# System likelihood for VAR's

April 23, 2015

# Conditional, marginal, and concentrated likelihoods

We begin with the model

$$\underset{1\times n}{y(t)} = \underset{1\times k}{X(t)}\mathrm{B} + \varepsilon(t)\,. \tag{1}$$

We assume $\varepsilon(t) \mid \{X(s), y(s-1),\, s \le t\} \sim N(0, \Sigma)$. The fact that this distribution for $\varepsilon(t)$ does not depend on $X$ or lagged $y$ is equivalent to the assumption that $X(t)$ is **predetermined** in this system, together with the assumption that $\varepsilon(t)$ is serially independent.

In this notation, each equation (column of the system) has the same $X(t)$ variable on the right-hand side and a distinct coefficient vector (column of $\mathrm{B}$). However, we can consider versions of the system with 0 constraints

on elements of $\mathrm{B}$, which create different lists of variables in different equations, or with other linear restrictions on $\mathrm{B}$, which might create links across $\mathrm{B}$'s in different equations.

Our assumptions let us write the pdf for the data at dates $t = 1, \ldots, T$, conditional on $\{X(s), s \leq 1\}$ as

$$\prod_{t=1}^{T} \phi\big(y(t) - X(t)\mathrm{B}; \Sigma\big) \cdot$$
$$q\big(X(t) \mid \{X(t-s), y(t-s),\, s > 0\}\,; \gamma\big)\, q_0\big(X(s), y(s),\, s \leq 0,; \gamma\big)\,, \quad (2)$$

where $\phi$ is the standard multivariate normal pdf, $q$ is the pdf of $X(t)$ given the past and $q_0$ is the marginal pdf for pre-sample values of $X$ and $y$. We need no assumptions on the form of $q$ and $q_0$, other than that the parameter vector $\gamma$ entering them is distinct from $\mathrm{B}, \Sigma$. Under these assumptions the

likelihood, as a function of $\mathrm{B}, \Sigma$ does not depend on $\gamma$ or on $q$ or $q_0$. In fact, in this case the likelihood is proportional to

$$|\Sigma|^{-T/2} \exp\left( -\frac{1}{2} \sum_{t=1}^{T} (y(t) - X(t)\mathrm{B}) \Sigma^{-1} (y(t) - X(t)\mathrm{B})' \right)$$

$$= |\Sigma|^{-T/2} \exp\left( -\frac{1}{2} \operatorname{tr}(\Sigma^{-1} S) \right) , \quad (3)$$

where $S = u'u$ and $u = y - X\mathrm{B}$ is the $T \times n$ matrix with typical row $y(t) - X(t)\mathrm{B}$. We can easily see that, as a function of $\mathrm{B}$ with $\Sigma$ held fixed, this expression is $e$ raised to a quadratic polynomial in $\mathrm{B}$, so it will be proportional to a Gaussian pdf with some mean and covariance matrix (assuming that the coefficient on the second-order term is negative definite, which it clearly is).

To see what is the mean and variance of the implied conditional distribution for $\mathrm{B}$, we introduce the notation $\mathrm{Vec}(\mathrm{B}) = \tilde{\beta}$ and $\mathrm{Vec}(y) = \tilde{y}$. Then we can write the exponent in (3) as

$$-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}S) = -\frac{1}{2}(\tilde{y}'(\Sigma^{-1}\otimes I)\tilde{y} - 2\tilde{y}'(\Sigma^{-1}\otimes X)\tilde{\beta} + \tilde{\beta}'(\Sigma^{-1}\otimes X'X)\tilde{\beta})\,. \tag{4}$$

This should be a somewhat familiar-looking quadratic form. By the usual completing-the-square exercise, we can rewrite it as

$$-\frac{1}{2}((\tilde{\beta}-\hat{\tilde{\beta}}_{OLS})'(\Sigma^{-1}\otimes X'X)(\tilde{\beta}-\hat{\tilde{\beta}}_{OLS}) + \mathrm{tr}(\Sigma^{-1}\hat{u}'\hat{u})\,, \tag{5}$$

where $\hat{u} = y - X\hat{\mathrm{B}}_{OLS}$, $\hat{\mathrm{B}}_{OLS} = (X'X)^{-1}X'y$, and $\hat{\tilde{\beta}}_{OLS} = \mathrm{Vec}(\hat{\mathrm{B}}_{OLS})$. From this expression it is clear that, conditional on $\Sigma$, the likelihood has the shape of a $N(\hat{\tilde{\beta}}_{OLS}, \Sigma\otimes(X'X)^{-1})$ distribution. Thus the conditional mean of $\mathrm{B}$ does not depend on $\Sigma$, so long as the prior is flat in $\mathrm{B}$.

To get the marginal posterior on $\mathrm{B}$, we need to integrate out $\Sigma$. For this, we return to the likelihood in the form on the right-hand side of (3). If we have an upper triangular Cholesky square root $Q$ of $S$ (i.e. an upper triangular $Q$ with $Q'Q = S$), we can write $V = Q\Sigma^{-1}Q'$ and then apply Propositions 2 and 1 from Appendix to rewrite (3) as

$$|V|^{T/2}\,|Q|^{-T}\exp\left(-\frac{1}{2}\operatorname{tr}(V)\right)\left|\frac{\partial\Sigma}{\partial\Sigma^{-1}}\right|\left|\frac{\partial\Sigma^{-1}}{\partial V}\right| dV$$

$$= |V|^{T/2}\,|Q|^{-T}\exp\left(-\frac{1}{2}\operatorname{tr}(V)\right)\left|\Sigma^{n+1}\right|\,|Q|^{-(n+1)}\,dV\,. \quad (6)$$

Using the fact that $|Q| = |S|^{-1/2}$, this can be rewritten as

$$\exp\left(-\frac{1}{2}\operatorname{tr}(V)\right)|V|^{T/2-n-1}\,|S|^{(-T+n+1)/2}\,dV\,. \quad (7)$$

This expression factors into a piece dependent on $V$ and another that depends only on $S$, so that when we integrate out $V$, we will be left with a term proportional to $|S|^{(-T+n+1)/2}$.

When the likelihood $f(z,\theta)$ is a function of a two-component parameter vector $\theta = (\theta_1,\theta_2)$, the likelihood **concentrated** with respect to $\theta_2$ is the function of $\theta_1$ obtained by maximizing $f$ with respect to $\theta_2$ for each value of $\theta_1$. It is easy to show, from (3) and the fact that, for any square matrix $Z$, $d\log|Z|/dz = Z^{-1}$, that concentrating likelihood with respect to $\Sigma$ produces a function of $\mathrm{B}$ proportional to $|S|^{-T/2}$. This does not match the marginal pdf for $\mathrm{B}$ with a flat prior on $\mathrm{B}$ and the upper triangle of $\Sigma$, which is what we computed above. However, if we take our prior on $\Sigma$ to be proportional to $|\Sigma|^{-(n+1)/2}$, it is easily seen that the marginal posterior on $\mathrm{B}$ and the concentrated likelihood match.

In the special case of $n = 1$, i.e. a single equation, this form is what

is known as a multivariate Student-$t$ distribution. In the multiple equation case, I'm not sure if there is a name for it.

The marginal distribution for $\Sigma$ is found by integrating $\mathrm{B}$ out of the likelihood. It is easy to see from (5) that the likelihood as a function of $\tilde{\beta}$ has the form of a normal pdf with mean $\hat{\tilde{\beta}}$ and variance matrix $\Sigma \otimes (X'X)^{-1}$. The normalizing constant for this distribution is then (ignoring a constant factor) $|\Sigma|^{k/2} |X'X|^{-n/2}$. The marginal pdf for $\Sigma$ is therefore proportional to

$$|\Sigma|^{-\frac{1}{2}(T-k)} e^{-\frac{1}{2}\operatorname{trace} \hat{S}\Sigma^{-1}}.$$

This is the kernel of the Inverse-Wishart distribution with scale parameter $S^{-1}$ and degrees of freedom (or shape) parameter $T - n - 1 - k$. If we use the $|\Sigma|^{-\frac{1}{2}(n+1)}$ improper prior, the degrees of freedom are instead just $T - k$. Since there are programs available to generate random draws from

the Wishart, the usual way to draw from the posterior pdf is to draw from the Inverse-Wishart marginal for $\Sigma$, then draw from the normal marginal distribution for $B \mid \Sigma$.

# Error bands for impulse responses

- In one sense, this is straightforward: make draws from the posterior pdf of $\Sigma$ and $A$ (or$\beta$) and for each draw calculate impulse responses $c_{ij}(t)$

- Plot the $c_{ij}t$ corresponding to the MLE (and/or the mean or median of the draws), together with the 5% upper and lower tails of the draws, for example. Could also plot HPD intervals.

## These are not confidence intervals

- It's a Bayesian calculation. It gives intervals with a clear interpretation, by a straightforward procedure, but they are not confidence intervals (except asymptotically!).

- True confidence intervals for individual $c_{ij}$'s are not possible.

- This is a special case of a general point. If $\theta$ is the complete parameter vector for the distribution of the data $X$, then we can *always* (in principle) produce a 90% (say) confidence *region* for $\theta$ by constructiong a $90\%$ significance level test for $\theta$ as $H_0$ for each $\theta$ in the parameter space. The set of $\theta$'s that are accepted in a given sample is an exact $90\%$ confidence set.

- But if we try to construct a confidence set for an individual $\theta_i$, the problem is that the distribution of any test statistic generally depends on all the parameters, not just $\theta_i$.

- The normal linear regression model is a special case where there is a set of test statistics that depend on single $\theta_i$'s.

- There are "asymptotic" confidence intervals. These have coverage probabilities that converge to the correct ones for parameter values in some neighborhood of the true parameter value. They can be constructed by linearizing the mapping from $B_{ij}$ (the AR coefficients) to $C_{ij}$, then transforming the normal asymptotic distribution for $B$ to the corresponding approximate normal distribution for $C$.

- But these have no more frequentist asymptotic justification than do the Bayesian intervals.

- As the forecast horizon expands, the nonlinearities rapidly become more extreme, so the intervals based on linearization are always inaccurate at distant horizons.

- The Bayesian intervals are accurate whether or not $A$ might have roots of one or larger in absolute value. The frequentist intervals are uninterpretable if that is true.

# Testing GCP

Suppose we have a special case of the system (1) of the form

$$\begin{bmatrix} y_1(t) & y_2(t) \end{bmatrix} = \begin{bmatrix} X_1(t) & X_2(t) \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} + \begin{bmatrix} \varepsilon_1(t) & \varepsilon_2(t) \end{bmatrix}. \quad (8)$$

We can choose a $\gamma$ to make $\nu(t) = \varepsilon_1(t) - \varepsilon_2(t)\gamma$ orthogonal to $\varepsilon_2(t)$ and define $C_1 = B_{11} - B_{12}\gamma$, $C_2 = B_{21} - B_{22}\gamma$, $\Omega = \mathrm{Var}([\nu(t)\ \varepsilon_2(t)])$ to allow us to rewrite the system as

$$\begin{bmatrix} y_1(t) - y_2(t)\gamma & y_2(t) \end{bmatrix} = \begin{bmatrix} X_1(t) & X_2(t) \end{bmatrix} \begin{bmatrix} C_1 & B_{12} \\ C_2 & B_{22} \end{bmatrix} + \begin{bmatrix} \nu(t) & \varepsilon_2(t) \end{bmatrix}. \quad (9)$$

Because the Jacobian of the transformation of parameters is the identity, and because the disturbances in the two blocks of equations in the transformed system are orthogonal, the likelihood factors into two pieces, one involving the parameters of the second equation only, the other involving the transformed first equation's parameters. Thus if our prior beliefs about the parameters of the transformed system make the parameters of the two blocks of equations independent, inference about the parameters of the second block of equations can be conducted by considering the likelihood for that block alone, as if there were no other equations in the system.

Of course in general, if we had started with prior beliefs expressed in terms of the original system's parameters, it would be a rare accident if our beliefs about the parameters of the two blocks in the transformed system were unrelated.

But testing for Granger causal priority (GCP) is a case where these results apply. GCP is the condition that, in a system in which $X(t)$ consists entirely of lagged values of $y$ and we have grouped all lagged values of the first block $y_1$ of $y$ into $X_1(t)$, $B_{12} = 0$. Thus a classical likelihood ratio test of the hypothesis that $y_2$ is GCP to $y_1$ is obtained by constructing twice the difference in log likelihoods and treating it as $\chi^2(df)$, where $df$, the degrees of freedom, is the number of elements in $B_{12}$. To be more specific, the classical LR test statistic is

$$T\left(\log(\left|S_{22}^R\right|) - \log(\left|S_{22}^U\right|)\right), \tag{10}$$

where $S_{22}^R$ and $S_{22}^U$ are the restricted and unrestricted cross-product matrices of residuals for the second block of equations.

# Predetermined, (Strictly) Exogenous, Weakly Exogenous, or Strongly Exogenous X's

The first two of these terms, "predetermined" and "exogenous", were in wide use in econometrics before Engle, Hendry and Richard 1983 (henceforth EHR) extended the terminology. The early usage characterized the relationship of right-hand side variables and disturbances in an equation system. In this usage, $X$ being **predetermined** in (1) means that

$$\varepsilon(t) \text{ is unrelated to } \{X(s), y(s-1),\, s \leq t\} \ .$$

The phrase "unrelated to" in this definition can be given a variety of specific meanings. Since we are working here with likelihood, we will take

predeterminedness of $X(t)$ to mean that

$$\varepsilon(t) \text{ is independent of } \{X(s), y(s-1),\, s \leq t\}\ .$$

Other versions of the assumption are, e.g.,

$$\mathcal{E}[\varepsilon(t) \mid \{X(s), y(s-1),\, s \leq t\}] = 0$$

or

$$E[\varepsilon(t) \mid \{X(s), y(s-1),\, s \leq t\}] = 0\,.$$

These weaker versions of the assumptions may be made as part of proofs that results we will derive for Gaussian likelihood are approximately correct under more general conditions.

The stronger assertion that $X$ is **exogenous** in (1) means that

$$\varepsilon(t) \text{ is unrelated to } \{X(s),\ -\infty < s < \infty\} \ .$$

In a likelihood framework again, “unrelated to" is taken to mean “independent of", and there are corresponding weakened version of this assumption based on the $\mathcal{E}$ and $E$ operators.

$X(t)$ predetermined implies the absence of serial dependence in $\varepsilon(t)$. Predetermined $X(t)$ is the main assumption required in proofs that in some sense least squares is the best estimator for $\mathrm{B}$.

$X(t)$ exogenous is generally impossible when $X(t)$ contains lagged dependent variables, and thus is an assumption usually made when the $X$ and $y$ variables are completely distinct. Exogeneity of $X$ is the main assumption required in proofs that GLS is the best estimator for $\mathrm{B}$.

The EHR notion of **weak exogeneity** in its general form makes no reference to equations or residuals. It is concerned entirely with conditions under which certain parameters of the joint distribution of two variables $y$ and $X$ can be estimated without loss of information from the conditional distribution of $y \mid X$ alone. But specialized to our case of linear regression with $\varepsilon(t) \sim N(0, \Sigma)$, weak exogeneity becomes the requirement that

i. $X(t)$ is predetermined in (1),

ii. the marginal distribution of $\{X(s), y(s-1),\ s \leq 1\}$ does not depend on any unknown parameters included in $\mathrm{B}, \Sigma$, and

iii. the pdf of $X(t) \mid \{X(s),\ y(s),\ s < t\}$ does not depend $\mathrm{B}$ or $\Sigma$.

In this case, the likelihood for all the data on $X$ and $y$ over the whole sample factors into two pieces, one of which is the usual Gaussian regression

likelihood, the other a function of $X$'s and of parameters other than $\mathrm{B}, \Sigma$. If the prior also makes $\mathrm{B}, \Sigma$ independent of the other parameters, the posterior pdf factors in the same way as the likelihood, justifying inference based on the regression likelihood alone.

**Strong exogeneity**, in EHR's terminology, applied to the multivariate Normal regression model, is exogeneity plus the requirement that the pdf of $\{X(s), s = 1, \ldots, T\}$ not depend on $\mathrm{B}, \Sigma$ and that the prior make $\mathrm{B}, \Sigma$ independent of other unknown parameters.

EHR argued correctly that these notions were important, because econometricians were used to justifying single-equation estimation methods based on claims about exogeneity and predeterminedness, without careful attention to model parameterization, which could lead to mistakes. However they presented their arguments in such a way that it might appear that predeterminedness and exogeneity, in their old senses, were mistaken

notions in themselves. This is not true. It is useful to be able to discuss whether, say, “price is predetermined in the demand equation" as a property of the actual demand equation, without regard to how we have parameterized it or how our beliefs about its parameters relate to beliefs about other parameters. EHR are correct in pointing out that after we have decided what we think about predeterminedness, we still have work to do in deciding whether single-equation estimation is justified. But these considerations can usefully separated from the question of whether the predeterminedness assumption itself is justified.

# Algebra of Symmetric Matrix Jacobians

# Notation

| symbol | dimensions | definition |
|---|---|---|
| $P$: | $n^2 \times n^2$ | $P\vec{A} = \overrightarrow{A'}$ |
| $F$: | $n^2 \times n^2$ | $F\vec{A} = \vec{B}$, where $b_{ij} = 0$ for $i \neq j$, $b_{ii} = a_{ii}$ |
| $H$: | $(n^2+n)/2 \times n^2$ | $H\vec{A}$ is the stacked upper triangle of $A$ |
| $S$: | $n^2 \times (n^2+n)/2$ | $H' + PH' - FH'$ |

In words, $F$ picks out of $\vec{A}$ the diagonal elements of $A$, $H$ picks out the upper triangle, and $HP$ picks out the lower triangle. $S'$ picks out the lower triangle, adds it to the upper triangle, then substracts the diagonal to avoid doubling it.

# Arguments

Here are some Lemmas, some of which are widely useful in doing calculus with matrices.

**Lemma 1.** $\overrightarrow{ABC} = (C' \otimes A)\vec{B}$ ,

*Proof.* This a straightforward, but perhaps tedious, exercise in wrting both sides of the equality out in sum notation. □

**Lemma 2.** *If $A$ and $B$ are square symmetric matrices, the eigenvectors of $A \otimes B$ are all the vectors of the form $v_i \otimes w_j$, where $v_i$ is an eigenvector of $A$ and $w_j$ is an eigenvector of $B$, and the corresponding eigenvalue is $\lambda_i \mu_j$, where $\lambda_i$ and $\mu_j$ are the corresponding eigenvalues of $A$ and $B$.*

*Proof.* Obviously $(A \otimes B)(v_i \otimes v_j) = \lambda_i \lambda_j (v_i \otimes v_j)$ and there are $n^2$ such

vectors. Here we are using the fact that an $n \otimes n$ square symmetric matrix always has $n$ distinct real eigenvectors. Since there are only $n^2$ eigenvectors for a matrix of this size, these clearly exhaust them. Note that since $v_i \otimes v_j$ has the same eigenvalue as $v_j \otimes v_i$, any linear combination of such a pair is also an eigenvector. Thus the assertion in the theorem has to be interpreted as meaning that these are all eigenvectors, and for repeated roots these vectors span the corresponding space of eigenvectors with the same eigenvalues. □

**Lemma 3.** *If $A$ and $B$ are symmetric matrices, and if we define $\hat{A} = H\vec{A}$ to be the stacked upper triangle of $A$, then*

$$\frac{\partial \hat{A}}{\partial \hat{B}} = H \frac{\partial \vec{A}}{\partial \vec{B}} S \,.$$

*Proof.* The $H$ on the left of the right-hand-side expression simply extracts

the upper triangle of $A$, where the rows of the partial derivative correspond to elements of $\vec{A}$. The $S$ on the right reflects our accounting for the fact that if we maintain symmetry in $B$, changing $b_{ij}$ when $i \neq j$ entails changing $b_{ji}$, so the derivative with respect to $b_{ij}$ when symmetry is maintained is the sum of the partial derivative with $b_{ji}$ held constant and the partial derivative w.r.t. $b_{ji}$ with $b_{ij}$ held constant. So we add $H'$, which picks out columns corresponding to the upper triangle, to $PH'$, which picks out columns corresponding to the lower triangle, and subtract $FH'$ to avoid double-counting the diagonal. □

**Lemma 4.**

$$H'H + PH'HP - F = I$$

*Proof.* $H'H$ is all zeros except for ones on the diagonal in the positions corresponding to the upper triangle of a stacked square matrix. $PH'HP$ rearranges the locations of the ones so that they are in the positions

corresponding to the lower triangle. Adding the two gives a matrix with ones on the diagonal in all positions except those corresponding to the diagonal in a stacked $n \times n$ matrix. Subtracting off $F$ then gives us the identity. □

**Lemma 5.** *If $A$ is symmetric, the eigenvalues of $H(A \otimes A)S$ are the vectors of the form*

$$H\big((v_i \otimes v_j) + (v_j \otimes v_i)\big)\,,$$

*with corresponding eigenvalues $\lambda_i \lambda_j$.*

*Proof.*

$$H\big((v_i \otimes v_j) + (v_j \otimes v_i)\big) = H(I+P)(v_i \otimes v_j)\,.$$

But

$$\begin{aligned} SH(I+P) &= H'H + H'HP + PH'H + PH'HP - FH'H - FH'HP \\ &= H'H + PH'HP - F + P(H'H + PH'HP - F) = I + P\,, \end{aligned}$$

using Lemma 4. In deriving this, we use the facts that $FH'H = F = FP$, which follow because $H'H$ is diagonal with ones in the upper triangular positions, while $F$ picks out the elements corresponding to the diagonal, and because the locations of the diagonal in $\vec{A}$ are the same as their locations in $\vec{A}' = P\vec{A}$. But with this result in hand the lemma follows immediately. □

**Proposition 1.**

$$\left|\frac{\partial \hat{A}^{-1}}{\partial \hat{A}}\right| = |A|^{-(n+1)} .$$

Here we are taking the $|\ |$ notation to indicate the absolute value of the determinant.

*Proof.* Applying Lemma 1, we can conclude that

$$\frac{\partial \vec{A}^{-1}}{\vec{A}} = -\left(A^{-1} \otimes A^{-1}\right).$$

Then applying Lemmas 3 and 5, the result follows from the fact that the determinant is the product of the eigenvalues. □

**Proposition 2.** *If $A = C' \cdot B \cdot C$. with $C$ upper triangular and $B$ symmetric, all $n \times n$, then*

$$\left|\frac{\partial A}{\partial B}\right| = |C|^{n+1}.$$

*Proof.* Again applying Lemma 1, we can obtain

$$\frac{\partial \vec{A}}{\partial \vec{B}} = C' \otimes C'.$$

This is a lower triangular matrix. If the diagonal elements of $C$ are all distinct, they are the eigenvalues of $C$ and we can apply essentially the same argument as for Proposition 1. But they need not be distinct, so we make a more direct argument.

$$H(C' \otimes C')S = H(C' \otimes C')H' + H(C' \otimes C')PH' - H(C' \otimes C')FH' .$$

The first term of the three on the right hand side is a lower triangular matrix with diagonal elements of the form $c_{ii}c_{jj}$, in fact with elements corresponding to all the distinct unordered pairs $i, j$. The second is also lower triangular, but the $P$ factor in it results in all diagonal elements except those corresponding to diagonal elements in a stacked matrix (i.e. the sequence $1, n, 2n, \ldots, n^2$) being zero, while the non-zero ones are in the same position as in the first factor. Finally the last factor is again a diagonal matrix with nonzero elements only in the same positions, $1, n, 2n, \ldots, n^2$. So the diagonal elements of the last two terms cancel to leave a zero diagonal, and the

determinant is just the product of the diagonal elements of the first term, which completes the proof. □

*

References

ENGLE, R. F., D. F.HENDRY, AND J.-F. RICHARD (1983): "Exogeneity," *Econometrica*, 51, 277–304.