Bayesian Vector Autoregressions

Lawrence J. Christiano

Bayesian Vector Autoregressions

- Vector Autoregressions are a flexible way to summarize the dynamics in the data, and use these to construct forecasts.
- Problem: vector autoregressions have an enormous number of parameters.
  - Individual parameters imprecisely estimated.
    - imprecision increases variance of forecast errors.
  - Doan, Litterman and Sims, working at the Federal Reserve Bank of Minneapolis, developed Bayesian methods to use Bayesian priors to reduced instability in estimated VAR parameters, and thus improve forecast accuracy.
- Initial work provided in Litterman's Phd dissertation, released as "A Bayesian Procedure for Forecasting with Vector Autoregression," Massachusetts Institute of Technology, Department of Economics Working Paper, 1980.
- Another important early paper: Doan, Litterman and Sims, 1984. "Forecasting and Conditional Projection Using Realistic Prior Distributions." Econometric Reviews 3:1–100.

Bayesian Vector Autoregressions

- Of course, much has been written to describe BVARs.
  - Classic treatment: Arnold Zellner, An Introduction to Bayesian Inference in Econometrics, John Wiley & Sons, 1971.
  - Hamilton's textbook, Time Series Analysis has a very good chapter.
  - Here is an accessible discussion: Robertson and Tallman, 'Vectors Autoregressions: Forecasting and Reality', Federal Reserve Bank of Atlanta, Economic Reviews, First Quarter, 1999.
  - Rigorous recent reviews of the subject: Del Negro and Schorfheide, 'Bayesian Macroeconometrics,' chapter in Handbook Bayesian Econometrics, Oxford University Press, 2011.

Outline

- Normal Likelihood, Illustrated with Simple AR(2) representation.
  - conditional versus unconditional likelihood.
  - maximum likelihood with level GDP data.
  - the Hurwicz bias.
- Three representations of a VAR.
  - Standard Representation
  - Matrix Representation
  - Vectorized Representation.
- Priors, posteriors and marginal likelihood
  - Dummy observations.
  - Conjugate Priors.
- Forecasting with BVARs
  - stochastic simulations, versus non-stochastic.
  - forecast probability intervals.

Scalar Autoregressive Representation

- p^{th} order autoregression:

$$\begin{aligned} y_t &= A_0 + A_1 y_{t-1} + A_2 y_{t-2} + u_t,\ u_t \sim \mathcal{N}(0,\Sigma) \\ Y &= \begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix}. \end{aligned}$$

- Observed data:

$$Y, \quad \overbrace{y_0, y_{-1}}^{\text{initial conditions}} \quad .$$

- Normal likelihood of observed data:

$$p\left(Y, y_0, y_{-1} | A, \Sigma\right) \sim \mathcal{N}\left(\mu, V\right),$$

where $V \sim (T+2) \times (T+2)$. To evaluate p, must invert V.

- Matrix inversion is expensive, $O\left(T^3\right)$.
- Express likelihood in recursive form to simplify inversion.

Recursive Representation of Likelihood

- Property of probabilities:

$$p(A,B) = p(A|B)\,p(B).$$

- Suppose $T = 1$.
  - Then, the joint likelihood of the data, y_1, y_0, y_{-1}, conditional on the model parameters:

$$p(y_1, y_0, y_{-1}|A, \Sigma)$$

$$= \overbrace{p(y_1|y_0, y_{-1}, A, \Sigma)}^{\text{likelihood of } y_1\text{, conditional on initial conditions}}$$

$$\times \overbrace{p(y_0, y_{-1}|A, \Sigma)}^{\text{marginal likelihood of initial conditions}}$$

Recursive Representation of Likelihood

- Consider $T = 2$:

$$p\left(y_2, y_1, y_0, y_{-1} | A, \Sigma\right) =$$

$$\overbrace{p\left(y_2 | y_1, y_0, y_{-1}, A, \Sigma\right) \times \overbrace{p\left(y_1 | y_0, y_{-1}, A, \Sigma\right) \times p\left(y_0, y_{-1} | A, \Sigma\right)}^{p(y_1, y_0, y_{-1} | A, \Sigma)}}^{p(y_2, y_1, y_0, y_{-1} | A, \Sigma)}$$

and so on for $T = 2, 3, \ldots$.

Recursive Representation of Likelihood

- Consider $T \geq 1$:

$$p\left(y_T, ..., y_1, y_0, y_{-1} | A, \Sigma\right) =$$

$$p\left(y_T | y_{T-1}, y_{T-2}, A, \Sigma\right)$$

$$\times p\left(y_{T-1} | y_{T-2}, y_{T-3}, A, \Sigma\right)$$

$$\times \cdots \times p\left(y_t | y_{t-1}, y_{t-2}, A, \Sigma\right)$$

$$\times \cdots \times p\left(y_2 | y_1, y_0, A, \Sigma\right) p\left(y_1 | y_0, y_{-1}, A, \Sigma\right) p\left(y_0, y_{-1} | A, \Sigma\right).$$

- Note how we have converted a single $(T+2) \times (T+2)$ inversion problem into a set of scalar inversions.

Conditional (Normal) Likelihood

- From Normality

$$
\begin{aligned}
& p\left(y_t | y_{t-1}, y_{t-2}, A, \Sigma\right) \\
= & \frac{1}{(2\pi\Sigma)^{1/2}} \exp\left[-\frac{1}{2} \frac{\left(y_t - A_0 - A_1 y_{t-1} - A_2 y_{t-2}\right)^2}{\Sigma}\right],
\end{aligned}
$$

for $t = 1, ..., T$.

- Likelihood of data, conditional on initial observations,

$$
p\left(Y | y_0, y_{-1}, A, \Sigma\right) = \prod_{t=1}^{T} p\left(y_t | y_{t-1}, y_{t-2}, A, \Sigma\right)
$$

$$
= \frac{1}{(2\pi\Sigma)^{T/2}} \exp\left[-\frac{1}{2} \sum_{t=1}^{T} \frac{\left(y_t - A_0 - A_1 y_{t-1} - A_2 y_{t-2}\right)^2}{\Sigma}\right].
$$

Maximum (Conditional) Likelihood

- Log-Likelihood conditional on initial observations:

$$
\begin{aligned}
& \log\left[p\left(Y|y_0, y_{-1}, A, \Sigma\right)\right] \\
= \ & -\frac{T}{2}\log\Sigma - \frac{T}{2}\log\left(2\pi\right) \\
& -\frac{1}{2}\sum_{t=1}^{T}\frac{\left(y_t - A_0 - A_1 y_{t-1} - A_2 y_{t-2}\right)^2}{\Sigma}
\end{aligned}
$$

- Conditional maximum likelihood: optimize w.r.t. A, Σ

- First order conditions for maximum provide four equations in four unknowns:

$$\hat{\Sigma}, \hat{A}_0, \hat{A}_1, \hat{A}_2.$$

First Order Conditions Associated with Conditional Maximum Likelihood

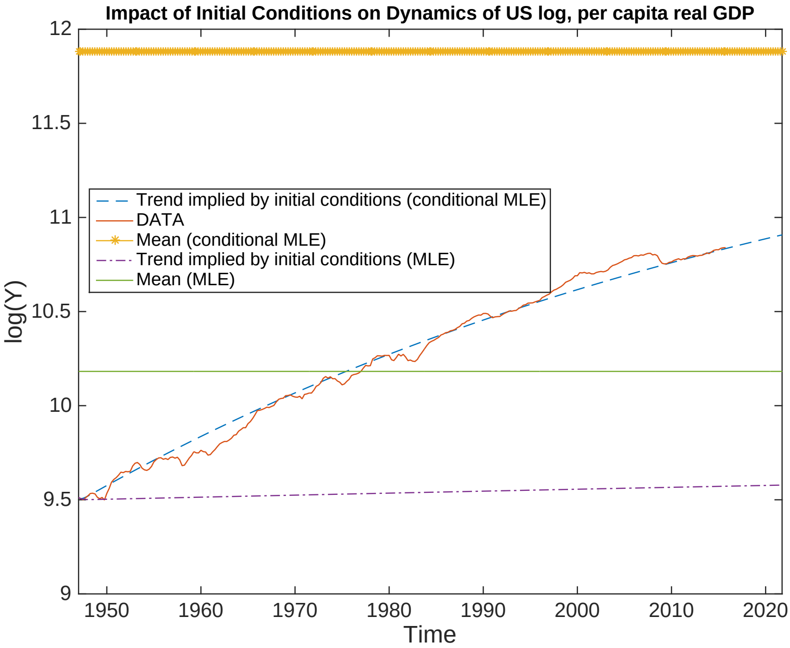
- Setting derivatives to zero:

$$
\begin{aligned}
\Sigma &: \hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \hat{A}_0 - \hat{A}_1 y_{t-1} - \hat{A}_2 y_{t-2}\right)^2 \\
A_0 &: \sum_{t=1}^{T}\left(y_t - \hat{A}_0 - \hat{A}_1 y_{t-1} - \hat{A}_2 y_{t-2}\right) = 0 \\
A_1 &: \sum_{t=1}^{T}\left(y_t - \hat{A}_0 - \hat{A}_1 y_{t-1} - \hat{A}_2 y_{t-2}\right) y_{t-1} = 0 \\
A_1 &: \sum_{t=1}^{T}\left(y_t - \hat{A}_0 - \hat{A}_1 y_{t-1} - \hat{A}_2 y_{t-2}\right) y_{t-2} = 0.
\end{aligned}
$$

- Yay....OLS!

Application

- US log, real per capita GDP, 1947Q1 - 2015Q4, $T = 274$
- Conditional maximum likelihood (OLS) estimates
  - Eigenvalues less than unity, so estimated model implies covariance stationarity
$$\lambda_i^2 - \hat{A}_1\lambda_i - \hat{A}_2 = 0 \rightarrow \lambda_1 = 0.9970,\ \lambda_2 = 0.3630.$$
  - Implied mean and standard deviation in u_t:
$$\frac{\hat{A}_0}{1 - \hat{A}_1 - \hat{A}_2} = 11.88,\ .$$
- Notice: if
$$y_t = 11.88 + a_1\lambda_1^t + a_2\lambda_2^t,\ \text{any } a_1, a_2,$$
then *trend implied by initial conditions*:
$$y_t = \hat{A}_0 + \hat{A}_1 y_{t-1} + \hat{A}_2 y_{t-2},\ t \geq 1.$$
Set a_1 and a_2 to be consistent with actual y_0 and y_{-1} ($\simeq 9.5$).


Impact of Initial Conditions on Dynamics of US log, per capita real GDP
Trend implied by initial conditions (conditional MLE)
DATA
Mean (conditional MLE)
Trend implied by initial conditions (MLE)
Mean (MLE)
log(Y)
Time
12
11.5
11
10.5
10
9.5
9
1950
1960
1970
1980
1990
2000
2010
2020

Message of Application

- Illustrates how maximum of conditional likelihood is computed by OLS.
- Maximum of conditional likelihood with growing data.
  - tends to 'explain' data as emerging from covariance stationary model (roots inside unit circle).
    - related to 'Hurwicz bias', tendency for roots of VAR to shrink towards zero.
  - interprets growth as reflecting transition from unusual initial conditions.
    - initial conditions account for a very large portion of data dynamics (see previous figure).
    - most researchers view this as implausible.

Unconditional Likelihood in the Application

- Alternative: go to (unconditional) maximum likelihood:

$$p\left(Y, y_0, y_{-1} | A, \Sigma\right) = p\left(Y | y_0, y_{-1}, A, \Sigma\right) p\left(y_0, y_{-1} | A, \Sigma\right),$$

where

$$\begin{aligned}
p\left(y_0, y_{-1} | A, \Sigma\right) &= \frac{1}{2\pi} |V|^{-1/2} \exp\left[-\frac{1}{2}\zeta' V^{-1} \zeta\right], \\
\zeta &= \begin{pmatrix} y_0 - \bar{y} \\ y_{-1} - \bar{y} \end{pmatrix}, \; V = \begin{bmatrix} c(0) & c(1) \\ c(1) & c(0) \end{bmatrix}, \\
c(\tau) &= E\left(y_t - \bar{y}\right)\left(y_{t-\tau} - \bar{y}\right), \; \bar{y} = \frac{A_0}{1 - A_1 - A_2} \\
c(1) &= \frac{A_1}{1 - A_2} c(0), \\
c(0) &= \frac{\Sigma}{1 - A_1^2 - A_2^2 - 2A_1 A_2 \frac{A_1}{1 - A_2}}
\end{aligned}$$

Unconditional Likelihood in the Application

- Unconditional likelihood:

$$p\left(Y, y_0, y_{-1} | A, \Sigma\right) = p\left(Y | y_0, y_{-1}, A, \Sigma\right) p\left(y_0, y_{-1} | A, \Sigma\right),$$

  where

$$p\left(y_0, y_{-1} | A, \Sigma\right) = \frac{1}{2\pi} \left|V\right|^{-1/2} \exp\left[-\frac{1}{2}\zeta' V^{-1} \zeta\right],$$

$$\zeta = \begin{pmatrix} y_0 - \bar{y} \\ y_{-1} - \bar{y} \end{pmatrix},\ V = \begin{bmatrix} c\left(0\right) & c\left(1\right) \\ c\left(1\right) & c\left(0\right) \end{bmatrix},$$

- presence of $\zeta' V^{-1} \zeta$ penalizes the OLS strategy of 'explaining' the data based on a trend that jumps off initial conditions.
  - in this application, trend virtually completely eliminated.

Results

Conditional versus Unconditional Likelihood		
Parameter	Cond. Likelihood (OLS)	Unconditional Likelihood
$\hat{A}_0$	0.023	0.0020
$\hat{A}_1$	1.36	1.499
$\hat{A}_2$	-0.3619	-0.4993
$\frac{\hat{A}_0}{1-\hat{A}_1-\hat{A}_2}$	11.88	10.18
λ_1	0.9970	0.9996
λ_2	0.3630	0.4995
$\sqrt{\hat{\Sigma}}$	0.00876	0.00916
$c\left(0\right)$	0.03154	0.41053
$c\left(1\right)$	0.03150	0.41048

Why Does OLS Like to Extrapolate Initial Conditions?

- Answer is related to the 'Hurwicz bias'.
- OLS estimator of ρ in $y_t = \rho y_{t-1} + u_t$, with $T = 2$ observations:

$$\begin{aligned}
\hat{\rho} &= \frac{y_2 y_1 + y_1 y_0}{y_1^2 + y_0^2} = \frac{(\rho y_1 + \varepsilon_2)\, y_1 + (\rho y_0 + \varepsilon_1)\, y_0}{y_1^2 + y_0^2} \\
&= \rho + \frac{\varepsilon_2 y_1 + \varepsilon_1 y_0}{y_1^2 + y_0^2} \\
&= \rho + \left(\frac{y_1}{y_1^2 + y_0^2} \right) \varepsilon_2 + \left(\frac{y_0}{y_1^2 + y_0^2} \right) \varepsilon_1 .
\end{aligned}$$

- Standard result that OLS is BLUE (Best Linear **Unbiased** Estimator) requires right hand variables independent of error terms.
  - Assumption fails in AR representations

Why Does OLS Like to Extrapolate Initial Conditions?

- The phenomenon reflects that y_1 and ε_1 are not independent

$$E\left(\frac{y_0}{y_1^2 + y_0^2}\right)\varepsilon_1 \neq E\left(\frac{y_0}{y_1^2 + y_0^2}\right)E\varepsilon_1.$$

- Problem gets smaller as $T \to \infty$ because there is less correlation between ε_t and denominator term:

$$E\left(\frac{y_{t-1}}{\sum_{j=1}^{T} y_{j-1}^2}\right)\varepsilon_t.$$

  Note that ε_t is dependent on only a relatively small number of y_j's in the denominator.

- This Hurwicz 'bias' is pervasive in VARs.

Initial Conditions

- General tendency in BVAR literature to work with level, growing data.
  - idea is incorporated in 'random walk prior' (i.e., Minnesota prior).
  - argument in Sims-Stock-Watson (Econometrica, 1990) suggests to many that working with level data is a good idea.
- Partly because of general tendency towards levels in the literature, literature is in the habit of working with the conditional likelihood.
  - Likelihood of initial conditions not defined when roots are unity or explosive.
- Still, some people worry about tendency of conditional likelihood to make implausibly high use of initial conditions.
  - could work with growth rates.
  - alternative strategies are suggested in Giannone, Lenza and Primiceri, 2015, 'Priors for the Long Run'.

Outline

- Normal Likelihood, Illustrated with Simple AR(2) representation. (done!)
  - conditional versus unconditional likelihood.
  - maximum likelihood with level GDP data.
  - the Hurwicz bias.
- Three representations of a VAR.
  - Standard Representation
  - Matrix Representation
  - Vectorized Representation.
- Priors, posteriors and marginal likelihood
  - Dummy observations.
  - Conjugate Priors.
- Forecasting with BVARs
  - stochastic simulations, versus non-stochastic.
  - forecast probability intervals.

VAR: Standard Representation

- Let

$$\begin{aligned} y_t &\sim m \times 1 \text{ vector of data} \\ \zeta_t &\sim q \times 1 \text{ vector of (unmodeled) exogenous variables} \\ &\qquad\qquad \text{(e.g., time trend, constant, World GDP)} \\ u_t &\sim m \times 1 \text{ vector of } iid \text{ disturbances, } u_t \sim N(0, \Sigma). \end{aligned}$$

- Vector Autoregression $VAR(p)$:

$$y_t = \underbrace{A_0}_{m\times q} \xi_t + \underbrace{A_1}_{m\times m} y_{t-1} + ... + \underbrace{A_p}_{m\times m} y_{t-p} + u_t, \; t = 1, ..., T,$$

$$u_t \text{ orthogonal to } \xi_{t-s}, y_{t-1-s}, \; s \geq 0.$$

- The available data:

$$y_{1-p}, ..., y_0, y_1, ..., y_T.$$

- Generally, take initial conditions as given

$$y_{1-p}, ..., y_0.$$

VAR: Likelihood

- Likelihood of data:

$$p\left(Y, y_{1-p}, ..., y_0 | A, \Sigma, \zeta\right) =$$

$$\overbrace{p\left(y_T | y_{T-1}, ..., y_{T-p}, A, \Sigma, \zeta\right) \times \cdots \times p\left(y_1 | y_0, ..., y_{-p}, A, \Sigma, \zeta\right)}^{\text{`conditional likelihood' (conditional on initial conditions and } \zeta\text{)}}$$

$$\times \qquad \overbrace{p\left(y_0, ..., y_{-p} | A, \Sigma, \zeta\right)}^{\text{likelihood of initial conditions (conditional on } \zeta\text{)}} \qquad ,$$

where the analysis is always conditioned on the exogenous variables, ζ :

$$\zeta = \begin{pmatrix} \zeta_T \\ \vdots \\ \zeta_{-p} \end{pmatrix}$$

- From here on, conditioning on ζ is taken for granted and not even included explicitly in the notation.

VAR: Likelihood

- First, let

$$\underbrace{x_t}_{k\times 1} = \begin{pmatrix} \zeta_t \\ y_{t-1} \\ \vdots \\ y_{t-p} \end{pmatrix}, \; t = 1, 2, ..., T, \; k \equiv q + pm$$

- Then,

$$y_t = A' x_t + u_t,$$

where

$$A' = \underbrace{\begin{bmatrix} A_0 & A_1 & \cdots & A_p \end{bmatrix}}_{m\times k}.$$

- Notice:

$$\begin{aligned} & p\left(y_t | x_t, A, \Sigma\right) \\ = & \frac{1}{(2\pi)^{\frac{m}{2}}} \left|\Sigma\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(y_t - A' x_t\right)' \Sigma^{-1} \left(y_t - A' x_t\right)\right] \end{aligned}$$

VAR: Likelihood

- Conditional likelihood of Y :

$$
\begin{aligned}
& p\left(Y|x_1, A, \Sigma\right) \\
= & \frac{1}{(2\pi)^{\frac{mT}{2}}} |\Sigma|^{-\frac{T}{2}} \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\left(y_t - A'x_t\right)'\Sigma^{-1}\left(y_t - A'x_t\right)\right]
\end{aligned}
$$

- From now on, drop the notation, x_1, to avoid clutter.
- Now, for a little matrix algebra....

Trace of a Matrix

- Trace of a square matrix, A :

$$tr\left[A\right] = \sum_{i} a_{ii}.$$

- Properties of trace:
  - *cyclic property* of trace: if A, B, C are (conformable) matrices, then

    $$tr\left(ABC\right) = tr\left(CAB\right) = tr\left(BCA\right).$$

    - example: if a is a $n \times 1$ vector and B is $n \times n$, then, by cyclic property,

      $$a'Ba = tr\left[a'Ba\right] = tr\left[aa'B\right] = tr\left[Baa'\right]$$

  - *linearity* property of trace:

    $$tr\left[A+B\right] = tr\left[A\right] + tr\left[B\right].$$

VAR: Likelihood

- Conditional likelihood of Y :

$$
\begin{aligned}
& p\left(Y \mid A, \Sigma\right) \\
&= \frac{1}{(2\pi)^{\frac{mT}{2}}} |\Sigma|^{-\frac{T}{2}} \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\left(y_t - A'x_t\right)'\Sigma^{-1}\left(y_t - A'x_t\right)\right] \\
&= \frac{1}{(2\pi)^{\frac{mT}{2}}} |\Sigma|^{-\frac{T}{2}} \exp\left[-\frac{1}{2}\sum_{t=1}^{T} tr\left[\left(y_t - A'x_t\right)'\Sigma^{-1}\left(y_t - A'x_t\right)\right]\right] \\
&= \frac{1}{(2\pi)^{\frac{mT}{2}}} |\Sigma|^{-\frac{T}{2}} \exp\left[-\frac{1}{2}\sum_{t=1}^{T} tr\left[\left(y_t - A'x_t\right)\left(y_t - A'x_t\right)'\Sigma^{-1}\right]\right] \\
&= \frac{1}{(2\pi)^{\frac{mT}{2}}} |\Sigma|^{-\frac{T}{2}} \exp\left[-\frac{1}{2}\sum_{t=1}^{T} tr\left[\Sigma^{-1}\left(y_t - A'x_t\right)\left(y_t - A'x_t\right)'\right]\right] \\
&= \frac{1}{(2\pi)^{\frac{mT}{2}}} |\Sigma|^{-\frac{T}{2}} \exp\left[-\frac{1}{2} tr\left[\Sigma^{-1}\sum_{t=1}^{T}\left(y_t - A'x_t\right)\left(y_t - A'x_t\right)'\right]\right]
\end{aligned}
$$

VAR: Matrix Representation

- Define

$$\underbrace{Y}_{T\times m} = \begin{pmatrix} y_1' \\ \vdots \\ y_T' \end{pmatrix}, \quad \underbrace{X}_{T\times k} = \begin{bmatrix} x_1' \\ \vdots \\ x_T' \end{bmatrix}.$$

- 'Standard VAR' representation:

$$y_t = A' x_t + u_t$$

- Transpose it:

$$y_t' = x_t' A + u_t'$$

- Then, 'stack':

$$Y = XA + U.$$

VAR: Likelihood in Matrices

- Representation in matrix form:

$$\sum_{t=1}^{T} \left(y_t - A'x_t\right)\left(y_t - A'x_t\right)' = \sum_{t=1}^{T} \left(y_t - A'x_t\right)\left(y_t' - x_t'A\right)$$

$$= \left[\begin{array}{ccc} \left(y_1 - A'x_1\right) & \cdots & \left(y_T - A'x_T\right) \end{array}\right] \left[\begin{array}{c} y_1' - x_1'A \\ \vdots \\ y_T' - x_T'A \end{array}\right]$$

$$= \left(Y - XA\right)'\left(Y - XA\right)$$

- So, matrix representation of VAR and (conditional) likelihood:

$$Y = XA + U$$

$$p\left(Y|A,\Sigma\right)$$

$$= \frac{1}{\left(2\pi\right)^{\frac{mT}{2}}} \left|\Sigma\right|^{-\frac{T}{2}} \exp\left[-\frac{1}{2} tr\left[\Sigma^{-1}\left(Y - XA\right)'\left(Y - XA\right)\right]\right]$$

- With the likelihood in hand, we now move on to the priors.

Priors and Posteriors

- Use Bayes' rule and priors to compute posterior distribution.
- Identities:

$$p(Y, \Sigma, A) = p(Y|\Sigma, A)\, p(\Sigma, A) = p(\Sigma, A|Y)\, p(Y),$$

so that

$$\text{Bayes' rule: } \overbrace{p(\Sigma, A|Y)}^{\text{posterior}} = \frac{\overbrace{p(Y|\Sigma, A)}^{\text{likelihood}}\overbrace{p(\Sigma, A)}^{\text{prior}}}{p(Y)}.$$

- Will work with 'conjugate prior', $p(\Sigma, A)$
  - $p(\Sigma, A|Y)$ is the same density as $p(\Sigma, A)$
- To find a conjugate prior, it is convenient to notice that a likelihood, $p(Y|\Sigma, A)$, can be rewritten so that it looks like a density function for A.

Rewriting the Likelihood

- OLS estimator of A and sum, squared residuals, $\hat{S}$:

$$\hat{A} \equiv (X'X)^{-1} X'Y,\ \hat{S} \equiv \left(Y - X\hat{A}\right)' \left(Y - X\hat{A}\right)$$

- Orthogonality property of OLS:

$$\begin{aligned} X' \left[Y - X\hat{A}\right] &= X' \left[I - X (X'X)^{-1} X'\right] Y \\ &= \left[X' - X'X (X'X)^{-1} X'\right] Y = 0 \end{aligned}$$

- Orthogonality implies:

$$\begin{aligned} & (Y - XA)' (Y - XA) \\ = \ & \left(Y - X\hat{A} + X\left(\hat{A} - A\right)\right)' \left(Y - X\hat{A} + X\left(\hat{A} - A\right)\right) \\ \overbrace{=}^{\text{orthogonality}} \ & \left(Y - X\hat{A}\right)' \left(Y - X\hat{A}\right) + \left(\hat{A} - A\right)' X'X \left(\hat{A} - A\right) \\ = \ & \hat{S} + \left(A - \hat{A}\right)' X'X \left(A - \hat{A}\right) \end{aligned}$$

Rewriting the Likelihood

- The previous results imply:

$$p\left(Y|A,\Sigma\right)$$

$$= \frac{1}{(2\pi)^{\frac{mT}{2}}} \left|\Sigma\right|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\left(Y-XA\right)'\left(Y-XA\right)\right]\right\}$$

$$= \frac{1}{(2\pi)^{\frac{mT}{2}}} \left|\Sigma\right|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\hat{S}\right]\right\}$$

$$\times \exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\left(A-\hat{A}\right)'X'X\left(A-\hat{A}\right)\right]\right\}$$

- The likelihood looks more and more like a distribution for A!
  - Just one more step...

Vectorization and Kronecker Product

- Kronecker product:

$$\begin{aligned} A &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \rightarrow A \otimes B \equiv \begin{bmatrix} a_{11}B & a_{12}B \\ a_{21}B & a_{22}B \end{bmatrix} \\ &\rightarrow \ (A \otimes B)' = A' \otimes B', \ (A \otimes B)^{-1} = A^{-1} \otimes B^{-1}. \end{aligned}$$

- Let the i^{th} column of the $m \times n$ matrix A be denoted by a_i, $i = 1, ..., n$:

$$A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \rightarrow vec\,(A) \equiv \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

Then,

$$\rightarrow tr\left[A'BCD'\right] = vec\,(A)'\,(D \otimes B)\,vec\,(C)$$

Rewriting the Likelihood

- Let

$$a = vec\left(A\right),\ \hat{a} = vec\left(\hat{A}\right)$$

- Then,

$$\begin{aligned}
& tr\left[\underbrace{\Sigma^{-1}}_{m\times m}\underbrace{\left(A-\hat{A}\right)'}_{m\times k}\underbrace{X'X}_{k\times k}\underbrace{\left(A-\hat{A}\right)}_{k\times m}\right] \\
= \;& tr\left[\left(A-\hat{A}\right)' X'X\left(A-\hat{A}\right)\Sigma^{-1}\right] \\
= \;& \left(a-\hat{a}\right)'\left(\Sigma^{-1}\otimes X'X\right)\left(a-\hat{a}\right) \\
= \;& \left(a-\hat{a}\right)'\left(\Sigma\otimes\left(X'X\right)^{-1}\right)^{-1}\left(a-\hat{a}\right)
\end{aligned}$$

- This looks a lot like the exponential term in the Normal distribution!

Rewriting the Likelihood

- Likelihood

$$p\left(Y|A,\Sigma\right)$$

$$= \frac{1}{(2\pi)^{\frac{mT}{2}}} |\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}\hat{S}\right]\right\}$$
$$\times \exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}\left(A-\hat{A}\right)' X'X \left(A-\hat{A}\right)\right]\right\}$$

$$= \frac{1}{(2\pi)^{\frac{mT}{2}}} |\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}\hat{S}\right]\right\}$$
$$\times \exp\left\{-\frac{1}{2}\left(a-\hat{a}\right)' \left(\Sigma \otimes \left(X'X\right)^{-1}\right)^{-1} \left(a-\hat{a}\right)\right\}$$

- Payoff: suppose that Σ is known and $p\left(A|\Sigma\right) =$ constant (**flat prior**), then posterior of a is $N\left(\hat{a}, \Sigma \otimes \left(X'X\right)^{-1}\right)$.

VAR: Vectorized Form

- VAR in Matrix form:

$$Y = XA + U.$$

- Matrix fact:

$$vec\left(ABC\right) = \left(C' \otimes A\right) vec\left(B\right),$$

so,

$$vec\left(XA\right) = vec\left(\underbrace{X}_{T\times k}\underbrace{A}_{k\times m} I_m\right) = \left(I_m \otimes X\right) a$$

- Then,

$$\boxed{y = \left(I_m \otimes X\right) a + u,\ u \sim N\left(0, \Sigma \otimes I_T\right),}$$

$$y \equiv vec\left(Y\right),\ u \equiv vec\left(U\right).$$

VAR: Vectorized Form

- VAR -

$$y = (I_m \otimes X)\, a + u,\ u \ \sim\ N\left(0, \Sigma \otimes I_T\right).$$

- OLS:

$$\begin{aligned} \hat{a} &= \left[(I_m \otimes X)' (I_m \otimes X)\right]^{-1} (I_m \otimes X)' y \\ &= a + \left[(I_m \otimes X)' (I_m \otimes X)\right]^{-1} (I_m \otimes X)' u. \end{aligned}$$

  – Classical (asymptotic) sampling theory for $\hat{a}$ is Normal with

$$\begin{array}{rl} \text{mean:} & a \\ \text{variance:} & \left[(I_m \otimes X)' (I_m \otimes X)\right]^{-1} (I_m \otimes X)' \\ & \times \overbrace{Euu'}^{=\Sigma \otimes I_T} (I_m \otimes X) \left(\left[(I_m \otimes X)' (I_m \otimes X)\right]^{-1}\right)' \end{array}$$

VAR: Vectorized Form

- Matrix facts:

$$\begin{aligned}(A\otimes B)' &= A'\otimes B',\ \ (A\otimes B)^{-1}=A^{-1}\otimes B^{-1},\\ (A\otimes B)(C\otimes D) &= (AC\otimes BD)\,,\ \ (AB)'=B'A'\end{aligned}$$

- Then,

$$\begin{aligned}&\left[(I_m\otimes X)'(I_m\otimes X)\right]^{-1}(I_m\otimes X)'(\Sigma\otimes I_T)\\ &\times(I_m\otimes X)\left(\left[(I_m\otimes X)'(I_m\otimes X)\right]^{-1}\right)'\\ =\ &\Sigma\otimes\left(X'X\right)^{-1}\end{aligned}$$

- This is a heuristic demonstration of the large sample result that

$$\hat{a}\ \sim\ \mathcal{N}\left(a,\Sigma\otimes\left(X'X\right)^{-1}\right)$$

- Interesting to compare with Bayesian posterior with flat prior:

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

Where we Now Stand: Three Representations of VAR

- Standard representation and likelihood:

$$y_t = A_0 \xi_t + A_1 y_{t-1} + ... + A_p y_{t-p} + u_t, \; E u_t u_t' = \Sigma.$$

$$p(Y|A,\Sigma)$$

$$= \frac{1}{(2\pi)^{\frac{mT}{2}}} |\Sigma|^{-\frac{T}{2}} \exp\left[-\frac{1}{2} tr \left[\Sigma^{-1} \sum_{t=1}^{T} (y_t - A' x_t)(y_t - A' x_t)' \right] \right]$$

- Matrix representation and likelihood:

$$Y = XA + U$$

$$p(Y|A,\Sigma)$$

$$= \frac{1}{(2\pi)^{\frac{mT}{2}}} |\Sigma|^{-\frac{T}{2}} \exp\left\{ -\frac{1}{2} tr \left[\Sigma^{-1} \hat{S} \right] \right\}$$

$$\times \exp\left\{ -\frac{1}{2} tr \left[\Sigma^{-1} (A - \hat{A})' X' X (A - \hat{A}) \right] \right\}$$

Where we Now Stand: Three Representations of VAR

- Finally: Vectorized representation and likelihood -

$$
\begin{gathered}
y = (I_m \otimes X)\, a + u,\ u \ \sim\ N\left(0, \Sigma \otimes I_T\right) \\
p\left(Y|A, \Sigma\right) \\
= \frac{1}{(2\pi)^{\frac{mT}{2}}} \left|\Sigma\right|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}\hat{S}\right]\right\} \\
\times \exp\left\{-\frac{1}{2}\left(a - \hat{a}\right)' \left(\Sigma \otimes \left(X'X\right)^{-1}\right)^{-1} \left(a - \hat{a}\right)\right\}
\end{gathered}
$$

- Key insight: this likelihood has the shape of $\mathcal{N}\left(\hat{a}, \Sigma \otimes \left(X'X\right)^{-1}\right)$.

Outline

- Normal Likelihood, Illustrated with Simple AR(2) representation. (done!)
  - conditional versus unconditional likelihood.
  - maximum likelihood with level GDP data.
  - the Hurwicz bias.
- Three representations of a VAR. (done!)
  - Standard Representation
  - Matrix Representation
  - Vectorized Representation.
- Priors, posteriors and marginal likelihood
  - Dummy observations.
  - Conjugate Priors.
- Forecasting with BVARs
  - stochastic simulations, versus non-stochastic.
  - forecast probability intervals.

Priors for VARs

- Priors designed based on insight in the vectorized representation of VAR.
- Example: suppose (for now) that Σ is known and $p(A|\Sigma) = c$ ('uninformative prior').
- Then,

$$p(A|Y,\Sigma) \propto p(Y|A,\Sigma)\, p(A|\Sigma)$$
$$= \frac{1}{(2\pi)^{\frac{mT}{2}}} |\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}\hat{S}\right]\right\}$$
$$\times \exp\left\{-\frac{1}{2}(a-\hat{a})'\left(\Sigma \otimes (X'X)^{-1}\right)^{-1}(a-\hat{a})\right\} c \quad ,$$

where $\propto$ means 'is proportional to'.

- We now turn to an influential class of priors for A, constructed using 'dummy observations'.

Priors and Dummy Observations

- Suppose we have $\bar{T}$ dummy observations, $(\bar{Y}, \bar{X})$.
- Consider the following 'likelihood' for the dummy observations:

$$p(\bar{Y}|A,\Sigma) = \frac{1}{(2\pi)^{\frac{m\bar{T}}{2}}} |\Sigma|^{-\frac{\bar{T}}{2}}$$

$$\times \exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\sum_{j=1}^{\bar{T}}\left(\bar{y}_j - A'\bar{x}_j\right)\left(\bar{y}_j - A'\bar{x}_j\right)'\right]\right\}.$$

- In vectorized form,

$$\bar{A} \equiv \left(\bar{X}'\bar{X}\right)^{-1}\bar{X}'\bar{Y},\ \bar{S} \equiv \left(\bar{Y} - \bar{X}\bar{A}\right)'\left(\bar{Y} - \bar{X}\bar{A}\right),$$

$$p(\bar{Y}|A,\Sigma) = \frac{1}{(2\pi)^{\frac{m\bar{T}}{2}}} |\Sigma|^{-\frac{\bar{T}}{2}} \exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\bar{S}\right]\right\}$$

$$\times \exp\left\{-\frac{1}{2}\left(a - \bar{a}\right)'\left(\Sigma \otimes \left(\bar{X}'\bar{X}\right)^{-1}\right)^{-1}\left(a - \bar{a}\right)\right\}.$$

- Prior distribution, $a \sim \mathcal{N}\left(\bar{a}, \Sigma \otimes \left(\bar{X}'\bar{X}\right)^{-1}\right).$

Dummy Observations and Posterior

Multiply $p(Y|A,\Sigma)$ (the likelihood of the data) times $p(\bar{Y}|A,\Sigma)$ (something proportional to a Normal prior for A) :

$$
\begin{aligned}
p(A|\Sigma, Y) &\propto p(Y|A,\Sigma)p(\bar{Y}|A,\Sigma) \\
&= \frac{1}{(2\pi)^{\frac{m(T+\bar{T})}{2}}} |\Sigma|^{-\frac{T+\bar{T}}{2}} \\
&\times \exp\left[-\frac{1}{2}tr\left[\Sigma^{-1}\sum_{t=1}^{T}\left(y_t - A'x_t\right)\left(y_t - A'x_t\right)'\right]\right] \\
&\times \exp\left[-\frac{1}{2}tr\left[\Sigma^{-1}\sum_{j=1}^{\bar{T}}\left(\bar{y}_j - A'\bar{x}_j\right)\left(\bar{y}_j - A'\bar{x}_j\right)'\right]\right]
\end{aligned}
$$

We have $\propto$ here because (i) we want the Normal prior for A which is only proportional to $p(\bar{Y}|A,\Sigma)$ and (ii) we need to divide by $p\left(Y|\Sigma\right)$.

Dummy Observations and Posterior

Collect terms in A (using linearity of $tr[\cdot]$)

$$tr[\Sigma^{-1}(\sum_{t=1}^{T}(y_t - A'x_t)(y_t - A'x_t)'$$

$$+ \sum_{j=1}^{\bar{T}}(\bar{y}_j - A'\bar{x}_j)(\bar{y}_j - A'\bar{x}_j)')]$$

$$= tr\left[\Sigma^{-1}(\underline{X} - \underline{Y}A)(\underline{X} - \underline{Y}A)'\right]$$

where

$$\underline{X} = \begin{bmatrix} x_1' \\ \vdots \\ x_T' \\ \bar{x}_1' \\ \vdots \\ \bar{x}_{\bar{T}}' \end{bmatrix} = \underbrace{\begin{bmatrix} X \\ \bar{X} \end{bmatrix}}_{(T+\bar{T})\times k}, \; \underline{Y} = \begin{bmatrix} y_1' \\ \vdots \\ y_T' \\ \bar{y}_1' \\ \vdots \\ \bar{y}_{\bar{T}}' \end{bmatrix} = \underbrace{\begin{bmatrix} Y \\ \bar{Y} \end{bmatrix}}_{(T+\bar{T})\times m}.$$

Dummy Observations and Posterior

- Mapping all the way to exponential representation:

$$
\begin{aligned}
& tr\left[\Sigma^{-1}\left(\underline{Y}-\underline{X}A\right)\left(\underline{Y}-\underline{X}A\right)'\right] \\
= \; & tr\left[\Sigma^{-1}\left(\underline{S}+\left(A-\underline{A}\right)'\underline{X}'\underline{X}\left(A-\underline{A}\right)\right)\right] \\
= \; & tr\left[\Sigma^{-1}\underline{S}\right]+tr\left[\left(A-\underline{A}\right)'\underline{X}'\underline{X}\left(A-\underline{A}\right)\right] \\
= \; & tr\left[\Sigma^{-1}\underline{S}\right]+\left(a-\underline{a}\right)'\left(\Sigma\otimes\left(\underline{X}'\underline{X}\right)^{-1}\right)^{-1}\left(a-\underline{a}\right)
\end{aligned}
$$

where

$$
\begin{aligned}
\underline{A} &= \left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\underline{Y} \\
\underline{S} &= \left(\underline{Y}-\underline{X}\underline{A}\right)'\left(\underline{Y}-\underline{X}\underline{A}\right).
\end{aligned}
$$

Dummy Observations and Posterior

- The posterior distribution is proportional to:

$$p(Y|A,\Sigma)p(\bar{Y}|A,\Sigma) = \frac{1}{(2\pi)^{\frac{m(T+\bar{T})}{2}}} |\Sigma|^{-\frac{T+\bar{T}}{2}} \exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\underline{S}\right]\right\}$$
$$\times \exp\left[-\frac{1}{2}(a-\underline{a})'\left(\Sigma\otimes\left(\underline{X}'\underline{X}\right)^{-1}\right)^{-1}(a-\underline{a})\right]$$

- Thus, we see that the posterior distribution of a, $p(A|\Sigma, Y)$, is Normal and:

$$p(A|\Sigma, Y) \propto p(Y|A,\Sigma)p(\bar{Y}|A,\Sigma).$$

Interpreting the Posterior

- Note

$$\begin{aligned}\underline{A} &= \left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\underline{Y} \\ &= \left(X'X+\bar{X}'\bar{X}\right)^{-1} \\ &\quad\times\left[X'X\overbrace{\left[\left(X'X\right)^{-1}X'Y\right]}^{\hat{A}}+\bar{X}'\bar{X}\overbrace{\left[\left(\bar{X}'\bar{X}\right)^{-1}\bar{X}'\bar{Y}\right]}^{\bar{A}}\right],\end{aligned}$$

so that the posterior mean of A, $\underline{A}$, is a weighted average of what the data say, $\hat{A}$, and the prior, $\bar{A}$.

Simple VAR(2), m=2

- Standard VAR representation:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \overbrace{\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}}^{A_0} + \overbrace{\begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix}}^{A_1} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix}$$

$$+ \overbrace{\begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}}^{A_2} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}$$

$$A = \begin{bmatrix} A_0' \\ A_1' \\ A_2' \end{bmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 \\ \beta_{11} & \beta_{21} \\ \beta_{12} & \beta_{22} \\ \gamma_{11} & \gamma_{21} \\ \gamma_{12} & \gamma_{22} \end{bmatrix}$$

Simple VAR(2), m=2

- Matrix representation:

$$\underbrace{x_t}_{5\times 1} = \begin{pmatrix} 1 \\ y_{1,t-1} \\ y_{2,t-1} \\ y_{1,t-2} \\ y_{2,t-2} \end{pmatrix}, \; X = \begin{bmatrix} x_1' \\ \vdots \\ x_T' \end{bmatrix}, \; Y = \begin{bmatrix} y_{1,1} & y_{2,1} \\ \vdots & \vdots \\ y_{1,T} & y_{2,T} \end{bmatrix}$$

- Then,

$$\underbrace{Y}_{T\times 2} = \underbrace{X}_{T\times 5} \begin{bmatrix} \alpha_1 & \alpha_2 \\ \beta_{11} & \beta_{21} \\ \beta_{12} & \beta_{22} \\ \gamma_{11} & \gamma_{21} \\ \gamma_{12} & \gamma_{22} \end{bmatrix} + \underbrace{U}_{T\times 2}.$$

Minnesota Prior

- Very clever!
- Basic idea: each variable is a scalar 1^{st} order autoregression:
  - $y_{i,t} = \beta_{ii} y_{i,t-1} + u_{i,t}$, $\beta_{ii} \sim \mathcal{N}\left(\phi_i, \frac{\Sigma_{ii}}{\lambda_1^2 s_i^2}\right)$
  - $y_{i,t} = \beta_{ij} y_{j,t-1} + u_{i,t}$, $\beta_{ij} \sim \mathcal{N}\left(0, \frac{\Sigma_{ii}}{\lambda_1^2 s_j^2}\right)$, $j \neq i$
  - $\lambda_1 \sim$ 'overall tightness parameter'
  - $s_i \sim$ 'scaling parameter on coefficient on $y_{j,t-1}$'
- Parameter, ϕ_i :
  - if $y_{i,t}$ is in levels, then $\phi_i = 1$ (random walk).
  - if $y_{i,t}$ is in first difference, then $\phi_i = 0$ (again, random walk).
  - could have $\phi_i \neq 1$.
- Analogous restrictions on lags $2, ..., p$ parameters.
  - Prior assumes that the data has less information on parameters at higher order lags.

Minnesota Prior

- Each variable follows a simple 1^{st} order scalar autoregression.
  - Motivation: it has been found that such models (especially, random walk) perform well in forecasting.
  - Although the prior is that data dynamics are quite simple, this need not be the case in the posterior when $\lambda_1, s_i < \infty$.
  - if the data *really* want a lot of interaction, the posterior will show that.
- Note: the variance of the prior is proportional to Σ_{ii}/s_j^2.
  - Motivation: the numerator is related to the volatility of $y_{i,t}$ and the denominator is (actually, *will* be) related to the volatility of $y_{j,t}$.
    - It is perhaps intuitively appealing that the confidence or strength of belief in the prior that β_{ij} is close to zero is stronger the more variable $y_{j,t}$ is, relative to $y_{i,t}$.
    - Imagine you feel β_{ij} is close to zero, $i \neq j$, and you see $y_{j,t}$ is highly variable while $y_{i,t}$ is not. This would reinforce your belief that $y_{j,t}$ has no impact on $y_{i,t}$.

Minnesota Prior

- Dummy observations for A :

$$\overbrace{\begin{bmatrix} \phi_1\lambda_1 s_1 & 0 \\ 0 & \phi_2\lambda_1 s_2 \end{bmatrix}}^{\bar{Y}_1} = \overbrace{\begin{bmatrix} 0 & \lambda_1 s_1 & 0 & 0 & 0 \\ 0 & 0 & \lambda_1 s_2 & 0 & 0 \end{bmatrix}}^{\bar{X}_1} \overbrace{\begin{bmatrix} \alpha_1 & \alpha_2 \\ \beta_{11} & \beta_{21} \\ \beta_{12} & \beta_{22} \\ \gamma_{11} & \gamma_{21} \\ \gamma_{12} & \gamma_{22} \end{bmatrix}}^{A}$$

$$+ \overbrace{\begin{bmatrix} u_{1,1} & u_{2,1} \\ u_{1,2} & u_{2,2} \end{bmatrix}}^{\bar{U}_1}$$

where λ_1 is an 'overall tightness' parameter; s_i is a tightness parameter that applies to the i^{th} equation; ϕ_i prior on parameter on own first lag of $y_{i,t}$.

Implications of Minnesota Prior

- 1,1 and 1,2 elements of system, $\bar{Y}_1 = \bar{X}_1 A + \bar{U}_1$:

$$\begin{aligned}
\phi_1 \lambda_1 s_1 &= \lambda_1 s_1 \beta_{11} + u_{1,1} \rightarrow \beta_{11} = \phi_1 - \frac{u_{1,1}}{\lambda_1 s_1} \\
&\rightarrow \beta_{11} \sim \mathcal{N}\left(\phi_1, \frac{\Sigma_{11}}{\lambda_1^2 s_1^2}\right) \\
0 &= \lambda_1 s_1 \beta_{21} + u_{2,1} \rightarrow \beta_{21} = 0 - \frac{u_{2,1}}{\lambda_1 s_1} \\
&\rightarrow \beta_{21} \sim \mathcal{N}\left(0, \frac{\Sigma_{22}}{\lambda_1^2 s_1^2}\right)
\end{aligned}$$

- Similarly, 2,2 and 2,1 elements imply:

$$\beta_{22} \sim \mathcal{N}\left(\phi_2, \frac{\Sigma_{22}}{\lambda_1^2 s_2^2}\right),\ \beta_{12} \sim \mathcal{N}\left(0, \frac{\Sigma_{11}}{\lambda_1^2 s_2^2}\right).$$

Minnesota Prior

- Dummy observations for A_l, $l > 1$:

$$\overbrace{\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}}^{\bar{Y}_2} = \overbrace{\begin{bmatrix} 0 & 0 & 0 & \lambda_1 s_1 l^{\lambda_2} & 0 \\ 0 & 0 & 0 & 0 & \lambda_1 s_2 l^{\lambda_2} \end{bmatrix}}^{\bar{X}_2} \overbrace{\begin{bmatrix} \alpha_1 & \alpha_2 \\ \beta_{11} & \beta_{21} \\ \beta_{12} & \beta_{22} \\ \gamma_{11} & \gamma_{21} \\ \gamma_{12} & \gamma_{22} \end{bmatrix}}^{A} + \bar{U}_2$$

$$\begin{aligned} \gamma_{11} &\sim \mathcal{N}\left(0, \frac{\Sigma_{11}}{\lambda_1^2 s_1^2 l^{2\lambda_2}}\right), \ \gamma_{21} \sim \mathcal{N}\left(0, \frac{\Sigma_{22}}{\lambda_1^2 s_1^2 l^{2\lambda_2}}\right), \\ \gamma_{12} &\sim \mathcal{N}\left(0, \frac{\Sigma_{11}}{\lambda_1^2 s_2^2 l^{2\lambda_2}}\right), \ \gamma_{22} \sim \mathcal{N}\left(0, \frac{\Sigma_{22}}{\lambda_1^2 s_2^2 l^{2\lambda_2}}\right), \end{aligned}$$

- Hyperparameter, $\lambda_2 > 0$, controls the amount of prior information at higher lags.
  - Bigger λ_2 or $l \sim$ more information in the prior at higher lags.
  - Prior is that there is relatively little info in data about high

Own-persistence Dummies

If $y_{i,t}$ has been stable at some level, $\bar{y}_i$, $i = 1, 2$, it tends to stay there:

$$\overbrace{\begin{bmatrix} \lambda_3\bar{y}_1 & 0 \\ 0 & \lambda_3\bar{y}_2 \end{bmatrix}}^{\bar{Y}_3} = \overbrace{\begin{bmatrix} 0 & \lambda_3\bar{y}_1 & 0 & \lambda_3\bar{y}_1 & 0 \\ 0 & 0 & \lambda_3\bar{y}_2 & 0 & \lambda_3\bar{y}_2 \end{bmatrix}}^{\bar{X}_3} \overbrace{\begin{bmatrix} \alpha_1 & \alpha_2 \\ \beta_{11} & \beta_{21} \\ \beta_{12} & \beta_{22} \\ \gamma_{11} & \gamma_{21} \\ \gamma_{12} & \gamma_{22} \end{bmatrix}}^{A}$$

$$+\bar{U}_3$$

$$\rightarrow \quad \lambda_3\bar{y}_1 = \lambda_3\bar{y}_1\beta_{11} + \lambda_3\bar{y}_1\gamma_{11} + u_{1,1}$$

$$\rightarrow \quad \beta_{11} + \gamma_{11} = 1 - \frac{u_{1,1}}{\lambda_3\bar{y}_1} \rightarrow (\beta_{11} + \gamma_{11}) \sim \mathcal{N}\left(1, \frac{\Sigma_{11}}{\lambda_3^2\bar{y}_1^2}\right)$$

Interpretation of Own-persistence Dummies

- Suppose
$$y_t = \beta_{11} y_{t-1} + \gamma_{11} y_{t-2} + u_t,$$
with
$$1 = \beta_{11} + \gamma_{11} \rightarrow \beta_{11} = 1 - \gamma_{11},$$
so that
$$y_t = (1 - \gamma_{11})\, y_{t-1} + \gamma_{11} y_{t-2} + u_t$$
or,
$$y_t - y_{t-1} = -\gamma_{11} \left(y_{t-1} - y_{t-2} \right) + u_t.$$
- Own-persistence is a generalization on random walk.
  - random walk: first differences not autocorrelated, but stationary.
  - sum of coefficients = unity: first differences are autocorrelated.
- example: US GDP looks like
$$\Delta y_t = 0.4 \Delta y_{-1} + u_t,\ \gamma_{11} = -0.4.$$

Co-persistence Dummies

- If $(y_{1,t}, y_{2,t})$ have been persistent at $(\bar{y}_1, \bar{y}_2)$ they tend to stay there:

$$\overbrace{\begin{bmatrix} \lambda_4\bar{y}_1 & \lambda_4\bar{y}_2 \end{bmatrix}}^{\bar{Y}_4} = \overbrace{\begin{bmatrix} \lambda_4 & \lambda_4\bar{y}_1 & \lambda_4\bar{y}_2 & \lambda_4\bar{y}_1 & \lambda_4\bar{y}_2 \end{bmatrix}}^{\bar{X}_4} A + \bar{U}_4$$

- This implies:

$$\begin{aligned}
\lambda_4\bar{y}_1 &= \lambda_4\bar{y}_1\beta_{11} + \lambda_4\bar{y}_2\beta_{12} + \lambda_4\bar{y}_1\gamma_{11} + \lambda_4\bar{y}_2\gamma_{12} + \lambda_4\alpha_1 + u_1 \\
&\rightarrow \bar{y}_1(1 - \beta_{11} - \gamma_{11}) = \alpha_1 + \bar{y}_2(\beta_{12} + \gamma_{12}) + \frac{u_1}{\lambda_4} \\
\lambda_4\bar{y}_2 &= \lambda_4\bar{y}_1\beta_{21} + \lambda_4\bar{y}_2\beta_{22} + \lambda_4\bar{y}_1\gamma_{21} + \lambda_4\bar{y}_2\gamma_{22} + \lambda_4\alpha_2 + u_2 \\
&\rightarrow \bar{y}_2(1 - \beta_{22} - \gamma_{22}) = \alpha_2 + \bar{y}_1(\beta_{21} + \gamma_{21}) + \frac{u_2}{\lambda_4}
\end{aligned}$$

Dummy Priors

- Set them up like this:

$$\bar{Y} = \begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \\ \bar{Y}_3 \\ \bar{Y}_4 \end{bmatrix}, \; \bar{X} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \bar{X}_3 \\ \bar{X}_4 \end{bmatrix}, \; \bar{U} = \begin{bmatrix} \bar{U}_1 \\ \bar{U}_2 \\ \bar{U}_3 \\ \bar{U}_4 \end{bmatrix}.$$

- Pad the Y and X vectors with the 'observations', $\bar{Y}$ and $\bar{X}$:

$$\underline{Y} = \begin{bmatrix} Y \\ \bar{Y} \end{bmatrix}, \underline{X} = \begin{bmatrix} X \\ \bar{X} \end{bmatrix}.$$

Prior for Variance-Covariance Matrix

- Up to now, we've focused on the prior and posterior for the VAR parameters in A.
- We've supposed that the analyst 'knows' the value of Σ.
- Next, we consider the more plausible case that the analyst also does not know Σ.

Inverse Wishart Prior for Variance-Covariance Matrix

- Trick is to find $p(\Sigma)$ that is 'sensible' and convenient, i.e., conjugate with the likelihood.
- Inverse Wishart distribution for Σ, $\mathcal{IW}(S, \nu)$:

$$p(\Sigma) = \frac{|S|^{\nu/2}}{2^{\nu m} \prod_{i=1}^{m} \Gamma\left[\frac{\nu+1-i}{2}\right]} |\Sigma|^{-\frac{\nu+m+1}{2}} \exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1} S\right]\right\},$$

where Γ denotes the gamma function.

  - Inverse Wishart distribution, $\mathcal{IW}(\nu, S)$, with 'degrees of freedom', ν, and 'scale matrix' S.

Properties of Inverse Wishart

- Looks like inverse of Chi-square distribution:
  - Draw ν vectors $Z_1, ..., Z_\nu$ from $\mathcal{N}\left(0, S^{-1}\right)$, and:

$$\Sigma = \left[Z_1 Z_1' + ... + Z_\nu Z_\nu'\right]^{-1}.$$

  Nice: (i) Σ is guaranteed to be positive definite for ν big enough, (ii) trace and determinant terms in $\mathcal{IW}\left(S, \nu\right)$ match up with analogous terms in rewritten Normal likelihood.

- Property:

$$\text{mean, } \Sigma = \frac{S}{\nu - (m+1)}, \text{ mode, } \Sigma = \frac{S}{\nu + (m+1)}$$

Recall

- We previously derived:

$$\overbrace{p(Y|A,\Sigma)p(\bar{Y}|A,\Sigma)}^{\propto p(A|Y,\Sigma)} = \frac{1}{(2\pi)^{\frac{m(T+\bar{T})}{2}}} |\Sigma|^{-\frac{T+\bar{T}}{2}} \exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}\underline{S}\right]\right\}$$
$$\times \exp\left[-\frac{1}{2}\left(a-\underline{a}\right)'\left(\Sigma \otimes \left(\underline{X}'\underline{X}\right)^{-1}\right)^{-1}\left(a-\underline{a}\right)\right],$$

where

$$\begin{aligned} \underline{A} &= \left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\underline{Y} \\ \underline{S} &= \left(\underline{Y}-\underline{XA}\right)'\left(\underline{Y}-\underline{XA}\right) \\ \underline{a} &= vec\left(\underline{A}\right). \end{aligned}$$

Prior and Posterior

- Want:

$$p\left(A,\Sigma|Y\right)\propto p(Y|A,\Sigma)p(\bar{Y}|A,\Sigma)\overbrace{p\left(\Sigma\right)}^{\mathcal{IW}(\nu,S^*)}.$$

- Plugging stuff in:

$$\begin{aligned}
&p\left(A,\Sigma|Y\right)\propto p(Y|A,\Sigma)p(\bar{Y}|A,\Sigma)p\left(\Sigma\right)\\
&=\frac{1}{\left(2\pi\right)^{\frac{m(T+\bar{T})}{2}}}\left|\Sigma\right|^{-\frac{T+\bar{T}}{2}}\exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\underline{S}\right]\right\}\\
&\times\exp\left[-\frac{1}{2}\left(a-\underline{a}\right)'\left(\Sigma\otimes\left(\underline{X}'\underline{X}\right)^{-1}\right)^{-1}\left(a-\underline{a}\right)\right]\\
&\times\frac{\left|S^*\right|^{\nu/2}}{2^{\nu m}\prod_{i=1}^{m}\Gamma\left[\frac{\nu+1-i}{2}\right]}\left|\Sigma\right|^{-\frac{\nu+m+1}{2}}\exp-\frac{1}{2}tr\left[\Sigma^{-1}S^*\right]
\end{aligned}$$

Prior and Posterior

- Collecting terms in A and Σ :

$$
\begin{aligned}
& p(Y|A,\Sigma)p(\bar{Y}|A,\Sigma)p\left(\Sigma\right) \\
= & \frac{1}{(2\pi)^{\frac{m(T+\bar{T})}{2}}}\left|\Sigma\right|^{-\frac{T+\bar{T}+\nu+m+1}{2}}\exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\left(\underline{S}+S^{*}\right)\right]\right\} \\
& \times\exp\left[-\frac{1}{2}\left(a-\underline{a}\right)'\left(\Sigma\otimes\left(\underline{X}'\underline{X}\right)^{-1}\right)^{-1}\left(a-\underline{a}\right)\right] \\
& \times\frac{\left|S^{*}\right|^{\nu/2}}{2^{\nu m}\prod_{i=1}^{m}\Gamma\left[\frac{\nu+1-i}{2}\right]}
\end{aligned}
$$

- We can sort of ‘see’ a Normal distribution in here and an inverse Wishart.
- Must dig a little to find it!

Prior and Posterior

- Multiply and divide non-exponential term in the Normal:

$$p(Y|A,\Sigma)p(\bar{Y}|A,\Sigma)p\left(\Sigma\right)$$

$$= \frac{1}{(2\pi)^{\frac{m(T+\bar{T})}{2}}} |\Sigma|^{-\frac{T+\bar{T}+\nu+m+1}{2}} \exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\left(\underline{S}+S^{*}\right)\right]\right\}$$

$$\times\mathcal{N}\left(\underline{a},\Sigma\otimes\left(\underline{X}'\underline{X}\right)^{-1}\right)(2\pi)^{\frac{mk}{2}}\left|\Sigma\otimes\left(\underline{X}'\underline{X}\right)^{-1}\right|^{\frac{1}{2}}$$

$$\times\frac{\left|S^{*}\right|^{\nu/2}}{2^{\nu m}\prod_{i=1}^{m}\Gamma\left[\frac{\nu+1-i}{2}\right]}$$

where

$$\mathcal{N}\left(\underline{a},\Sigma\otimes\left(\underline{X}'\underline{X}\right)^{-1}\right)=(2\pi)^{-\frac{mk}{2}}\left|\Sigma\otimes\left(\underline{X}'\underline{X}\right)^{-1}\right|^{-\frac{1}{2}}$$

$$\times\exp\left[-\frac{1}{2}\left(a-\underline{a}\right)'\left(\Sigma\otimes\left(\underline{X}'\underline{X}\right)^{-1}\right)^{-1}\left(a-\underline{a}\right)\right]$$

Fact About Determinant of Kronecker Product

- Suppose A is $m \times m$ and B is $n \times n$.
- Then,
$$|A \otimes B| = |A|^n |B|^m .$$
  - Special case where A is a scalar:
$$|A \otimes B| = A^n |B|$$
- So,
$$\left|\Sigma \otimes \left(\underline{X}'\underline{X}\right)^{-1}\right| = |\Sigma|^k \left|\underline{X}'\underline{X}\right|^{-m}$$

Prior and Posterior

Multiply and divide non-exponential term in the Normal:

$$
\begin{aligned}
& p(Y|A,\Sigma)p(\bar{Y}|A,\Sigma)p\left(\Sigma\right) \\
= & \frac{1}{(2\pi)^{\frac{m(T+\bar{T})}{2}}}\left|\Sigma\right|^{-\frac{T+\bar{T}-k+\nu+m+1}{2}}\exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\left(\underline{S}+S^{*}\right)\right]\right\} \\
& \times\mathcal{N}\left(\underline{a},\Sigma\otimes\left(\underline{X}'\underline{X}\right)^{-1}\right)(2\pi)^{\frac{mk}{2}}\left|\underline{X}'\underline{X}\right|^{-m} \\
& \times\frac{\left|S^{*}\right|^{\nu/2}}{2^{\nu m}\prod_{i=1}^{m}\Gamma\left[\frac{\nu+1-i}{2}\right]} \\
\propto & \mathcal{N}\left(\underline{a},\Sigma\otimes\left(\underline{X}'\underline{X}\right)^{-1}\right)\mathcal{IW}\left(T+\bar{T}-k+\nu,\underline{S}+S^{*}\right)
\end{aligned}
$$

Prior and Posterior

- Conclude:

$$\begin{aligned} p\left(A, \Sigma | Y\right) &= \mathcal{N}\left(\underline{a}, \Sigma \otimes \left(\underline{X}'\underline{X}\right)^{-1}\right) \\ &\quad \times \mathcal{IW}\left(T + \bar{T} - k + \nu, \underline{S} + S^{*}\right) \\ &= p\left(A | Y, \Sigma\right) p\left(\Sigma\right). \end{aligned}$$

- Drawing A, Σ from posterior:
  - Draw Σ from $\mathcal{IW}\left(T + \bar{T} - q - pm + \nu, S^{*} + \underline{S}\right)$.
  - Then, draw a from $\mathcal{N}\left(\underline{a}, \Sigma \otimes \left(\underline{X}'\underline{X}\right)^{-1}\right)$.

Hyperparameters for Priors

- Inverse Wishart prior: degrees of freedom, ν, and scale, S^*.
  - In practice, S^* is a diagonal matrix constructed by (i) constructing a diagonal matrix using the variance of fitted disturbances in univariate autoregressive representations of the variables in y_t fit to a pre-sample and (ii) multiplying that matrix by ν.
  - Sometimes, $S^* = 0$ and priors for Σ are instead captured with dummies (see Del Negro and Schorfheide, 2011).
- Dummies: $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, other parameters - $s_1, s_2, \bar{y}_1, \bar{y}_2$.

Marginal Likelihood

- Marginal likelihood of data (see, e.g., Del Negro and Schorfheide, 2011, equation 15):

$$p(Y) = \int_{A,\Sigma} p(Y|A,\Sigma)\, p(A|\Sigma)\, p(\Sigma)\, dA d\Sigma$$

$$= (2\pi)^{-\frac{mT}{2}} \frac{\left|\underline{X}'\underline{X}\right|^{-\frac{m}{2}} \left|\underline{S}\right|^{-\frac{T+\bar{T}-k}{2}}}{\left|\bar{X}'\bar{X}\right|^{-\frac{m}{2}} \left|S^*\right|^{-\frac{\bar{T}-k}{2}}} \times \frac{2^{\frac{m(T+\bar{T}-k)}{2}} \prod_{i=1}^{m} \Gamma\left(\frac{T+\bar{T}-k+1-i}{2}\right)}{2^{\frac{m(\bar{T}-k)}{2}} \prod_{i=1}^{m} \Gamma\left(\frac{\bar{T}-k+1-i}{2}\right)},$$

where Γ is the gamma function, independent of the value of hyperparameters,

$$\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4).$$

  - The hyperparameters could be selected to maximize $p(Y)$.

Outline

- Normal Likelihood, Illustrated with Simple AR(2) representation. (done!)
  - conditional versus unconditional likelihood.
  - maximum likelihood with level GDP data.
  - the Hurwicz bias.
- Three representations of a VAR. (done!)
  - Standard Representation
  - Matrix Representation
  - Vectorized Representation.
- Priors, posteriors and marginal likelihood (done!)
  - Dummy observations.
  - Conjugate Priors.
- Forecasting with BVARs
  - stochastic simulations, versus non-stochastic.
  - forecast probability intervals.

Forecasting

- Repeated draws from $p\left(y_{T+1},...,y_{T+F}|Y,\xi_{T+1},...,\xi_{T+F}\right)$, where F is the forecast horizon.
- Stochastic simulation algorithm. For $l=1,...,N$,
  - Draw $A^{(l)},\Sigma^{(l)}$ from
$$p\left(A,\Sigma|Y\right)=\mathcal{N}\left(\underline{a},\Sigma\otimes\left(\underline{X}'\underline{X}\right)^{-1}\right)\times\mathcal{IW}\left(T+\bar{T}-k+\nu,S^{*}+\underline{S}\right)$$
  - Draw, for $t=T+1,...,T+F$:
$$u_t^{(l)}\sim\mathcal{N}\left(0,\Sigma^{(l)}\right).$$
  - Solve, recursively, for $y_t^{(l)}$, $t=T+1,...,T+F$:
$$y_t^{(l)}=A_0^{(l)}\xi_t+A_1^{(l)}y_{t-1}^{(l)}+...+A_p^{(l)}y_{t-p}^{(l)}+u_t^{(l)},$$
    where
$$y_t^{(l)}=y_t,\ \text{for } t\leq T.$$

Forecasting

- The sequence,
$$y_{T+1}^{(l)}, \ldots, y_{T+F}^{(l)},$$
for $l = 1, \ldots, N$, is a single draw from
$p\left(y_{T+1}, \ldots, y_{T+F} | Y, \xi_{T+1}, \ldots, \xi_{T+F}\right)$.
- For each i, $i = 1, \ldots, m$, we have
$$\underbrace{M_i}_{N \times F} = \left[\begin{array}{ccc} y_{i,T+1}^{(1)} & \cdots & y_{i,T+F}^{(1)} \\ \vdots & \ddots & \vdots \\ y_{i,T+1}^{(N)} & \cdots & y_{i,T+F}^{(N)} \end{array}\right].$$
- Then, for example, letting
$$\begin{aligned} \underbrace{\tau}_{1 \times N} &= \frac{1}{N}\left[\begin{array}{ccc} 1 & \cdots & 1 \end{array}\right], \\ & E_T\left[y_{i,T+1}, \ldots, y_{i,T+F}\right] \\ &\equiv E\left[y_{i,T+1}, \ldots, y_{i,T+F} | Y, \xi_{T+1}, \ldots, \xi_{T+F}\right] = \tau M_i. \end{aligned}$$

Mean Forecast, AR(1), T+1, T+2

$$
\begin{aligned}
y_{T+1}^{(l)} &= A_0^{(l)} + A_1^{(l)} y_T^{(l)} + u_{T+1}^{(l)} \\
y_{T+2}^{(l)} &= A_0^{(l)} + A_1^{(l)} \left[A_0^{(l)} + A_1^{(l)} y_T^{(l)} + u_{T+1}^{(l)} \right] + u_{T+1}^{(l)} \\
&= \left[A_0^{(l)} + A_1^{(l)} A_0^{(l)} \right] + \left(A_1^{(l)} \right)^2 y_T^{(l)} + A_1^{(l)} u_{T+1}^{(l)} + u_{T+1}^{(l)},
\end{aligned}
$$

for $l = 1, .., N$. Then, if $\hat{A}_i \equiv E_T A_i$, $i \geq 0$:

$$
\begin{aligned}
E_T y_{T+2} &= E_T \left[A_0 + A_1 A_0 \right] + y_T E_T \left(A_1 \right)^2 \\
&\quad + \overbrace{E_T \left[A_1 u_{T+1} \right]}^{= E_T A_1 E_T u_{T+1} = 0} + \overbrace{E_T \left[u_{T+1} \right]}^{=0} \\
&= E_T A_0 + Cov_T \left(A_1, A_0 \right) + E_T A_0 E_T A_1 \\
&\quad + y_T \left[var_T \left(A_1 \right) + \left(E_T \left(A_1 \right) \right)^2 \right] \\
&\neq \hat{A}_0 + \hat{A}_0 \hat{A}_1 + \hat{A}_1^2 y_T .
\end{aligned}
$$

Message of Previous Slide

- To obtain mathematically correct mean forecast, $E_T y_{T+i}$, $i = 1, ..., F$,
  - must do stochastic simulations of future.
  - simple *non-stochastic simulations* not enough:

$$y_t^{(l)} = \hat{A}_0 \xi_t + \hat{A}_1 y_{t-1} + ... + \hat{A}_p y_{t-p},$$

    setting $E_T u_{T+i} = 0$ for $i = 1, ..., F$.

- Problem with non-stochastic simulation procedure is quantitatively large if there is a lot of uncertainty at T about A and Σ (e.g., posterior second moments of A are large).
  - Whether it is worth the extra time to do stochastic simulation must be assessed on case by case basis.

Forecast Probability Interval

- After stochastic simulation, we have:

$$\underbrace{M_i}_{N\times F} = \begin{bmatrix} y_{i,T+1}^{(1)} & \cdots & y_{i,T+F}^{(1)} \\ \vdots & \ddots & \vdots \\ y_{i,T+1}^{(N)} & \cdots & y_{i,T+F}^{(N)} \end{bmatrix},$$

  for $i = 1, ..., m$.

- To obtain the date T conditional distribution of $y_{i,T+j}$ display histogram of j^{th} column of M_i.
- 90% probability interval for $y_{i,T+j}$ obtained by:
  - sorting contents of i^{th} column of M_i from smallest to largest
  - reporting 50th and 950th elements (say, $N = 1,000$).