Models with General Error Structure / Model Comparison

(K Ch. 6, KPT Ch. 5,13)

AAEC 6564
Instructor: KLAUS MOELTNER

Matlab scripts: `mod4_data, mod4_sur, mod4_ppp, mod4_outage, mod4_outage_hpdi, mod4_sddr, mod4_outage_sddr, mod4_sur_chib,`
Matlab functions: `gs_sur, gs_sur_chib`

The Seemingly Unrelated Regression (SUR) Model

We will now examine regression models of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim n(\mathbf{0}, \boldsymbol{\Omega}) \tag{1}$$

Thus, we no longer restrict the error terms to be uncorrelated and share the same variance. This general structure nests many popular deviations from the basic case, such as heteroskedasticity, autocorrelation, and multi-equation systems. Here we will focus on the Seemingly Unrelated Regression (SUR) model, which forms the basis for some of the latent variable models we'll discuss later in this course.

Lets' consider a sample of $i = 1 \ldots n$ individuals. For each individual we have one observation for each of $m = 1 \ldots M$ equations. Each equation is represented by a linear regression model of the form

$$y_{mi} = \underset{k_m x 1}{\mathbf{x}'_{\mathbf{mi}}} \boldsymbol{\beta}_{\mathbf{m}} + \varepsilon_{mi} \qquad i = 1 \ldots n, m = 1 \ldots M \tag{2}$$

Thus, we allow ex ante for a different set of regressors for each equation (thus the subscript "*m*" to **x**) and, as is standard for SUR models, a different set of coefficients per equation (thus the subscript "*m*" to $\boldsymbol{\beta}$). For a given individual, we can write her *M* observations as a block, i.e.

$$\underset{Mx1}{\mathbf{y}_\mathbf{i}} = \underset{Mxk_m}{\mathbf{X}_\mathbf{i}} \underset{kx1}{\boldsymbol{\beta}} + \underset{Mx1}{\boldsymbol{\varepsilon}_\mathbf{i}} \quad \text{with} \quad \boldsymbol{\varepsilon}_\mathbf{i} \sim n\left(\mathbf{0}, \underset{MxM}{\boldsymbol{\Sigma}}\right),$$

$$\mathbf{X}_\mathbf{i} = \begin{bmatrix} \mathbf{x}'_{\mathbf{1i}} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}'_{\mathbf{2i}} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}'_{\mathbf{Mi}} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_\mathbf{1} \\ \boldsymbol{\beta}_\mathbf{2} \\ \vdots \\ \boldsymbol{\beta}_\mathbf{M} \end{bmatrix} \qquad \boldsymbol{\varepsilon}_\mathbf{i} = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{Mi} \end{bmatrix} \qquad k = \sum_{m=1}^{M} k_m \tag{3}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1M} & \sigma_{2M} & \cdots & \sigma_{MM} \end{bmatrix}$$

As shown above, the key new feature of this model is its full variance-covariance matrix (VCOV) for $\boldsymbol{\varepsilon}_\mathbf{i}$. In other words, we ex ante allow the error terms associated with a given person or other cross-sectional

unit to be correlated across equations. The VCOV $\boldsymbol{\Sigma}$ can have up to $\frac{M*(M+1)}{2}$ free elements. To keep estimation tractable, we also make the standard assumption that all individuals "share" the same $\boldsymbol{\Sigma}$. The full model across all individuals can be written as in (1), with

$$\mathbf{y}=\underset{(nxM)x1}{\begin{bmatrix}\mathbf{y_1}\\ \mathbf{y_2}\\ \vdots\\ \mathbf{y_n}\end{bmatrix}} \qquad \mathbf{X}=\underset{(nxM)xk}{\begin{bmatrix}\mathbf{X_1}\\ \mathbf{X_2}\\ \vdots\\ \mathbf{X_n}\end{bmatrix}} \qquad \boldsymbol{\beta}=\begin{bmatrix}\boldsymbol{\beta_1}\\ \boldsymbol{\beta_2}\\ \vdots\\ \boldsymbol{\beta_M}\end{bmatrix} \qquad \boldsymbol{\varepsilon}=\begin{bmatrix}\boldsymbol{\varepsilon_1}\\ \boldsymbol{\varepsilon_2}\\ \vdots\\ \boldsymbol{\varepsilon_n}\end{bmatrix}\sim n(\mathbf{0},\boldsymbol{\Omega}) \qquad \text{with}$$

$$\boldsymbol{\Omega}=\underset{(nxM)x(nxM)}{\begin{bmatrix}\boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma} & \cdots & \mathbf{0}\\ \mathbf{0} & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}\end{bmatrix}} \tag{4}$$

The likelihood is given as

$$p(\mathbf{y}\mid\boldsymbol{\beta},\boldsymbol{\Sigma},\mathbf{X})=\prod_{i=1}^{n}(2\pi)^{-M/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\tfrac{1}{2}(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta})'\,\boldsymbol{\Sigma}^{-1}(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta})\right)=$$

$$(2\pi)^{-Mn/2}|\boldsymbol{\Sigma}|^{-n/2}\exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta})'\,\boldsymbol{\Sigma}^{-1}(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta})\right)= \tag{5}$$

$$(2\pi)^{-Mn/2}|\boldsymbol{\Sigma}|^{-n/2}\exp\left(-\tfrac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\,\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right) \quad \text{since} \quad \boldsymbol{\Omega}^{-1}=\begin{bmatrix}\boldsymbol{\Sigma}^{-1} & & & \\ & \boldsymbol{\Sigma}^{-1} & & \\ & & \ddots & \\ & & & \boldsymbol{\Sigma}^{-1}\end{bmatrix}$$

We use again independent priors for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. A convenient prior for the latter is the Inverse Wishart (*IW*) density, which can be interpreted as the multi-variate version of the inverse-gamma. The priors are thus given as follows:

$$p(\boldsymbol{\beta},\boldsymbol{\Sigma})=p(\boldsymbol{\beta})\,p(\boldsymbol{\Sigma}) \qquad \text{where}$$

$$\boldsymbol{\beta}\sim n(\boldsymbol{\mu_0},\mathbf{V_0}), \quad \boldsymbol{\Sigma}\sim IW(\nu_0,S_0)$$

$$p(\boldsymbol{\beta})=(2\pi)^{-k/2}|\mathbf{V_0}|^{-1/2}\exp\left(-\tfrac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu_0})'\,\mathbf{V_0^{-1}}(\boldsymbol{\beta}-\boldsymbol{\mu_0})\right) \tag{6}$$

$$p(\boldsymbol{\Sigma})=\left(2^{\nu_0 M/2}\pi^{M(M-1)/4}\prod_{i=1}^{M}\Gamma\left(\frac{\nu_0+1-i}{2}\right)\right)^{-1}*|\mathbf{S_0}|^{\nu_0/2}|\boldsymbol{\Sigma}|^{-(\nu_0+M+1)/2}\exp\left(-\frac{1}{2}tr\left(\mathbf{S_0}\cdot\boldsymbol{\Sigma}^{-1}\right)\right)$$

where "*tr*" is the trace operator, i.e. the sum of the diagonal elements of a matrix. We choose again the form given in Gelman et al. (2004), for the *IW* density, where ν_0 is the degrees-of-freedom parameter and $\mathbf{S_0}$ is a symmetric, positive-definite scale matrix of dimension *M* by *M*. We need to set $\nu_0\geq M+2$ for the density to be well-defined (i.e. to have a finite integral). In this parameterization, the *IW* has the following expectation:

$$E(\boldsymbol{\Sigma}) = (\nu_0 - M - 1)^{-1}\mathbf{S_0} \tag{7}$$

(This is useful to report in a paper so the reader immediately knows which parameterization was chosen.)

Combining the priors with the likelihood, and dropping all terms that are multiplicatively unrelated to our parameters of interest yields the posterior kernel

$$p(\boldsymbol{\beta},\boldsymbol{\Sigma}\,|\,\mathbf{y},\mathbf{X}) \propto \\ |\boldsymbol{\Sigma}|^{-(\nu_0+M+1+n)/2}\exp\left(-\frac{1}{2}tr\left(\mathbf{S_0}\cdot\boldsymbol{\Sigma}^{-1}\right)\right)\exp\left(-\tfrac{1}{2}\left((\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\,\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+(\boldsymbol{\beta}-\boldsymbol{\mu}_0)'\,\mathbf{V}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right)\right). \tag{8}$$

As before, we first aim to find the posterior density for $\boldsymbol{\beta}$, conditional on $\boldsymbol{\Sigma}$. Thus, we will fist focus on the components of the posterior kernel that cannot be multiplicatively separated from $\boldsymbol{\beta}$. This leaves

$$p(\boldsymbol{\beta}\,|\,\boldsymbol{\Sigma},\mathbf{y},\mathbf{X}) \propto \exp\left(-\tfrac{1}{2}\left((\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\,\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+(\boldsymbol{\beta}-\boldsymbol{\mu}_0)'\,\mathbf{V}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right)\right). \tag{9}$$

Using transformations analogous to the models discussed so far, we obtain:

$$\boldsymbol{\beta}\,|\,\boldsymbol{\Sigma},\mathbf{y},\mathbf{X} \sim n(\boldsymbol{\mu}_1,\mathbf{V}_1) \quad \text{with} \quad \mathbf{V}_1 = \left(\mathbf{V}_0^{-1}+\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1} \text{ and } \quad \boldsymbol{\mu}_1 = \mathbf{V}_1\left(\mathbf{V}_0^{-1}\boldsymbol{\mu}_0+\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}\right) \tag{10}$$

A computational hint: If *M* and / or *n* are large, taking the inverse of $\boldsymbol{\Omega}$ can be very slow in Matlab. You may instead work with sums over individual-specific blocks, i.e. use

$$\mathbf{V}_1 = \left(\mathbf{V}_0^{-1}+\sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right)^{-1} \text{ and } \quad \boldsymbol{\mu}_1 = \mathbf{V}_1\left(\mathbf{V}_0^{-1}\boldsymbol{\mu}_0+\sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{y}_i\right) \tag{11}$$

To derive the conditional posterior density for $\boldsymbol{\Sigma}$, we return to our original form for the joint posterior given in (8). Ignoring terms that are not related to $\boldsymbol{\Sigma}$, and replacing $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\,\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ with $\sum_{i=1}^{n}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'\,\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})$ in the likelihood we have

$$p(\boldsymbol{\Sigma}\,|\,\boldsymbol{\beta},\mathbf{y},\mathbf{X}) \propto |\boldsymbol{\Sigma}|^{-(\nu_0+M+1+n)/2}\exp\left(-\frac{1}{2}\left(tr\left(\mathbf{S_0}\cdot\boldsymbol{\Sigma}^{-1}\right)+\sum_{i=1}^{n}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'\,\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})\right)\right). \tag{12}$$

Before we can recognize this clearly as another *IW* kernel, we need to implement a few transformations involving "trace rules" (see e.g. Greene, 5th ed., p.829):

$$
\begin{aligned}
&\exp\left(-\frac{1}{2}\left(tr\left(\mathbf{S_0}\cdot\boldsymbol{\Sigma}^{-1}\right)+\sum_{i=1}^{n}\left(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)\right)= \\
&\exp\left(-\frac{1}{2}\left(tr\left(\mathbf{S_0}\cdot\boldsymbol{\Sigma}^{-1}\right)+tr\left(\sum_{i=1}^{n}\left(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)\right)\right)= \quad \text{since } \sum_{i=1}^{n}(\bullet) \text{ is a scalar} \\
&\exp\left(-\frac{1}{2}\left(tr\left(\mathbf{S_0}\cdot\boldsymbol{\Sigma}^{-1}\right)+\sum_{i=1}^{n}tr\left(\left(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)\right)\right)= \quad \text{since } tr\left(\mathbf{A}+\mathbf{B}\right)=tr\left(\mathbf{A}\right)+tr\left(\mathbf{B}\right) \\
&\exp\left(-\frac{1}{2}\left(tr\left(\mathbf{S_0}\cdot\boldsymbol{\Sigma}^{-1}\right)+\sum_{i=1}^{n}tr\left(\left(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta}\right)\left(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\right)\right)\right)= \quad \text{since } tr\left(\mathbf{ABC}\right)=tr\left(\mathbf{CBA}\right) \\
&\exp\left(-\frac{1}{2}\left(tr\left(\mathbf{S_0}\cdot\boldsymbol{\Sigma}^{-1}\right)+tr\left(\left(\sum_{i=1}^{n}\left(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta}\right)\left(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta}\right)'\right)\boldsymbol{\Sigma}^{-1}\right)\right)\right)= \quad \text{take } \boldsymbol{\Sigma} \text{ outside the summation} \\
&\exp\left(-\frac{1}{2}\left(tr\left(\left(\mathbf{S_0}+\left(\sum_{i=1}^{n}\left(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta}\right)\left(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta}\right)'\right)\right)\boldsymbol{\Sigma}^{-1}\right)\right)\right)
\end{aligned}
\tag{13}
$$

We can now see that

$$
\begin{aligned}
&\boldsymbol{\Sigma}\,|\,\boldsymbol{\beta},\mathbf{y},\mathbf{X}\sim IW\left(\nu_1,\mathbf{S_1}\right) \quad \text{with} \\
&\nu_1=\nu_0+n \qquad \mathbf{S_1}=\mathbf{S_0}+\sum_{i=1}^{n}\left(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta}\right)\left(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta}\right)'
\end{aligned}
\tag{14}
$$

Thus, the SUR model can be estimated via a generic Gibbs Sampler, similar to the one we have used for the normal regression model with independent priors.

To draw from the *IW* in Matlab, use Matlab's built-in `iwishrnd(S1,v1)`. Script `mod4_sur` with function `gs_sur` illustrates the methodology using simulated data (implemented via `mod4_data`)

Application: SUR model for outage cost data

matlab script `mod4_outage`
Data file: `outage.txt`

These data were collected in 1998 / 1999 by a Seattle area Utility (= electricity provider) to assess the cost of power outages of different timing & duration on commercial/ industrial customers. Each firm was presented with the same six outage scenarios & asked to estimate the costs it would suffer from such a hypothetical outage (in foregone production, spoiled goods, delayed deliveries, etc), using a spreadsheet.

The details of these data are given in

Moeltner, K., D. F. Layton. (2002). A Censored Random Coefficients Model for Pooled Survey Data with Application to the Estimation of Power Outage Costs. *Review of Economics and Statistics*, 84 (3), p. 552-561.

The firms were grouped into 4 size categories according to their annual power consumption ("1" = small, like a pub or a small store, "4" = huge, like a large manufacturing plant or a refinery). We will drop the largest category for this analysis. We estimate a simple SUR model with 3 equations, relating outage costs (in dollars) to the three remaining size groups. Equation 1 uses only data for outage scenario 1 (*1hr, weekday, daytime*), equation 2 uses data for outage scenario 5 (*1hr, weekday, nighttime*), and equation 3 uses data for scenario 6 (*1hr, weekend, daytime*). We ignore scenarios 2, 3, and 4 for this exercise.

Let's estimate the SUR model & look at the log file.

- Do the sampler diagnostics check out?
- Is there a reasonable pattern for a specific outage over size groups?
- Which outage causes the most / least costs? Does this result hold for all size groups?

Examining linear restrictions using HPDI

Let's assume you want to examine the hypothesis that outage scenarios 5 and 6 have the same effect on a firm of size 2. Given the ordering of betas in your Gibbs Sampler output, this implies

$\beta_5 = \beta_8$, or $\beta_5 - \beta_8 = 0$

We can now construct a HPDI for this difference. If "0" is outside the bounds, we would have strong doubts that our hypothesis holds. We can also examine hypotheses based on inequalities. For example you're interested if a scenario 1 – type outage has a *smaller* effect on size1 firms than size 2 firms. Econometrically, you would implement this via

$\beta_2 - \beta_1 = 0$

A location of "0" outside the *right* HPDI bound would lend credibility to your hypothesis.
See script `mod4_outage_hpdi` for implementation of these "tests".

The Savage-Dickey Density Ratio (SDDR)

The SDDR is another very simple and straightforward method to compare *nested* models. In contrast to HPDI's, the SDDR is firmly anchored in probability theory, and thus perfectly legitimate. The only restriction is that the two models under comparison must be nested. There is another restriction on priors which we'll discuss in a moment.

Consider an *un*restricted model M_2 with parameters $\boldsymbol{\theta} = \left[\boldsymbol{\omega}' \quad \boldsymbol{\psi}'\right]'$. The likelihood and prior are given by $p\left(\mathbf{y} \mid \boldsymbol{\omega}, \boldsymbol{\psi}, M_2\right)$ and $p\left(\boldsymbol{\omega}, \boldsymbol{\psi}, \mid M_2\right)$, where we explicitly include the model index as conditioner for clarity of exposition. The prior is written in its generic form and doesn't necessarily imply that $\boldsymbol{\omega}$ and $\boldsymbol{\psi}$ are jointly distributed. Consider a restricted version of the model, say model M_1 that sets $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. The likelihood and prior for this model are $p\left(\mathbf{y} \mid \boldsymbol{\psi}, M_1\right)$ and $p\left(\boldsymbol{\psi} \mid M_1\right)$.

Suppose the priors under the two models satisfy $p\left(\boldsymbol{\psi} \mid \boldsymbol{\omega} = \boldsymbol{\omega}_0, M_2\right) = p\left(\boldsymbol{\psi} \mid M_1\right)$. This is the second requirement for using the SDDR. It is trivially satisfied if $\boldsymbol{\omega}$ and $\boldsymbol{\psi}$ have independent priors under M_2, and the same prior density is chosen for $\boldsymbol{\psi}$ for both models. This is a very common case.

Then, the SDDR, which is simply the Bayes factor for comparing M_1 to M_2, can be written as

$$SDDR_{12} = BF_{12} = \frac{p(\omega = \omega_0 \mid \mathbf{y}, M_2)}{p(\omega = \omega_0 \mid M_2)} \tag{15}$$

The proof is given in KPT, p. 69 and attached to these notes. The first thing to note is that we only need to work with the unconstrained model. The denominator is simply the prior of $\boldsymbol{\omega}$ evaluated at $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. The numerator is a bit more tricky – it's the marginal posterior of $\boldsymbol{\omega}$ evaluated at $\boldsymbol{\omega} = \boldsymbol{\omega}_0$.

In most applications, we know the conditional posterior $p(\boldsymbol{\omega} \mid \boldsymbol{\psi}, \mathbf{y}, M_2)$. However, we can use the same intuition as for Monte Carlo integration and posterior predictions, and derive an approximation of the marginal posterior as

$$p(\omega = \omega_0 \mid \mathbf{y}, M_2) \approx \frac{1}{R}\sum_{r=1}^{R} p(\boldsymbol{\omega} = \boldsymbol{\omega}_0 \mid \boldsymbol{\psi}_\mathbf{r}, \mathbf{y}, M_2), \tag{16}$$

i.e. we evaluate the conditional posterior of $\boldsymbol{\omega}$ at $\boldsymbol{\omega}_0$ at all values of $\boldsymbol{\psi}_\mathbf{r}$ from the original Gibbs Sampler, then average over the density evaluations. See script `mod4_sddr` for an example.

Application: Using the SD-method to test restrictions for the outage cost SUR model

Assume we want to test the hypothesis that the difference in cost between outage scenarios 5 and 6 is equal across all firm sizes. Econometrically:

$\beta_4 - \beta_7 - (\beta_5 - \beta_8) = 0$ and $\beta_4 - \beta_7 - (\beta_6 - \beta_9) = 0$ or

$\mathbf{R\beta}=\mathbf{q}$

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & -1 & 0 & 1 \end{bmatrix} \quad \mathbf{q} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Thus, we need to work with the prior and posterior density of $\mathbf{R\beta}$. Given the known form of the prior and conditional posterior of $\boldsymbol{\beta}$, by matrix rules we get:

$\mathbf{R\beta} \sim n(\mathbf{R\mu_0}, \mathbf{RV_0R}')$ and $\mathbf{R\beta} \mid \sigma^2, \mathbf{y} \sim n(\mathbf{R\mu_1}, \mathbf{RV_1R}')$

See script `mod4_outage_sddr` for an implementation.

Proof for the Savage-Dickey ratio

Let $g(\boldsymbol{\omega}, \boldsymbol{\psi}) = g(\boldsymbol{\psi} \mid \boldsymbol{\omega})\, g(\boldsymbol{\omega})$ be a continuous density defined over $\Re^k$ and express the priors for the two models as $p(\boldsymbol{\omega}, \boldsymbol{\psi} \mid M_2) = g(\boldsymbol{\omega}, \boldsymbol{\psi})$ and $p(\boldsymbol{\psi} \mid M_1) = g(\boldsymbol{\psi} \mid \boldsymbol{\omega} = \boldsymbol{\omega}_0)$. This implies that the prior of the unconstrained parameter $\boldsymbol{\psi}$ is the same at $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ in both models. Naturally, if the two priors are independent, i.e. $g(\boldsymbol{\omega}, \boldsymbol{\psi}) = g(\boldsymbol{\psi})\, g(\boldsymbol{\omega})$ and we use the same prior for $\boldsymbol{\psi}$ in both models this is trivially satisfied.

Also express the likelihood for the unconstrained model as $L(\boldsymbol{\omega},\boldsymbol{\psi})$ and for the constrained model as $L(\boldsymbol{\omega}=\boldsymbol{\omega}_0,\boldsymbol{\psi})$. This implies that the models are *nested* – they have the same likelihood function except for the linear constraint $\boldsymbol{\omega}=\boldsymbol{\omega}_0$.

We now need to show that the Bayes Factor for M_1 versus M_2 can be written as the ratio of the marginal posterior density of $\boldsymbol{\omega}$ to the marginal prior density of $\boldsymbol{\omega}$, each evaluated at $\boldsymbol{\omega}=\boldsymbol{\omega}_0$.

Start with the expression for the posterior of the unconstrained model:

$$p(\boldsymbol{\omega},\boldsymbol{\psi}\mid\mathbf{y},M_2)=\frac{p(\boldsymbol{\omega},\boldsymbol{\psi}\mid M_2)L(\boldsymbol{\omega},\boldsymbol{\psi})}{p(\mathbf{y}\mid M_2)}=\frac{g(\boldsymbol{\omega},\boldsymbol{\psi}\mid M_2)L(\boldsymbol{\omega},\boldsymbol{\psi})}{\int_{\boldsymbol{\psi}}\int_{\boldsymbol{\omega}} g(\boldsymbol{\omega},\boldsymbol{\psi}\mid M_2)L(\boldsymbol{\omega},\boldsymbol{\psi})\,d\boldsymbol{\omega}\,d\boldsymbol{\psi}} \tag{17}$$

Integrate the nominator and denominator over $\boldsymbol{\psi}$ to get the marginal posterior for $\boldsymbol{\omega}$:

$$p(\boldsymbol{\omega}\mid\mathbf{y},M_2)=\int_{\boldsymbol{\psi}}\left(\frac{g(\boldsymbol{\psi}\mid\boldsymbol{\omega},M_2)g(\boldsymbol{\omega})L(\boldsymbol{\omega},\boldsymbol{\psi})}{\int_{\boldsymbol{\psi}}\int_{\boldsymbol{\omega}} g(\boldsymbol{\omega},\boldsymbol{\psi}\mid M_2)L(\boldsymbol{\omega},\boldsymbol{\psi})\,d\boldsymbol{\omega}\,d\boldsymbol{\psi}}\right)d\boldsymbol{\psi} \tag{18}$$

Now divide both sides by $p(\boldsymbol{\omega}\mid M_2)$ and evaluate both sides at $\boldsymbol{\omega}=\boldsymbol{\omega}_0$:

$$\begin{aligned}&\frac{p(\boldsymbol{\omega}=\boldsymbol{\omega}_0\mid\mathbf{y},M_2)}{p(\boldsymbol{\omega}=\boldsymbol{\omega}_0\mid M_2)}=\int_{\boldsymbol{\psi}}\left(\frac{g(\boldsymbol{\psi}\mid\boldsymbol{\omega}=\boldsymbol{\omega}_0,M_2)g(\boldsymbol{\omega}=\boldsymbol{\omega}_0)L(\boldsymbol{\omega}=\boldsymbol{\omega}_0,\boldsymbol{\psi})}{g(\boldsymbol{\omega}=\boldsymbol{\omega}_0)\int_{\boldsymbol{\psi}}\int_{\boldsymbol{\omega}} g(\boldsymbol{\omega},\boldsymbol{\psi}\mid M_2)L(\boldsymbol{\omega},\boldsymbol{\psi})\,d\boldsymbol{\omega}\,d\boldsymbol{\psi}}\right)d\boldsymbol{\psi}=\\ &\int_{\boldsymbol{\psi}}\left(\frac{p(\boldsymbol{\psi}\mid M_1)L(\boldsymbol{\omega}=\boldsymbol{\omega}_0,\boldsymbol{\psi})}{\int_{\boldsymbol{\psi}}\int_{\boldsymbol{\omega}} g(\boldsymbol{\omega},\boldsymbol{\psi}\mid M_2)L(\boldsymbol{\omega},\boldsymbol{\psi})\,d\boldsymbol{\omega}\,d\boldsymbol{\psi}}\right)d\boldsymbol{\psi}=\frac{p(\mathbf{y}\mid M_1)}{p(\mathbf{y}\mid M_2)}=BF_{12}\end{aligned} \tag{19}$$

So we're indeed left with a generic ratio of marginal likelihoods – the general expression for a Bayes Factor.

Chib's (1995) Method of Computing the Marginal Likelihood (mLH)

Matlab scripts: `mod4_sur_chib,`
Matlab functions: `gs_sur_chib`

(Following Koop p.157 ff)

Once again, the generic expression for the marginal likelihood (mLH) is given by

$$p(y)=\int p(y\mid\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\boldsymbol{\theta} \tag{20}$$

The mLH describes how your data should look like before you collect it, given your priors and your structural modeling assumptions (implicitly captured by your likelihood function). For this reason it also referred to as "*prior predictive density*". In rare cases, the analytical form of $p(y)$ is known. Alternatively, plotting this density is straightforward in general: Simply take a draw from the (known) prior $p(\theta)$, plug this draw into your likelihood function (LHF), then draw a value of *y* from $p(y|\boldsymbol{\theta})$. Repeat this process *R* times and plot the resulting *R* draws of *y*. Such plots are often useful in model comparison and model checking.

Now assume you have collected data **y** and estimated the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$. We would like to know if the collected data "confirmed" your original modeling assumption. In other words, you would like to evaluate the *prior* predictive density for the collected data. The better your data agree with your priors and model, the larger (i.e. closer to "1") $p(\mathbf{y})$ will be.

If the analytical form of $p(y)$ is known, you can proceed as for the normal regression model with conjugate priors and directly compute the value of the mLH for your collected data, i.e. the value of $p(\mathbf{y})$ [1]. However, as pointed out earlier, the analytical form of $p(y)$ will be unknown for most applications. Chib (1995) suggested an approach for evaluating $p(\mathbf{y})$ via simulation steps. His approach works well where other Methods (such as Savage-Dickey or Gelfand-Dey) come short – specifically for models with many parameters (= many elements in $\boldsymbol{\theta}$).

We start by writing our basic Bayes' Rule as

$$p(\mathbf{y}) = \frac{p(\boldsymbol{\theta})\, p(\mathbf{y}|\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})} \tag{21}$$

Chib (1995) noted that the left hand side is not an explicit function of $\boldsymbol{\theta}$, so the equality has to hold for *any* $\boldsymbol{\theta}$. For example, we can use the posterior mean $\bar{\boldsymbol{\theta}}$ and write

$$p(\mathbf{y}) = \frac{p(\bar{\boldsymbol{\theta}})\, p(\mathbf{y}|\bar{\boldsymbol{\theta}})}{p(\bar{\boldsymbol{\theta}}|\mathbf{y})} \quad \text{or} \quad \log p(\mathbf{y}) = \log p(\mathbf{y}|\bar{\boldsymbol{\theta}}) + \left(\log p(\bar{\boldsymbol{\theta}}) - \log p(\bar{\boldsymbol{\theta}}|\mathbf{y})\right) \tag{22}$$

Chib refers to this expression as the *basic marginal likelihood identity*. Thus, the log mLH for a given model will be high if the sample likelihood is high at the posterior mean (i.e. if the chosen likelihood function is supported by the data), and if the posterior density at $\bar{\boldsymbol{\theta}}$ is not substantially larger than the prior ordinate at $\bar{\boldsymbol{\theta}}$, controlling for the information content in **y**. The latter condition implies that the prior was ex ante well-chosen. Under vague priors the difference between posterior and prior ordinate in (8) will likely be considerable, especially if the parameter space is large. In that case a high mLH score will largely rest on the appropriateness of the likelihood function.

If we knew the exact form (not just the kernels) of all components on the right hand side, the computation of $p(\mathbf{y})$ would be simple. In most applications we do know the exact analytical form of the prior and the

[1] I switched to a boldface vector **y** to indicate that we're now talking about a specific set of observations, or data.

LHF, but not of the posterior $p(\bar{\boldsymbol{\theta}}\,|\,\mathbf{y})$.[2] Chib proposes a method to evaluate $p(\bar{\boldsymbol{\theta}}\,|\,\mathbf{y})$ via simulation. Once we have this value, we can plug it into (22) and compute $p(\mathbf{y})$.

Suppose in our Bayesian estimation of $\boldsymbol{\theta}$ we broke $\boldsymbol{\theta}$ into 2 components and used a Gibbs Sampler (GS) to sequentially draw from $p(\boldsymbol{\theta}_1\,|\,\boldsymbol{\theta}_2,\mathbf{y})$ and $p(\boldsymbol{\theta}_2\,|\,\boldsymbol{\theta}_1,\mathbf{y})$, where $\boldsymbol{\theta}=[\boldsymbol{\theta}_1' \quad \boldsymbol{\theta}_2']'$. Thus, we have posterior simulator output $\boldsymbol{\theta}_1^r$ and $\boldsymbol{\theta}_2^r$ for r = 1…R. We seek to evaluate $p(\bar{\boldsymbol{\theta}}\,|\,\mathbf{y})=p(\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_2\,|\,\mathbf{y})$. Using – once again – the basic conditioning rule of probability we can write

$$p(\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_2\,|\,\mathbf{y})=p(\bar{\boldsymbol{\theta}}_1\,|\,\mathbf{y})\,p(\bar{\boldsymbol{\theta}}_2\,|\,\bar{\boldsymbol{\theta}}_1,\mathbf{y}) \tag{23}$$

The first term on the right hand side can be formally expressed as

$$p(\bar{\boldsymbol{\theta}}_1\,|\,\mathbf{y})=\int p(\bar{\boldsymbol{\theta}}_1\,|\,\boldsymbol{\theta}_2,\mathbf{y})p(\boldsymbol{\theta}_2\,|\,\mathbf{y})\,d\,\boldsymbol{\theta}_2 \tag{24}$$

We know by now that we can often approximate integrals via Monte Carlo simulation. In this case:

$$p(\bar{\boldsymbol{\theta}}_1\,|\,\mathbf{y})\approx\frac{1}{R}\sum_{r=1}^{R}p(\bar{\boldsymbol{\theta}}_1\,|\,\boldsymbol{\theta}_2^r,\mathbf{y}) \tag{25}$$

Thus, for this first step we can directly use the draws of $\boldsymbol{\theta}_2^r$ generated by our original GS. For each draw, we evaluate the conditional posterior $p(\boldsymbol{\theta}_1\,|\,\boldsymbol{\theta}_2^r,\mathbf{y})$, holding $\boldsymbol{\theta}_1$ at $\bar{\boldsymbol{\theta}}_1$ throughout. We then take the average over the resulting *density* values of $p(\bar{\boldsymbol{\theta}}_1\,|\,\boldsymbol{\theta}_2^r,\mathbf{y})$ (NOT draws!). This is our approximation of $p(\bar{\boldsymbol{\theta}}_1\,|\,\mathbf{y})$. Note that this is the exact same procedure we used for the SDDR method.

The next step is even faster: We simply evaluate the known conditional posterior $p(\boldsymbol{\theta}_2\,|\,\boldsymbol{\theta}_1,\mathbf{y})$ at $\boldsymbol{\theta}_1=\bar{\boldsymbol{\theta}}_1$ and $\boldsymbol{\theta}_2=\bar{\boldsymbol{\theta}}_2$. Thus, the final result for our mLH (in log form) becomes

$$\log p(\mathbf{y})=\log p(\mathbf{y}\,|\,\bar{\boldsymbol{\theta}})+\log p(\bar{\boldsymbol{\theta}})-\left(\log p(\bar{\boldsymbol{\theta}}_1\,|\,\mathbf{y})+\log p(\bar{\boldsymbol{\theta}}_2\,|\,\bar{\boldsymbol{\theta}}_1,\mathbf{y})\right) \tag{26}$$

Chib method with three or more parameter blocks

The Chib method becomes a bit more involved when our original GS breaks $\boldsymbol{\theta}$ into more than two blocks. For three blocks, we have

$$p(\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_2,\bar{\boldsymbol{\theta}}_3\,|\,\mathbf{y})=p(\bar{\boldsymbol{\theta}}_3\,|\,\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_2,\mathbf{y})\,p(\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_2\,|\,\mathbf{y})=p(\bar{\boldsymbol{\theta}}_1\,|\,\mathbf{y})\,p(\bar{\boldsymbol{\theta}}_2\,|\,\bar{\boldsymbol{\theta}}_1,\mathbf{y})\,p(\bar{\boldsymbol{\theta}}_3\,|\,\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_2,\mathbf{y}) \tag{27}$$

We start analogous to the previous case by deriving

---

[2] Note that we may well be able to simulate and (for dimensions of $\boldsymbol{\theta}$ not exceeding 2) plot $p(\boldsymbol{\theta}\,|\,\mathbf{y})$, but to *evaluate* this posterior density at any $\boldsymbol{\theta}$, the exact analytical form of the posterior is needed.

$$p\left(\bar{\boldsymbol{\theta}}_1 \mid \mathbf{y}\right) \approx \frac{1}{R}\sum_{r=1}^{R} p\left(\bar{\boldsymbol{\theta}}_1 \mid \boldsymbol{\theta}_2^r, \boldsymbol{\theta}_3^r, \mathbf{y}\right) \tag{28}$$

using our original draws of $\boldsymbol{\theta}_2^r$ and $\boldsymbol{\theta}_3^r$. Next, we need to approximate

$$p\left(\bar{\boldsymbol{\theta}}_2 \mid \bar{\boldsymbol{\theta}}_1, \mathbf{y}\right) = \int p\left(\bar{\boldsymbol{\theta}}_2 \mid \bar{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_3, \mathbf{y}\right) p\left(\boldsymbol{\theta}_3 \mid \bar{\boldsymbol{\theta}}_1, \mathbf{y}\right) d\,\boldsymbol{\theta}_3 \tag{29}$$

There is now a slight but important difference compared to (24). Specifically, we don't have draws of $\boldsymbol{\theta}_3 \mid \bar{\boldsymbol{\theta}}_1, \mathbf{y}$ from the original GS. In fact, we usually don't even know the exact form of the "partial conditional" density $p\left(\boldsymbol{\theta}_3 \mid \bar{\boldsymbol{\theta}}_1, \mathbf{y}\right)$, and thus can't directly draw $\boldsymbol{\theta}_3$ from it. However, we do know the full conditional density $p\left(\boldsymbol{\theta}_3 \mid \bar{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2, \mathbf{y}\right)$. The two expressions are implicitly related via

$$p\left(\boldsymbol{\theta}_2, \boldsymbol{\theta}_3 \mid \bar{\boldsymbol{\theta}}_1, \mathbf{y}\right) = p\left(\boldsymbol{\theta}_3 \mid \bar{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2, \mathbf{y}\right) p\left(\boldsymbol{\theta}_2 \mid \bar{\boldsymbol{\theta}}_1, \mathbf{y}\right) = p\left(\boldsymbol{\theta}_2 \mid \bar{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_3, \mathbf{y}\right) p\left(\boldsymbol{\theta}_3 \mid \bar{\boldsymbol{\theta}}_1, \mathbf{y}\right) \tag{30}$$

Thus, we can use a ***second (separate) GS***, which draws sequentially from $p\left(\boldsymbol{\theta}_3 \mid \bar{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2, \mathbf{y}\right)$ and $p\left(\boldsymbol{\theta}_2 \mid \bar{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_3, \mathbf{y}\right)$. Thus, it's like the original GS except with $\boldsymbol{\theta}_1$ set to $\bar{\boldsymbol{\theta}}_1$, and – consequentially – no step for drawing $\boldsymbol{\theta}_1$. To be specific, we first draw $\boldsymbol{\theta}_3$ from $p\left(\boldsymbol{\theta}_3 \mid \bar{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2, \mathbf{y}\right)$, using some starting value for $\boldsymbol{\theta}_2$ - for example the last $\boldsymbol{\theta}_2^r$ from the original GS output. Call this first draw of $\boldsymbol{\theta}_3$ " $\boldsymbol{\theta}_3^1$ ". We then draw $\boldsymbol{\theta}_2$ from $p\left(\boldsymbol{\theta}_2 \mid \bar{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_3^1, \mathbf{y}\right)$. Call this draw $\boldsymbol{\theta}_2^1$. Then continue with drawing $\boldsymbol{\theta}_3^2$ from $p\left(\boldsymbol{\theta}_3 \mid \bar{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2^1, \mathbf{y}\right)$, $\boldsymbol{\theta}_2$ from $p\left(\boldsymbol{\theta}_2 \mid \bar{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_3^2, \mathbf{y}\right)$ and so on. At each round *s*, evaluate $p\left(\boldsymbol{\theta}_2 \mid \bar{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_3^s, \mathbf{y}\right)$ at $\bar{\boldsymbol{\theta}}_2$. The sought density in (29) can then be approximated by

$$p\left(\bar{\boldsymbol{\theta}}_2 \mid \bar{\boldsymbol{\theta}}_1, \mathbf{y}\right) \approx \frac{1}{R}\sum_{r=1}^{R} p\left(\bar{\boldsymbol{\theta}}_2 \mid \bar{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_3^r, \mathbf{y}\right) \tag{31}$$

The final step is, again, straightforward. We simply evaluate the known conditional posterior $p\left(\boldsymbol{\theta}_3 \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{y}\right)$ at $\boldsymbol{\theta}_1 = \bar{\boldsymbol{\theta}}_1$, $\boldsymbol{\theta}_2 = \bar{\boldsymbol{\theta}}_2$, and $\boldsymbol{\theta}_3 = \bar{\boldsymbol{\theta}}_3$. Thus, the final result for our mLH (say in log form) becomes

$$\log p\left(\mathbf{y}\right) = \log p\left(\mathbf{y} \mid \bar{\boldsymbol{\theta}}\right) + \log p\left(\bar{\boldsymbol{\theta}}\right) - \left(\log p\left(\bar{\boldsymbol{\theta}}_1 \mid \mathbf{y}\right) + \log p\left(\bar{\boldsymbol{\theta}}_2 \mid \bar{\boldsymbol{\theta}}_1, \mathbf{y}\right) + \log p\left(\bar{\boldsymbol{\theta}}_3 \mid \bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2, \mathbf{y}\right)\right) \tag{32}$$

This concept extends in straightforward fashion to models with >3 blocks of $\boldsymbol{\theta}$. For each additional block, a new separate GS is needed to derive the desired expressions feeding into $p\left(\bar{\boldsymbol{\theta}} \mid \mathbf{y}\right)$.