Advanced Macroeconometrics

Max Heinze, Summer Term 2023

Contents

0 Preliminaries

Statistics

Expectation, Variance, Moments

Informally, the **expected value** of a **random variable** is the mean of a large number of independent draws of that variable, or in short, the **mean** of the random variable. We can think of the expected value of a **distribution** as a measure of the distribution's center.

The n-th **central moment** is defined as $\mu_n = \mathrm{E}(X-\mu)^n$. Thus, the variance, the second central moment, equals $\mathrm{Var}(X) = \mathrm{E}((X-\mu)^2) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$.

For constants $a, b, c \in \mathbb{R}$ functions $g(\cdot)$ and $h(\cdot)$ as well as random variables X and Y,

$$\mathrm{E}(ag(X) + bh(Y) + c) = a\mathrm{E}(g(X)) + b\mathrm{E}(h(Y)) + c \tag{0.1}$$

$$\mathrm{E}(XY) \neq \mathrm{E}(X)\mathrm{E}(Y) \qquad \textbf{except} \text{ if } X \text{ and } Y \text{ are independent} \tag{0.2}$$

$$\mathrm{Var}(ag(X) + bh(Y) + c) = a^2\,\mathrm{Var}(g(X)) + b^2\,\mathrm{Var}(h(Y)) + 2ab\,\mathrm{Cov}(g(x), h(Y)) \tag{0.3}$$

$$\mathrm{Cov}(X, Y) = \mathrm{E}\big((X - \mathrm{E}(X))\ (Y - \mathrm{E}(Y))\big) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \tag{0.4}$$

Notation

We can denote a time series as a **stochastic process**:

$$\{y_t : t = 1, 2, \ldots, T\}.$$

To denote earlier values, we can use the **lag operator**:

$$\mathrm{L}y_t = y_{t-1}, \quad \mathrm{L}^2 y_t = y_{t-2}.$$

To denote differences, we can use the **difference operator**:

$$\Delta y_t = y_t - y_{t-1}.$$

Variance-Covariance Matrix

Consider a univariate time series. Its variance is given by

$$\mathrm{Var}\big(\{y_t\}_{t=1}^T\big) = \sum_{t=0}^{T} \frac{(y_t - \bar{y})^2}{N} = \sum_{t=0}^{T} \frac{(y_t - \bar{y})(y_t - \bar{y})}{N}. \tag{0.5}$$

If $\boldsymbol{y}_t$ is a vector of M time series, the formula yields a variance-covariance matrix of the following form:

$$\begin{bmatrix} \mathrm{Var}(y_{1t}) & \mathrm{Cov}(y_{1t}, y_{2t}) & \ldots & \mathrm{Cov}(y_{1t}, y_{Mt}) \\ \mathrm{Cov}(y_{2t}, y_{1t}) & \mathrm{Var}(y_{1t}) & \ldots & \mathrm{Cov}(y_{2t}, y_{Mt}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(y_{Mt}, y_{1t}) & \mathrm{Cov}(y_{Mt}, y_{2t}) & \ldots & \mathrm{Var}(y_{Mt}) \end{bmatrix} \tag{0.6}$$

Table of Probability Distributions

Discrete Distributions					
Distribution	**Parameters**	**Mean**	**Variance**	**Probability Mass Function**	**Conjugate Prior**
Bernoulli	p	p	$p(1-p)$	$f(k \mid p) = p^k(1-p)^{n-k}$	Beta
Binomial	n, p	np	$np(1-p)$	$f(k \mid n, p) = \binom{n}{k} p^k(1-p)^{n-k}$	Beta
Poisson	λ	λ	λ	$f(k \mid \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$	Gamma
Continuous Distributions					
Distribution	**Parameters**	**Mean**	**Variance**	**Probability Density Function**	**Conjugate Prior**
Normal	μ, σ^2	μ	σ^2	$f(\theta \mid \mu, \sigma) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(\theta-\mu)^2\right)$	μ: N σ^2: G^{-1} μ, σ^2: —
Beta	α, β	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$f(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$	
Gamma	α, β	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$f(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1} e^{-\beta\theta} \beta^\alpha}{\Gamma(\alpha)}$	Gamma
Inverted Gamma	α, β	$\frac{\beta}{\alpha-1}$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$	$f(\theta \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-\alpha-1} \exp\left(-\frac{\beta}{\theta}\right)$	Gamma
Multivar. Normal	$\boldsymbol{\mu}, \boldsymbol{\Sigma}$				N.–Inv. Wishart

1 Univariate Time Series

Time Series Data

Characteristics of Time Series

Time series are observations of quantities that carry a time index:

$$x_t.$$

Unlike in cross-sectional analysis, where we treat observations of individuals, x_i, this index carries meaning and indices are not arbitrarily interchangable. We cannot simply assume that observations are independent, they are usually dependent on prior realizations of themselves.

In economics, we often encounter time series data that is measured at annual, quarterly, or monthly frequency. Examples for this include GDP, the unemployment rate, inflation, and many more. We can analyze this data with different purposes in mind, namely, description, causal inference, and prediction.

Consider the following time series:

$$\begin{aligned} &\{y_t : t = 1, \dots, T\}, \\ &\{x_t : t = 1, \dots, T\}, \\ &\{\varepsilon_t : t = 1, \dots, T\}, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2). \end{aligned} \tag{1.1}$$

Imagine we applied standard linear OLS methods and estimated

$$y_t = x_t\beta + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2),$$

applying the assumptions that $\mathrm{E}(\varepsilon_t) = 0$, $\mathrm{Var}(\varepsilon_t) = \sigma^2$, and $\mathrm{Cov}(\varepsilon_t, \varepsilon_s) = 0 \quad \forall t \neq s$. The third assump-

tion, assuming independence of individual realizations of the error, would be unreasonable in time series contexts, since adjacent values of the error term are often correlated.

Using regular OLS methods to regress one time series on another will very often lead to the coefficient β_j being significantly different from zero, no matter whether there is a relation or not. This is referred to as the **spurious regression problem**.

Instead of using the model

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i,$$

where we can answer questions about ceteris paribus effects and rely on exogeneity of x_{ij}, i.e. $\mathrm{E}(\varepsilon_i \mid x_{i1}, \ldots, x_{iK}) = 0$, we can use the model

$$y_t = \beta_0 + \delta y_{t-1} + x_{t1}\beta_1 + x_{t2}\beta_2 + \cdots + x_{tK}\beta_K + \varepsilon_t, \tag{1.2}$$

where we explicitly model the development of the time series as well. Note that a change in one of the x_{ij} not only has an effect on y_t, but also on all other subsequent realizations of y.

Components of Time Series

For a given time series, variation can come from different sources:

$$y_t = \mu_t + c_t + \varepsilon_t, \tag{1.3}$$

where μ_t is a **trend** component, c_t is a **cyclical** component, and ε_t is an **irregular** component. The trend component covers long term changes in the mean of the time series over time, the cyclical component covers seasonal (e.g., weekday effects) and non-seasonal (e.g., business cycles) oscillations, and the irregular component covers random fluctuations when all other components have been removed.

Usually, we are interested in the random part of the time series. For analyzing this, we have to remove the fixed parts, i.e., **detrend** the time series. We can do this a number of ways, such as:

(1) Using **growth rates** such as year-on-year or month-on-month growth.

(2) Linear or non-linear **de-trending**, deterministic **de-seasonalizing**,

(3) Applying a **filter** (such as the Bandpass Filter or the Hodrick-Prescott Filter).

Time Series Concepts and Properties

Autocovariance and Autocorrelation

The **autocovariance function** at lag j is given by

$$\gamma(j) = \mathrm{Cov}\big(y_t, y_{t-j}\big) = \mathrm{E}\big((y_t - \mu_t)(y_{t-j} - \mu_{t-j})\big), \tag{1.4}$$

that is, the covariance of a value y_t and the value y_{t-j} j periods before it.

The **autocorrelation function** (ACF) at lag j is then given by

$$\rho(j) = \text{Cor}\left(y_t, y_{t-j}\right) = \frac{\text{Cov}\left(y_t, y_{t-j}\right)}{\sigma^2} = \text{E}\left(\frac{(y_t - \mu_t)(y_{t-j} - \mu_{t-j})}{\sigma^2}\right). \tag{1.5}$$

Both the autocovariance and the autocorrelation function provide information about the memory of the process, that is, how much information from past observation persists through to the current observation at time t. The values of the autocorrelation function can, like any correlation, never have an absolute value greater than 1:

$$-1 \leq \rho(j) \leq 1.$$

In addition, by definition, $\rho(0) = 1$.

When dealing with real data, we can assess a time series's **empirical autocorrelation** given by

$$\hat{\rho}(j) = \frac{1}{T} \sum_{t=j+1}^{T} \frac{y_t - \bar{y}_t}{s_y} \frac{y_{t-j} - \bar{y}_t}{s_y}, \tag{1.6}$$

where s_y is the standard deviation of y_t. We often use plots of the empirical autocorrelation in order to assess the adjacency of time series value for different time series. It can also be used to assess the residuals of a regression.

Durbin-Watson Test

We can use the Durbin-Watson test to check whether first-order serial correlation of OLS residuals is present. The test statistic is given by:

$$d = \frac{\sum_{t=2}^{T} (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{T} \hat{\varepsilon}_t^2}. \tag{1.7}$$

In the case of no serial correlation, the expected value of the test statistic is $\text{E}(d) = \frac{\text{E}\left(\varepsilon_t^2\right) + \text{E}\left(\varepsilon_{t-1}^2\right)}{\text{E}\left(\varepsilon_t^2\right)} = \frac{2\sigma^2}{\sigma^2} = 2$. If the value of the test statistic is $0 \leq d < 2$, autocorrelation at lag 1 is positive, if the value of the test statistic is $2 < d \leq 4$, there is negative autocorrelation at lag 1.

Time Series Econometrics

A **stochastic process** is a collection of random variables defined on a common probability space. A **time series** $\{y_t : t = 1, \ldots, T\}$ is a realization of a stochastic process in the time dimension.

We observe the data, a time series, and try to infer the data-generating process, that is, the specific stochastic process that the time series is a realization of. The unconditional mean of the time series is $\text{E}(y_t) = \mu$ and its variance is $\text{Var}(y_t) = \sigma^2$. In general, the random variables Y_t and Y_s at different time points t and s are not independent of each other.

Stationarity

Stationarity is one important property of stochastic processes. Note that stationarity is a property of the stochastic process, and not of the time series, which is a realization of that process.

A stochastic process $\{y_t : t, 1, \ldots, T\}$ is said to be **covariance stationary** or **weakly stationary** if it fulfills the following three conditions:

(1) The **mean** is constant for all t:

$$\mathrm{E}(y_t) = \mu \quad \forall t$$

(2) The **variance** is constant for all t:

$$\mathrm{Var}(y_t) = \sigma^2 \quad \forall t$$

(3) The **covariance** at two time periods depends only on the difference between the two time periods, that is, the lag, and not on the actual time at which the covariance is computed. In other words, the autocorrelation is time invariant:

$$\mathrm{E}\big((y_t - \mu)(y_{t-j} - \mu)\big) = \gamma_j \quad \forall t \text{ and for any } j.$$

A stochastic process $\{y_t : t, 1, \ldots, T\}$ is said to be **strictly stationary** if the joint distribution of $(y_{t1}, y_{t2}, \ldots, y_{tk})$ is the same as that of $(y_{t1+h}, y_{t2+h}, \ldots, y_{tk+h})$. That is, the joint distribution depends only on the difference h, not on time t.

Weak stationarity does not imply strict stationarity, since it only requires the first two unconditional moments to be time invariant, while strict stationarity requires the whole joint distribution to be independent of time. Even the converse does not necessarily hold: Strict stationarity does not necessarily imply weak stationarity. However, it does imply weak stationarity when the first two unconditional moments exist and are finite. If the mean or variance do not exist or are not finite, then the process can be strictly stationary without being weakly stationary.

A process is called **trend stationary** if removing an underlying trend from the process results in a weakly stationary process.

Stationary Models

White Noise Process

A white noise process is described by

$$y_t = \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2). \tag{1.8}$$

It is thus a sequence of random numbers. It is not autocorrelated beyond lag $j = 0$. Any white noise process is weakly stationary. A white noise process as defined above is also strictly stationary.

Moving Average Process

A moving average (1), or MA(1) in short, process is defined as

$$y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}. \tag{1.9}$$

Thus, y_t is described as the weighted sum of the current error and past errors. If the coefficient $\theta = 0$, the process is reduced to a white noise process.

The unconditional moments of an MA(1) process are

$$\begin{aligned}
&\mathrm{E}(y_t) = \mu, \\
&\quad \gamma_0 = \mathrm{E}\big((y_t-\mu)^2\big) = (1+\theta^2)\sigma^2, \\
&\quad \gamma_1 = \mathrm{E}((y_t-\mu)(y_{t-1}-\mu)) = \theta\sigma^2, \\
&\quad \gamma_j = \mathrm{E}\big((y_t-\mu)(y_{t-j}-\mu)\big) = 0 \text{ for } j>1.
\end{aligned}$$

An MA(1) process is therefore weakly stationary. The autocorrelation for lags $j > 1$ is zero.

A moving average (q), or MA(q), process is defined as

$$y_t = \mu + \sum_{j=0}^{q} \theta_j \varepsilon_{t-j}, \qquad \varepsilon_t \sim \mathcal{N}(0,\sigma^2), \quad \theta_0 = 1. \tag{1.10}$$

The unconditional moments of an MA(q) process are

$$\begin{aligned}
&\mathrm{E}(y_t) = \mu, \\
&\quad \gamma_0 = \mathrm{E}\big((y_t-\mu)^2\big) = (\theta_0^2+\theta_1^2+\theta_2^2+\cdots+\theta_q^2)\sigma^2, \\
&\quad \gamma_1 = \mathrm{E}((y_t-\mu)(y_{t-1}-\mu)) = \sigma^2(\theta_j\theta_0+\theta_{j+1}\theta_1+\theta_{j+2}\theta_2+\cdots) \text{ for } j>0, \\
&\quad \gamma_j = 0 \text{ for } j>q.
\end{aligned}$$

Thus, an MA(q) process is weakly stationary if the sums $\sum_{i=1}^{q}\theta_i^2$ and $(\theta_s+\theta_{j+1}\theta_1+\theta_{j+2}\theta_2+\cdots)$ are finite.

Autoregressive Process

In an **autoregressive model**, y_t is regressed onto its past values $y_{t-1}, y_{t-2}, \ldots$. Thus, y_t is explained by its own past, plus a random error term.

An **autoregressive process** of order 1, or AR(1), is defined as

$$y_t = c + \varphi y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,\sigma^2). \tag{1.11}$$

The parameter c is called drift and influences the mean of the process. The process is stationary if $|\varphi| < 1$, reduced to white noise if $\varphi = 0$, a random walk if $\varphi = 1$, explosive if $\varphi > 1$, and oscillating if $\varphi < 0$.

Assuming that $y_0 = \varepsilon_0$, we find for y_t (by substituting):

$$y_1 = \varphi\varepsilon_0 + \varepsilon_1.$$

Similarly, we find for y_2:

$$y_2 = \varphi(\varphi\varepsilon_0 + \varepsilon_1) + \varepsilon_2 = \varphi^2\varepsilon_0 + \varphi\varepsilon_1 + \varepsilon_2.$$

By doing this again and again, which is called the **Wold transformation**, we find that

$$y_t = \sum_{j=0}^{\infty} \varphi^j \varepsilon_{t-j}, \tag{1.12}$$

which is an MA(∞) process. We find for its variance and covariance:

$$\begin{aligned}
\gamma_0 &= \mathrm{Var}(y_t) = (1 + \varphi^2 + \varphi^4 + \varphi^6 + \cdots)\sigma^2, \\
\gamma_1 &= \mathrm{Cov}\big(y_t, y_{t-j}\big) = (1 + \varphi^2 + \varphi^4 + \varphi^6 + \cdots)\sigma^2\varphi^j,
\end{aligned}$$

which are finite if $|\varphi| < 1$. Then,

$$\begin{aligned}
\gamma_0 &= \frac{\sigma^2}{1-\varphi^2}, \\
\gamma_1 &= \frac{\sigma^2\varphi^j}{1-\varphi^2}.
\end{aligned}$$

Since for a stationary AR(1) process, the autocovariance at lag j is given by $\frac{\varphi^j\sigma^2}{1-\varphi^2}$, and the variance is given by $\frac{\sigma^2}{1-\varphi^2}$, the autocorrelation function is given by

$$\rho(j) = \frac{\gamma(j)}{\mathrm{Var}(y_t)} = \varphi^j, \tag{1.13}$$

which decays to zero for a stationary AR(1) process where $|\varphi| < 1$.

An AR(1) process with intercept, $y_t = c + \varphi y_{t-1} + \varepsilon_t,\ \varepsilon \sim \mathcal{N}(0, \sigma^2)$, can be characterized as

$$y_t = \frac{c}{1-\varphi} + \sum_{j=0}^{\infty} \varphi^j \varepsilon_{t-j},$$

and is stationary if $|\varphi| < 1$ because then $\sum_{j=0}^{\infty} |\theta_j| = \sum_{j=0}^{\infty} |\varphi|^j < \infty$.

The unconditional moments of an AR(1) process with intercept are:

$$
\begin{aligned}
\mathrm{E}(y_t) = \mu &= \frac{c}{1-\varphi}, \\
\gamma_0 &= \mathrm{E}\big((y_t-\mu)^2\big) = \frac{\sigma^2}{1-\varphi^2}, \\
\gamma_j &= \mathrm{E}\big((y_t-\mu)(y_{t-j}-\mu)\big) = \frac{\varphi^j}{1-\varphi^2}\sigma^2, \\
\rho_j &= \frac{\gamma_j}{\gamma_0} = \varphi^j.
\end{aligned}
$$

As long as $\varphi \neq 0$, the subsequent values of y_t are dependent on the preceding values. That means that the conditional distribution of y_t given y_{t-1} is different from the unconditional distribution of y_t. In forecasting, knowing the immediate past can therefore be used to forecast y_t. For an AR(1) process, the long-run mean is equal to $\mathrm{E}(y_t) = \mu$, whereas the conditional mean is

$$
\mathrm{E}(y_t \mid y_{t-1}) = c + \varphi y_{t-1}.
$$

For zero-mean processes, that is, AR(1) processes where $c = 0$, the model is linear in the unknown parameter φ and standard OLS estimation is therefore unbiased and consistent. However, if $c \neq 0$, this is not the case anymore and using OLS yields problems such as small sample bias and (a positive) φ getting underestimated. Therefore, there is a need to either use maximum likelihood estimation or use this knowledge as prior information in a Bayesian approach.

An **autoregressive process of order** p or AR(p) process can be written as

$$
y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2). \tag{1.14}
$$

Stationarity conditions are more complicated and depend on the order p of the process. For an AR(2) process, stationarity conditions are given as

$$
\begin{aligned}
\varphi_1 + \varphi_2 &< 1, \\
\varphi_2 - \varphi_1 &< 1, \text{ and} \\
|\varphi_2| &< 1.
\end{aligned}
$$

The unconditional moments of an AR(p) process are:

$$
\begin{aligned}
\mathrm{E}(y_t) = \mu = \psi(L)c &= \frac{c}{1-\varphi_1-\varphi_2-\ldots-\varphi_p} \\
\gamma_j &= \begin{cases} \varphi_1\gamma_1 + \varphi_2\gamma_2 + \cdots + \varphi_p\gamma_p + \sigma^2 & \text{if } j = 0 \\ \varphi_1\gamma_{j-1} + \varphi_2\gamma_{j-2} + \cdots + \varphi_p\gamma_{j-p} & \text{if } j > 0 \end{cases} \\
\rho_j &= \varphi_1\rho_{j-1} + \varphi_2\rho_{j-2} + \cdots + \varphi_p\rho_{j-p} \quad \text{for } j = 1, 2, \ldots
\end{aligned}
$$

The autocorrelation function is also known as the Yule-Walker Equations.

Autoregressive Moving Average Process

A process y_t that contains both p AR terms and q MA terms is called an ARMA(p,q) process. It is given by

$$y_t = c + \sum_{i=0}^{p} \varphi_i y_{t-i} + \sum_{j=0}^{q} \theta_j \varepsilon_{t-j}, \qquad \varepsilon \sim \mathcal{N}(0,\sigma^2), \quad \theta_0 = 1. \tag{1.15}$$

We can think of an ARMA model as a multiple regression model with past observations and past errors as predictors. An ARMA process is stationary if and only if the AR part defines a stationary process.

Nonstationary Models

Nonstationarity may occur in many ways, such as non-constant means due to deterministic trends, non-constant variances, seasonal patterns, structural breaks, or unit roots. Depending on the source, we need different **transformations** to achieve stationarity, such as detrending for trend-stationary processes, removing heteroskedasticity, or using the d-th difference so that $\Delta^d y_t$ is stationary.

The **random walk** process is given by

$$y_t = y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,\sigma^2), \tag{1.16}$$

that is, an AR(1) process with $\varphi = 1$. A random walk is not stationary as its variance tends to infinity and the mean is dependent on time. Its autocorrelation decays only very slowly. Straightforwardly, taking first differences yields a white noise process, which is stationary.

Unit Root Nonstationarity

Consider the AR(1) process given by:

$$y_t = \varphi y_{t-1} + \varepsilon_t.$$

Rewriting the process using the lag operator yields:

$$y_t = \varphi y_{t-1} + \varepsilon_t \Leftrightarrow (1-\varphi L) y_t = \varepsilon_t.$$

Then, the characteristic polynomial of the process is $(1-\varphi L)$ and the process is stationary if its root is above unity:

$$(1-\varphi L) = 0 \quad \Rightarrow \quad L = \varphi^{-1},$$

thus, $|L| > 1 \Leftrightarrow |\varphi| < 1$. If y_t is a random walk process, then $L = 1$, i.e. y_t has a unit root.

A random walk process of order p has the following characteristic polynomial of order p:

$$(1 - \varphi_1 L - \varphi_2 L^2 - \ldots - \varphi_p L^P) = 0.$$

The characteristic roots are the values of the lag operator that solve the polynomial. y_t is stationary if all roots lie outside the unit circle.

Integrated Time Series

Stationary time series are said to be integrated of order zero, or $I(0)$. Time series whose d-th differences are stationary are called integrated of order d, or $I(d)$. $I(1)$ processes are referred to as unit root processes. In economics, we usually deal with time series that are either $I(0)$ or $I(1)$.

Testing for Unit Roots

Using a t test to test for a unit root in an AR(1) process is not valid, since the t-statistic is asymptotically t-distributed only if the true value of the coefficient we hypothesize to be 1 is not actually 1. We therefore need to use a different test.

The **Dickey-Fuller Test** is a statistical test to test for unit roots. To construct the test statistic, we rewrite the AR(1) process like this:

$$\Delta y_t = y_t - y_{t-1} = (\varphi - 1) y_{t-1} + \varepsilon_t.$$

Letting $\kappa = \varphi - 1$, we can write

$$\Delta y_t = \kappa y_{t-1} + \varepsilon_t \tag{1.17}$$

and test for

$$H_0 : \kappa = 0,$$
$$H_1 : \kappa < 0.$$

The null hypothesis is rejected if the test statistic is smaller than a critical value (one-sided test). Rejecting the null hypothesis means that the process has no unit root and that $\varphi < 1$. There is no well-known asymptotic distribution of the test statistic, critical values depend on sample size and whether the model includes a drift term and further predictors (Augmented Dickey-Fuller Test, ADF Test). However, if the null hypothesis is not rejected, there is the possibility that nonstationarity is due to other forms of nonstationarity (particularly if the test statistic is positive). Also, misspecification of the underlying model can lead to the test not working properly.

The ARIMA Model

In practice, many time series are nonstationary, but can be made stationary by taking first differences. In addition, the series of first differences Δy_t of a nonstationary time series often exhibits autocorrelation. We can then model Δy_t as an ARMA(p, q) process. We call y_t an ARIMA$(p, 1, q)$ process if

$$y_t = y_{t-1} + \varepsilon_t, \qquad \text{where } \varepsilon_t \text{ follows an ARMA}(p, q) \text{ process,} \tag{1.18}$$

since then $\Delta y_t = \varepsilon_t$ is modeled as an ARMA(p, q) process. The process is called ARIMA (with the I standing for "integrated") since we sum up, or integrate, the Δy_t.

The ARIMA$(p, 1, q)$ model can be generalized to an ARIMA(p, d, q) model, where d defines the number of differences taken on y_t before an ARMA(p, q) model is applied.

2 Multivariate Time Series

Introduction

The models described in the previous section help in explaining the current value of a variable by lags of itself and previous shocks. However, we often want to investigate the interdependency of two or more time series, for which we can use **vector autoregressive (VAR)** models. These allow us to describe macroeconomic time series, predict future realizations, explain their (structural) behavior and use this information to give policy advice.

The Vector Autoregressive Model

Consider the three macroeconomic time series GDP (Δy), inflation (π) and a policy interest rate (r). A vector autoregressive model allows us to answer questions about the dynamic behavior and interactions of the variables, effects of shocks in one variable on the others, and trajectories of one variable conditional on some future path for another.

To write a VAR model, we first stack the time series like this:

$$\boldsymbol{y} = \begin{bmatrix} \Delta y_1 & \Delta y_2 & \dots & \Delta y_T \\ \pi_1 & \pi_2 & \dots & \pi_T \\ r_1 & r_2 & \dots & r_T \end{bmatrix},$$

where we denote as $\boldsymbol{y}_t$ the 3×1 vector s.t. $\boldsymbol{y}_t' = (\Delta y_t, \pi_t, r_t)$. A potential representation of the VAR model then is given by

$$\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{y}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}). \tag{2.1}$$

This is, using the present example, a more compact form of

$$\begin{bmatrix} \Delta y_t \\ \pi_t \\ r_t \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} \Delta y_{t-1} \\ \pi_{t-1} \\ r_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{\Delta y_t} \\ \varepsilon_{\pi_t} \\ \varepsilon_{r_t} \end{bmatrix},$$

which corresponds to the following system of linear equations:

$$\begin{aligned} \Delta y_t &= a_{11}\Delta y_{t-1} + a_{12}\pi_{t-1} + a_{13}r_{t-1} + \varepsilon_{\Delta y_t}, \\ \pi_t &= a_{21}\Delta y_{t-1} + a_{22}\pi_{t-1} + a_{23}r_{t-1} + \varepsilon_{\pi_t}, \\ r_t &= a_{31}\Delta y_{t-1} + a_{32}\pi_{t-1} + a_{33}r_{t-1} + \varepsilon_{r_t}. \end{aligned}$$

Note that the variance-covariance matrix of the error terms, $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, has non-zero off-diagonal elements which capture contemporaneous relations.

Suppose we are interested in the relationship between a set of M time series variables $\boldsymbol{y}_t = (y_{1t}, y_{2t}, \cdots, y_{Mt})$ and that the DGP consists of a deterministic part $\boldsymbol{\mu}_t$ and a purely stochastic part $\boldsymbol{x}_t$ with mean zero,

$$\boldsymbol{y}_t = \boldsymbol{\mu}_t + \boldsymbol{x}_t, \tag{2.2}$$

where $\boldsymbol{x}_t$ is assumed to follow a linear VAR process of order p, that is,

$$\boldsymbol{x}_t = \boldsymbol{A}_1 \boldsymbol{x}_{t-1} + \cdots + \boldsymbol{A}_p \boldsymbol{x}_{t-p} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}),$$

where $\boldsymbol{A}_i$ denote $M \times M$ coefficient matrices and $\boldsymbol{\varepsilon}_t$ denotes a zero-mean Gaussian error vector with am $M \times M$ variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. In this model, $\mathrm{E}\left(\boldsymbol{y}_t\right) = \boldsymbol{\mu}_t$ is the unconditional mean of the DGP, which we consider to be constant for simplicity, that is, $\boldsymbol{\mu}_t = \boldsymbol{\mu}_0 = \boldsymbol{\mu}$. Using the lag polynomial $\boldsymbol{A}(L) = \boldsymbol{I} - \boldsymbol{A}_1 L - \ldots - \boldsymbol{A}_p L^p$, where L is the lag operator as before, we can rewrite the above VAR process as

$$\boldsymbol{A}(L)\boldsymbol{x}_t = \boldsymbol{\varepsilon}_t,$$

which we can pre-multiply to Equation 2.2 to get

$$\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \ldots + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}_M(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{2.3}$$

where $\boldsymbol{c} = \boldsymbol{A}(L)\boldsymbol{\mu}_0$ denotes a constant. This is the **reduced form** of a VAR model. More compactly, we can write it using the lag polynomial as

$$\boldsymbol{A}(L)\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{\varepsilon}_t.$$

A **stable** VAR process has time-invariant means, time-invariant variances and a time-invariant covariance structure and is hence stationary. A VAR process is stable if all roots of the determinantal polynomial of the VAR operator lie outside the unit circle, i.e.

$$\det(A(z)) = \det(\boldsymbol{I} - \boldsymbol{A}_1 L - \ldots - \boldsymbol{A}_p L^p) = 0 \quad \forall L \in \mathbb{C}, \quad |L| \geq 1. \tag{2.4}$$

We can rewrite an AR(p) process as an AR(1) process. Consider the following AR(p) process with $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$:

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + \varepsilon_t.$$

We cabn stack the y values into the vector $\boldsymbol{Y}$ and rewrite the process in **companion form**:

$$\underbrace{\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \end{bmatrix}}_{\boldsymbol{Y}_t} = \underbrace{\begin{bmatrix} \varphi_1 & \varphi_2 & \cdots & \varphi_{p-1} & \varphi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}}_{\Phi} \underbrace{\begin{bmatrix} y_{t-1} \\ y_{t-2} \\ y_{t-3} \\ \vdots \\ y_{t-p} \end{bmatrix}}_{\boldsymbol{Y}_{t-1}} + \underbrace{\begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{\mathrm{E}_t},$$

which gives us the AR(1) process

$$\underset{p\times 1}{\boldsymbol{Y}_t} = \underset{p\times p}{\boldsymbol{\Phi}} \underset{p\times 1}{\boldsymbol{Y}_{t-1}} + \underset{p\times 1}{\boldsymbol{E}_t}$$

Similarly, we can rewrite any VAR(p) process as a VAR(1) process:

$$
\begin{aligned}
\boldsymbol{Y}_t &= \boldsymbol{\nu} + \boldsymbol{A}\boldsymbol{Y}_{t-1} + \boldsymbol{E}_t \\
\boldsymbol{J}'\boldsymbol{Y}_t &= \boldsymbol{y}_t = \boldsymbol{J}'(\boldsymbol{\nu} + \boldsymbol{A}\boldsymbol{Y}_{t-1} + \boldsymbol{E}_t) \\
\boldsymbol{J}'\boldsymbol{Y}_t &= \boldsymbol{y}_t = \boldsymbol{c} + [\boldsymbol{A}_1\, \boldsymbol{A}_2\, \cdots\, \boldsymbol{A}_p]\boldsymbol{Y}_{t-1} + \boldsymbol{\varepsilon}_t,
\end{aligned}
$$

where

$$
\underset{Mp\times 1}{\boldsymbol{Y}_t} \equiv \begin{bmatrix} \boldsymbol{y}_t \\ \boldsymbol{y}_{t-1} \\ \vdots \\ \boldsymbol{y}_{t-p+1} \end{bmatrix}, \underset{Mp\times Mp}{\boldsymbol{A}} \equiv \begin{bmatrix} \boldsymbol{A_1} & \boldsymbol{A_2} & \dots & \boldsymbol{A_{p-1}} & \boldsymbol{A_p} \\ \boldsymbol{I_M} & \boldsymbol{0_M} & \dots & \boldsymbol{0_M} & \boldsymbol{0_M} \\ \boldsymbol{0_M} & \boldsymbol{I_M} & & \boldsymbol{0_M} & \boldsymbol{0_M} \\ \vdots & & \ddots & \vdots & \vdots \\ \boldsymbol{0_M} & \boldsymbol{0_M} & \dots & \boldsymbol{I_M} & \boldsymbol{0_M} \end{bmatrix},
$$

$$
\underset{Mp\times 1}{\boldsymbol{\nu}} \equiv \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{0_M} \\ \vdots \\ \boldsymbol{0_M} \end{bmatrix}, \quad \underset{Mp\times 1}{\boldsymbol{E_t}} \equiv \begin{bmatrix} \varepsilon_t \\ \boldsymbol{0_M} \\ \vdots \\ \boldsymbol{0_M} \end{bmatrix} = \boldsymbol{J}\boldsymbol{\varepsilon}_t, \underset{Mp\times M}{\boldsymbol{J}} \equiv \begin{bmatrix} \boldsymbol{I_M} \\ \boldsymbol{0_M} \\ \vdots \\ \boldsymbol{0_M} \end{bmatrix}.
$$

Such a VAR process is stable if

$$
\det(\boldsymbol{I_{Mp}} - \boldsymbol{A}(\boldsymbol{L})) = 0 \quad \forall L \in \mathbb{C}, \quad |L| \geq 1.
$$

By construction, the eigenvalues of $\boldsymbol{A}$ are the reciprocals of the roots of the VAR lag polynomial. The condition is thus equivalent to the real part of all eigenvalues of $\boldsymbol{A}$ having modulus (absolute value) less than one.

A VAR(p) model can be estimated using OLS, generalized LS, maximum likelihood, or Bayesian estimation techniques. The number of free parameters in a VAR, that is, the number of elements in $\boldsymbol{c}$, $\boldsymbol{A}_j$ ($j = 1, \dots, p$), and $\boldsymbol{\Sigma}$ equals

$$
M(Mp + 1) + \frac{(M+1)M}{2},
$$

which grows rapidly as M and p increase. This is also referred to as the **curse of dimensionality**. Frequentist approaches struggle with a such large number of parameters and can therefore produce imprecise results.

Model Specification, Model Diagnostics and Predictions

When specifying a VAR model, decisions have to be taken on which variables to include, which transformations to use and how many lags to use. It is important to check the M time series for stationarity and trend properties. In case of suspected cointegration, a Vector Error Correction Model (VECM) or Bayesian approaches can be used.

In a frequentist setting, there are different procedures that can be used to **determine the lag order** p. These

include sequential testing procedures (top-down procedures where models with different p are estimated and it is checked whether or not $\boldsymbol{A}_p = 0$; or bottom-up procedures that start with the smallest model and test for residual autocorrelation) and usage of information criteria like AIC, HQC, and BIC. In a Bayesian setting, shrinkage is often induced on coefficients associated with higher lag orders, which pushes them towards zero. In addition, economic intuition can be used.

There is a large set of tools for checking whether a given VAR model represents the DGP of the variables adequately, such as tests for autocorrelation in the innovations, nonnormality of the innovations, heteroskedasticity in the error terms, or time invariance of the model (i.e. stationarity).

Predictions

Rewriting any VAR(p) process as a VAR(1) process, as explained before, helps for some applications:

$$\begin{aligned}
\boldsymbol{Y}_t &= \boldsymbol{\nu} + \boldsymbol{A}\boldsymbol{Y}_{t-1} + \boldsymbol{E}_t \\
\boldsymbol{J}'\boldsymbol{Y}_t = \boldsymbol{y}_t &= \boldsymbol{J}'(\boldsymbol{\nu} + \boldsymbol{A}\boldsymbol{Y}_{t-1} + \boldsymbol{E}_t) \\
\boldsymbol{J}'\boldsymbol{Y}_t = \boldsymbol{y}_t &= \boldsymbol{c} + [\boldsymbol{A}_1\, \boldsymbol{A}_2\, \cdots\, \boldsymbol{A}_p]\boldsymbol{Y}_{t-1} + \boldsymbol{\varepsilon}_t,
\end{aligned}$$

where

$$\underset{Mp\times 1}{\boldsymbol{Y}_t} \equiv \begin{bmatrix} \boldsymbol{y}_t \\ \boldsymbol{y}_{t-1} \\ \vdots \\ \boldsymbol{y}_{t-p+1} \end{bmatrix}, \underset{Mp\times Mp}{\boldsymbol{A}} \equiv \begin{bmatrix} \boldsymbol{A_1} & \boldsymbol{A_2} & \dots & \boldsymbol{A_{p-1}} & \boldsymbol{A_p} \\ \boldsymbol{I_M} & \boldsymbol{0_M} & \dots & \boldsymbol{0_M} & \boldsymbol{0_M} \\ \boldsymbol{0_M} & \boldsymbol{I_M} & & \boldsymbol{0_M} & \boldsymbol{0_M} \\ \vdots & & \ddots & \vdots & \vdots \\ \boldsymbol{0_M} & \boldsymbol{0_M} & \dots & \boldsymbol{I_M} & \boldsymbol{0_M} \end{bmatrix},$$

$$\underset{Mp\times 1}{\boldsymbol{\nu}} \equiv \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{0_M} \\ \vdots \\ \boldsymbol{0_M} \end{bmatrix}, \quad \underset{Mp\times 1}{\boldsymbol{E_t}} \equiv \begin{bmatrix} \boldsymbol{\varepsilon_t} \\ \boldsymbol{0_M} \\ \vdots \\ \boldsymbol{0_M} \end{bmatrix} = \boldsymbol{J}\boldsymbol{\varepsilon}_t, \underset{Mp\times M}{\boldsymbol{J}} \equiv \begin{bmatrix} \boldsymbol{I_M} \\ \boldsymbol{0_M} \\ \vdots \\ \boldsymbol{0_M} \end{bmatrix}.$$

The **companion form** allows us to derive straightforward expressions for the h-step ahead forecast:

$$\boldsymbol{y}_{t+h|t} = \boldsymbol{J}'\boldsymbol{A}^h\boldsymbol{Y}_t = \boldsymbol{J}' \begin{bmatrix} \boldsymbol{A_1} & \boldsymbol{A_2} & \dots & \boldsymbol{A_{p-1}} & \boldsymbol{A_p} \\ \boldsymbol{I_M} & \boldsymbol{0_M} & \dots & \boldsymbol{0_M} & \boldsymbol{0_M} \\ \boldsymbol{0_M} & \boldsymbol{I_M} & & \boldsymbol{0_M} & \boldsymbol{0_M} \\ \vdots & & \ddots & \vdots & \vdots \\ \boldsymbol{0_M} & \boldsymbol{0_M} & \dots & \boldsymbol{I_M} & \boldsymbol{0_M} \end{bmatrix}^h \boldsymbol{Y}_t. \tag{2.5}$$

The prediction error associated with an h-step ahead prediction is

$$\boldsymbol{e}_{t+h|t} = \boldsymbol{y}_{t+h} - \boldsymbol{y}_{t+h|t} = \boldsymbol{J}'(\boldsymbol{Y}_{t+h} - \boldsymbol{Y}_{t+h|t}). \tag{2.6}$$

Mean and variance of the prediction error are then

$$\mathrm{E}\left(\boldsymbol{e}_{t+h|t}\right) = \mathbf{0} \quad \forall t, h,$$
$$\mathrm{Var}\left(\boldsymbol{e}_{t+h|t}\right) = \sum_{j=0}^{\infty} \boldsymbol{\Phi}_j \boldsymbol{\Sigma}_\varepsilon \boldsymbol{\Phi}_j',$$

where $\boldsymbol{\Phi}_j = \boldsymbol{J}'\boldsymbol{A}^j\boldsymbol{J}$. However, this measures only the uncertainty associated with the forecast errors and ignores parameter uncertainty.

There are several metrics available to evaluate prediction accuracy, such as the root-mean squared error,

$$\mathrm{RMSE} = \sqrt{\mathrm{E}\left((\boldsymbol{y}_{t+h} - \boldsymbol{y}_{t+h|t})^2\right)}, \tag{2.7}$$

or log-predictive density scores,

$$\mathrm{LPDS} = \log p(\boldsymbol{y}_{t+h} \mid \boldsymbol{\theta}), \tag{2.8}$$

where $p(\cdot)$ denotes the predictive density of $\boldsymbol{y}_{t+h}$.

Structural Vector Autoregressions

With the above, we can describe and summarize macroeconomic time series and compute forecasts. However, it does not allow us to understand how variables interact and what the effect of a shock in one variable is on the behavior of another variable.

Similarly to the Wold transformation in a univariate setting, we can represent a stable VAR(p) process as the weighted sum of past and present innovations, i.e., as a VMA(∞) process:

We have

$$\boldsymbol{Y}_t = \boldsymbol{\nu} + \boldsymbol{A}\boldsymbol{Y}_{t-1} + \boldsymbol{E}_t$$

Recursive Substitution yields: bm

$$\begin{aligned}
\boldsymbol{Y}_t &= \boldsymbol{\nu} + \boldsymbol{A}(\boldsymbol{\nu} + \boldsymbol{A}\boldsymbol{Y}_{t-2} + \boldsymbol{E}_{t-1}) + \boldsymbol{E}_t \\
&= \sum_{j=0}^{1} \boldsymbol{A}^j \boldsymbol{\nu} + \boldsymbol{A}^2 \boldsymbol{Y}_{t-2} + \sum_{j=0}^{1} \boldsymbol{A}^j \boldsymbol{E}_{t-j} \\
&\vdots \\
&= \sum_{j=0}^{k} \boldsymbol{A}^j \boldsymbol{\nu} + \boldsymbol{A}^{k+1} \boldsymbol{Y}_{t-(k+1)} + \sum_{j=0}^{k} \boldsymbol{A}^j \boldsymbol{E}_{t-j}.
\end{aligned}$$

Taking the limit as k approaches ∞ yields

$$
\begin{aligned}
\boldsymbol{Y_t} &= \lim_{k\to\infty} \sum_{j=0}^{k} \boldsymbol{A}^j \boldsymbol{\nu} + \boldsymbol{A}^{k+1} \boldsymbol{Y}_{t-(k+1)} + \sum_{j=0}^{k} \boldsymbol{A}^j \boldsymbol{E}_{t-j} \\
&= \sum_{j=0}^{\infty} \boldsymbol{A}^j \boldsymbol{\nu} + \sum_{j=0}^{\infty} \boldsymbol{A}^j \boldsymbol{E}_{t-j} \\
&= \left(\boldsymbol{I}_{Mp} - \boldsymbol{A}\right)^{-1} \boldsymbol{\nu} + \sum_{j=0}^{\infty} \boldsymbol{A}^j \boldsymbol{E}_{t-j}
\end{aligned}
$$

Thus, when estimating a VAR model we think of our (economic) system being driven by weighted current and past (exogenous) shocks.

The correlation of the residuals reflects contemporaneous relations between the variables. Thus, they are a mixture of underlying structural shocks, which is the reason why we cannot interpret reduced form errors $\boldsymbol{\varepsilon}_t$ as structural shocks. Reduced-form VARs do not tell us anything about the structure of the economy.

Since reduced form errors $\boldsymbol{\varepsilon}_t$ are a linear combination of the structural errors, which in the above example would have been $\boldsymbol{e}_{\Delta y}$, $\boldsymbol{e}_{\pi}$ and $\boldsymbol{e}_{r}$, it is hard to find the nature of the shock. We run into an identification problem, since we do not know which of the structural errors causes the changes we observe. In order to perform policy analysis, we need orthogonal shocks with economic meaning, that is, we need a structural representation.

The reduced form VAR was given as

$$
\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{y}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon_t} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}).
$$

We need to recover

$$
\boldsymbol{B}_0 \boldsymbol{y}_t = \underbrace{\boldsymbol{B}}_{\boldsymbol{B}_0\boldsymbol{A}} \boldsymbol{y}_{t-1} + \boldsymbol{e}_t, \qquad \boldsymbol{e}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}).
$$

We cannot estimate the structural form using OLS since the regressors are correlated with the error term. In addition, the matrix $\boldsymbol{B_0}$ is problematic, since it includes all the contemporaneous relations among the endogenous variables.

Consider the following structural VAR model:

$$
\begin{bmatrix} b_{0,11} & b_{0,12} & b_{0,13} \\ b_{0,21} & b_{0,22} & b_{0,23} \\ b_{0,31} & b_{0,32} & b_{0,33} \end{bmatrix} \begin{bmatrix} \Delta y_t \\ \pi_t \\ r_t \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} \begin{bmatrix} \Delta y_{t-1} \\ \pi_{t-1} \\ r_{t-1} \end{bmatrix} + \begin{bmatrix} e_{\Delta y_t} \\ e_{\pi_t} \\ e_{r_t} \end{bmatrix}
$$

The coefficient $b_{0,31}$ is the impact multiplier of monetary policy shocks on GDP, while the coefficient $b_{0,32}$ is the impact multiplier of monetary policy shocks on inflation. With a model simulation, we can evaluate the time profile of a monetary policy shock on GDP or inflation.

Consider again a VAR(p) model in its reduced form:

$$
\boldsymbol{y_t} = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \ldots + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma_\varepsilon)
$$

Now, suppose we can find a matrix $\boldsymbol{B}_0^{-1}$ s.t. $\boldsymbol{\Sigma}_\varepsilon = \boldsymbol{B}_0^{-1}\left(\boldsymbol{B}_0^{-1}\right)'$, i.e., $\boldsymbol{\varepsilon}_t = \boldsymbol{B}_0^{-1}\boldsymbol{e_t}$ with $\boldsymbol{e_t} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I_M})$, that is, we have orthogonal shocks.

Then, the VAR(p) model can be expressed in its structural form as

$$\underbrace{\boldsymbol{B}_0^{-1}\boldsymbol{B}_0}_{I}\boldsymbol{y}_t = \underbrace{\boldsymbol{B}_0^{-1}\boldsymbol{B}_1}_{A_1}\boldsymbol{y}_{t-1} + \ldots + \underbrace{\boldsymbol{B}_0^{-1}\boldsymbol{B}_p}_{A_p}\boldsymbol{y}_{t-p} + \overbrace{\boldsymbol{B}_0^{-1}\boldsymbol{e}_t}^{\varepsilon_t}, \quad \boldsymbol{e}_t \sim \mathcal{N}(0, \boldsymbol{I_M}) \tag{2.9}$$

- Note that we normalized the variance-covariance matrix of the structural errors, $\mathrm{E}\left(\boldsymbol{e_t}\boldsymbol{e_t'}\right) \equiv \boldsymbol{I_M}$. The orthogonality of $\boldsymbol{e_t}$ allows interpreting each shock individually, something not possible for ε_t with its full variance-covariance matrix $\Sigma_\varepsilon = \left(B_0^{-1}\boldsymbol{B_0^{-1'}}\right)$.

We need information about $\boldsymbol{B}_0$ (or $\boldsymbol{B}_0^{-1}$) governing the contemporaneous relations between the variables. The question of how to find such a matrix $\boldsymbol{B_0}$, i.e. how to identify the structural model or shocks, is difficult. For now, we just take it as given that there is a hidden mechanism, such that

$$\boldsymbol{B}_0^{-1}\varepsilon_t = \boldsymbol{e_t} \quad \text{where} \quad \boldsymbol{e_t} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I_M})$$

holds and we have truly exogenous shocks. With a structural VAR model, we can finally interpret the errors, since the structural errors are mutually uncorrelated.

Impulse Response Analysis

When we have uncorrelated errors from a structural VAR model, we can analyze he response of current and future values of each of the variables to a one-unit increase in the current value of one of the structural errors (assuming all other errors and the error in question in all subsequent periods are zero).

We define a vector of exogenous impulses, $\boldsymbol{s}_\tau$, s.t. the impulse to e_1, i.e. $s_{1,1} = 1$, and all other elements are zero. We are interested in the dynamic responses of all variables in the system to the shocks,

$$\frac{\partial \boldsymbol{y}_{t+j}}{\partial \boldsymbol{e}_t} = \underset{M\times M}{\boldsymbol{\Theta}_j}, \qquad j = 0, 1, 2, \ldots, h, \tag{2.10}$$

where $\boldsymbol{\Theta}_j \equiv \boldsymbol{\Phi}_j\boldsymbol{B}_0^{-1}$ (where $\boldsymbol{\Phi}_j = \boldsymbol{J'A}^j\boldsymbol{J}$) denotes the matrix of impulse responses for period j.

The elements of $\boldsymbol{\Theta}_j$ are denoted as follows:

$$\frac{\partial y_{it+j}}{\partial e_{kt}} = \theta_{ik,j}, \qquad i, k = 1, \ldots, M.$$

We usually want to plot the response of some variables to specific shocks over time to assess how those variables react to such a shock. Under the assumption that the model is stable, the effect of the shocks to $\boldsymbol{e}_t$ fade out, i.e., there are no long term effects.

We can switch between reduced form and structural IRFs by multiplying the reduced-form responses to unit shocks with $\boldsymbol{B}_0^{-1}$, i.e.

$$
\begin{aligned}
\boldsymbol{\Theta}_0 &= \boldsymbol{\Phi}_0 \boldsymbol{B}_0^{-1} = \boldsymbol{J}' \boldsymbol{A}^0 \boldsymbol{J} \boldsymbol{B}_0^{-1} = \boldsymbol{I}_K \boldsymbol{B}_0^{-1} = \boldsymbol{B}_0^{-1} \\
\boldsymbol{\Theta}_1 &= \boldsymbol{\Phi}_1 \boldsymbol{B}_0^{-1} = \boldsymbol{J}' \boldsymbol{A}^1 \boldsymbol{J} \boldsymbol{B}_0^{-1} \\
\boldsymbol{\Theta}_2 &= \boldsymbol{\Phi}_2 \boldsymbol{B}_0^{-1} = \boldsymbol{J}' \boldsymbol{A}^2 \boldsymbol{J} \boldsymbol{B}_0^{-1} \\
&\vdots
\end{aligned}
$$

Forecast Error Variance Decomposition

Forecast error variance decomposition answers the question about what portion of the variance of the forecast error in predicting $y_{i,t+h}$ is due to a structural shock $\boldsymbol{e}_i$. It thus provides information on the relative importance of each structural shock in affecting the variables in the VAR. Formally, it answers how much of the forecast error variance of y_{t+h} at horizon h is accounted for by each structural shock $e_{kt}, k = 1, \ldots, M$. The h-step ahead forecast error is

$$
f_{t+h|t} = \boldsymbol{y}_{t+h} - \boldsymbol{y}_{t+h|t} = \sum_{j=0}^{h-1} \boldsymbol{\Phi}_j \boldsymbol{\varepsilon}_{t+h-j} = \sum_{j=0}^{h-1} \boldsymbol{\Theta}_j \boldsymbol{e}_{t+h-j}, \tag{2.11}
$$

since $\boldsymbol{\varepsilon}_t = \boldsymbol{B}_0^{-1} \boldsymbol{e}_t$ and $\boldsymbol{\Theta}_j = \boldsymbol{B}_0^{-1} = \boldsymbol{J}' \boldsymbol{A}^j \boldsymbol{J} \boldsymbol{B}_0^{-1}$. The forecast error is the yet unobserved realization of the shocks.

The mean squared prediction error (MSPE) at horizon h is

$$
\begin{aligned}
\text{MSPE}(h) &= \text{E}\left((\boldsymbol{y}_{t+h} - \boldsymbol{y}_{t+h|t})(\boldsymbol{y}_{t+h} - \boldsymbol{y}_{t+h|t})' \right) = \sum_{j=0}^{h-1} \boldsymbol{\Phi}_j \boldsymbol{\Sigma} \boldsymbol{\Phi}_j' \\
&= \sum_{j=0}^{h-1} \boldsymbol{\Phi}_j \boldsymbol{B}_0^{-1} (\boldsymbol{B}_0^{-1})' \boldsymbol{\Phi}_j' = \sum_{j=0}^{h-1} (\boldsymbol{\Phi}_j \boldsymbol{B}_0^{-1})(\boldsymbol{\Phi}_j \boldsymbol{B}_0^{-1})' \\
&= \sum_{j=0}^{h-1} \boldsymbol{\Theta}_j \boldsymbol{\Theta}_j'.
\end{aligned} \tag{2.12}
$$

Historical Decomposition

Historical decomposition allows us to answer the question about what portion of the deviation of $y_{i,t}$ from its unconditional mean is due to the structural shock $\boldsymbol{e}_i$. We can think of structural shocks as pushing the variables away from their equilibrium values. Historical decomposition then allows us to answer how much a given structural shock explains of the historically observed fluctuations in the VAR variables.

For each point in time, the cumulative effect of a given structural shock can be computed as

$$
\boldsymbol{y}_t = \sum_{j=0}^{t-1} \boldsymbol{\Theta}_j \boldsymbol{e}_{t-j} + \sum_{j=t}^{\infty} \boldsymbol{\Theta}_j \boldsymbol{e}_{t-j} \approx \sum_{j=0}^{t-1} \boldsymbol{\Theta}_j \boldsymbol{e}_{t-j}. \tag{2.13}
$$

Historical decomposition consists of two parts, shocks that predate the sample, and shocks that happen in the sample. Since the MA coefficients die out, the second term has a steadily diminishing effect and we can disregard it.

Additional Remarks

Forecast scenarios assess the sensitivity of reduced form VAR forecasts to hypothetical future events. Policymakers are often interested in hypothetical what-if questions.

Simulating counterfactual outcomes concerns itself with simulating a path of the VAR variables under a different sequence of structural shocks than observed in the data.

Policy Counterfactuals concern themselves with changing the policy reaction function to construct policy counterfactuals, for example, analyzing what would have happened after a oil price shock if the central bank holds the interest constant.

Nonfundamentalness describes the problem that an SVAR econometrician typically analyzes a much narrower information set than economic agents such as central banks. In such a situation, an SVAR model, in general, cannot consistently estimate the IRFs of the structural shocks. Because the present and past values of the series considered are not sufficient informative as the innovation of the econometrician does not coincide with that of the agents. It is therefore important to double-check econometric results with economic intuition and theory.

3 Introduction to Bayesian Econometrics

Bayesian Updating

The interpretation that probability expresses a belief that an event occurs is labeled **subjective or Bayesian probability**. This is different from **frequentist probability**. When using Bayesian methods, we start with existing **prior** beliefs that we update using observed data to obtain **posterior** beliefs.

Formally, we update our prior belief about a latent value θ using the observed data $\mathscr{D}$ to get an updated belief:

$$p(\theta) \times p(\mathscr{D} \mid \theta) \rightarrow p(\theta \mid \mathscr{D}). \tag{3.1}$$

Bayes's Rule and Bayes's Theorem

Bayes's Rule

Assume there are two events A and B. We can observe event B directly, but not A, and want to learn about the probability of A using B. Bayes's rule states that

$$\begin{aligned} P(A \mid B) &= \frac{P(B \mid A)P(A)}{P(B)} \propto P(B \mid A)P(A), \\ P(\neg A \mid B) &= \frac{P(B \mid \neg A)P(\neg A)}{P(B)}. \end{aligned} \tag{3.2}$$

To prove Bayes's rule, we can express the joint probability of both A and B occurring as

$$\begin{aligned} P(A \cap B) &= P(A \mid B)P(B) \\ &= P(B \mid A)P(A). \end{aligned}$$

Rearranging then yields

$$P(A \mid B)P(B) = P(A \cap B) = P(B \mid A)P(A),$$
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

When applying Bayes's rule, we rely on a prior probability for A, which means that we need to specify $P(A)$. To do that, we can use historical data, theoretical insights or external information; or we can use an uninformative prior, for example, setting $P(A) = \frac{1}{2}$.

The denominator in Bayes's rule is given by the law of total probability: $P(B) = P(B \mid A) + P(B \mid \neg A)P(\neg A)$. It acts as a normalizing constant such that $P(A \mid B) + P(\neg A \mid B) = 1$. In addition, it is expensive to compute. However, we only need the proportional posterior in the numerator, and can normalize afterwards. This yields for the posterior

$$P(A \mid B) \propto P(B \mid A)P(A). \tag{3.3}$$

Bayes's Theorem

Working with probabilities, or with events that either occur or do not occur, means that we have implicitly worked with the Bernoulli distribution, where

$$p(x \mid p) = \begin{cases} 1-p & \text{if } x = 0, \\ p & \text{if } x = 1. \end{cases} \tag{3.4}$$

When using probability distributions in general, the prior will be a probability distribution and data is generated by a stochastic model depending on the latent. Bayes's theorem then gives us the posterior distribution:

$$p(\theta \mid \mathscr{D}) \propto p(\mathscr{D} \mid \theta)p(\theta), \tag{3.5}$$

where θ is a latent and $p(\cdot)$ denotes the density of a probability distribution.

Bayesian Inference, Priors and Posteriors

Steps for Bayesian Inference

To make statements about the latent given the observed data, we need to

(1) Specify a **sampling distribution**, or likelihood, in terms of unknown parameters for the data $\mathscr{D}$:

$$p(\mathscr{D} \mid \theta).$$

(2) Specify a **prior distribution** for the unknown parameters:

$$p(\theta).$$

(3) Use Bayes's theorem to derive the posterior distribution:

$$p(\theta \mid \mathscr{D}) \propto p(\mathscr{D} \mid \theta) p(\theta).$$

The Likelihood

The likelihood, $p(\mathscr{D} \mid \theta)$ provides a framework for the data to do the talking. In frequentist analysis, it is the sole factor. Arguably, the choice of a likelihood is as subjective as choosing a prior. The structure of the data can help with a suitable choice – we can ask ourselves what values the outcome can assume, whether outcomes are independent of one another, and whether they are identically distributed. A likelihood is not a probability since it is a function of the parameter, not the data. It need not integrate to one.

If we treat a Bernoulli-distributed example, our latent is the probability θ. We can model the outcome with a binary random variable Y. A single realization is equal to one with a probability of θ, $P(y_i = 1 \mid \theta) = \theta$. The probability mass function for a Bernoulli distribution is given by

$$f(k \mid p) = p^k (1-p)^{1-k} \text{ for } k \in \{0, 1\}. \tag{3.6}$$

The probability density of a single realization in the example is thus

$$p(y_i \mid \theta) = \theta^{y_i} (1-\theta)^{1-y_i} \text{ for } y_i \in \{0, 1\}.$$

If observations are independent, their joint density is

$$\begin{aligned} f(\boldsymbol{y} \mid \theta) &= \prod_{i=1}^{n} p(y_i \mid \theta), \\ &= \theta^{S_n} (1-\theta)^{n-S_n}, \end{aligned}$$

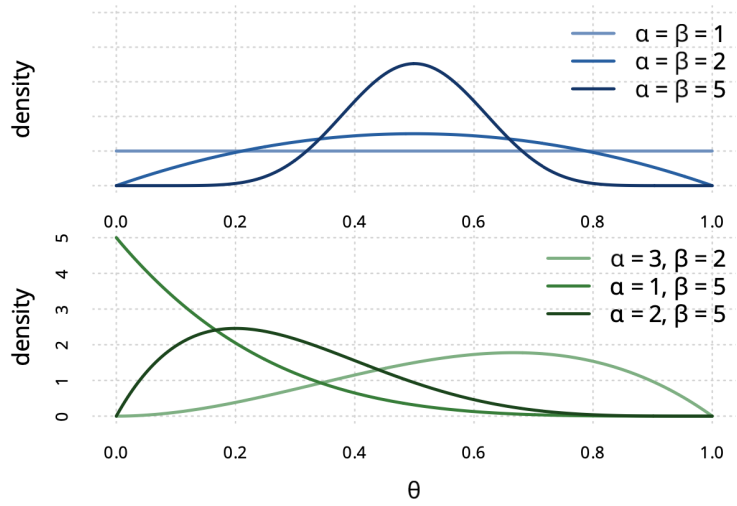
where $S_n = \sum_{i=1}^{n} y_i$ is the observed number of occurrences. Getting a likelihood for dependent data, e.g. in time series contexts, is more complicated.

The Prior

As a prior for the latent probability $\theta \in [0, 1]$, we can use a beta distribution. It is a continuous distribution defined on the unit interval. A random variable $\theta \sim \text{Beta}(\alpha, \beta)$ has the probability density function

$$f(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1} (1-\theta)^{\beta-1}}{\text{B}(\alpha, \beta)}, \tag{3.7}$$

where $B(\cdot)$ is the beta function and α, β are parameters. The following are some PDFs of different Beta distributed θ with different parameters α, β:



The prior $p(\theta)$ determines which regions of the framework we believe to be relevant. Its choice is subjective, similarly to the likelihood. There are many approaches to choosing a sensible prior. Conjugate priors make the computation much easier. In this case, we used a conjugate Beta prior. A conjugate prior is a prior probability distribution that, when combined with a likelihood function through Bayes's theorem, produces a posterior distribution that is of the same family as the prior distribution. Setting $a_0 = b_0 = 1$ produces a uniform prior across the unit interval.

Instead of a_0 and b_0, we can consider as parameters success probability $m_0 \approx \hat{\theta}_n = \frac{S_n}{n}$ and prior information $s_0 = a_0 + b_0$. These parameters are tied to the moments of the Beta distribution:

$$\mathrm{E}(\theta) = \frac{a_0}{a_0 + b_0} = m_0, \qquad \mathrm{Var}(\theta) = \frac{\mathrm{E}(\theta)(1 - \mathrm{E}(\theta))}{a_0 + b_0 + 1}.$$

Our prior is then given by $\mathrm{Beta}(m_0 s_0, (1 - m_0) s_0)$.

The Posterior

Obtaining the posterior from combining the Bernoulli likelihood and the Beta prior, we get:

$$\begin{aligned} p(\theta \mid \boldsymbol{y}) &\sim \mathrm{Beta}(a_n, b_n), \\ a_n &= a_0 + S_n, \\ b_n &= b_0 + n - S_n. \end{aligned}$$

The data enters only via two sufficient statistics, the number of successes, S_n, and the number of failures, $n - S_n$.

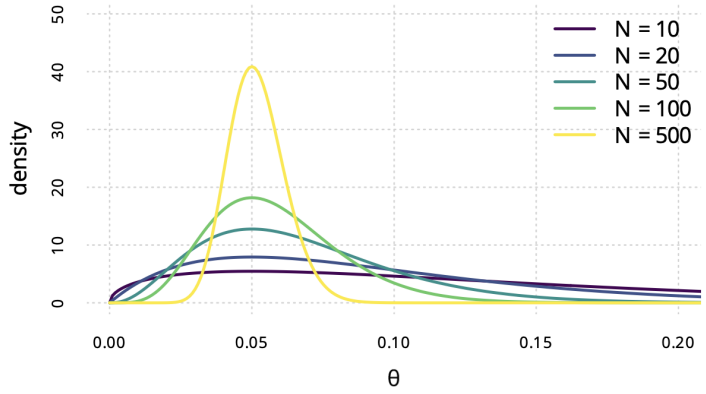
We get this result by the following computation: A $\mathrm{Beta}(\alpha, \beta)$ density is, up to a constant, given by

$$f(\theta \mid \alpha, \beta) \propto \theta^{\alpha - 1} (1 - \theta)^{\beta - 1}.$$

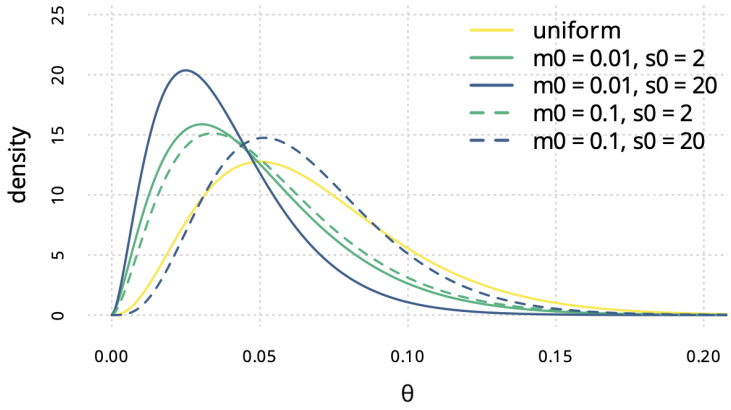
We then get the posterior from Bayes's theorem, dropping factors independent of θ and matching moments to the Beta density:

$$
\begin{aligned}
p(\theta \mid \boldsymbol{y}) &\propto p(\boldsymbol{y} \mid \theta) p(\theta) \\
&\propto \theta^{S_n}(1-\theta)^{n-S_n} \quad \cdot \quad \theta^{a_0-1}(1-\theta)^{b_0-1} \\
&= \theta^{S_n+a_0-1}(1-\theta)^{n-S_n+b_0-1} \\
\Rightarrow p(\theta \mid \boldsymbol{y}) &\sim \text{Beta}(a_0 + S_n, b_0 + n - S_n).
\end{aligned}
$$

Using a uniform prior, this is how the posterior looks using different values of n and $\frac{S_n}{n} = 0.05$:



Using m_0 and s_0 as specified above as priors yields the following posteriors for $n = 50$ and $\frac{S_n}{n} = 0.05$:

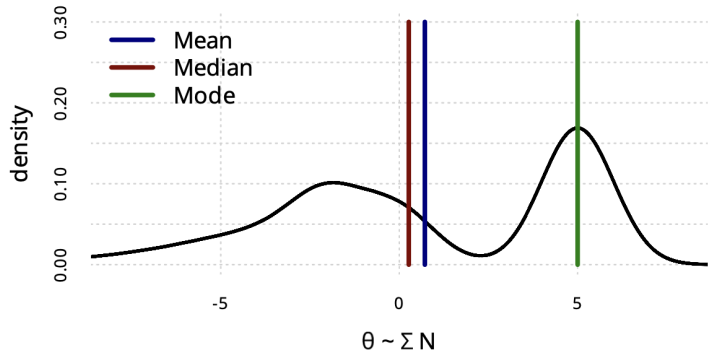


From a full Bayesian analysis, we get a posterior density instead of just a point estimate. We can therefore obtain posterior moments and quantiles as well as any number of summary measures.

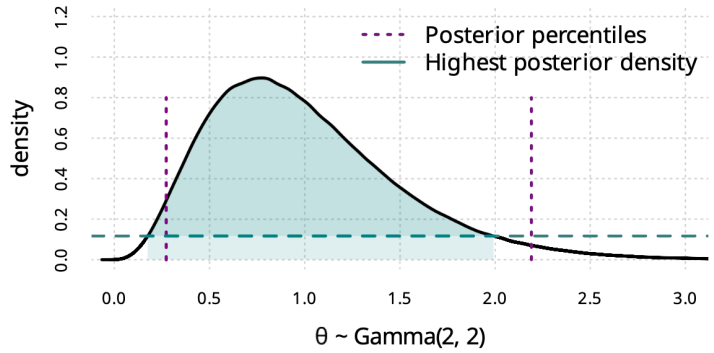
We can also get credible regions or credible intervals C_α, which contain the parameter θ with probability α. The highest posterior density region is the smallest region C_α for a given α.

Summarizing the Posterior

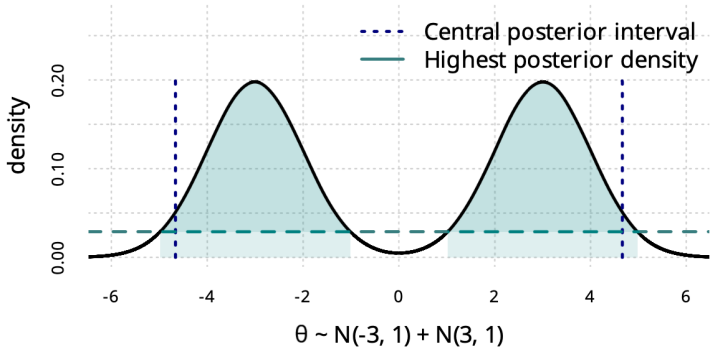
Summarizing the posterior using a simple number requires us to choose from multiple options, such as mode, median and mean.



Even choosing an interval to cover 95% probability is challenging. See this comparison between choosing the interval based on percentiles and choosing the HPD region:



The HPD region can even be non-contiguous:



Posterior summaries are the core of Bayesian inference. In the case of the Beta posterior, moments are well-known and can be derived analytically. However, sometimes, we cannot derive them analytically or the posterior might not be of a well-known form.

For a posterior summary, we essentially need the integral of a function $h(\cdot)$ of a parameter θ:

$$\mathrm{E}(h(\theta)) = \int h(\theta) f(\theta) d\theta,$$

where θ follows a distribution with density function $f(\theta)$. The posterior variance is given by:

$$\text{Var}(\theta \mid \mathscr{D}) = \text{E}\big(\theta\theta' \mid \mathscr{D}\big) - \text{E}(\theta \mid \mathscr{D})\text{E}(\theta \mid \mathscr{D})',$$
$$\text{E}\big(\theta\theta' \mid \mathscr{D}\big) = \int (\theta\theta')p(\theta \mid \mathscr{D})d\theta,$$
$$\text{E}(\theta \mid \mathscr{D}) = \int \theta p(\theta \mid \mathscr{D})d\theta.$$

We can either compute this integral analytically if a closed-form solution exists, or we can use Monte Carlo integration if we can obtain samples $\theta^{(1)}, \dots, \theta^{(S)}$ from f like this:

$$\text{E}(h(\theta) \mid \mathscr{D}) \approx \frac{1}{S}\sum_{s=1}^{S} h(\theta^{(s)}).$$

If we can obtain samples, then the law of large numbers guarantees convergence of empirical moments to the moments of the probability distribution as $S \to \infty$. If neither computing the integral analytically nor sampling from the distribution is possible, we can use Markov chain Monte Carlo methods.

4 Bayesian Regression Analysis

Simple Linear Model

Setup

The standard linear regression model is given by

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \dots, n. \tag{4.1}$$

Our distibutional assumption is with respect to the error term: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \forall i$, and $\text{E}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = 0 \quad \Rightarrow \text{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$. The latent quantities in this model are $\boldsymbol{\beta}$, the regression coefficients, and σ^2, the error variance.

Assuming that there are no covariates, that is, $\boldsymbol{X} = \boldsymbol{1}$, we can express the model as

$$y_i \sim \mathcal{N}(\beta, \sigma^2).$$

Recall that a random variable $\theta \sim \mathcal{N}(\beta, \sigma^2)$ has the probability density function

$$f(\theta \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\left(-\frac{1}{2\sigma^2}(\theta - \mu)^2\right), \tag{4.2}$$

where $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$ are parameters and the support is $\theta \in \mathbb{R}$.

We want to learn about two free parameters, the mean β and the variance σ^2. The likelihood function of our data is given by

$$
\begin{aligned}
p(\boldsymbol{y} \mid \beta, \sigma^2) &= \prod_{i=1}^{n} p(y_i \mid \beta, \sigma^2) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta)^2 \right).
\end{aligned}
$$

Inference for σ^2 with β known

To get the conditional posterior $p(\sigma^2 \mid \boldsymbol{y}, \beta)$, we use Bayes's theorem. First, we consider the likelihood as a function of σ^2, meaning that we can drop constant values that do not depend on σ^2. We get

$$
p(\boldsymbol{y} \mid \sigma^2, \beta) \propto (\sigma^2)^{n/2+1-1} \exp\left(-\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta)^2 / 2 \right).
$$

This is proportional to the PDF of an inverted Gamma distribution, $\theta \sim G^{-1}(c, d)$ with shape $c = n/2 + 1$ and rate $d = \sum_{i=1}^{n} (y_i - \beta)^2 / 2$. The density is given by

$$
p(\theta \mid c, d) \propto (\theta)^{-c-1} \exp\left(-\frac{d}{\theta} \right).
$$

If we use an inverted Gamma prior for σ^2, that is, $\sigma^2 \mid \beta \sim G^{-1}(c_0, d_0)$, then the prior and the likelihood are conjugate, and the posterior will also have the form of an inverted Gamma distribution.

A random variable that is inverted-Gamma distributed, that is, $\theta \sim G^{-1}(\alpha, \beta)$, has the PDF

$$
f(\theta \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{-\alpha-1} \exp\left(-\frac{\beta}{\theta} \right), \tag{4.3}
$$

where $\Gamma(\cdot)$ is the Gamma function and $\alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+$ are parameters. The support is $\theta \in \mathbb{R}^+$.

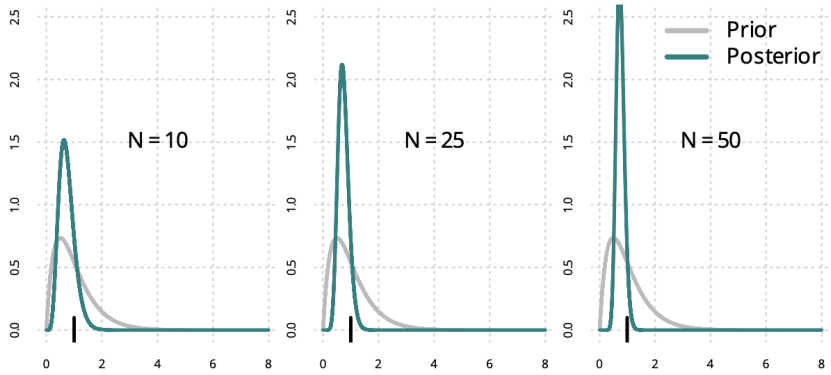
The posterior of σ^2 with β known is then

$$
\begin{aligned}
p(\sigma^2 \mid \beta) &\propto p(\boldsymbol{y} \mid \sigma^2, \beta) \times p(\sigma^2 \mid \beta) \\
&\propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta)^2 \right) \times (\sigma^2)^{-c_0-1} \exp\left(-\frac{d_0}{\sigma^2} \right) \\
&\propto (\sigma^2)^{-(n/2+c_0)-1} \exp\left(-\frac{1}{\sigma^2} \left(d_0 + \sum_{i=1}^{n} (y_i - \beta)^2 / 2 \right) \right)
\end{aligned}
$$

This gives us the conditional posterior $p(\sigma^2 \mid \boldsymbol{y}, \beta) \sim G^{-1}(c_n, d_n)$, where

$$
c_n = c_0 + \frac{n}{2}, \qquad d_n = d_0 + \sum_{i=1}^{n} (y_i - \beta)^2 / 2.
$$

There are two sufficient statistics, n and the variation around the mean, β. The following figure shows prior and posterior densities of σ^2 with increasing observations under an inverse Gamma prior, $G^{-1}(2, 2)$, and data from a Normal with mean two and variance one:



Inference for β with σ^2 known

The likelihood $p(\boldsymbol{y} \mid \beta, \sigma^2)$, considered as a function of β, is proportional to:

$$p(\boldsymbol{y} \mid \beta, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta)^2\right).$$

We can simplify this by using a variance decomposition to express the sum as

$$\sum_{i=1}^{n}(y_i - \beta)^2 = n(\beta - \bar{y})^2 + (n-1)s^2,$$

where $\bar{y}$ is the sample mean and s^2 is the bias-corrected sample variance. We can do this since

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \beta)^2 &= \sum_{i=1}^{n}(y_i - \bar{y} + \bar{y} - \beta)^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 + 2(\bar{y} - \beta)\left(\sum_{i=1}^{n}(y_i - \bar{y})\right) + n(\bar{y} - \beta)^2 2 \\ &= (n-1)s^2 + 0 + n(\beta - \bar{y})^2.\end{aligned}$$

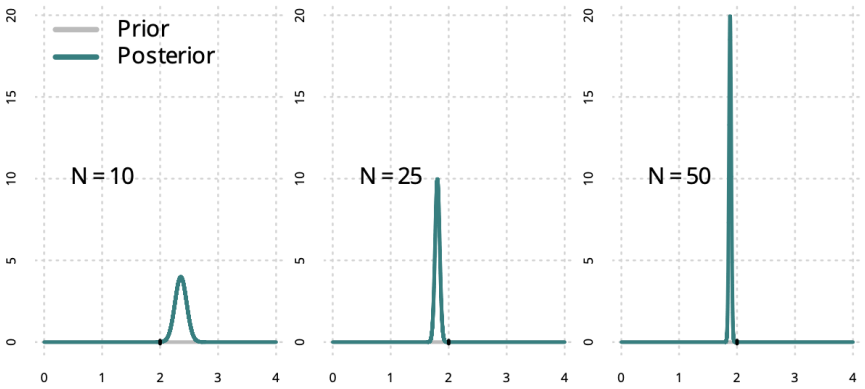
Using this variance decomposition, we can express the likelihood as

$$p(\boldsymbol{y} \mid \beta, \sigma^2) \propto \exp\left(-\frac{(\beta - \bar{y})^2}{2\sigma^2/n}\right).$$

This is proportional to the PDF of a normal distribution since

$$p(\theta \mid \mu, \varsigma) \propto \exp\left(\frac{(\theta - \mu)^2}{2\varsigma}\right).$$

Under a flat prior $p(\beta) \propto \kappa$, the posterior is then given by $p(\beta \mid \boldsymbol{y}, \sigma^2) \sim \mathcal{N}(\bar{y}, \sigma^2/n)$. The sufficient statistics are n and the sample mean, $\bar{y}$. The following figure shows prior and posterior densities of μ with increasing observations under a flat prior and data from a Normal with mean two and variance one.



Inference on the Full Linear Regression Model

In the previous, we derived the conditional posteriors $p(\sigma^2 \mid \beta, \boldsymbol{y})$ and $p(\beta \mid \sigma^2, \boldsymbol{y})$ under the assumption that we know the other latent parameter. Now we need to derive the joint posterior distribution of (β, σ^2) as well as the marginal posteriors $p(\sigma^2 \mid \boldsymbol{y})$ and $p(\beta \mid \boldsymbol{y})$.

Using the variance decomposition, we can write the likelihood as

$$p(\boldsymbol{y} \mid \beta, \sigma^2) = (2\pi\sigma^2)^{n/2} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right) \exp\left(\frac{(\beta - \bar{y})^2}{2\sigma^2/n}\right)$$

If we use the improper prior $p(\beta, \sigma^2) \propto \sigma^{-2}$, an improper prior being a prior that does not integrate to one, we can derive the posterior:

$$p(\beta, \sigma^2 \mid \boldsymbol{y}) \propto (\sigma^2)^{-\frac{n+1}{2}} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right) \sigma^{-1} \exp\left(\frac{(\beta - \bar{y})^2}{2\sigma^2/n}\right)$$

We can factor this posterior into two well-known densities by splitting it into the product of a conditional Normal density for $\beta \mid \sigma^2$, and a marginal inverted Gamma density for σ^2. There are three sufficient statistics, n, $\bar{y}$, and s^2. We have:

$$\beta \mid \boldsymbol{y}, \sigma^2 \sim \mathcal{N}(\mu_n, \Sigma_n \sigma^2),$$

where $\mu_n = \bar{y}$ and $\Sigma_n = \frac{1}{n}$, and

$$\sigma^2 \mid \boldsymbol{y} \sim G^{-1}(c_n, d_n),$$

where $c_n = (n-1)/2$ and $d_n = (n-1)s^2/2$.

To find the marginal posterior of β, we need to solve the following integral:

$$\begin{aligned} p(\beta \mid \boldsymbol{y}) &= \int_{\mathbb{R}^+} p(\beta, \sigma^2 \mid \boldsymbol{y}) d\sigma^2 \\ &= \int_{\mathbb{R}^+} p(\beta, \sigma^2 \mid \boldsymbol{y}) p(\sigma^2 \mid \boldsymbol{y}) d\sigma^2. \end{aligned}$$

By integrating out σ^2, we find the following t density:

$$p(\beta \mid \boldsymbol{y}) \sim t_{2c_n}(\mu_n, \Sigma_n d_n / c_n).$$

With the prior from before, $p(\beta, \sigma^2) \propto \sigma^{-2}$, this simplifies to $t_{n-1}(\bar{y}, s^2/n)$.

Frequentist versus Bayesian Inference

Frequentist confidence regions are based on

$$(\beta - \hat{\beta}) \sim t_v(0, \hat{\sigma}^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}),$$

where $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ and $\hat{\sigma}^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/(n-k)$. Bayesians use the posterior:

$$p(\beta \mid \boldsymbol{y}) \sim t_v(\hat{\beta}, \hat{\sigma}^2 (\boldsymbol{X}'\boldsymbol{X}^{-1}) \quad \Rightarrow \quad (\beta - \hat{\beta}) \mid \boldsymbol{y} \sim t_v(0, \hat{\sigma}^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

This implies that frequentist confidence regions can often (but not generally) be interpreted in a Bayesian sense.

Standard Regression Analysis

Priors and Posteriors

The standard linear regression model with multiple explanatory variables and spherical variance-covariance can be expressed as

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n),$$

where $\mathcal{N}_n$ denotes an n-dimensional multivariate Normal distribution, n is the number of observations and k is the number of explanatory variables. To draw inference on $\boldsymbol{\beta} \in \mathbb{R}^k$ and $\sigma^2 \in \mathbb{R}^+$, we require the posterior distribution

$$p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2).$$

The likelihood of the standard linear regression model is given by

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right).$$

As a function of σ^2, it equals the kernel of an inverted Gamma distribution. Thus, the (conditionally) conjugate prior for σ^2 is the IG distribution. As a function of $\boldsymbol{\beta}$, it is a quadratic form and mirrors the kernel of a multivariate Normal (MVN) distribution. The (conditionally) conjugate prior for $\boldsymbol{\beta}$ is thus the MVN distribution. A kernel, intuitively speaking, is the PDF with factors that ensure that the PDF integrates to one and are not functions of free variables omitted.

We can obtain a closed form posterior with the following prior:

$$\beta \mid \sigma^2 \sim \mathcal{N}_k\left(\boldsymbol{\mu}_0, \sigma^2 \boldsymbol{\Sigma}_0\right), \quad \sigma^2 \sim \mathrm{G}^{-1}\left(c_0, d_0\right).$$

The joint posterior of $\left(\beta, \sigma^2\right)$ is then

$$\beta \left| \sigma^2, \boldsymbol{y} \sim \mathcal{N}_k\left(\boldsymbol{\mu}_n, \sigma^2 \boldsymbol{\Sigma}_n\right), \quad \sigma^2 \right| \boldsymbol{y} \sim \mathrm{G}^{-1}\left(c_n, d_n\right),$$

where:

$$\begin{aligned} \boldsymbol{\mu}_n &= \boldsymbol{\Sigma}_n\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{X}'\boldsymbol{y}\right), & c_n &= c_0 + n/2 \\ \boldsymbol{\Sigma}_n &= \left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{X}'\boldsymbol{X}\right)^{-1}, & d_n &= d_0 + \boldsymbol{S}_\varepsilon/2, \\ \boldsymbol{S}_\varepsilon &= \boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\mu}_0'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_n'\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\mu}_n. \end{aligned}$$

Sampling from the Joint Posterior Distribution

If the joint posterior is available in closed form, we can obtain independent samples to visualize and draw inference. To do this and obtain $m = 1, \ldots, M$ samples $(\beta_{(m)}, \sigma^2_{(m)})$, we first sample $\sigma^2_{(m)}$ from the marginal posterior and then use this to sample $\beta_{(m)}$ from the conditional posterior.

We can express the posterior mean of our model as

$$\begin{aligned} \mathrm{E}(\beta \mid \boldsymbol{y}) &= (\Sigma_0^{-1} + \boldsymbol{X}'\boldsymbol{X})^{-1}(\Sigma_0^{-1}\mu_0 + \boldsymbol{X}'\boldsymbol{X}\hat{\beta}) \\ &= \boldsymbol{W}\Sigma_0^{-1}\mathrm{E}(\beta) + \boldsymbol{W}\boldsymbol{X}'\boldsymbol{X}\hat{\beta}, \end{aligned}$$

where $\boldsymbol{W} = (\Sigma_0^{-1} + \boldsymbol{X}'\boldsymbol{X})^{-1}$ is the denominator of a weight matrix that depends on the prior information (or precision) Σ_0^{-1} and the information matrix $\boldsymbol{X}'\boldsymbol{X}$. The posterior mean is therefore a weighted average of the prior mean $\mathrm{E}(\beta) = \mu_0$ and the OLS estimator $\hat{\beta}$.

Consider again the case where $\boldsymbol{X} = \boldsymbol{1}$ and therefore $\boldsymbol{X}'\boldsymbol{X} = n$. The posterior expectation of $\boldsymbol{\beta}$ is a weighted average of the sample mean $\bar{y}$ and the prior mean $\mathrm{E}(\beta) = \mu_0$:

$$\mathrm{E}(\beta \mid \boldsymbol{y}) = \frac{\Sigma_0^{-1}}{n + \Sigma_0^{-1}}\mathrm{E}(\beta) + \frac{n}{n + \Sigma_0^{-1}}\bar{y}.$$

With increasing n, the effect of the prior disappears. The prior $p(\beta)$ imposes **shrinkage**, that is, the posterior expectation is pulled towards the prior expectation. Shrinkage priors can be understood as a Bayesian analogue to penalized estimators. The general idea is to shrink when there is not enough information in the data, and be agnostic otherwise.

Multiple Regression with a Non-Conjugate Prior

Before, we assumed that the prior covariance matrix of β depends on σ^2. If we assume that β and σ^2 are independent a priori, that is,

$$\beta \sim \mathcal{N}(\mu_0, \Sigma_0), \qquad \sigma^2 \sim G^{-1}(c_0, d_0),$$

where the prior covariance matrix Σ_0 directly controls shrinkage towards the prior mean, then we get a joint posterior that does not have a well-known closed form. There is, however, a closed form for the conditional posteriors:

$$\beta \mid \sigma^2, \boldsymbol{y} \sim \mathcal{N}_k(\mu_n, \Sigma_n), \qquad \sigma^2 \mid \beta \boldsymbol{y} \sim G^{-1}(c_n, d_n),$$

where

$$\begin{aligned}
\mu_n &= \Sigma_n(\Sigma_0^{-1}\mu_0 + \boldsymbol{X}'\boldsymbol{y}/\sigma^2), \\
\Sigma_n &= (\Sigma_0^{-1} + \boldsymbol{X}'\boldsymbol{X}/\sigma^2)^{-1}, \\
c_n &= c_0 + n/2, \\
d_n &= d_0 + \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/2, \\
\boldsymbol{\varepsilon} &= \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}.
\end{aligned}$$

Previously, we have used conjugate priors, such that the posterior is from the same distribution family as the prior. This meant that we could directly compute many quantities of interest, and obtain independent samples from the posterior, which we used to compute arbitrary quantities of interest. Now, the posterior has no simple analytical form and we cannot produce independent samples from $p(\beta, \sigma^2 \mid \boldsymbol{y})$. However, we can produce dependent samples.

5 Bayesian Estimation

Numerical Integration

We want to learn about the posterior $p(\theta \mid \mathscr{D})$ from

$$p(\theta \mid \mathscr{D}) \propto p(\mathscr{D} \mid \theta) p(\theta).$$

In some cases, we know the posterior distribution and some summaries of interest, generally however, we do not. We therefore need numerical integration methods to summarize the posterior. If we can obtain samples from our posterior density, $f(\theta)$, we can use probabilistic **Monte Carlo integration**, where we do the following:

(1) Draw samples $(\theta^{(1)}, \ldots, \theta^{(S)})$,

(2) Compute the desired function, $h(\theta^{(s)})$, for each sample,

(3) Use the results to compute a summary, such as the expected value.

The sampling error decreases as S increases.

Markov Chain Monte Carlo

MCMC Methods allow sampling from any distribution. The idea is that we construct a Markov chain that has the desired distribution as its stationary distribution. We can then obtain samples from the desired distribution by taking states from the Markov chain.

A Markov chain process is a memoryless stochastic process. We can define it as a sequence of random variables that satisfies this property:

$$p(x_{t+1} \mid x_t, x_{t-1}, \ldots, x_0) = p(x_{t+1} \mid x_t) \forall t,$$

where x_t is the state of the chain at time t and $p(x_{t+1} \mid x_t)$ is the transition probability. A Markov chain can be fully characterized by an initial state x_0, a state space $\mathscr{X}$, and transition probabilities. The function $K(a, b)$ that gives the probability of transition to b given a state a is called the transition kernel:

$$K(a, b) = p(x_{t+1} = b \mid x_t = a), \qquad a, b \in \mathscr{X}.$$

Markov Chains with discrete state space have a transition kernel that can be represented as a matrix. The Markov chain is stationary if $\rho((K)) < 1$, where $\rho(\cdot)$ denotes the spectral radius, that is, the maximal absolute value of the eigenvalues. However, in a simple autoregressive model like this:

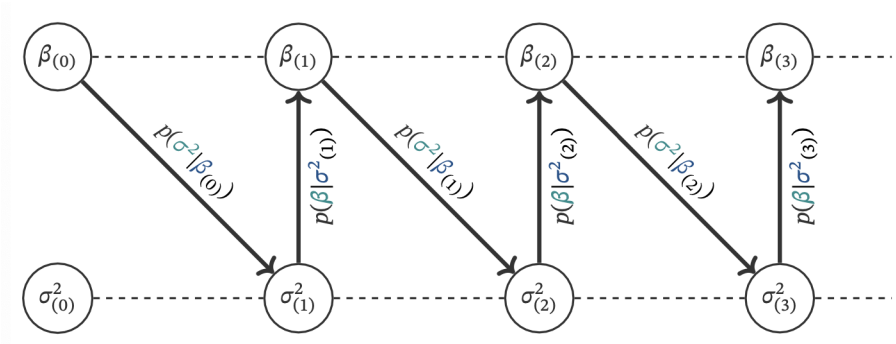
$$y_t = \rho y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, 1),$$

the state space is $\mathscr{X} \in \mathbb{R}$ and the transition kernel is

$$K(a, b) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(b - \rho a)^2}{2}\right).$$

Gibbs Sampler

Before, we imposed independent Normal and Inverse Gamma priors. We need to know the joint posterior $p(\beta, \sigma^2 \mid \boldsymbol{y})$, but only know the conditional posteriors $p(\beta \mid \sigma^2, \boldsymbol{y})$ and $p(\sigma^2 \mid \beta, \boldsymbol{y})$. The Gibbs sampler, an MCMC algorithm, sequentially generates values from each of the conditional distributions, conditional on current other values. This sequence is a Markov chain with the desired joint distribution as its stationary distribution. It can be visualized like this:



In general, for a latent that can be divided into two blocks $\theta = (\theta^1, \theta^2)$, we can use a **two-block Gibbs sampling algorithm**. We start with a starting value $\theta^2_{(0)}$. Then, we iterate for $s = 1, \ldots, S$:

(1) We sample $\theta^1_{(s)}$ from the conditional distribution $p(\theta^1 \mid \theta^2_{(s-1)}, \mathscr{D})$,

(2) We sample $\theta^2_{(s)}$ from the conditional distribution $p(\theta^2 \mid \theta^1_{(s)}, \mathscr{D})$.

We then discard the first S_0 draws as burn-in, such that the sampler has converged to the joint distribution. The transition kernel is $K(\theta_{\text{new}}, \theta_{\text{old}}) = p(\theta^1_{\text{new}} \mid \theta^2_{\text{new}}, \mathscr{D}) p(\theta^2_{\text{new}} \mid \theta^1_{\text{old}})$.

That Gibbs sampler is a special case of the Metropolis-Hastings algorithm where draws are always accepted. The idea of the MH algorithm is to

(1) Propose a new draw from $q(\theta_{\text{new}} \mid \theta_{\text{old}})$,

(2) Accept the new value with probability $\alpha(\theta_{\text{new}} \mid \theta_{\text{old}})$:

$$\alpha(\theta_{\text{new}} \mid \theta_{\text{old}}) = \min\left(1, \frac{p(\theta_{\text{new}} \mid \mathscr{D})q(\theta_{\text{old}} \mid \theta_{\text{new}})}{p(\theta_{\text{old}} \mid \mathscr{D})q(\theta_{\text{new}} \mid \theta_{\text{old}})}\right)$$

MH samplers do not rely on well-known conditional distributions. They only need a function that is proportional to the target distribution.

Convergence

The stationary distribution of the Gibbs sampler **converges** to the joint distribution under relatively mild assumptions. The choice of the starting value is deterministic, but can only distort inference if we obtain only few samples. The aforementioned burn-in period helps in limiting the impact of the starting value. To assess convergence, we can use convergence checks and multiple chains with different starting values. Plots we can use for convergence checks include trace plots, density plots and QQ plots.

MCMC Algorithms are approximate. We want to know whether the samples are representative of the distribution and how effective the sampler is at exploring the distribution. These issues refer to the **mixing** of the Markov chain. A sampler mixes well if it explores the full distribution (i.e., not only certain areas) effectively (i.e., with low autocorrelation).

MCMC draws are autocorrelated since they are dependent on earlier draws. Dependent samples are less informative than independent ones. To assess this inefficiency, we can use the effective sample size:

$$\text{ESS} = \frac{m}{\tau},$$

where m is the sample size, and τ is a scaling factor, such as $1 + 2\sum \rho(l)$, where $\rho(l)$ is the autocorrelation at lag l.

To address poor mixing, we can increase the number of draws, change the sampler, improve the sampler, use block sampling, reparameterize models or work out a closed form.

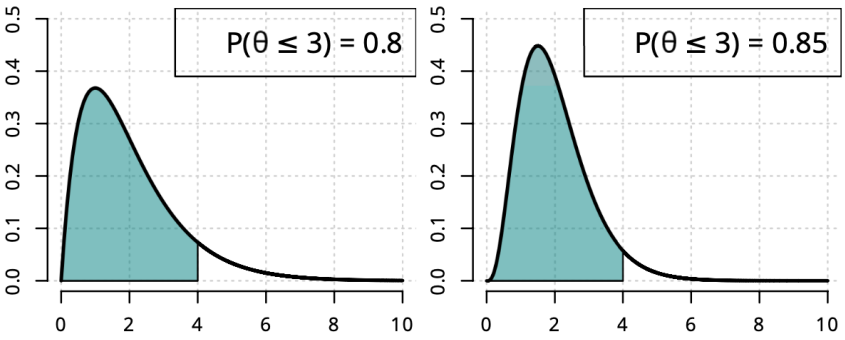
6 Model Selection and Priors

Bayesian Hypothesis Testing

For Bayesian "testing," we can use the probability that the latent θ lies in a subset of the parameter space C given the data $\mathscr{D}$:

$$\int_C p(\theta \mid \mathscr{D}) d\theta. \tag{6.1}$$

Thus, we can directly determine the posterior probability of a hypothesis by computing the integral:



We have strong evidence for a hypothesis $H_a : \theta \leq \theta_0 \mid \mathscr{D}$ if $P(\theta \leq \theta_0 \mid \mathscr{D}) \gg P(\theta > \theta_0 \mid \mathscr{D})$. We select a null hypothesis, thereby expressing prior belief in the null, and reject it if the data is in conflict with it. With Bayesian testing, we get a probability directly. With frequentist testing, we get a "yes" or "no" answer that is correct with a certain probability.

Model Selection

Assume we have m models $M_1, \ldots, M_k$ to explain our data $\mathscr{D}$. These models are characterized by their latent parameters θ_j, the conditional likelihood $p(\mathscr{D} \mid \theta_j, M_j)$, and the conditional prior $p(\theta_j \mid M_j)$. We can go about different ways in the selection of a model to settle on.

Bayesian Model Selection

We consider each model to arise from a discrete model space and assign prior probabilities $P(M_j)$ to each model. We then compute the posterior probabilities $P(M_j \mid \mathscr{D})$ over the model space:

$$P(M_j \mid \mathscr{D}) \propto P(\mathscr{D} \mid M_j) P(M_j) \qquad \forall j \in 1, \ldots, k, \tag{6.2}$$

where $P(\mathscr{D} \mid M_j)$ is the marginal likelihood of model M_j, $P(\mathscr{D} \mid M_j) = \int_{\Theta_m} p(\mathscr{D}, \theta_j \mid M_j) d\theta_j = \int_{\Theta_m} p(\mathscr{D} \mid \theta_j, M_j) p(\theta_m \mid M_j) d\theta_j$.

Let $R(M_j, M)$ be a loss function for selecting the model M_j when M is the true model. We can determine the expected loss of selecting a model for each candidate:

$$\mathrm{E}\big(R(M_j) \mid \mathscr{D}\big) = \sum_{j=1}^{k} R(M_j, M) P(M \mid \mathscr{D}). \tag{6.3}$$

We can then select the model with the lowest expected loss. The $0/1$ loss function is given by

$$r(M_j, M) = \begin{cases} 0, & \text{if } M_j = M, \\ 1, & \text{otherwise.} \end{cases}$$

Let M_{j^*} be the true model. Then the expected risk is

$$\mathrm{E}\big(R(M_j) \mid \mathscr{D}\big) = \sum_{j=1, j \neq j^*}^{k} P(M_{j^*} \mid \mathscr{D}) = 1 - P(M_j \mid \mathscr{D}).$$

We minimize risk (the expected loss) by choosing the model with the highest posterior probability. If all models have the same prior probability, this loss function selects the model with highest marginal likelihood.

The **candidate's formula** is given by

$$p(\mathscr{D} \mid M) = \frac{p(\mathscr{D} \mid \theta) p(\theta \mid M)}{p(\theta \mid \mathscr{D}, M)}. \tag{6.4}$$

If the posterior density comes from a well-known distribution family, we can compute the marginal likelihood explicitly – it is equal to the normalizing constant of the posterior density:

$$p(\theta \mid \mathscr{D}) = \frac{p(\mathscr{D} \mid \theta) p(\theta)}{p(\mathscr{D})}.$$

For a computational example, consider data from the Poisson distribution, $y_i \overset{\text{iid}}{\sim} P(\mu)$, and the conditionally conjugate Gamma prior, $\mu \sim G(a_0, b_0)$. We can solve the integration for the marginal likelihood of a model M_1 analytically:

$$\begin{aligned} p(\boldsymbol{y} \mid M_1) &= \int_0^\infty p(\boldsymbol{y} \mid \mu, M_1)\, p(\mu \mid M_1)\, d\mu, \\ &= \frac{(\boldsymbol{y} \mid \mu, M_1)\, p(\mu \mid M_1)}{p(\mu \mid \boldsymbol{y}, M_1)} = \frac{b_0^{a_0} \Gamma(a_n)}{b_n^{a_n} \Gamma(a_0) \prod_{i=1}^n \Gamma(y_i + 1)} \end{aligned}$$

where $a_n = a_0 + \sum y_i$ and $b_n = b_0 + n$ are the sufficient statistics for the posterior of $\mu \mid \boldsymbol{y}, M_1 \sim \mathrm{G}(a_n, b_n)$.

Now consider an alternative model with a structural break, where we have $y_i \overset{\text{iid}}{\sim} P\left(\mu_j\right)$, with

$$\mu_j = \begin{cases} \mu_1 \text{ if } i < i_0 \\ \mu_2 \text{ if } i \geq i_0 \end{cases}$$

With conditionally conjugate priors, $\mu_1 \sim G(a_0, b_0), \mu_2 \sim G(a_0, b_0)$, such that $\mathrm{E}(\mu_1 - \mu_2) = 0$, the marginal likelihood is:

$$\begin{aligned} p(\boldsymbol{y} \mid M_2) &= \frac{(\boldsymbol{y} \mid \mu_1, \mu_2, M_2)\, p(\mu_1, \mu_2 \mid M_2)}{p(\mu_1, \mu_2 \mid \boldsymbol{y}, M_2)} \\ &= \frac{b_0^{2a_0} \Gamma\left(a_{n,1}\right) \Gamma\left(a_{n,2}\right)}{b_{n,1}^{a_{n,1}} b_{n,2}^{a_{n,2}} \Gamma(a_0) \prod_{i=1}^n \Gamma(y_i + 1)}, \end{aligned}$$

where $a_{n,1} = a_0 + n_1 \overline{\boldsymbol{y}}_1, b_{n,1} = b_0 + n_1$ (and analogously $a_{n,2}, b_{n,2}$) are sufficient statistics for the posterior of $\mu_1, \mu_2 \mid \boldsymbol{y}, M_2$.

The **Bayes factor** is the odds ratio of two marginal likelihoods:

$$\begin{aligned}\text{BF} &= \frac{p(\mathscr{D} \mid M_1)}{p(\mathscr{D} \mid M_2)}, \\ &= \frac{\int p(\mathscr{D} \mid \theta_1, M_1)\, p(\theta_1 \mid M_1)\, d\theta_1}{\int p(\mathscr{D} \mid \theta_2, M_2)\, p(\theta_2 \mid M_1)\, d\theta_2} = \frac{p(M_1 \mid \mathscr{D})\, p(M_2)}{p(M_2 \mid \mathscr{D})\, p(M_1)}\end{aligned}$$

For the example from before, we have:

$$\text{BF} = \frac{\Gamma(a_0)}{b_0^{a_0}} \times \frac{\Gamma(a_n)}{\Gamma\left(a_{n,1}\right)\Gamma\left(a_{n,2}\right)} \times \frac{b_{n,1}^{a_{n,1}}\, b_{n,2}^{a_{n,2}}}{b_n^{1_n}}.$$

Exact computation of the marginal likelihood is not possible for many interesting models, and numerical integration can be very challenging computationally. Models to approximate it are bridge sampling and importance sampling. Model space MCMC methods, such as stochastic search variable selection (SSVS) are an alternative approach. When comparing models using the marginal likelihood, it is important to consider the potentially strong influence of priors and their parameters.

The standard linear regression model, $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$ with conjugate priors and closed form posteriors

$$\begin{aligned}\beta \mid \sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2 \Sigma_0), \qquad & \sigma^2 \sim G^{-1}(c_0, d_0), \\ \beta \mid \sigma^2, \boldsymbol{y} \sim \mathcal{N}(\mu_n, \sigma^2 \Sigma_n), \qquad & \sigma^2 \mid \boldsymbol{y} \sim G^{-1}(c_n, d_n),\end{aligned}$$

has, by the candidate's formula, the following marginal likelihood:

$$\begin{aligned}p(\boldsymbol{y} \mid M) &= \frac{\Gamma(c_n)(d_0)^{c_0}|\Sigma_n|^{0.5}}{\Gamma(c_0)(d_n)^{c_n}|\Sigma_0|^{0.5}(2\pi)^{n/2}} \\ &= |\Sigma_0^{-1}|^{0.5} \times \frac{d_0^{c_0}}{\Gamma(c_0)} \times \frac{\Gamma(c_n)|\Sigma_n|^{0.5}}{(d_n)^{c_n}} \times \frac{1}{(2\pi)^{n/2}}.\end{aligned}$$

By **Lindley's paradox**, as the prior is becoming more and more vague, the simpler model will be chosen with probability 1, regardless of the data $\mathscr{D}$, the sample size n, and the true model. Prior distributions for parameters that appear only in some models should therefore not be vague.

Variable Selection for Regression Models

Consider the standard regression model

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n)$$

with k covariates. To answer which of the k covariates we should include, we can either compute all $P(M_j \mid \mathscr{D})$ and choose the model with the highest posterior probability, or use an approximation (which penalizes models with many parameters), sich as the Bayesian Information Criterion:

$$\text{BIC} = -2\ell + k \ln n. \tag{6.5}$$

For standard regression models, the marginal likelihood and the Bayesian Information Criterion are closely related if the number of observations is large.

It would be highly computationally intensive to compute the ML or BIC for all possible models since there are $m = 2^{k-1}$ possible models to compare. The idea of **Model space MCMC** is therefore to sample models according to their posterior probability. We search for models stochastically, and model probabilities are estimated by their relative frequency. More promising models are visited more often, while models with low probability are usually not visited. One method for this is **stochastic search variable selection** (SSVS). The idea of SSVS is to extend the standard linear model with indicators δ_j for each element of $\boldsymbol{\beta}$ such that

$$\beta_j = \begin{cases} 0 & \text{if } \delta_j = 0 \\ \text{unrestricted} & \text{if } \delta_j = 1. \end{cases}$$

The full model is then given by

$$y_i = \delta_1 x_{i1}\beta_1 + \cdots + \delta_k x_{ik}\beta_k + \varepsilon_i.$$

MCMC estimation of the SSVS model is then conceptually straightforward. These samplers are sometimes called model choice MCMC (MC3). We proceed as follows:

(1) **Model Search**: Sample a new indicator δ. Sample δ_j from $p(\delta_j \mid \delta_{-j}, \boldsymbol{y})$ for all j.

(2) **Parameter Estimation**: Sample non-zero elements of β and σ^2 conditional on δ from $p(\beta, \sigma^2 \mid \delta, \boldsymbol{y})$ using standard methods.

With a Gibbs sampler, we sample elements δ_j in blocks from their discrete, conditional posteriors,

$$P(\delta_j = 0 \mid \delta_{-j}, \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \delta_j = 0, \delta_{-j}) p(\delta_j = 0 \mid \delta_{-j}),$$
$$P(\delta_j = 1 \mid \delta_{-j}, \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \delta_j = 1, \delta_{-j}) p(\delta_j = 1 \mid \delta_{-j}).$$

The probabilities $P(\delta_j = 0 \mid \delta_{-j}, \boldsymbol{y})$ and $P(\delta_j = 1 \mid \delta_{-j}, \boldsymbol{y})$ need to be normalized before sampling.

A simple prior for the indicator δ is to assume a fixed proportion π of non-zero coefficients:

$$P(\delta_j = 1 \mid \pi) = \pi.$$

This implies a binomial distribution on the model size k_δ:

$$k_\delta = \sum_{i=1}^{k} \delta_i \sim \text{Binom}(k, \pi).$$

With $\pi = 0.5$, we have a uniform prior over all models, i.e., we are uninformative with respect to the model. However, this prior is highly informative on the model size k_δ. To induce a uniform prior over the model *size*, we can use a hierarchical prior, where $\pi \sim U[0, 1]$. For large models, this is preferable to fixing π.

Assume, for example, that there is one good model with δ^* out of 2^{20}, such that $p(\boldsymbol{y} \mid \delta^*) = \text{BF} p(\boldsymbol{y} \mid \delta_m)$, where $\text{BF} > 1$:

$$
\begin{aligned}
p(\delta^* \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{y} \mid \delta^*)p(\delta^*)}{p(\boldsymbol{y} \mid \delta^*)p(\delta^*) + \sum_{\delta_j \neq \delta^*} p(\boldsymbol{y} \mid \delta_j)p(\delta_j)} \\
&= \frac{\text{BF}}{\text{BF} + (2^{20} - 1)} \approx \frac{\text{BF}}{\text{BF} + 1{,}000{,}000}.
\end{aligned}
$$

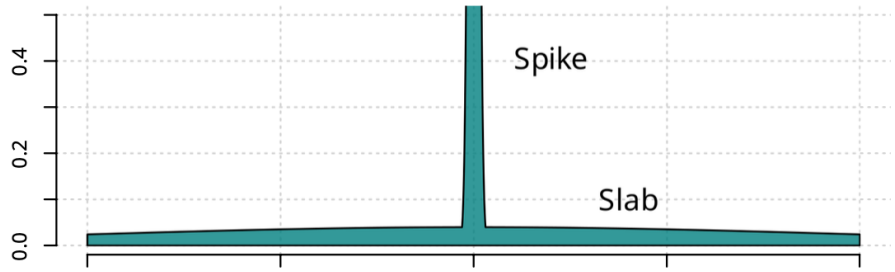
Bayesian Model Averaging

An alternative to looking for the "best model" and thereby ignoring uncertainty around the model selection is to consider all candidate models, weighted by their posterior probability. This approach is called Bayesian model averaging (BMA). Instead of conditioning parameters on one model, we now keep all parameter estimates and weigh them.

Priors that zero-out unwanted parameters are called spike-and-slab priors. We can express them using a mixture of normals,

$$
p(\beta \mid \delta) \sim \delta \mathcal{N}(0, v_1) + (1 - \delta)\mathcal{N}(0, v_2),
$$

where $v_1 \gg v_2$, meaning that we pull β towards zero if $\delta = 0$. The mixture of normals can be visualized like this:



Other Approaches

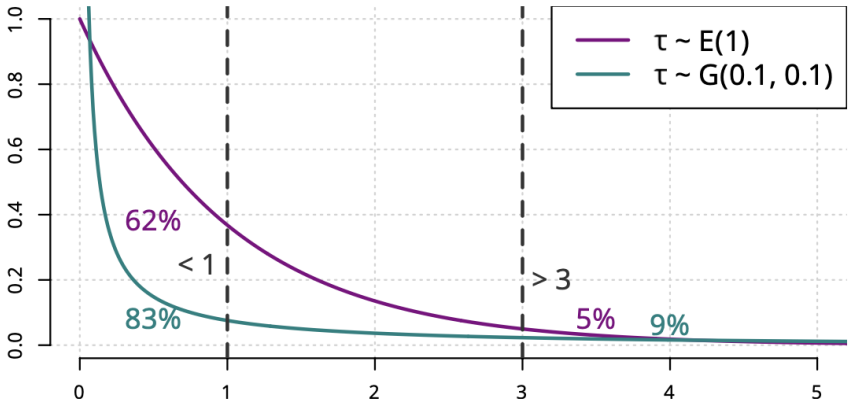
Another approach are continuous shrinkage priors, which are similar to frequentist regularized estimators. Examples are Bayesian LASSO, where

$$
p(\beta \mid \tau_j^2) \sim \mathcal{N}\left(0, \frac{2}{\lambda^2}\tau_j^2\right), \qquad \tau_j^2 \sim \mathcal{E}(1),
$$

or the Normal-Gamma prior, where

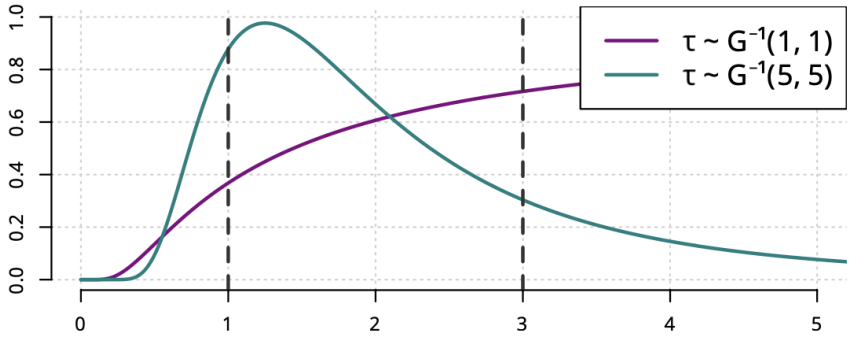
$$
p(\beta \mid \tau_j^2) \sim \mathcal{N}\left(0, \frac{2}{\lambda^2}\tau_j^2\right), \qquad \tau_j^2 \sim G(a_\tau, a_\tau).
$$

In priors like these, the term $2/\lambda^2$ applies global shrinkage, while τ_j^2 apples local shrinkage for the coefficients.



Not every scale-mixture prior induces shrinkage, such as $\beta \mid \eta \sim t_{2\eta}(0, \lambda^2)$:

$$\beta_j \mid \tau_j^2 \sim \mathcal{N}(0, \lambda^2 \tau_j^2), \qquad \tau_j^2 \sim G^{-1}(\eta, \eta).$$



7 Bayesian Vector Autoregressions

Bayesian Estimation of a VAR

A reduced-form VAR(p) model is given by

$$\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \ldots + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}_M(\boldsymbol{0}, \boldsymbol{\Sigma}). \tag{7.1}$$

More compactly, we can write it using the lag polynomial as

$$\boldsymbol{A}(L)\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{\varepsilon}_t.$$

We can rewrite any VAR(p) process as a VAR(1) process:

$$\begin{aligned}
\boldsymbol{Y}_t &= \boldsymbol{\nu} + \boldsymbol{A}\boldsymbol{Y}_{t-1} + \boldsymbol{E}_t \\
\boldsymbol{J}'\boldsymbol{Y}_t = \boldsymbol{y}_t &= \boldsymbol{J}'(\boldsymbol{\nu} + \boldsymbol{A}\boldsymbol{Y}_{t-1} + \boldsymbol{E}_t) \\
\boldsymbol{J}'\boldsymbol{Y}_t = \boldsymbol{y}_t &= \boldsymbol{c} + [\boldsymbol{A}_1\, \boldsymbol{A}_2\, \cdots\, \boldsymbol{A}_p]\boldsymbol{Y}_{t-1} + \boldsymbol{\varepsilon}_t,
\end{aligned}$$

where

$$\underset{Mp\times 1}{\boldsymbol{Y_t}} \equiv \begin{bmatrix} \boldsymbol{y_t} \\ \boldsymbol{y_{t-1}} \\ \vdots \\ \boldsymbol{y_{t-p+1}} \end{bmatrix}, \underset{Mp\times Mp}{\boldsymbol{A}} \equiv \begin{bmatrix} \boldsymbol{A_1} & \boldsymbol{A_2} & \dots & \boldsymbol{A_{p-1}} & \boldsymbol{A_p} \\ \boldsymbol{I_M} & \boldsymbol{0_M} & \dots & \boldsymbol{0_M} & \boldsymbol{0_M} \\ \boldsymbol{0_M} & \boldsymbol{I_M} & & \boldsymbol{0_M} & \boldsymbol{0_M} \\ \vdots & & \ddots & \vdots & \vdots \\ \boldsymbol{0_M} & \boldsymbol{0_M} & \dots & \boldsymbol{I_M} & \boldsymbol{0_M} \end{bmatrix},$$

$$\underset{Mp\times 1}{\boldsymbol{v}} \equiv \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{0_M} \\ \vdots \\ \boldsymbol{0_M} \end{bmatrix}, \quad \underset{Mp\times 1}{\boldsymbol{E_t}} \equiv \begin{bmatrix} \boldsymbol{\varepsilon_t} \\ \boldsymbol{0_M} \\ \vdots \\ \boldsymbol{0_M} \end{bmatrix} = \boldsymbol{J}\boldsymbol{\varepsilon}_t, \underset{Mp\times M}{\boldsymbol{J}} \equiv \begin{bmatrix} \boldsymbol{I_M} \\ \boldsymbol{0_M} \\ \vdots \\ \boldsymbol{0_M} \end{bmatrix}.$$

We can estimate this model using OLS, generalized LS, maximum likelihood, or Bayesian techniques. As mentioned before, the curse of dimensionality says that if there are M equations, $M+pM^2$ parameters have to be estimated, where M is the number of variables and p is the lag order. The number of free parameters (elements in $\boldsymbol{c}, \boldsymbol{A}_j$ and $\boldsymbol{\Sigma}$) is given by

$$M(Mp+1)+\frac{(M+1)M}{2}.$$

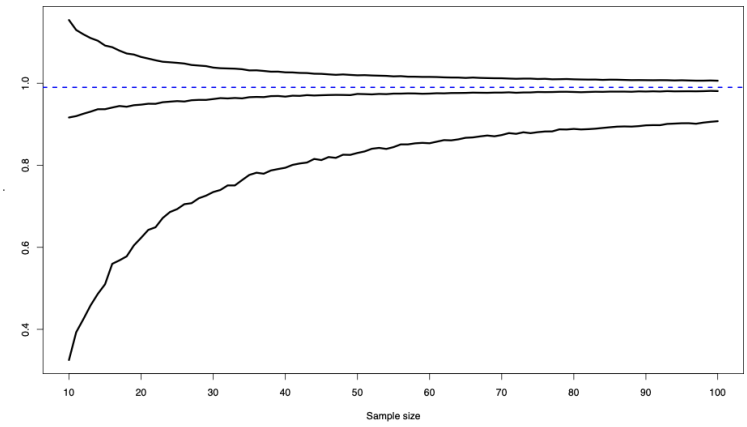
Writing the VAR(p) model as

$$\boldsymbol{y}_t = (\underbrace{[1, \boldsymbol{y}'_{t-1}, \dots, \boldsymbol{y}'_{t-p}]}_{\boldsymbol{x}_t}\underbrace{[\boldsymbol{c}', \boldsymbol{A}'_1, \dots, \boldsymbol{A}'_p]}_{\boldsymbol{A}})' + \boldsymbol{\varepsilon}_t,$$

OLS gives estimates $\hat{\boldsymbol{A}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ and $\hat{\boldsymbol{\Sigma}} = (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{A})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{A})/(T-Mp-1)$. These estimates are consistent and are asymptotically normally distributed with $\hat{\boldsymbol{\alpha}} = \text{vec}(\boldsymbol{A})$, where vec($\boldsymbol{A}$) is the vectorization of a matrix A obtained by stacking the columns of the matrix on top of one another:

$$\hat{\boldsymbol{\alpha}} \overset{LLN}{\sim} \mathcal{N}(\boldsymbol{\alpha}, (\boldsymbol{X}'\boldsymbol{X})^{-1} \otimes \hat{\boldsymbol{\Sigma}}).$$

However, OLS has a small sample bias:



OLS has an implicit prior on lower autoregressive values.

Bayesian methods are often used in estimating VARs. Reasons for this include proliferation of parameters (curse of dimensionality, addressed by regularization), and conditioning on initial values (via suitable prior means). To perform estimation like this, we need the likelihood of a VAR, the prior setup for the autoregressive coefficients and deterministics, and the prior setup for the variance parameters.

The Likelihood

In the VAR model from before,

$$\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \ldots + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

we can define $\boldsymbol{\alpha} = \text{vec}(\boldsymbol{A}) = \text{vec}((\boldsymbol{c}', \boldsymbol{A}_1', \ldots, \boldsymbol{A}_p'))$ and $\boldsymbol{x}_t = (1, \boldsymbol{y}_{t-1}', \ldots, \boldsymbol{y}_{t-p}')$ and rewrite the model as

$$\boldsymbol{y}_t = (\boldsymbol{I}_M \otimes \boldsymbol{x}_t)\boldsymbol{\alpha} + \boldsymbol{\varepsilon}_t.$$

The conditional likelihood is then given by

$$p(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_{t-p}, \boldsymbol{A}, \boldsymbol{\Sigma}) \sim \mathcal{N}_M([\boldsymbol{x}_t \boldsymbol{A}]', \boldsymbol{\Sigma}), \tag{7.2}$$

or

$$p(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_{t-p}, \boldsymbol{A}, \boldsymbol{\Sigma}) \sim \mathcal{N}_M((\boldsymbol{I}_M \otimes \boldsymbol{x}_t)\boldsymbol{\alpha}, \boldsymbol{\Sigma}).$$

The Prior

As prior for $\boldsymbol{A}$ (or $\boldsymbol{\alpha} = \text{vec}(\boldsymbol{A})$), we can use

$$p(\boldsymbol{\alpha}) \sim \mathcal{N}(\underline{\boldsymbol{\alpha}}, \underline{\boldsymbol{V}}). \tag{7.3}$$

As prior for $\boldsymbol{\Sigma}$, we can use an Inverse Wishart distribution, which is something like a multivariate version of the inverse Gamma distribution. Together, this forms an independent NiW prior for our VAR.

If the sample size is small, two problems arise in estimation of a VAR: the number of parameters, and the tendency to underestimate persistence since the VAR is analyzed with the conditional likelihood. Unrestricted VARs often forecast worse than univariate random walk models.

As prior mean, we propose

$$\underline{\boldsymbol{A}} = \mathrm{E}(\boldsymbol{A}) = (\boldsymbol{I}_M, \boldsymbol{0}, \boldsymbol{0}). \tag{7.4}$$

That is, we set the coefficient associated with the first lag to unity, and all other coefficients to zero (which we call regularization). This induces high persistence and pushes the system towards random walk behavior.

As prior for the variance, we can use the Minnesota prior given by

$$\underline{\boldsymbol{V}} = \mathrm{Var}(\boldsymbol{\alpha}) = \begin{cases} \left(\frac{\lambda_1}{k}\right)^2 & \text{for } i = j \text{ and the } k\text{-th lag,} \\ \left(\frac{\sigma_i^2}{\sigma_j^2}\right)\left(\frac{\lambda_1\lambda_2}{k}\right)^2 & \text{for } i \neq j \text{ and the } k\text{-th lag,} \\ \lambda_3\sigma_i^2 & \text{for the deterministic part of the model.} \end{cases} \tag{7.5}$$

This allows us to provide more shrinkage to more distant lags. The hyperparameter λ_1 governs shrinkage for own lags, λ_2 governs shrinkage for cross-variable lags, and λ_3 governs shrinkage of deterministics, and σ_i^2 is the standard deviation of an autoregressive model of variable i to account for different scalings.

For the variance $\boldsymbol{\Sigma}$, we choose an Inverse Wishart prior, which is similar to a multivariate form of the Inverse Gamma distribution:

$$\boldsymbol{\Sigma} \mid \boldsymbol{Y} \sim iW(\underline{s}, \underline{\boldsymbol{S}}), \tag{7.6}$$

where $\underline{s}$ are the prior degrees of freedom and $\underline{\boldsymbol{S}}$ is the prior scaling matrix. We can, for example, set the following hyperparameter values:

$$\begin{aligned} \underline{s} &= M + 2, \\ \underline{\boldsymbol{S}} &= (\underline{s} - M - 1)\mathrm{diag}(\sigma_1^2, \ldots, \sigma_M^2). \end{aligned}$$

We can also use the LDL decomposition of $\boldsymbol{\Sigma}$,

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}',$$

where $\boldsymbol{L}$ is a lower triangular matrix and $\boldsymbol{D}$ is a diagonal matrix. On the off-diagonal elements of $\boldsymbol{L}$, we place Gaussian priors and on the diagonal elements of $\boldsymbol{D}$, we place inverse Gamma priors. The posterior is then very similar to the Inverse Wishart setup, but allows in contrast to the latter for equation-by-equation estimation of the autoregressive coefficients. It is also straightforward to account for heteroskedastic errors, e.g., through stochastic volatility.

The Posterior

We then receive the following posterior for the autoregressive parameter:

$$\boldsymbol{\alpha} \mid \boldsymbol{\Sigma}, \boldsymbol{Y} \sim \mathcal{N}(\overline{\boldsymbol{a}}, \overline{\boldsymbol{V}}), \tag{7.7}$$

where

$$\begin{aligned}\overline{\boldsymbol{V}} &= (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}'\boldsymbol{X} + \underline{\boldsymbol{V}}^{-1})^{-1}, \\ \overline{\boldsymbol{a}} &= \overline{\boldsymbol{V}}(\underline{\boldsymbol{V}}^{-1}\underline{\boldsymbol{a}} + (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}')\boldsymbol{y}).\end{aligned}$$

The posterior for the variance is:

$$\boldsymbol{\Sigma} \mid \boldsymbol{Y} \sim iW(\overline{s}, \overline{\boldsymbol{S}}), \tag{7.8}$$

where

$$\begin{aligned}\overline{s} &= T + \underline{s}, \\ \overline{\boldsymbol{S}} &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{A})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{A}) + \underline{\boldsymbol{S}}.\end{aligned}$$

8 Identification of Vector Autoregressions

Identification in Macroeconomics

Before, we looked at tools that can be used with a VAR in its structural form, like structural impulse responses, forecast error variance, or historical decompositions. For this, we needed the structural form of a VAR(p) model, given by

$$\underbrace{\boldsymbol{B}_0^{-1}\boldsymbol{B}_0}_{\boldsymbol{I}}\boldsymbol{y}_t = \underbrace{\boldsymbol{B}_0^{-1}\boldsymbol{B}_1}_{\boldsymbol{A}_1}\boldsymbol{y}_{t-1} + \ldots + \underbrace{\boldsymbol{B}_0^{-1}\boldsymbol{B}_p}_{\boldsymbol{A}_p}\boldsymbol{y}_{t-p} + \overbrace{\boldsymbol{B}_0^{-1}\boldsymbol{e}_t}^{\varepsilon_t}, \quad \boldsymbol{e}_t \sim \mathcal{N}(0, \boldsymbol{I}_M)$$

Before, we assumed $\boldsymbol{B}_0^{-1}$ to be known. In this section, we will investigate how to find $\boldsymbol{B}_0^{-1}$, that is, how to identify a VAR.

Identification means to draw causal or structural conclusions from the correlations we observe in the data. In macroeconomics, this is difficult, since we need to find exogenous variation in the shock we would like to analyze. Experiments in macroeconomics are generally not possible, and even quasi-experiments are very difficult, since macroeconomic policies tend to hit the entire economy and not one separable part of it.

The reduced form of a VAR model (for simplicity a VAR(1) model is shown) is given by

$$\boldsymbol{y}_t = \boldsymbol{A}_1\boldsymbol{y}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t = \boldsymbol{B}_0^{-1}\boldsymbol{e}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\varepsilon), \tag{8.1}$$

and the structural form is given by

$$\boldsymbol{B}_0 \boldsymbol{y}_t = \boldsymbol{B}_1 \boldsymbol{y}_{t-1} + \boldsymbol{e}_t, \qquad \boldsymbol{A}_1 = \boldsymbol{B}_0^{-1} \boldsymbol{B}_1. \tag{8.2}$$

In the reduced form, the number of free parameters is

$$\underbrace{M^2}_{\text{in } \boldsymbol{A}_1} + \underbrace{\frac{M \times (M+1)}{2}}_{\text{in } \boldsymbol{\Sigma}_e \text{ (symmetric)}},$$

and in the structural form, it is

$$\underbrace{M^2}_{\text{in } \boldsymbol{B}_0} + \underbrace{M^2}_{\text{in } \boldsymbol{B}_1}.$$

The difference, that is, the number of restrictions we need to impose on $\boldsymbol{B}_0^{-1}$, is therefore

$$M^2 - \frac{M^2}{2} - \frac{M}{2} = \frac{M \times (M-1)}{2}.$$

Macroeconomic Shocks

We are interested in identifying an economically meaningful structural shock. It should be exogenous with respect to other current and lagged endogenous variables in the model, uncorrelated with other exogenous shocks (which is hard to defend), and represent either unanticipated movements in exogenous variables or news about future movements in exogenous variables. Examples in macroeconomics include monetary policy shocks, fiscal policy shocks, supply side shocks, demand side shocks, uncertainty shocks, etc.

Consider, for example, a monetary policy shock (that should in principle be free of political influence). The monetary policy shock e_t^{MP} reflects the preferences of the monetary authority and is subject to its strategic considerations and technical factors.

Identification Schemes

We can impose long-run restrictions, short-run restrictions and sign restrictions. Other identification strategies include narrative methods, external instruments / Proxy VARs (where we assume to have a relevant and contemporaneously exogenous instrument for a shock, which means that projecting it on the reduced form errors, which are linear functions of the structural errors, will generate a (scaled) estimate of the structural errors under the identifying assumptions), or high-frequency identification (e.g., looking only at a 30-minute window around policy announcements).

Short-Run Restrictions

We can define that a shock has no contemporaneous impact element on a certain variable. We do this by setting the corresponding elements in $\boldsymbol{B}_0^{-1}$ to zero, such that

$$\boldsymbol{B}_0^{-1} = \begin{bmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{bmatrix} \quad \text{or} \quad \boldsymbol{B}_0^{-1} = \begin{bmatrix} * & 0 & 0 \\ * & b_{22} & b_{23} \\ * & * & * \end{bmatrix}$$

We could also impose linear restrictions on some elements of $\boldsymbol{B}_0^{-1}$ instead of setting them to zero. However, we need to argue for any identification scheme to be economically meaningful.

Consider as an example the macroeconomic model where $\boldsymbol{y}_t = [\tilde{y}_t, \pi_t, r_t]$, where the elements are output growth, inflation, and the short-term interest rate. We can estimate a reduced-form VAR(1) process

$$\boldsymbol{y}_t = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}_M(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

Finding the structural form of the process, we get

$$\begin{aligned} \boldsymbol{B}_0^{-1} \boldsymbol{y}_t &= \boldsymbol{B}_1 \boldsymbol{y}_{t-1} + \boldsymbol{e}_t, \qquad \boldsymbol{e}_t \sim \mathcal{N}_M(\boldsymbol{0}, \boldsymbol{I}_M) \\ \boldsymbol{y}_t &= \boldsymbol{B}_0^{-1} \boldsymbol{B}_1 \boldsymbol{y}_{t-1} + \boldsymbol{B}_0^{-1} \boldsymbol{e}_t \\ \boldsymbol{y}_t &= \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \boldsymbol{\varepsilon}_t. \end{aligned}$$

From this, we know that

$$\underbrace{\begin{bmatrix} \varepsilon_t^{\tilde{y}} \\ \varepsilon_t^{\pi} \\ \varepsilon_t^{r} \end{bmatrix}}_{\boldsymbol{\varepsilon}_t} = \underbrace{\begin{bmatrix} b_0^{11} & b_0^{12} & b_0^{13} \\ b_0^{21} & b_0^{22} & b_0^{23} \\ b_0^{31} & b_0^{32} & b_0^{33} \end{bmatrix}}_{\boldsymbol{B}_0^{-1}} \underbrace{\begin{bmatrix} e_t^{\tilde{y}} \\ e_t^{\pi} \\ e_t^{r} \end{bmatrix}}_{\boldsymbol{e}_t}$$

This is a system of three linear equations in nine unknowns. The variance,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \sigma_{13}^2 \\ \sigma_{12}^2 & \sigma_{22}^2 & \sigma_{23}^2 \\ \sigma_{13}^2 & \sigma_{23}^2 & \sigma_{33}^2 \end{bmatrix}$$

provides six parameters, meaning that we need to impose three additional restrictions to solve the system.

A simple choice is to set $\boldsymbol{B}_0^{-1} = \boldsymbol{P}$, where $\boldsymbol{P}$ is the lower-triangular Cholesky decomposition of $\boldsymbol{\Sigma}_\varepsilon$. This is because by construction, $\boldsymbol{\varepsilon}_t = \boldsymbol{B}_0^{-1} \boldsymbol{e}_t \Leftrightarrow \boldsymbol{e}_t = \boldsymbol{B}_0 \boldsymbol{\varepsilon}_t$. Hence, the variance-covariance matrix of the reduced form errors is

$$\boldsymbol{\Sigma}_\varepsilon = \mathrm{E}\left(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t'\right) = \boldsymbol{B}_0^{-1} \mathrm{E}\left(\boldsymbol{e}_t \boldsymbol{e}_t'\right) \boldsymbol{B}_0^{-1\prime} = \boldsymbol{B}_0^{-1} \boldsymbol{\Sigma}_e \boldsymbol{B}_0^{-1\prime} = \boldsymbol{B}_0^{-1} \boldsymbol{B}_0^{-1\prime}.$$

The Cholesky decomposition can in a way be understood as "taking the square root" in matrix form. By applying a Cholesky decomposition of $\boldsymbol{\Sigma}_\varepsilon$, we therefore obtain $\boldsymbol{\Sigma}_\varepsilon = \boldsymbol{P}\boldsymbol{P}'$, but also $\boldsymbol{\Sigma}_\varepsilon = \boldsymbol{B}_0^{-1} \boldsymbol{B}_0^{-1\prime}$, where both $\boldsymbol{P}$ and $\boldsymbol{B}_0^{-1}$ are lower triangular. Therefore,

$$\boldsymbol{P} = \boldsymbol{B}_0^{-1},$$

which means identification.

To identify $\boldsymbol{B}_0^{-1}$, we need to impose $M(M-1)/2$ restrictions. In the case of $M=2$, the Cholesky decomposition is

$$\boldsymbol{P} = \begin{pmatrix} \sqrt{\sigma_{11}} & 0 \\ \frac{\sigma_{12}}{\sqrt{\sigma_{11}}} & \sqrt{\sigma_{22} - \frac{1}{\sigma_{11}}\sigma_{12}^2}, \end{pmatrix}$$

where imposing one single zero restriction yields exact identification. We have recursive ordering for instantaneous responses: y_{1t} depends only on e_{1t}, y_{2t} depends on e_{1t} and e_{2t}. Given a certain ordering of variables, the decomposition is unique. However, it is not unique w.r.t. the ordering of the variables as there exist $M!$ possible orderings. Permuting the order of elements in $\boldsymbol{y}$ yields different orthogonal VMA representations. Therefore, the ordering of variables matters.

In the example from before, defining $\boldsymbol{B}_0^{-1} = \boldsymbol{P}$ as the lower triangular Cholesky part of $\boldsymbol{\Sigma}_\varepsilon$, we get

$$\begin{bmatrix} \varepsilon_t^{\tilde{y}} \\ \varepsilon_t^{\pi} \\ \varepsilon_t^{r} \end{bmatrix} = \begin{bmatrix} p^{11} & 0 & 0 \\ p^{21} & p^{22} & 0 \\ p^{31} & p^{32} & p^{33} \end{bmatrix} \begin{bmatrix} e_t^{\tilde{y}} \\ e_t^{\pi} \\ e_t^{r} \end{bmatrix}.$$

The system is now identified. From an interpretation perspective, the restrictions imply the following three equations:

$$\begin{aligned} \Delta y_t &= \cdots + p_{11} e_t^{\tilde{y}_t} \\ \pi_t &= \cdots + p_{21} e_t^{\tilde{y}_t} + p_{22} e_t^{\pi_t} \\ r_t &= \cdots + p_{31} e_t^{\tilde{y}_t} + p_{32} e_t^{\pi_t} + p_{33} e_t^{r_t} \end{aligned}$$

We can think of the first two equations as being aggregate demand and aggregate supply equations. AD shocks then move both output and prices on impact, whereas AS shocks move prices only. This can be visualized by a horizontal AD curve and an upward-sloping AS curve. The monetary policy authority then responds systematically to innovations in prices and output, and any change not accounted by these responses marks an exogenous monetary policy shock.

It is important to note that we assume that output and prices are not responding to interest rate shocks within a quarter/month, which may be plausible for a month, but not for a quarter.

Sign Restrictions

For short-run restrictions we set the contemporaneous reactions of variables to certain shocks to a fixed value, usually to zero, a priori. Economic theory often suggests rather how and not when variables react to shocks; not necessarily giving an exact response. Hence, an alternative approach would be to restrict reactions of variables to increase/decrease given a certain shock → sign restrictions.

Consider a simple bivariate model of a goods market with a demand shock (e_t^{demand}) and a supply shock (e_t^{supply}). As before, the relationship between the reduced-form residuals $\boldsymbol{\varepsilon}_t$ and the structural shocks is then

$$\boldsymbol{\varepsilon}_t = \boldsymbol{B}_0^{-1}\boldsymbol{e}_t,$$

where $\boldsymbol{\varepsilon}_t = (\varepsilon_t^q, \varepsilon_t^p)$ and $\boldsymbol{e}_t = (e_t^{\text{supply}}, e_t^{\text{demand}})$. Identifying this model using short-run zero restrictions would imply, for example, production not responding contemporaneously to a price increase, meaning that the supply curve would be horizontal. This would be hard to argue for economically.

In general, a supply shock will increase quantity and reduce the price, while a demand shock raises both of them, which tells us something about the sign of the effect. We can therefore impose

$$\begin{pmatrix} \varepsilon_t^q \\ \varepsilon_t^p \end{pmatrix} = \begin{bmatrix} + & + \\ - & + \end{bmatrix} \begin{pmatrix} e_t^{\text{supply}} \\ e_t^{\text{demand}} \end{pmatrix},$$

where $+$ and $-$ indicate strictly positive or negative signs of the parameters in the structural multiplier matrix. In this, shocks cannot share the same sign pattern because then we could not identify them separately. Parameters of $\boldsymbol{B}_0^{-1}$ are no longer point identified, but set identified. This complicates estimation. Sign restrictions can also be combined with zero restrictions. The sign-identified model does not nest the recursively identified model. It relaxes the exclusion restriction but does it at the cost of restricting other parameters that were unrestricted.

Consider, for example, a VAR(p) model in its structural form,

$$\boldsymbol{B}_0\boldsymbol{y}_t = \boldsymbol{B}_1\boldsymbol{y}_{t-1} + \ldots + \boldsymbol{B}_p\boldsymbol{y}_{t-p} + \boldsymbol{e}_t, \quad \boldsymbol{e}_t \sim \mathcal{N}(0, \boldsymbol{I}_M).$$

Suppose we have our structural multiplier matrix $\boldsymbol{B}_0^{-1}$ (with $\boldsymbol{B}_0^{-1}\boldsymbol{B}_0^{-1\prime} = \boldsymbol{\Sigma}_\varepsilon$) and we can find a full matrix $\boldsymbol{Q}$ such that

$$\boldsymbol{\varepsilon}_t = \boldsymbol{B}_0^{-1}\boldsymbol{Q}\boldsymbol{e}_t, \qquad \text{implying that } \boldsymbol{\Sigma}_\varepsilon = \boldsymbol{B}_0^{-1}\boldsymbol{Q}\boldsymbol{Q}'\boldsymbol{B}_0^{-1\prime}.$$

This holds if $\boldsymbol{Q}$ is an orthonormal rotation matrix, $\boldsymbol{Q}\boldsymbol{Q}' = \boldsymbol{I}_M$. We can search for $\boldsymbol{Q}$ stochastically or deterministically such that $\boldsymbol{C} = \boldsymbol{B}_0^{-1}\boldsymbol{Q}$ fulfills the sign restriction. For each value of $\boldsymbol{\Sigma}_\varepsilon$, there is a continuum of orthonormal $\boldsymbol{Q}$ that all satisfy the constrains, which means that we have set identification.

There are several algorithms to find $\boldsymbol{Q}$. The most efficient one for larger systems is based on the Householder transformation:

(1) Draw each element of $\tilde{\boldsymbol{S}}(M \times M)$ from $\mathcal{N}(0, 1)$.

(2) Compute the QR decomposition of $\tilde{\boldsymbol{S}}$. The $\boldsymbol{Q}$ factor is a candidate rotation.

(3) Compute $\boldsymbol{C} = \boldsymbol{B}_0^{-1}\boldsymbol{Q}$ and the associated IRFs and check whether the constraints are fulfilled.

(4) If so, store the IRFs. If not, discard them.

(5) Perform N replications and report the median impulse response along with some quantiles.

Often, we only have knowledge about the effects of some shocks, but not others. In other cases, only one shock is of interest. In the example of above, not knowing about the effects of a supply shock on price and quantity would mean that

$$\begin{pmatrix} \varepsilon_t^q \\ \varepsilon_t^p \end{pmatrix} = \begin{bmatrix} ? & + \\ ? & + \end{bmatrix} \begin{pmatrix} e_t^{\text{supply}} \\ e_t^{\text{demand}} \end{pmatrix},$$

Checking only the second column of $\boldsymbol{C}$ would in this case not be enough, as a possible solution for the signs of it could be

$$\begin{bmatrix} + & + \\ + & + \end{bmatrix},$$

which would not allow for the distinction of the shocks. Admissible draws would be

$$\begin{bmatrix} + & + \\ - & + \end{bmatrix} \text{ or } \begin{bmatrix} - & + \\ + & + \end{bmatrix}.$$

Under sign restriction based IRFs, we do not obtain unique point estimates of the impulse response function. The structural models are then only set identified. Computing classical confidence intervals in a frequentist setting is then complicated, but Bayesian methods allow for more precise and reliable confidence bounds while also incorporating parameter uncertainty. To achieve this, we can get posterior draws of the parameters $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$, compute IRFS with a $\boldsymbol{Q}$, and then compute quantiles.