Supplementary Material

Bayesian seemingly unrelated regression

Bayesian Seemingly Unrelated Regression (SUR) is a Bayesian approach to estimating a system of regression equations with correlated error terms **[17]**. SUR extends the classical Seemingly Unrelated Regression (SUR) model to incorporate Bayesian inference, allowing for the estimation of parameters and uncertainty assessment in a more flexible and robust manner.

Overview of seemingly unrelated regression (SUR) model

In the SUR model, multiple regression equations are estimated simultaneously, where the error terms in each equation are allowed to be correlated. This accounts for the potential interdependence or correlation between the dependent variables across the equations **[18]**.

There are two main motivations for the development of SUR model. The first one is to gain efficiency in estimation by combining information on different equations. The second motivation is to impose and/or test restrictions that involve parameters in different equations.Suppose that y_{it} is a dependent variable, $x_{it} = \left(1, x_{it,1}, x_{it,2}, \ldots, x_{it,K_{i-1}}\right)'$ is a K_i vector of explanatory variables for observational unit i, and u_{it} an unobservable error term, where the double index is it denotes the t^{th} observation of the i^{th} equation in the system **[18]**. Often t denotes time and we will refer to this as the time dimension, but in some applications, t could have other interpretations, for example as a location in space. A classical linear SUR model is a system of linear regression equations,

$$y_{1t} = \beta_1' x_{1t} + u_{1t}$$
$$\vdots$$
$$y_{Nt} = \beta_N' x_{Nt} + u_{Nt}$$

where $i = 1, \ldots, N$, and $t = 1, \ldots, T$. Denote $L = K_1 + \cdots + K_N$. Further simplification in notation can be accomplished by stacking observations either in the t dimension or for each i. For example, if we stack for each observation t, let $Y_t = [y_{it}, \ldots, y_{Nt}]'$, $\widetilde{X_t} = diag(x_{1t}, x_{2t}, \ldots, x_{NT})$, a block-diagonal matrix with $(x_{1t}, x_{2t}, \ldots, x_{NT})$on its diagonal, $U_t = [u_{it}, \ldots, u_{Nt}]'$and $\beta = [\beta_1', \ldots, \beta_N']$ [23]. Then

$$Y_t = \widetilde{X_t'}\beta + U_t \tag{2}$$

Another way to present the SUR model is to write it in the form of multivariate regression with parameter restrictions. For this, define $X_t = [x_{1t}', x_{2t}', \ldots, x_{Nt}']'$ and $A(\beta) = diag(\beta_1, \ldots, \beta_N)$ to be a (LxN) block diagonal coefficient matrix **[18]**. Then, the SUR model in Equation 2 can be written as,

$$Y_t = A(\beta)'X_t + U_t \tag{3}$$

and the coefficient $A(\beta)$ satisfies

$$vec\big(A(\beta)\big) = G \tag{4}$$

for some $(NLxL)$ full rank matrix G. In the special case where $K_1 = \cdots = K_N = K$, we have $G = diag(i_1, \ldots, i_N) \otimes I_k$, where i_j denotes te j_{th} column of the NxN identity matrix I_N.

Bayesian estimation procedure

The Bayesian SUR approach involves specifying prior distributions for the parameters of interest, such as regression coefficients and error variances **[17]**. The joint posterior distribution is then obtained by combining the prior distributions with the likelihood function, which represents the probability of observing the data given the parameters.

Model specification

The model can be seen as a set of regressions for multivariate responses $Y = (y_1, \ldots, y_s), y_k = (y_{1k}, \ldots, y_{nk})^T$, for $k = 1, \ldots, s$ and corresponding covariate matrices X_k with dimensions $n \times p$. We assume independence between samples but allow for dependence across responses **[19]**. Moreover, we assume that the same set of predictors is available for all responses.

Variable selection is performed on the predictors using binary indicators vector $\gamma_k = (\gamma_{k1}, \ldots, \gamma_{kp})^T$, where γ_{kj} is 1 if covariate j is included in the regression for response k and 0 if not. We use the shorthand notation X_{γ_k} for the columns of X_k selected by the vector γ_k and similar for β_{γ_k}. Thus, we can write the set of linked regressions such as,

$$y_k = X_{\gamma_k}\beta_{\gamma_k} + u_k, k = 1, \ldots, s \tag{5}$$

but most importantly the residuals will be correlated, that is, $ui = (u_{i1}, \ldots, u_{is}) \sim N(0, C)$.

We can also write the likelihood of this model as,

$$Y \mid X, \{\beta_1, \ldots, \beta_s\}, \{\gamma_1, \ldots, \gamma_s\}, C \sim N\left(X_\gamma \beta_\gamma, C \otimes \mathbb{I}_n\right) \tag{6}$$

where $Y = vec(Y), vec(\cdot)$ being the vectorization operator $\gamma = vec(\gamma_1, \ldots, \gamma_s), \beta_\gamma = vec\left(\beta_{\gamma_1}, \ldots, \beta_{\gamma_s}\right)$ and X_γ is a block diagonal matrix with X_{γ_k} as the k^{th} diagonal element **[19]**.

Likelihood

Under standard assumptions of diagonal C or row sparsity for γ_k, combined with conjugate priors for C and β_γ, it is typically possible to integrate out these parameters analytically. However, the priors utilized in our specific model do not

maintain this conjugacy, rendering analytical integration impossible. Despite this limitation, the full conditional distributions remain mathematically simple form, which facilitates the straightforward implementation of a Gibbs sampler for posterior estimation. **[19]**. The computational time needed is, however, prohibitive for most high-dimensional settings.

To overcome this issue, we decompose the covariance matrix C iteratively as,

$$C_{(k)} = \begin{pmatrix} C_{(k-1)} & c_k \\ c_k^T & c_k \end{pmatrix}$$

for all $k = 2, \ldots, s$, with $C_{(s)} = C$ and $C_{(1)} = c_1 = C11$ (the scalar variance of response 1) and c_1 is null. Thus, each $C_{(k)}$ is the marginal covariance matrix for responses $1, \ldots, k, c_k$ is the variance of response k, and c_k is the vector of covariances between response k and responses $1, \ldots, (k-1)$.

With this decomposition, the likelihood can be factorized as,

$$p(Y|X, \beta, \gamma, C) = \prod_{k=1}^{s} N\left(y_k \middle| X_{\gamma_k}\beta_{\gamma_k} + U_{(k-1)}\rho_k, \sigma_k^2 \mathbb{I}_n\right) \tag{7}$$

where $U_{(k-1)} = Y_{(k-1)} - \left(X_{\gamma_1}\beta_{\gamma_1}, X_{\gamma_2}\beta_{\gamma_2}, \ldots, X_{\gamma_{k_1}}\beta_{\gamma_{k-1}}\right)$ is a matrix consisting of the first $(k-1)$ residuals from the original linked regressions where $U_{(0)}$ is null. For $k = 1$, the likelihood simplifies to $N\left(y_1 \middle| X_{\gamma_1}\beta_{\gamma_1}, \sigma_1^2 \mathbb{I}_n\right)$. The parameters σ_k^2 and ρ_k are also defined through the reparameterization of the residual covariance matrix, that is $\sigma_1^2 = c_1$, and

$$\left.\begin{aligned} \sigma_k^2 &= c_k - c_k^T C_{(k-1)}^{(-1)} c_k, \\ \rho_k &= C_{(k-1)}^{(-1)} c_k \end{aligned}\right\} k = 2, \ldots, s \tag{8}$$

Full model

To account for sparsity within the residual dependency structure, we employ a decomposable graph, denoted as G. In this framework, any two variables are treated as

conditionally independent if they are not connected by a direct edge **[19]**. Conditional on the graph, we assign a hyper inverse-Wishart prior to the original covariance matrix $C \sim HIW_G(\nu, M)$.

The new variables σ_k^2 and ρ_k k are defined within the prime components, that is,

$$\sigma_k^2 = c_k - c_{k,P_q}^T \left(C_{P_q}^{(k-1)}\right)^{-1} c_{k,P_q} \tag{9}$$

$$\rho_{k,P_q} = \left(C_{P_q}^{(k-1)}\right)^{-1} c_{k,P_q} \tag{10}$$

where $C_{P_q}^{(k-1)}$ is the submatrix of C_{P_q} with variable k removed and c_{k,P_q} is the final column of C_{P_q} without the last element. All other elements of ρ_k are zero.

Here we summarize the full model with all its conditional dependencies and provide the version using the sparse covariance structure. The joint distribution is:

$$\begin{gathered}\prod_{k=1}^{s} p\left(y_k|\beta_k, \gamma_k, \sigma_k^2, \rho_k\right) p(o_k) \times \\ \prod_{j=1}^{p} p\left(\beta_{kj}|\gamma_{kj}, w\right) p\left(\gamma_{kj}|o_k, \pi_j\right) p\left(\pi_j\right) p(w) p(J|\eta) p(\eta) p(\tau) p\left(\sigma^2_{\xi_j^{-1}(1)}\right) \times \\ \prod_{k \varepsilon P_1^{(J)}} p\left(\rho_k|\sigma_k^2, \tau, J\right) \times \prod_{q=2}^{Q} \prod_{k \varepsilon R_q^{(J)}} p\left(\sigma_k^2|\tau, J\right) p\left(\rho_k|\sigma_k^2, \tau, J\right)\end{gathered} \tag{11}$$

where,

$$y_k \mid \beta_k, \gamma_k, \sigma_k^2, \rho_k \sim N\left(X_{\gamma_k}\beta_{\gamma_k} + U_{(k-1)}\rho_k, \sigma_k^2 \mathbb{I}_n\right)$$

$$\beta_{kj} \mid \gamma_{kj}, w \sim \gamma_{kj} N(0, w^{-1}) + \left(1 - \gamma_{kj}\right)\delta_0$$

$$\gamma_{kj} \mid o_k, \pi_j \sim Ber\left(o_k \times \pi_j\right)$$

$$o_k \sim Beta(a_0, b_0)$$

$$\pi_j \sim Ga(a_\pi, b_\pi)$$

$$w \sim IGa(a_w, b_w)$$

$$\tau \sim Ga(a_\tau, b_\tau)$$

$$\eta \sim Beta(a_\eta, b_\eta)$$

with $U_{(k-1)}$ defined as in Equation 7 and δ_0 is the Dirac delta function centered at 0. The parameter J denotes the junction tree corresponding to the graph, while $|J|$ represents the total count of edges within that graph structure. The variable ξ_J indicates a perfect elimination ordering of the nodes. Both $S_q^{(J)}$ and $H_q^{(J)}$ are contingent on the graph and formally defined in the "Likelihood" section. Furthermore, the index $q_{(k)}$ identifies the specific prime residual $R_q^{(J)}$ associated with node k under the current ordering for graph J. Finally, the set of hyperparameters a_o, b_o, a_π, b_π, a_w, b_w, a_τ, b_τ, a_η, b_η , as well as the degrees of freedom $\nu > s - 1$ in the inverse-Wishart distribution, are treated as fixed constants **[19]**.