deterministic component may absorb too much low frequency variation of the time series.
Most economists would be skeptical of this likely spurious explanatory power of deterministic trends, and may want to downplay it when conducting inference. In principle, "one way to accomplish this is to use priors favoring pure unit-root low frequency behavior" (Sims, 2000, pp. 451), according to which implausibly precise long-term forecasts are unlikely. However, it is not obvious how to formulate such a prior. For example, the undesirable properties of the deterministic component persist even when using the popular Minnesota prior, which is centered on the assumption that all variables in the VAR are random walks with drift (Litterman, 1979, see also appendix B for a detailed description). When the tightness of this prior is set to conventional values in the literature (see appendix C), the implied deterministic components are similar to those of the flat-prior case, as shown by the dashed lines in figure 2.1. In the next section we detail our specific proposal regarding how to address this problem.

**Figure 2.1.** Deterministic component 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.
• Write VAR in VECM form:

\[ \Delta y_t = \Pi_0 + \Pi_* y_{t-1} + \sum_{j=1}^{p-1} \Pi_i \Delta y_{t-j} + u_t \]

where \( \Pi_* = \alpha \beta' \).

• Reasonable prior for columns of \( \alpha \) will depend on the rows of \( \beta' \):

  • if \( i' \)th row of \( \beta' \) corresponds to a linear combination that is stationary, then it makes sense to choose a prior for \( i' \)th column of \( \alpha \) with mass away from zero.

  • if \( i' \)th row of \( \beta' \) corresponds to a linear combination that is non-stationary, then it makes sense to choose a prior for \( i' \)th column of \( \alpha \) with mass away from zero.

• See paper for details on how to implement this.
Figure 5.1. 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.
Extension 2: Sparse versus Dense Models


- Sparse models: only a few predictors are relevant.

- Dense models: many predictors are relevant but only have small individual effects.

- Model:
  \[ y_t = x_t'\phi + z_t'\beta + u_t. \]

  Here \( x_t \)'s are included in all specifications (low dimensional), \( z_t \)'s are optional (high dimensional).

- Prior – part 1:
  \[ p(\sigma^2) \propto \frac{1}{\sigma^2}, \quad \phi \propto c. \]
Extension 2: Sparse versus Dense Models

• Prior – part 2: “spike and slab”

\[ \beta_i | (\sigma^2, \gamma^2, q) \sim \begin{cases} 
N(0, \sigma^2 \gamma^2) & \text{with prob. } q \\
0 & \text{with prob. } 1 - q 
\end{cases} \]

• For \( q = 1 \) we obtain our “standard” prior (“Ridge Regression”)

• Rewrite prior as

\[ \beta_i | (\sigma^2, \gamma^2, \nu_i) \sim N(0, \sigma^2 \gamma^2, \nu_i), \quad \nu_i \sim \text{Bernoulli}(q). \]

• By changing the mixing distribution, we can generate a wide variety of priors, including a Bayesian version of LASSO.
Extension 2: Sparse versus Dense Models

• In problems of this form it is often good to standardize and orthogonalize the regressors $x_t$ prior to the estimation.

• To specify a prior on the hyperparameters $(q, \gamma^2)$ they suggest to define

$$R^2(\gamma^2, q) = \frac{q k \gamma^2 \bar{\sigma}_z^2}{q k \gamma^2 \bar{\sigma}_z^2 + 1}$$

where $k$ is the number of regressors $z$ and $\bar{\sigma}_z^2$ is the average sample variance of the $z_j$'s.

• The prior takes the form

$$q \sim \text{Beta}(a, b), \quad R^2 \sim \text{Beta}(A, B).$$

• The paper works out the posterior.
Table 1. Description of the datasets.

<table>
<thead>
<tr>
<th></th>
<th>Dependent variable</th>
<th>Possible predictors</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Macro 1</strong></td>
<td>Monthly growth rate of US industrial production</td>
<td>130 lagged macroeconomic indicators</td>
<td>659 monthly time-series observations, from February 1960 to December 2014</td>
</tr>
<tr>
<td><strong>Macro 2</strong></td>
<td>Average growth rate of GDP over the sample 1960-1985</td>
<td>60 socio-economic, institutional and geographical characteristics, measured at pre-60s value</td>
<td>90 cross-sectional country observations</td>
</tr>
<tr>
<td><strong>Finance 1</strong></td>
<td>US equity premium (S&amp;P 500)</td>
<td>16 lagged financial and macroeconomic indicators</td>
<td>58 annual time-series observations, from 1948 to 2015</td>
</tr>
<tr>
<td><strong>Finance 2</strong></td>
<td>Stock returns of US firms</td>
<td>144 dummies classifying stock as very low, low, high or very high in terms of 36 lagged characteristics</td>
<td>1400k panel observations for an average of 2250 stocks over a span of 624 months, from July 1963 to June 2015</td>
</tr>
<tr>
<td><strong>Micro 1</strong></td>
<td>Per-capita crime (murder) rates</td>
<td>Effective abortion rate and 284 controls including possible covariate of crime and their transformations</td>
<td>576 panel observations for 48 US states over a span of 144 months, from January 1986 to December 1997</td>
</tr>
<tr>
<td><strong>Micro 2</strong></td>
<td>Number of pro-plaintiff eminent domain decisions in a specific circuit and in a specific year</td>
<td>Characteristics of judicial panels capturing aspects related to gender, race, religion, political affiliation, education and professional history of the judges, together with some interactions among the latter, for a total of 138 regressors</td>
<td>312 panel circuit/year observations, from 1975 to 2008</td>
</tr>
</tbody>
</table>
artificially recover sparse model representations simply as a device to reduce estimation error. Our findings indicate that these extreme strategies might perhaps be appropriate only for our micro-1 application, given that its posterior in figure 4.2 is tightly concentrated.
around extremely low values of $q$. More generally, however, our results suggest that the best predictive models are those that optimally combine probability of inclusion and shrinkage.

4.2. Probability of inclusion and out-of-sample predictive accuracy. What is then the appropriate probability of inclusion, considering that models with different sizes require differential shrinkage? To answer this question, figure 4.3 plots the marginal posterior of $q$, obtained by integrating out $\sigma^2$ from the joint posterior distribution of figures 4.1 and 4.2. Notice that the densities in figure 4.3 behave quite differently across applications. For
Figure 4.5. Heat map of the probabilities of inclusion of each predictor (best viewed in color).

percent of the times, although this is more difficult to see from the plot. All other predictors are included in the model much more rarely.

The important message of figure 4.5, however, is that the remaining five applications do not exhibit a distinct pattern of sparsity, in the sense that none of the predictors appear to be systematically excluded. This finding was probably expected for macro 2 and finance 1, since the posterior of $q$ peaks around very high values in these two applications. The absence of clear sparsity patterns, however, should be more surprising when the posterior of $q$ has most of its mass on lower values. For example, let us consider the case of macro 1, in which the best fitting models are those with $q$ around 0.25, according to figure 4.3. This