Generalized Dummy Observation Priors

Christopher A. Sims
Princeton University
sims@princeton.edu

February 15, 2018

Theil Mixed Estimation

Model (SNLM): $$\underset{T\times 1}{Y} = X \underset{k\times 1}{\beta} + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2 I)$$

Dummy observations: $$\underset{N\times 1}{Y^*} = X^* \beta + \varepsilon^* \,.$$

Stack real and dummy observations, apply OLS and use likelihood as posterior. *Almost* equivalent to using as prior

$$\{\beta \mid \sigma^2\} \sim N((X^{*\prime} X^*)^{-1} X^{*\prime} Y^*, \sigma^2 (X^{*\prime} X^*)^{-1})$$

Why “Almost”?

- If we are doing model comparison, integrating the likelihood with dummy observations, even after multiplying by a proper prior on σ^2, does not produce a correct posterior weight.
- We have to calculate the integral of the prior on σ^2 times the dummy-observation component of the likelihood, and divide the dummy-obs likelihood by the resulting scale factor.
- This can be done analytically, and R and matlab code that does this analytically for VAR’s and regression models is available on my website.

Is mixed estimation just a variant notation?

- Dummy observation priors in the SNLM are always equivalent to some conjugate prior — this is in fact the definition of a conjugate prior — though the correspondence is not one-one.
- In the SNLM a sequence of k dummy observations such that X^* is square and full rank is equivalent to the prior constructed by asserting independent $N(Y_i^*, \sigma^2)$ marginal priors on the $X_i^* \beta$ linear combinations of β's.
- So is the dummy observation idea nearly trivial, mainly a matter of notation?

Building priors incrementally from mental "observations"

- Even in the SNLM and with linear "observations", the equivalence breaks down if we consider more than k dummy observations. This creates no difficulties with the dummy observation framework, but is impossible if we think of building a prior directly from k independent marginals.

- There is no reason we can't use dummy observations on nonlinear functions of the parameters, or more generally dummy observations that do not take the form of a conjugate prior.

- For nonlinear functions $f_i(\beta)$, asserting k independent marginal distributions with densities $p_i(f_i - \hat{f}_i)$ for k f_i's is not equivalent to asserting k dummy observations on these same f_i's with error densities p_i.

- The difference is in the treatment of Jacobian terms. The product of the k marginals becomes a joint pdf for β only after multiplication by $|\partial f / \partial \beta)|$.

- The corresponding step for the k dummy observations is calculating the integral w.r.t. β of the product of the p_i's and scaling by it. No Jacobians are involved.

Advantages of building priors from dummy observations

- We may have more than k "ideas" about β. Figuring out what k of them imply for the $k+n$'th in the direct approach can be a nasty task.

- There may be constraints on the f's that are difficult to characterize analytically. Error-ridden dummy observations need not satisfy these constraints, whereas directly asserted priors must be formulated to satisfy the constraints.

- In a large model, with complicated f_i's, calculating Jacobians can be a challenge, and may be difficult to automate in a computer program. Calculating scaling factors by integrating a product of dummy observation factors can always be attacked by brute force numerical integration, if necessary. In MCMC contexts, the technology for doing this will be readily at hand.

Examples: transition matrices

- Often we have an idea about what the ergodic distribution should look like. Specifying a Dirichlet-type prior directly on the p_{ij}'s so as to embody our ideas about the ergodic distribution (the left eigenvector of the P matrix corresponding to the unit eigenvalue, a highly nonlinear function of P) is difficult. A dummy observation factor expressing such beliefs is easy to formulate.

- We often try to use intuition about what the "mean and variance" of p_{ii} should be in setting up a Markov model. Choosing a Dirichlet or Beta prior that has a target mean and variance is probably often not what we really intend, as it can easily result, when the mean of p_{ii} is close to one, in a pdf that has a pole at one. A Gaussian dummy observation on p_{ii} with the target mean and variance may be closer to what we intend.

Simultaneous equations

- Apparently innocuous Gaussian priors on the coefficients of a simultaneous equations model may imply bizarre beliefs about the effects of exogenous shifters.

- E.g., Gaussian priors on elasticities of demand and supply may imply a lot of probability on extremely large impacts of exogenous demand or supply shifts.

- Effects of the exogenous shifters are the reduced form parameters $\Pi = \beta\Gamma^{-1}$, and the inversion mapping translates nontrivial density at points where $|\Gamma| = 0$ into fat tails on the Π distribution.

- If we combine dummy observations on Γ with dummy observations on Π we may avoid these problems.

VAR impulse responses

- There are a lot of them, generally more in graphs of estimation output than there are parameters in the model.
- They depend on the model parameters in a highly nonlinear way.
- There are constraints on them that are difficult to characterize and impose analytically.
- A lot of economists think they have an idea of what they should look like, at least in monetary SVAR's.
- There are many attempts at formalizing use of prior information about impulse response shapes: Uhlig, Faust, Canova, Dwyer, Kocięcki, and probably others I am missing.

Constraints and Jacobians

- With the model stacked into first-order form $y(t) = Ay(t-1) + \varepsilon(t)$, the matrix of impulse responses at horizon h is A^h.

- The derivative matrix of the mapping $A \to A^h$ is

$$A^{h-1} \otimes I + A^{h-2} \otimes A + \cdots + I \otimes A^{h-1}$$

- The determinant of this is not so bad to calculate, but if we have put priors on a scattering of responses at a scattering of horizons, or on nonlinear functions of responses, the Jacobian can be a mess.

- The real eigenvalues of A^h for even values of h must either be positive or occur in pairs of negative values of equal absolute value.

Keeping track of these constraints in imposing priors directly is challenging.

- We are likely to want to put priors on nonlinear functions of impulse responses. E.g. to suppress excessively fancy long run predictions from initial conditions, we might want to impose a prior belief that

$$\sum_{s=H}^{T} (a_{ij}(s) - a_{ij}(s-1))^2$$

is small.

- Besides the aim of focusing on "reasonable monetary SVAR" impulse responses, there is the aim of suppressing the tendency of estimates that condition on initial observations to attribute unreasonable predictive power to initial conditions very far from steady state.

- This is basically what the Minnesota Prior is trying to do. It includes a single dummy observation in which for each variable y_j in the model, $y_j(t-s) = \bar{y}_j$, $s = 0, \ldots, \ell$, where ℓ is lag length, and n dummy observations in which, for the j'th, $y_j(t-s) = \bar{y}_j$, $s = 0, \ldots, \ell$ while $y_i(t-s) = 0$ for all $i \neq j$, $s = 0, \ldots, \ell$.

- These pull toward a model with n unit roots. In a model with many lags and/or many variables, however, there are $n\ell$ roots, any of which can be complex and near one in absolute value. Such roots can intereact with initial conditions to produce exactly the kind of implausible estimates we would like to suppress.

- Dummy observations more directly place on the shapes of impulse responses might perform better and allow reasonable estimates without such tight Minnesota priors on sums of coefficients.