Specialist Courses

Daniele Zago

August 21, 2022

Contents

Time series

Instructor: prof. L. Bisaglia

This is a short course aimed at giving an introduction to time series models and their application to modelling and prediction of data collected sequentially in time. The aim is to provide specific techniques for handling data and at the same time to provide some understanding of the theoretical basis for the techniques. Topics covered will include univariate linear and non linear models (both in mean and variance) and some basics of spectral analysis. Finally, we will cover some aspects of long-memory and integer autoregressive models for count data.

Textbook references

Brockwell and Davis (2016)	Introduction to Time Series and Forecasting
Fan and Yao (2005)	Nonlinear Time Series: Nonparametric and Parametric Methods
Shumway and Stoffer (2017)	Time Series Analysis and Its Applications: With R Examples
Tsay (2013)	Multivariate Time Series Analysis: With R and Financial Applications
Wei (2019)	Multivariate Time Series Analysis and Applications
Brockwell (2009)	Time Series: Theory and Methods
Douc et al. (2014)	Nonlinear Time Series: Theory, Methods and Applications with R Examples

Lecture 1: Introduction to time series

2021-11-12

In general, in time series we are interested in *a)* understanding the stochastic mechanism that gives rise to an observed series and *b)* to forecast future values of a series based on the observed history. As this course is introductory, we will restrict our analysis to *univariate* time series.

Assumption We assume that future behaviour is equal to previous behaviour, i.e. we are able to forecast the future based on the information about the observed past data.

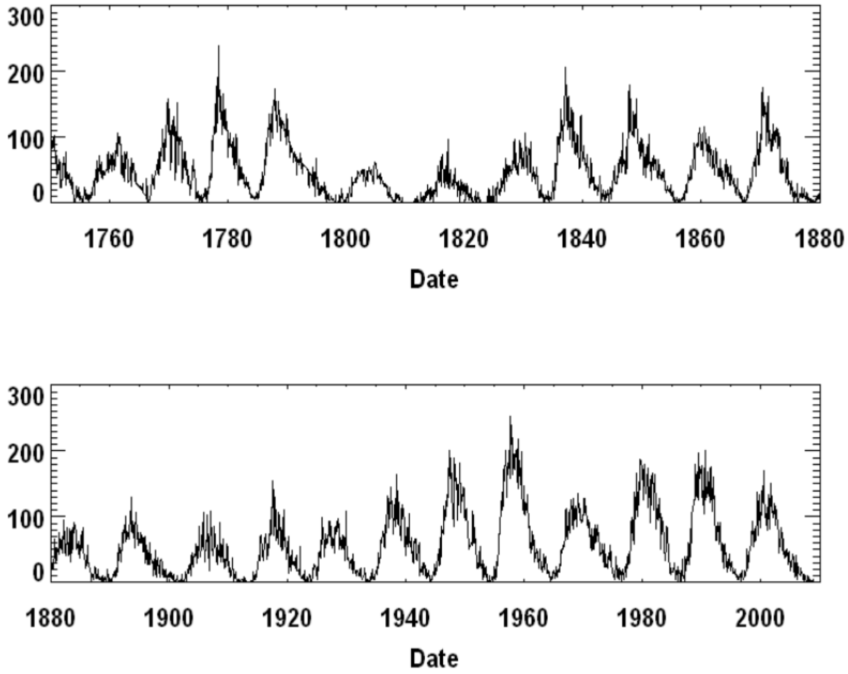

Figure 1: Classical linear time series models are not able to explain this behaviour, since cyclic components are not constant in amplitude over time.

There are several approaches in modern time series, namely

- *Classical approach*: trend + cycle + seasonality.
- *Modern approach*: Box and Jenkins procedure with ARIMA models.
- *State-space approach*: follows Durbin and Koopman (2012), we will not treat it here.

1.1 Classical approach

We assume a data-generating process given by a basic deterministic function of time plus additive noise,

$$Y_t = f(t) + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{WN}(0, \sigma_\varepsilon^2),$$

such that $\mathbb{E}[\varepsilon_t] = 0$, $\mathbb{V}[\varepsilon_t] = \sigma_\varepsilon^2$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$. Assuming different shapes of $f(t)$ lets us obtain different types of time series:

- Additive: TREND + SEASONALITY + CYCLES: $f(t) = T_t + S_t + C_t$

› *Multiplicative*: TREND · SEASONALITY · CYCLES: $f(t) = T_t \cdot S_t \cdot C_t$

The classical approach establishes that trend, seasonal, and cyclic components should be *estimated separately* with simple models and then *combined*. For example, we can use a linear model for the trend such as

$$T_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \ldots + \alpha_g t^g.$$

On the other hand, in order to model S_t we could use dummy variables with sine/cosine transform to promote cyclic behaviour.

Problem Empirical time series contain both deterministic trends and stochastic trends, which cannot be modeled by stationary processes.

$$\text{DETERMINISTIC TREND} \quad \mathbb{E}[X_t] = f(t)$$

$$\text{STOCHASTIC TREND} \quad \sum_{i=1}^{t} \varepsilon_t$$

1.2 Modern approach

We can consider the data-generating process (DGP) as a stochastic process which yields the observed time series as a sample path over time. To perform statistical inference, we need to assume that at least some features of the underlying probability law are *stationary* over the time period of interest.

Def. (Stochastic process)
A collection of random variables $X = (X_t)_t$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called a ***stochastic process***

Remark A stochastic process is therefore a function of two arguments $X : \mathcal{T} \times \Omega \to X$, $(t, \omega) \mapsto X_t(\omega)$ and for a fixed value of ω we obtain a *path* from the stochatstic process.

Sample We only observe a portion of the infinite path of the stochastic process,

$$\ldots, X_{-t}, X_{-t-1}, \ldots, X_0, \underbrace{X_1, X_2, \ldots, X_t}_{x_1, x_2, \ldots, x_t}, \ldots,$$

therefore if we want to make inference over the DGP we must make some strong assumptions.

Def. (Mean function)
For a stochastic process X_t, the ***mean function*** is

$$\mu_t = \mathbb{E}[X_t] \quad \text{for } t \in \mathcal{T}.$$

Def. (Autocovariance function)
For a stochastic process X_t, the ***autocovariance function*** is

$$\gamma_{t,s} = \mathrm{Cov}(X_t, X_s) = \mathbb{E}[(X_t - \mu_t)(X_s - \mu_s)], \quad \text{for } t, s = 0, \pm 1, \pm 2, \ldots$$

The autocorrelation function is then defined as

$$\mathrm{Corr}(X_t, X_s) = \frac{\mathrm{Cov}(X_t, X_s)}{\sqrt{\mathbb{V}[X_t]}\sqrt{\mathbb{V}[X_s]}}.$$

We need to make some strong assumptions on the structure of the process in order to make inference possible.

Def. (Strong stationarity)
A process $(X_t)_t$ is ***strictly stationary*** if it is invariant under time shifts, i.e. if

$$(X_{t_1}, \dots, X_{t_n}) \overset{\mathrm{d}}{=} (X_{t_1+k}, \dots, X_{t_n+k})$$

for any $n \geq 1$, any choice of $t_1, \dots, t_n$ and al time shifts $k \in \mathbb{Z}$.

Marginals Choosing for instance $n = 1$ means that the marginal distribution of X_t the same as that of X_{t-k} for all t and k.

Def. (Weak stationarity)
A process $(X_t)_t$ is ***weakly stationary*** if

1. $\mathbb{E}[X_t] = \mu < \infty$ for all t.
2. $\mathbb{V}[X_t] = \sigma^2 < \infty$ for all t.
3. $\mathrm{Cov}(X_t, X_{t-k}) = \gamma(k)$ is independent of t for each k.

Weaker Rather than imposing conditions on all possible distributions, we impose conditions only on the first two moments of the series.

Implications

› Strong stationarity + $\mathbb{E}[X_t]^2 < \infty \implies$ weak stationarity.

› Weak stationarity $\not\Longrightarrow$ strong stationarity.

› Weak stationarity + Gaussian $\implies$ Strong stationarity.

Example (Random walk)
For a random walk $Y_t = Y_{t-1} + \varepsilon_t$, we have that $\mathbb{V}[Y_t] = t\sigma^2$ and the process is therefore non-stationary.

Since our objective is to find a model which is able to take into account the ***linear*** dependence between the observations, two very important functions are the autocorrelation and autocovariance functions.

Since for a stationary time series X_t we have $\mathrm{Cov}(X_t, X_{t-k}) = \gamma(k)$ for all k, we can therefore define the ACF as

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)}, \quad k = 0, \pm 1, \pm 2, \dots$$

from which we can see that γ and ρ are even functions, namely

$$\gamma(-k) = \gamma(k), \quad \rho(-k) = \rho(k).$$

Def. (Sample autocorrelation function)

We define the ***sample autocorrelation function*** (ACF) as

$$\widehat{\rho}(k) = \frac{\widehat{\gamma}(k)}{\widehat{\gamma}(0)},$$

where

$$\widehat{\gamma}(k) = \frac{1}{n} \sum_{t=1}^{n-|k|} (X_t - \overline{X})(X_{t+|k|} - \overline{X}).$$

Bias Even if this estimator is biased in finite samples, this is preferred to the unbiased estimator since when dividing by n we have a nonnegative-definite estimator.

In addition to autocorrelation, we also consider the correlation between X_t and X_{t+k} after controlling for the effect of the intermediate values $X_{t+1}, \dots, X_{t+k-1}$ using a linear regression (projection). We call this dependence the *partial autocorrelation* of X.

Def. (Partial autocorrelation)

The ***partial autocorrelation*** at lag k is the autocorrelation between z_t and z_{t+k} with the linear dependence of z_t on $z_{t+1}, \dots, z_{t+k-1}$ removed. Namely,

$$\begin{cases} \alpha(1) = \mathrm{Corr}(z_{t+1}, z_t) \\ \alpha(k) = \mathrm{Corr}\left(z_{t+k} - \pi_{t,k}(z_{t+k}), z_t - \pi_{t,k}(z_t)\right) & \text{if } k \geq 2 \end{cases}$$

where $\pi_{t,k}(x)$ is the orthogonal projection (regression) of x onto $z_{t+1}, \dots, z_{t+k-1}$.

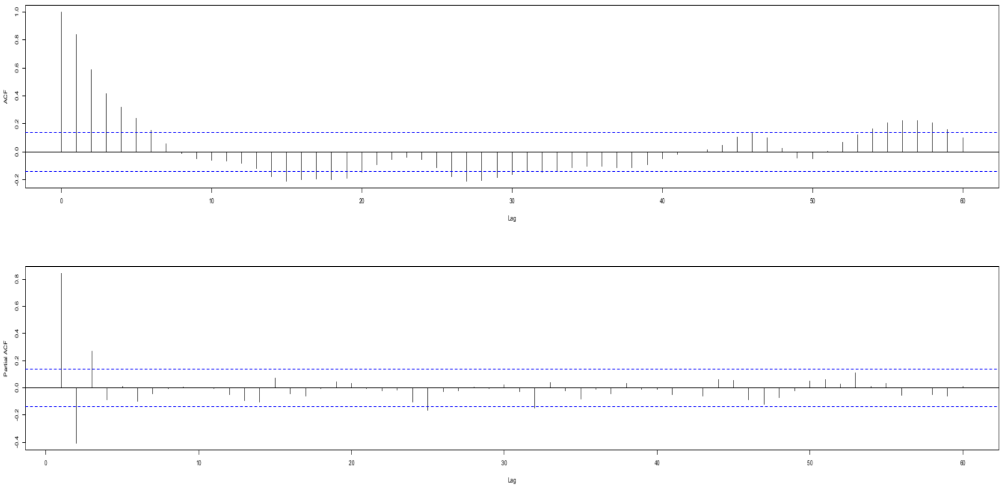

Figure 2: Autocorrelation (top) and partial autocorrelation (bottom) for a simulated time series.

The ACF and PACF are the main instruments that we use for choosing the most appropriate model for the DGP under the modern approach to time series (Box-Jenkins procedure).

Lecture 2: Stochastic processes in time-series analysis

2021-11-29

In this lecture we review some of the fundamental processes used in time-series analysis, starting from the simplest process (white noise) and then moving towards standard but more complicated construction (ARIMA).

2.1 White noise process

The white-noise process serves as the building block for defining more complex linear time series processes and reflects information that is not directly observable. In general, any sequence X_t of i.i.d random variables such that $\mathbb{E}[X_t] = 0$ and $\mathbb{V}[X_t] = \sigma^2 < \infty$ is a white noise process.

In general it's convenient to write a stochastic process as a sum of white noise terms,

$$X_t = \sum_{j=1}^{\infty} \psi_j \varepsilon_j,$$

since we can leverage standard proof techniques to prove theorems related to the process behaviour.

In the white-noise case, the probability behavior (law) of X is completely determined by all of its finite-dimensional distributions. When all of the finite-dimensional distributions are Gaussian, the process is called a Gaussian process.

Since uncorrelated normal random variables are also independent, a Gaussian white-noise process is, in fact, a sequence of i.i.d normal random variables.

2.2 Random walk

Whereas the white noise is a simple process with no memory, the random walk has infinite memory and is nonstationary,

$$X_t = \mu + X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma^2),$$

where $X_0 = 0$ by convention. The process is such that by recursion,

$$\begin{aligned} X_t &= \mu + (\mu + X_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \dots \\ &= \underbrace{t\mu}_{\text{drift}} + \underbrace{\sum_{i=1}^{t} \varepsilon_i}_{\text{stoch. trend}}. \end{aligned}$$

The last sum is called ***stochastic trend***, since every error ε enters with the same weight both from new and from past observations. Moreover, $\mathbb{E}[X_t] = t\mu$ and $\mathbb{V}[X_t] = t\sigma^2$. Applying the first-difference operator $(1 - B)$, where B is such that $BX_t = X_{t-1}$, to the process yields

$$(1 - B)X_t = X_t - X_{t-1} = \mu + \varepsilon_t,$$

which is a stationary model.

2.3 Linear time series

We introduce the ARMA model, the most famous type of linear time-series model which is used even for non-linear data. Forecasts from these models in these case have been empirically shown to be more accurate than forecasts from more complicated models.

A general linear process is of the form

$$X_t = \varepsilon_t + \sum_{i=1}^{\infty} \psi_i \varepsilon_{t-i},$$

where $\sum_{i=1}^{\infty} \psi_i^2 < \infty$. This type of process is such that

i. $\mathbb{E}[X_t] = 0$ for each t

ii. $\mathrm{Cov}(X_t, t_{t-k}) = \sigma_\varepsilon^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+k}$ for $k \geq 0$ and $\psi_0 = 1$.

An example is when the weights are an exponentially decaying sequence $\psi_j = \varphi^j$ with $|\varphi| < 1$, and in this case

$$X_t = \varepsilon_t + \varphi \varepsilon_{t-1} + \varphi^2 \varepsilon_{t-2} + \dots$$

The variance of this process can be written as a geometric series

$$\mathbb{V}[X_t] = \sigma_\varepsilon^2 \cdot \sum_{i=0}^{\infty} \varphi^k = \frac{\sigma_\varepsilon^2}{1 - \varphi^2},$$

moreover, the covariance and correlation functions are

$$\mathrm{Cov}(X_t, X_{t-k}) = \frac{\varphi^k \sigma_\varepsilon^2}{1 - \varphi^2}$$

$$\mathrm{Corr}(X_t, X_{t-k}) = \varphi^k$$

A moving average model of the form MA(q) is the above model truncated to the first q components:

$$X_t = \vartheta_1 \varepsilon_{t-1} + \dots + \vartheta_q \varepsilon_{t-q} + \varepsilon_t.$$

These models are easily tractable since they are stationary by definition and estimation is very simple in the Gaussian case.

On the other hand, an autoregressive model AR is such that

$$X_t = c + \varphi_1 X_1 + \dots + \varphi_p X_{t-p} + \varepsilon_t,$$

where X_t's could also be random variables each uncorrelated with the next value X_{t+1}. The expected value of the process can be calculated in terms of the autoregressive coefficients, by assuming the process to be stationary

$$\mathbb{E}[X_t] = \mathbb{E}[c + \varphi_1 X_{t-1} + \dots + \varphi_p X_{t-p} + \varepsilon_t] \implies \mathbb{E}[X_t] = \frac{c}{1 - \sum_{i=1}^{p} X_{t-i}}.$$

Again, by assuming the process to be stationary we observe an autocovariance of the form

$$\gamma_k = \begin{cases} \varphi_1\gamma_1 + \varphi_2\gamma_2 + \ldots + \varphi\gamma_p + \sigma_\varepsilon^2 & k = 0 \\ \varphi_1\gamma_1 + \varphi_2\gamma_2 + \ldots + \varphi\gamma_p & k > 0 \end{cases}$$

$$\rho_k = \varphi_1\rho_{k-1} + \varphi_2\rho_{k-2} + \ldots + \varphi_p\rho_{k-p}, \quad k > 0,$$

which yield the Yule-Walker equations when considering them for $k = 1, \ldots, p$. These equations can be used to compute the model coefficients when solving them in terms of the unknown $\boldsymbol{\varphi}$ and sample autocorrelation $\widehat{\rho}_k$.

For an AR model we have that the sample autocorrelation is exponentially decaying in k, depending on the model parameters, and its partial autocorrelation function is null for $k > p$.

2.4 ARMA model

We introduce the combined ARMA model in order to model more complicated dynamics of time series, yielding the ARMA(p, q) defined as

$$X_t = \varphi_1 X_{t-1} + \ldots + \varphi_p X_{t-p} + \vartheta_1\varepsilon_{t-1} + \ldots + \vartheta_q\varepsilon_{t-q} + \varepsilon_t, \tag{1}$$

which usually allows us to model more complicated correlation structures using a smaller number of parameters.

Using a backshift operator $B^k X_t = X_{t-k}$ we can write this model as

$$\varphi(B)X_t = \vartheta(B)\varepsilon_t,$$

where the polynomials in B are defined as

$$\varphi(B) = 1 - \varphi_1 B - \ldots - \varphi_p B^p$$

$$\vartheta(B) = 1 + \vartheta B + \ldots + \vartheta_q B^q$$

and are extremely important since we can determine the properties of an ARMA model in terms of $\vartheta(\cdot)$ and $\varphi(\cdot)$. Moreover, if the ARMA model (1) is stationary, we can find an AR(∞) representation for it by solving the equality

$$\vartheta(B)^{-1}\varphi(B)X_t = \varepsilon_t,$$

and a $MA(\infty)$ representation by solving

$$X_t = \varphi(B)^{-1}\vartheta(B)\varepsilon_t.$$

Example (AR(1))

Consider the AR(1) model, then

$$\begin{aligned} Y_t &= \varphi Y_{t-1} + \varepsilon_t \\ &= \varphi(\varphi Y_{t-2} + \varepsilon_{t+1}) + \varepsilon_t \\ &= \dots \\ &= \varepsilon_t + \varphi\varepsilon_{t-1} + \varphi^2\varepsilon_{t-2} + \dots \end{aligned}$$

Example (General procedure)
We write the relationship

$$(1-\varphi B)y_t = (1-\vartheta B)\varepsilon_t,$$

for which we can write

$$y_t = \frac{1-\vartheta B}{1-\varphi B}\varepsilon_t.$$

Our goal now is to obtain a relationship of the form

$$Y_t = \Psi(B)\varepsilon_t = \sum_{i=0}^{\infty}\psi_i B^i,$$

and therefore $\Psi(B) = \frac{1-\vartheta B}{1-\varphi B}$, from which

$$(1-\varphi)\Psi(B) = 1-\vartheta B$$

$$\Updownarrow$$

$$(1-\varphi B)(1+\psi_2 B+\psi_2 B^2+\dots) = 1-\vartheta B$$

$$\Updownarrow$$

Example (General AR(∞))
The procedure is the same, except we now have to find a relationship of the form

$$\varphi(B)Y_t = \vartheta(B)\varepsilon_t \longrightarrow$$

$$\frac{\varphi(B)}{\vartheta(B)} = \Xi(B),$$

and find the parameters in terms of φ and ϑ.

In order to check for invertibility of the process, we need to invert the MA operator which is doable if the characteristic equation $\vartheta(B) = 0$ has solutions $|B_i| > 1$.
In order to check for of the process, we need to invert the AR operator which is doable if the characteristic equation $\vartheta(B) = 0$ has solutions $|B_i| > 1$.

Time series models only work if the data is stationary, therefore in general it's recommended to check for the evidence of trend or seasonality before applying an ARMA model. We can remove nonstationarity either via regression or via simple differentiation.

Even though the model might not be invertible, it's still better to have it stationary and not invertible. In general, it's advised to differentiate the series rather than risking for the time series to be nonstationary.

Testing whether the trend is deterministic or stochastic can be performed via a unit root test, which is usually not very powerful.
We obtain the ARIMA(p, d, q) class of models by applying a d-order to X_t and modeling the result as an ARMA(p, q) model, i.e.

$$\varphi(B)(1-B)^d Y_t = \vartheta(B)\varepsilon_t.$$

› In general, this process is simply an ARMA$(p+d, q)$ with d unit roots in the autoregressive polynomial.

› In general, we don't see time series such that $d > 2$.

Lecture 3: Transfer function models

2021-12-06

Theorem 1 (Wold decomposition)
Let $(X_t)_t$ be a non-deterministic stationary time series with $\mathbb{E}[X_t] = 0$, then

$$X_t = \sum_{j=0}^{\infty} \psi_j a_{t-j} + V_t,$$

where V_t is deterministic and

1. $\psi_0 = 1$ *and* $\sum_{j=1}^{\infty} \psi_j^2 < \infty$.
2. $a_t = WN(0, \sigma^2)$.
3. $\mathbb{E}[a_t a_s] = 0$ *for all* $s, t = 0, \pm 1, \pm 2, \ldots$

With this decomposition we can approximate any stationary time series using a linear process of the form

$$X_t = \mu + \sum_{j=0}^{\infty} \psi_j a_{t-j},$$

where $\sum_{j=1}^{\infty} |\psi_j| < \infty$.

3.1 Transfer function models

Transfer function models are models where an output series y_t is related to one or more input series x_t. We link the two series by the following ***transfer function model*** (TFM)

$$y_i = \nu(B) x_t + \eta_t, \tag{2}$$

where the relationship is linear, $\nu(B) = \sum_{j=-\infty}^{\infty} \nu_j B^j$, and is called the *transfer function of the linear filter* that transform x_t into y_t. η_t is a noise series independent of x_t. The weights ν_j are called ***impulse response weights*** and the TFM is called ***stable*** if

$$\sum_{j=-\infty}^{\infty} |\nu_j| < \infty,$$

and in particular this yields a BIBO (Bounded Input Bounded Output) relationship. The TFM is said to be ***causal*** if $\nu_j = 0$ for $j < 0$, since the present output is affected only by the system current and past values,

$$y_t = \sum_{j=0}^{\infty} \nu_j B^j.$$

The purpose of TF models is to identify the TF $\nu(B)$ and the noise model, possibly using a simpler representation which is similar to an ARIMA model

$$\delta(B) y_t = \omega(B) B^b x_t,$$

where

$$\delta(B) = 1 - \delta_1 B - \delta_2 B^2 - \ldots \delta_r B^r$$

$$\omega(B) = \omega_0 - \omega_1 B - \ldots - \omega_s B^s$$

and b is a delay parameter that tells us the lag that elapses before the impulse of the input variable produces an effect on the output variable.

With the above representation, we can rearrange the terms so that y_t has an explicit representation in terms of x_t,

$$y_t = \frac{\omega(B)}{\delta(B)} x_{t-b} + \eta_t, \tag{3}$$

and by equating Equation (3) to Equation (2) we can write the transfer function $\nu(B)$ as

$$\nu(B) = \frac{\omega(B)B^b}{\delta(B)}. \tag{4}$$

and the orders s, r, b of the model in Equation (3) can be found by equating the coefficients of B^j to both sides in Equation (4)

$$\delta(B)\nu(B) = \omega(B)B^b,$$

which yields the following equation

$$(1 - \delta_1 B - \delta_2 B^2 - \ldots - \delta_r B^r)(\nu_0 + \nu_1 B + \ldots) = (\omega - \omega_1 B - \ldots - \omega_s B^s)B^b,$$

and we obtain the following relationships between the components of the model

$$\begin{aligned}
&\nu_j = 0 && \text{if } j < b \\
&\nu_j = \delta_1 \nu_{j-1} + \delta_2 \nu_{j-2} + \ldots + \delta_r \nu_{j-r} + \omega_0 && \text{if } j = b \\
&\nu_j = \delta_1 \nu_{j-1} + \delta_2 \nu_{j-2} + \ldots + \delta_r \nu_{j-r} - \omega_{j-b} && \text{if } j = b+1, \ldots, b+s \\
&\nu_j = \delta_1 \nu_{j-1} + \delta_2 \nu_{j-2} + \ldots + \delta_r \nu_{j-r} && \text{if } j > b+s
\end{aligned}$$

By observing the behaviour of the cross-correlation function between x_t and y_t – similarly to what we do with ACF and PACF for estimating p, d, q in an ARIMA model) – we can find the appropriate values of s, r, b.

Def. (Cross-correlation function)

We say that X_t and Y_t are ***jointly stationary*** if they are univariate stationary and $\mathrm{Cov}(X_t, Y_s) = f(|s-t|)$, and in this case we define the ***cross-correlation function*** between X_t and Y_t as the function

$$\gamma_{XY}(k) = \mathbb{E}[(X_t - \mu_X)(Y_{t+k} - \mu_Y)],$$

Marginals By definition we have that $\rho_{XX}(k) = \rho_X(k)$.

Symmetry It's relevant the order in which we compute the cross-correlation function, since unlike the ACF the CCF is not symmetric around the origin,

$$\rho_{XY}(k) \neq \rho_{XY}(-k),$$

instead we have that

$$\rho_{XY}(k) \neq \rho_{YX}(-k).$$

However, we have a way of obtaining the direction of association between the time series by inspecting the graph of the ACF. The direction depends on the software implementation of the function.

Example (AR(1) model)
Let $Y_t \sim \mathrm{AR}(1)$, then we have $(1-\varphi B)Y_t = X_t$ and for time $t+k$ we can write

$$Y_{t+k} = \frac{1}{1-\varphi B} X_{t+k} = X_{t+k} + \varphi X_{t+k-1} + \varphi^2 X_{t+k-2} + \dots,$$

therefore the cross-covariance function between X_t and Y_t are

$$\gamma_{XY}(k) = \mathbb{E}[X_t Y_{t+k}] = \begin{cases} \varphi^k \sigma_k^2 & \text{if } k \geq 0 \\ 0 & \text{if } k \leq 0 \end{cases}$$

ARMA model In general, the ARMA(p,q) model can be written as a transfer function model without the white noise term η_t, and where X_t is a white noise itself uncorrelated with Y_t.

3.2 Cross-correlation function and TF models

Let x_t and y_t be stationary series with $\mu_x = \mu_y = 0$, then the transfer function at time $t+k$ is

$$y_{t+k} = \nu_0 x_{t+k} + \nu_1 x_{t+k-1} + \nu_2 x_{t+k-2} + \dots + \eta_{t+k},$$

therefore if we multiply both left and right by x_t and take expectations we have

$$\gamma_{xy}(k) = \nu_0 \gamma_x(k) + \nu_1 \gamma_x(k-1) + \nu_2 \gamma_x(k-2) + \dots,$$

hence the CCF in the doubly stationary case has the following simple representation:

$$\rho_{xy}(k) = \frac{\sigma_x}{\sigma_y} \left[\nu_0 \rho_x(k) + \nu_1 \rho_x(k-1) + \nu_2 \rho_x(k-2) + \dots \right]. \tag{5}$$

Therefore, by Equation (5) we observe that the relationship between the CCF and IRF ν_j is contaminated by the fact that they are not white noise, and therefore display the correlations at previous times. However, for a ***white noise model*** x_t we would see $\rho_x(k) = 0$ for all $k \neq 0$ and therefore we would have a direct way of estimating ν_k by letting

$$\gamma_{xy}(k) = \nu_k \sigma_k^2,$$

hence we can estimate the covariance function nd obtain an impulse response function which is directly proportional to the CCF,

$$\rho_{xy}(k) = \frac{\sigma_x}{\sigma_y}\nu_k \implies \nu_k = \frac{\sigma_y}{\sigma_x}\rho_{xy}(k). \tag{6}$$

Idea Therefore, our goal for estimating a TF model is to reduce the problem to a whitened series for x_t, and then apply the estimation procedure above.

In the general TF model given by

$$y_t = \nu(B)x_t + \eta_t,$$

if we assume $x_t \sim \mathrm{ARMA}(p,q)$ we can calculate the ***pre-whitened input series***

$$\alpha_t = \frac{\varphi_x(B)}{\vartheta_x(B)}x_t,$$

and applying this transformation to both y_t and η_t we can obtain the ***filtered series***

$$\begin{cases} \beta_t = \frac{\varphi_x(B)}{\vartheta_x(B)}y_t \\ \varepsilon_t = \frac{\varphi_x(B)}{\vartheta_x(B)}\eta_t \end{cases}$$

Finally, the TF model becomes

$$\beta_t = \nu(B)\alpha_t + \varepsilon_t,$$

where the input series is $\alpha_t \sim \mathrm{WN}(0,\sigma^2)$ and we can estimate the transfer function using Equation (6) between β_t and α_t.

3.2.1 General procedure for the identification of a TF model

1. Identify an $\mathrm{ARMA}(p,q)$ model for the input x_t,

$$\varphi_x(B)x_t = \vartheta_x(B)\alpha_t$$

2. Prewhiten $x_t \to \alpha_t = \frac{\varphi_x(B)}{\vartheta_x(B)}x_t$ and apply the same filter to $y_t \to \beta_t = \frac{\varphi_x(B)}{\vartheta_x(B)}y_t$.

3. Calculate the CCF between the whitened input series and the residuals of the model for y_t,

$$\widehat{\nu}_k = \frac{\widehat{\sigma}_\beta}{\widehat{\sigma}_\alpha}\ldots,$$

to get a preliminary estimation of the transfer function ν_k.

4. Identify the order b, r, s of the TF model by inspecting the estimated TF (or equivalently the CCF) and estimate the transfer function using the fact that

$$\widehat{\nu}_j = \frac{\widehat{\omega}(B)}{\widehat{\delta}(B)}B^b,$$

which is of course done by nonlinear least squares or other methods.

5. Identify a model for the estimated residuals $\widehat{\eta}_t$ given by

$$\widehat{\eta}_t = y_t - \widehat{\nu}(B)x_t.$$

6. Estimate the model and check goodness-of-fit, generally by checking both $\widehat{\varepsilon}_t$ and $\widehat{\alpha}_t$ are white noise. Moreover, since we assume that $\varepsilon_t \sim \mathrm{WN}$ and $\eta_t \perp\!\!\!\perp x_t$. we need to check that $\widehat{\rho}_{\alpha,\widehat{\varepsilon}}(k)$ is non significant.

For checking the last step, there are test statistics which are based on Portmanteau tests.

Lecture 4: Spectral analysis

2021-12-13

In this lecture we introduce spectral analysis, which transforms the data from the time domain to the frequency domain by decomposing the time series into a Fourier basis of coefficients. The idea is to decompose X_t in terms of combination of sinusoids with random and uncorrelated coefficients.

4.1 Periodicity

Consider a periodic process of the form

$$X_t = C \cdot \cos(2\pi\omega t + \varphi), \quad t = \pm 1, \pm 2, \ldots,$$

where ω is a *frequency* index, C the *amplitude* and φ the *phase* of the process. We can introduce random variation in X_t by allowing the amplitude and phase to vary, since by the usual sine and cosine rules we can write X_t as

$$X_t = A\cos(2\pi\omega t) + B\sin(2\pi\omega t). \tag{7}$$

In the above equation, $A = C\cos\varphi$ and $B = -C\sin\varphi$, and $A, B \sim \mathcal{N}(0, \sigma^2)$. We have that

1. $C = \sqrt{A^2 + B^2}$
2. $\varphi = \tan^{-1}(-B/A)$
3. X_t is a stationary process with $\mu_t = 0$ and

$$\gamma(h) = \text{Cov}(X_t, X_{t+h}) = \sigma^2 \cos(2\pi\omega h).$$

We consider a generalization of (7) given by a mixture of periodic series with multiple frequencies and amplitudes,

$$X_t = A_0 + \sum_{i=1}^{q} \left\{A_i \cos(2\pi\omega_i t) + B_i \sin(2\pi\omega_i t)\right\}, \tag{8}$$

where $A_i, B_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ and the ω_i are distinct frequencies. In this case,

$$\gamma(0) = \sum_{i=1}^{q} \sigma_i^2, \quad \gamma(h) = \sum_{i=1}^{q} \sigma_i^2 \cos(2\pi\omega_i t).$$

The main objective of spectral-based time-series analysis is to sort out the essential frequency components ω_i of a time series, including their relative contribution to the total power of the signal.

For a sample $x_1, \ldots, x_n$ from X_t we can write the following representation

$$X_t = A_0 + \sum_{j=1}^{\frac{n-1}{2}} \left\{A_j \cos(2\pi tj/n) + B_j \sin(2\pi tj/n)\right\}, \tag{9}$$

for $t = 1, 2, \ldots, n$ and suitably chosen coefficients. If n is even we can modify the above equation and an additional component. Equation (9) holds for any sample and can be interpreted as an approximation to (8) with some coefficients possibly close to zero.

Our problem is now to estimate the A_j's and B_j's using a linear model given the frequencies which are relevant to the observed model. We do so by plotting the ***periodogram***, i.e. the estimates of the variance explained by the j^{th} component $P(j/n) = \frac{1}{2}(\widehat{A}_j^2 + \widehat{B}_j^2)$.

By inspecting the periodogram we can observe which frequencies $\omega_j = j/n$ are predominant over the others and eventually observe frequencies which are "hidden" inside the time series.

Theorem 2 (Parseval's theorem)
The sample variance is the sum of the contribution of the observed periodogram

$$\frac{1}{n}\sum_{t=1}^{n}(X_t - \overline{X})^2 = \frac{1}{2}\sum_{j=1}^{\frac{n-1}{2}}(A_j^2 + B_j^2) = \sum_{j=1}^{\frac{n-1}{2}} P(j/n)$$

Lecture 5: Spectral analysis (cont.)

2021-12-20

If a stationary process X_t has auto covariance function which is absolutely summable,

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty,$$

then it has the representation in terms of its Fourier transform given by

$$\gamma(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} f(\omega) e^{2\pi i \omega h} f(\omega) \,\mathrm{d}\omega, \quad h = 0, \pm 1, \pm 2, \dots,$$

and we can define the spectral density of the process as

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h}, \quad -\frac{1}{2} \leq \omega \leq \frac{1}{2}.$$

Properties of f

1. $f(\omega) \geq 0$ for all ω
2. $f(\omega) = f(-\omega)$, therefore we only consider $\omega > 0$.
3. $\gamma(0) = \mathbb{V}[X_t] = \int_{-1/2}^{1/2} f(\omega) \,\mathrm{d}\omega$, which is the total variance of the process.

Since the ARMA processes satisfy absolute summability, we can represent them in terms of their spectral densities.

Example (White noise)
White noise is such that $X_t \sim \mathrm{WN}(0, \sigma^2)$ has a constant frequency spectrum, since no frequency dominates over any other.

Example (AR(1))
We can write

$$f(\omega) = \sum_{-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h} = \frac{\sigma_\varepsilon^2}{1 - 2\varphi \cos(2\pi\omega) + \varphi^2},$$

and if $\varphi > 0$ the spectrum is dominated by low frequencies, whereas if $\varphi < 0$ the dominating frequencies are high.

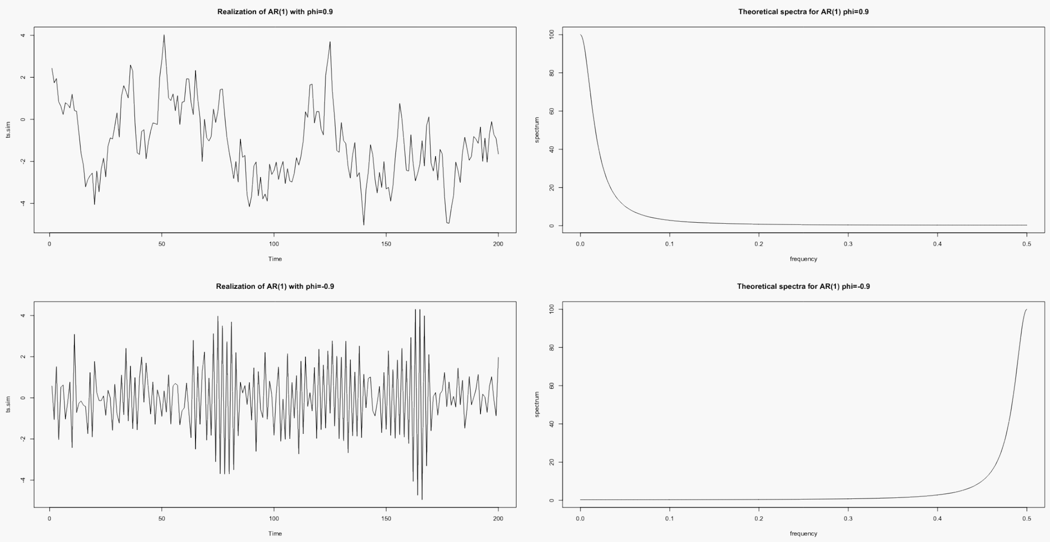

Figure 3: Frequency spectrum for AR(1) models with $\varphi > 0$ (above) and $\varphi < 0$ (below).

Example (MA(1))

The same behaviour can be seen for a moving average model, which is however less visible than the AR process. Indeed, the MA component is useful to model the component of the process which has not been explained by the autoregressive part.

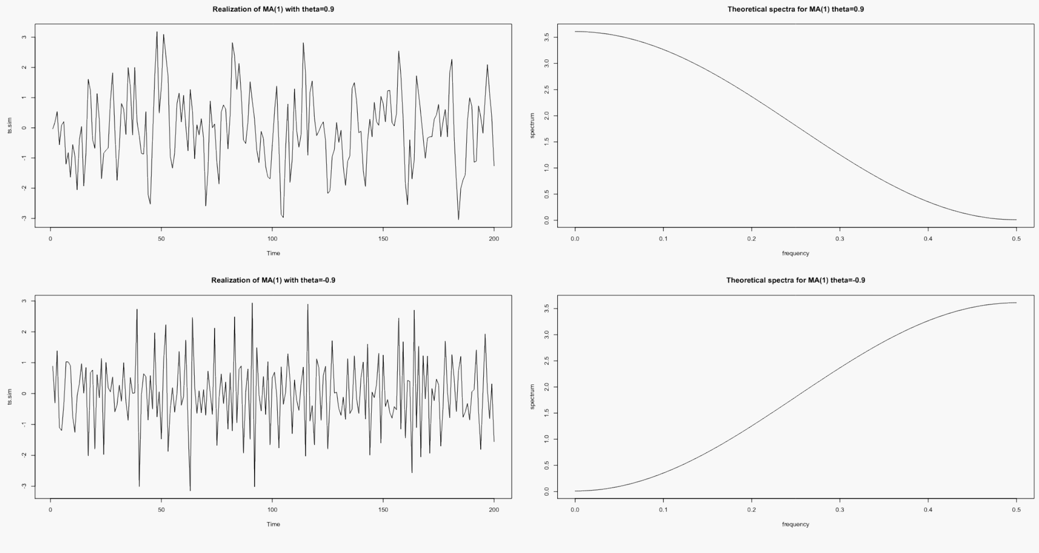

Figure 4: Frequency spectrum for a MA(1) model with $\vartheta > 0$ (above) and $\vartheta < 0$ (below).

Example ($\texttt{ARMA}(p, q)$)

For a general $X_t \sim \text{ARMA}(p,q)$ we can prove that

$$f(\omega) = \sigma_\varepsilon^2 \frac{|\vartheta(e^{-2\pi i\omega})|^2}{|\varphi(e^{-2\pi i\omega})|^2} \tag{10}$$

5.1 Estimation of the spectral density

Estimating the spectral density can be done similarly to what we do for the histogram of a continuous density. Let X_t be a zero-mean stochastic process, we define

$$\widehat{f}(\omega) = \widehat{\gamma}_0 + 2\sum_{k=1}^{n-1} \widehat{\gamma}(k)\cos(2\pi\omega k),$$

which is a way in which we can write the spectral density of X_t. The periodogram as we defined above is an inconsistent estimate of the true spectral density, which has poor sample properties.

For a periodogram we can see that

$$\frac{\widehat{A}_\omega^2 + \widehat{B}_\omega^2}{\gamma_0} = \frac{2\widehat{f}(\omega)}{f(\omega)} \sim \chi_2^2,$$

hence $\mathbb{E}[\widehat{f}(\omega)] = f(\omega)$ and $\mathbb{V}[\widehat{f}(\omega)] = f^2(\omega)$. Hence, the variance of the estimator $\not\to 0$ as $n \to \infty$, which means that the estimation strategy yields a bad result.

Alternative methods for constructing estimators include the following approaches:

1. *Nonparametric estimators*: using a moving average to smooth the estimate,

$$\bar{f}(\omega) = \sum_{k=-m}^{m} w_m(k)\widehat{f}(\omega_j + \frac{k}{n}),$$

where $w_m(k) \geq 0$, $w_m(k) = w_m(-k)$ and $\sum_{k=-m}^{m} w_m(k) = 1$. Using the smoothed spectrum we have

$$\frac{2(2m+1)\bar{f}(\omega)}{f(\omega)} \sim \chi^2_{2(2m+1)},$$

consequently $\mathbb{E}[\bar{f}(\omega)] \approx f(\omega)$ and $\mathbb{V}[\bar{f}(\omega)] \approx f^2(\omega)/(2m+1)$ and we have consistency if $m \to \infty$ while $m/n \to 0$. We usually choose something like $m = \sqrt{n}$ to get some insights into the shape of the true spectrum.

2. *Parametric estimators* based on a fitted AR model. Since any ARMA time series admits an AR(∞) representation, we can use the theoretical estimate of such a model estimated using AIC, AICC, BIC, ...

   The disadvantage in this case can be the fact that p can be very large. Instead, we can use an ARMA model and use the theoretical spectral density in Equation (10) to represent the estimated $\widehat{f}$.

Lecture 6: Nonlinear time series models

2022-01-11

A lot of research in the last 30 years has been spent to develop time-series models which can detect nonlinear patterns in the data. We refer again to Wold's theorem (**??**) and consider a time series of the form

$$X_t = \sum_{j=0}^{\infty} \psi_j a_{t-j} + V_t,$$

where $\psi_0 = 1$ and $\sum_{j=1}^{\infty} \psi_j^2 < \infty$, $a_t \sim \mathrm{WN}(0, \sigma^2)$ and a_j's are uncorrelated.

Def. (General linear process)
X_t is said to be ***linear*** if if can be written as

$$X_t = \mu + \sum_{j=0}^{\infty} \psi_j a_{t-j},$$

such that $\sum |\psi_j| < \infty$ and $a_j \overset{\text{iid}}{\sim} \mathrm{WN}(0, \sigma^2)$. The above process is sometimes called ***general linear process***.

Limitations Linear models are limited in the sense that they cannot model ***strong asymmetries*** in data, ***irregular jumps***, and ***switching regimes***.

6.1 Nonlinear framework

Whereas linearity is well-defined, non-linearity is hardly definite and the early development of nonlinear time series focused on various parametric forms: ARCH, GARCH, threshold models, ...

In this lectures we emphasize simple parametric models which are applicable without overly-complex specifications. There are examples of *explicit* and *implicit* approaches, which differ in the way they are represented:

› *Implicit*: ARMA model with non-gaussian innovations.

› *Explicit*: $X_t = h(X_{t-1}, \dots, X_{t-k})$.

Explicit models have surpassed implicit modelling since it is in general difficult to identify the correct distribution of the white-noise terms.

Attention In the nonlinear setting, the tools of standard analysis (ACF, PCF, ...) are not helpful since they only detect linear patterns.

Def. (Nonlinear model)
The genreal representation is

$$X_t = f(a_t, a_{t-1}, a_{t-2}),$$

where $a_t \overset{\text{iid}}{\sim} \mathrm{WN}(0, \sigma^2)$ and $f(\cdot)$ is some nonlinear function.

Remark We could linearize the above model by considering the Taylor series around zero (Volterra series)

$$X_t = \mu + \sum_{i,j} b_{ij} a_{i,t-j} + \sum_{ijkl} b_{ij,kl} a_{i,t-j} a_{k,t-l} + \dots \tag{11}$$

The Volterra series (11) is usually too complicated unless severely truncated.

Other very general models are ARMA models with time-dependent parameters.

The model for nonlinear time series can be written in terms of the conditional mean and variance,

$$\mu_t = \mathbb{E}[X_t|\mathcal{F}_{t-1}] = g(F_{t-1})$$

$$\sigma_t^2 = \mathbb{V}[X_t|\mathcal{F}_{t-1}] = h(F_{t-1})$$

where g and h are well-defined nonlinear function with $h(\cdot) > 0$. If $g(\cdot)$ is nonlinear and $h(\cdot)$ is constant, then X_t is nonlinear in mean, for example

$$X_t = \varepsilon_t + \alpha\varepsilon_{t-1}^2.$$

Otherwise, if $h(\cdot)$ is time-variant then X_t is nonlinear in variance. All GARCH models are of this type.

6.2 NLAR(1)

We consider the simplest conditional mean model given by

$$X_t = g(X_{t-1}, \vartheta) + a_t,$$

where ϑ is a vector of parameters and $a_t \sim$ IID. It's natural to consider functions which are nearly linear as first candidates, but also more extreme nonlinear functions if needed.

There are few papers on nonlinear autoregressive processes, since the fact that there are too many nonlinear functions that we can consider renders this class of processes unusable for statistical analysis.

6.3 Reversibility

Def. (Reversibility)
The stationary sequence X_t is ***time-reversible*** if the finite-dimensional distributions of $(X_1, X_2, \dots, X_n)$ and $(X_n, \dots, X_2, X_1)$ is the same for any n.

Usage Since i.i.d sequences and ARMA models are time-reversible, we can use time-reversibility to detect deviations from the Gaussianity-linearity hypothesis. For example, Chen et al (2000) look at the test statistic

$$\mathbb{E}[\sin(\omega(X_t - X_{t-k}))],$$

and if the value is zero then the process is time-reversible. Other tests rely for example on differences between backward and forward autocorrelation.

6.4 Threshold AR models

Consider the change-point model

$$X_t = \begin{cases} \varphi_1 X_{t-1} + a_t & \text{if } X_{t-1} < r \\ \varphi_2 X_{t-1} + a_t & \text{if } X_{t-1} \geq r \end{cases}$$

we call this model the *self-exciting* threshold-AR model. If X_{t-1} is replaced by an exogenous variable Z_{t-d}, then this model is called *threshold-autoregressive* (TAR). Using piecewise linear models we can obtain a better approximation of the conditional mean equation.

Def. (SETAR model)
We define the ***self-exciting TAR model*** (***SETAR***) with threshold X_{t-d} if

$$X_t = \varphi_0^{(j)} + \varphi_1^{(j)} X_{t-1} + \ldots + \varphi_p^{(j)} X_{t-p} + a_t^{(j)} \quad \text{if } \gamma_{j-1} \leq X_{t-d} \leq \gamma_j,$$

where $j = 1, \ldots, k$.

Remark In the above model, γ_j's are the thresholds and X_{t-d} is the threshold variable.

6.4.1 Estimation

Suppose that we have an observed time series $X_1, \ldots, X_n$ and we fix the order k of the SETAR model. We alternate a two-step procedure:

- First we assume that the partition A_j and orders p_j are known, so that we can use the least squares procedure to minimize the loss function

$$\mathcal{S}(\vartheta) = \sum_{j=1}^{k} S^{(j)} = \sum_{\substack{X_{t-d \in A_j} \\ p<t\leq n}} \Big[X_t - (\varphi_0^{(j)} + \varphi_1^{(j)} X_{t-1} + \ldots + \varphi_p^{(j)} X_{t-p}) \Big]^2$$

- We find the partition $\widehat{A}_j$ such that $\mathcal{S}(\widehat{\vartheta})$ is minimized.

Otherwise, to determine the autoregressive orders $p^{(j)}$'s we might use an information criteria such as BIC, AICC, ...

Testing for linearity becomes essential to check non-linearity when fitting nonlinear models. In general, there are

a) Tests for departure from linear models towards general nonlinear models (less powerful).

b) Tests for departure from linear models towards threshold autoregressive (more powerful).

6.4.2 Smooth Transition Autoregressive models

A generalization of the TAR model is the ***smooth transition autoregressive model*** (STAR) which allows for a smoother transition between the two regimes via a cumulative distribution function instead of a jump function.

Lecture 7: Markov-switching model

2022-01-13

A Markov-switching (MS) model changes the rules of the switching regimes from the threshold-autoregressive model, which are not deterministic anymore. A MS(p) model with two regimes can be defined as

$$X_t = \begin{cases} \alpha_1 + \sum_{i=1}^p \varphi_{1,i} X_{t-i} + a_{1,t} & \text{if } s_t = 1 \\ \alpha_2 + \sum_{i=1}^p \varphi_{2,i} X_{t-i} + a_{2,t} & \text{if } s_t = 2 \end{cases} \tag{12}$$

where $a_{i,t} \sim \text{IID}(0, \sigma_i^2)$. The state variable s_t is unobservable and we assume that it follows a first-order Markov chain with transition probabilities

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix},$$

and a particular choice of P drives the model behaviour. This is crucially different from the SETAR model, since the regimes are defined by the Markov chain and not simply determined by the past values of X_t. Thus, *forecast* of a Markov-switching model are linear combinations of forecasts produced by the sub-models.

Estimating a MS model is much harder since the states are not observable, therefore we need a ***filtering*** approach. The log-likelihood can be constructing recursively from some initial conditions since

$$f_{it} = f(X_t | s_t = i, X_{t-1}, \vartheta_i), \quad i = 1, 2,$$

and under normality this is a Gaussian density with parameters from Equation (12). Now, we can write the contribution by marginalizing s_t as

$$g(X_t | X_{t-1}, \vartheta_1, \vartheta_2) = f_{1t}\rho_{1t|t-1} + f_{2t}\rho_{2t|t-1}, \tag{13}$$

and Hamilton have shown that the optimal inference and forecast can be determined from the conditional likelihood in (13).

7.1 Bilinear models

This class of models is not very used/useful even though they are a natural extension of the ARMA model. A general bilinear model BL(p, q, r, s) can be written as

$$X_t = c + \sum_{i=1}^p \varphi_i X_{t-i} + a_t + \sum_{i=1}^q \vartheta_j a_{t-j} + \underbrace{\sum_{i=1}^r \sum_{j=1}^s \beta_{ij} X_{t-i} a_{t-j}}_{\text{bilinear component}},$$

where $a_t \sim \text{IID}(0, \sigma_a^2)$. This model is too complex and therefore people usually study the Lower Triangular Bilinear Model

$$X_t = c + \sum_{i=1}^p \varphi_i X_{t-i} + a_t + \sum_{i=1}^q \vartheta_j a_{t-j} + \underbrace{\sum_{i=1}^r \sum_{j=1}^s \beta_{ij} X_{t-i-j} a_{t-i}}_{\text{bilinear component}},$$

where X has only past values w.r. to a in the bilinear component.

- These models can model occasional outbursts in time series.
- BL model can have conditional heteroscedasticity, although GARCH-type models are better in this regard.
- ML procedures are used but asymptotic distribution is unknown.
- Probabilistic properties are often derived using the state-space representation.

7.2 Long-memory models

ARMA models are said to be short-term models, since their autocorrelation function usually tends to zero with an exponential decrease. Thus the $d \in \{0, 1\}$ parameter in an ARIMA(p, d, q) model controls the transition from short-memory to infinite memory.

Granger introduced an extension to the ARIMA model by considering $d \in [0, 1]$ yielding ARFIMA models. These models are necessary if we want to take into account series that show memory between $I(0)$ and $I(1)$ processes.

Properties ARFIMA models can take into account

- Presence of long-range cycles
- Slowly-decaying autocorrelation structures.

There are different definitions of long-memory processes:

1. In the time domain, a long-memory process is such that its autocorrelation function decays like a power function, i.e. if $\alpha \in (0, 1)$ and $c_\rho > 0$
$$\rho(k) = c_\rho k^{-\alpha}, \quad k \longrightarrow \infty.$$
2. In the frequency domain, a long-memory process is such that its spectral density is unbounded at zero, i.e.
$$f(\omega) \sim c_f \omega^{-\alpha}, \quad \omega \longrightarrow 0^+.$$

We can define the fractional difference operator using the Gamma function as

$$(1 - B)^d = \sum_{j=0}^{\infty} \frac{\Gamma(j - d)}{\Gamma(j + 1)\Gamma(-d)} B^j,$$

and when $d \in (0, 0.5)$ the ARFIMA(p, d, q) process is stationary with $\rho(k) \sim k^{2d-1}$. When $d \in (-0.5, 0)$ the process is stationary with intermediate memory, although in practice this is never used. For $d \in [0.5, 1)$ the process is mean-reverting even though it is not covariance-stationary.

In the following we will focus on ARFIMA(p, d, q) processes with $d \in (0, 0.5)$ which yields the most interesting type of process.

7.2.1 Estimation

Estimation approaches are mainly divided into two broad classes:

1. ML estimation, which requires specifying both p and q.
2. Semi-parametric or nonparametric approaches, where we assume that the ARMA component is relatively unimportant.

One of the best-known method is the semi-parametric GPH estimator introduced by Geweke and Porter-Hudack and developed by Robinson. This method approximates the spectral density near the origin,

$$f(\omega) \sim c_f \big(4\sin^2(\omega/2)\big)^{-d},$$

and therefore we can apply the least-squares method to

$$\log l(\omega_j) = \log c_f - d \log \big(4\sin^2(\omega_j/2)\big) + u_j,$$

where u_j are i.i.d error terms and ω_j are the Fourier frequencies. The problem with this method is its high variance in the estimates.

7.3 Integer autoregressive (INAR) models

Integer autoregressive (INAR) models can be used to model time series of counts, which are of particular interest in practice. In some cases, the discrete values can be approximated by Gaussian models, however for small values we need a more proper model.

Notation We introduce the ***thinning operator*** $\circ$ which substitutes the multiplication operator. Let $\alpha \in [0,1]$, then we define

$$\alpha \circ X = \sum_{i=1}^{X} Y_i, \quad Y_i \text{ i.i.d r.v.'s with } \mathbb{E}[Y_i] = \alpha. \tag{14}$$

Typically, Y_i's are assumed to be i.i.d Ber(α), and therefore we have

$$Y_i \overset{\text{iid}}{\sim} \text{Ber}(\alpha) \implies \alpha \circ X | X \sim \text{Bin}(X, \alpha).$$

The INAR(1) model is defined as

$$X_t = \alpha \circ X_{t-1} + \varepsilon_t,$$

where $\alpha \in [0,1)$ and ε_t are i.i.d discrete random variables with mean $\mu_\varepsilon > 0$ and variance σ^2_ε. Usually we consider Poisson-distributed errors, but more flexible discrete distributions are possible.

The INAR(1) process is non-linear due to the thinning operator, but it's a member of the conditional linear first-order AR models,

$$\mathbb{E}[X_t|X_{t-1}] = \alpha X_{t-1} + \mu_\varepsilon,$$

$$\mathbb{V}[X_t|X_{t-1}] = \alpha(1-\alpha)X_{t-1} + \sigma^2_\varepsilon.$$

If $\varepsilon_t \sim \text{Pois}(\lambda)$, then $X_t \sim \text{Pois}\left(\frac{\lambda}{1-\alpha}\right)$.

Generalizing to the INAR(p) is not straightforward and depends on the definition of the thinning operator, e.g.

$$X_t = \alpha_1 \circ X_{t-1} + \ldots + \alpha_p \circ X_{t-p} + \varepsilon_t,$$

where the thinning operators are applied with independent Y_j's from (14).

Estimation is simple for Poisson innovations and $p = 1$, whereas for all other cases we have some problems. As for the forecast we use the median, since we want an integer value for our predictions and the mean is usually a real number.

References

Brockwell, P. J. (2009). *Time Series: Theory and Methods.* 2° edizione. New York, N.Y: Springer.

Brockwell, P. J. and Davis, R. A. (2016). *Introduction to Time Series and Forecasting.* Third. Springer Texts in Statistics. New York: Springer-Verlag.

Douc, R. et al. (2014). *Nonlinear Time Series: Theory, Methods and Applications with R Examples.* 1° edizione. Boca Raton: Chapman and Hall/CRC.

Durbin, T. l. J. and Koopman, S. J. (2012). *Time Series Analysis by State Space Methods.* 2 edizione. Oxford: OUP Oxford.

Fan, J. and Yao, Q. (2005). *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer.

Shumway, R. H. and Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R Examples.* 4th ed. New York, NY: Springer.

Tsay, R. S. (2013). *Multivariate Time Series Analysis: With R and Financial Applications.* 1. edizione. Hoboken, New Jersey: John Wiley & Sons Inc.

Wei, W. W. S. (2019). *Multivariate Time Series Analysis and Applications.* 1. edizione. Hoboken, NJ: John Wiley & Sons Inc.

Sampling Theory

Instructor: prof. P.F. Perri

The idea of this short course is to give a general overview of some concepts and topics in sampling theory. Some topics are broad, whereas others are relatively technical and more specific. We will discuss mostly basic topics in estimation of population parameters, stratified sampling and optimal allocation of strata, inclusion of auxiliary information via regression and calibration, surveys of sensitive questions, and adjustments for non-response.

References: Tille (2020)
Valliant et al. (2018)
Särndal et al. (2003)

Lecture 8: Basic concepts in sampling theory

2022-02-09

We will start by considering the estimator of the population total from a finite-size population N. We are interested in the variance of the estimator which is itself a parameter that depends on the value in the population. Hence, we are also interested in the estimator of the variance. In sampling theory we have general three ingredients for designing a sample survey:

1. definition of the estimator $\widehat{\vartheta}$;
2. definition of the variance of the estimator $\mathbb{V}[\widehat{\vartheta}]$;
3. definition of the estimator for the variance of the estimator $\widehat{\mathbb{V}}[\widehat{\vartheta}]$;

8.1 Estimation of the total

Def. (Population)

We define a ***population*** as a finite set of N identifiable units $U = \{1, 2, \ldots, N\}$.

Identifiability Identifiability is very important, since for instance we are not able to assign a label to a virtual population.

Fixed size In general we assume N to be fixed, which is an assumption that is not often satisfied in practice. Consider a population of homeless people, of which we do not know the total number of units. In this case we use a *capture-recapture* method to estimate N.

Def. (Population parameter)

A ***population parameter*** is a constant describing the salient features of U w.r. to some variable, $\mathcal{Y} \to \vartheta = \vartheta(Y_1, \ldots, Y_N)$.

Remark Note that Y_i's are not random variables, since they represent the value of the variable for the i^{th} unit.

Example (of some interesting parameters)

Some examples of interesting parameters from a population are

› Total of the survey variable $\mathcal{Y}$, $Y = \sum_{i=1}^{N} Y_i$

› Mean of $\mathcal{Y}$, $\overline{Y} = Y/N$.

› Ration between variables, $R = \overline{Y}/\overline{X}$

Def. (Sample)

A ***sample*** $s = \{i_1, \ldots, i_n\}$ is a subset of U that is selected using a *probabilistic mechanism.*

Def. (Inclusion probability)

The ***inclusion probability*** of a unit i is the probability of unit i being included in the sample before running the sampling procedure.

Remark For a probabilistic sampling, *all* units must have a positive probability of being included. If even one unit has a null probability of being sampled, then the procedure is a ***non-probabilistic sampling scheme***.

Remark When we have non-probabilistic surveys our inference in uncertain, unless we use some corrections to bring them back to being similar to the population.

Example (Online surveys)

Online surveys exclude people that do not have internet connection from surveys, hence they have no possibility of being selected in a sample. This is not a probabilistic sample, and care should be put in the inferential conclusions.

Def. (Sampling space)

The ***sampling space*** is the set S of all *possible* and *distinct* samples which can be selected from U using a selection procedure.

Def. (Sampling design)

A ***sampling design*** is a real-valued set function

$$p : S \longrightarrow [0,1]$$

$$s \longmapsto p(s)$$

such that $p(s) \geq 0$ and $\sum_{s \in S} p(s) = 1$.

Duality In general we have a duality between $p(s)$ and the ***selection scheme***: if we assign a probability to each sampling design, there is an algorithm to obtain the selection of a single unit, and vice versa (Hedayat and Sinha, 1991, p. 5). Hence,

$$p(s) \iff \text{selection scheme}$$

Def. (Estimator)

An ***estimator*** is a function of the sampled observations $\widehat{\vartheta}(y_1, \ldots, y_n)$.

Def. (Sampling distribution)

The ***sampling distribution of*** $\boldsymbol{\widehat{\vartheta}}$ is the probability of $\widehat{\vartheta}$ attaining each of these values

$$\mathbb{P}(\widehat{\vartheta} = c) = \sum_{s : \widehat{\vartheta} = c} p(s).$$

Remark The sampling distribution of $\widehat{\vartheta}$ is defined by the sampling design that we choose. Indeed, we observe that

$$\mathbb{E}[\widehat{\vartheta}] = \sum_{s\in S} p(s)\widehat{\vartheta}(s)$$

$$\mathbb{V}[\widehat{\vartheta}] = \sum_{s\in S} p(s)\Big[\widehat{\vartheta}(s) - \mathbb{E}[\widehat{\vartheta}]\Big]^2$$

$$\text{MSE}(\widehat{\vartheta}) = \mathbb{E}[\widehat{\vartheta} - \vartheta]^2 = \sum_{s\in S} p(s)\Big[\widehat{\vartheta}(s) - \vartheta\Big]^2$$

Approaches in model surveys

1. For the reasons denoted above, in this course we will take the ***design-based approach*** to sampling survey.
2. In model-based surveys we have instead a superpopulation modelled by a multivariate random variable, from which we find our population.
3. In model-assisted approach we combine the two approaches to get a mixed approach.

Theorem 3 (Probability rules)
Let A, B be sets, then

$$(A\cup B)^c = (A^c \cap B^c), \quad (A\cap B)^c = (A^c \cup B^c).$$

$$\mathbb{P}(A\cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A\cap B)$$

If the A_i's are mutually exclusive events, then

$$\mathbb{P}\Big(\bigcup_{i\in I} A_i\Big) = \sum_{i\in I} \mathbb{P}(A_i).$$

If the A_i's are independent events, then

$$\mathbb{P}\Big(\bigcap_{i\in I} A_i\Big) = \prod_{i\in I} \mathbb{P}(A_i).$$

Example (Bank of Italy)
Ignoring the sampling design and using standard MLE techniques under i.i.d assumptions yields wrong inferences. This is true for datasets collected by the Bank of Italy, which uses a very complicated sampling design in order to carry out their sample surveys.

Let $A_i \subseteq S = \{s \in S : i \in S\}$, then the inclusion probability of first order can be written as

$$\mathbb{P}(i \in s) = \pi_i = \sum_{s\in A_i} p(s), \quad i = 1, \dots, N.$$

Alternatively, let δ_i be the indicator of whether the i^{th} is included in the sample, then

$$\delta_i = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{if } i \notin s \end{cases}$$

then $\delta_i \sim \text{Ber}(\pi_i)$ and

$$\pi_i = \sum_{s \in S} p(s)\delta_i = \mathbb{E}[\delta_i]$$

Finally, let $A_{ij} \subseteq S = \{s \in S : (i,j) \in S\}$ be the set of samples in which both i and j appear, then the joint probability of inclusion is

$$\pi_{ij} = \sum_{s \in A_{ij}} p(s) = \mathbb{E}[\delta_i \delta_j].$$

If a sample of size n is drawn ***with replacement***, then we have that the probability of inclusion π_i and π_{ij} are, respectively,

$$\begin{aligned} \pi_i &= 1 - (1 - P_i)^n, \\ \pi_{ij} &= 1 - (1 - P_i)^n - (1 - P_j)^n + (1 - P_i - P_j)^n, \end{aligned}$$

where P_i is the ***selection probability*** of unit i at each selection draw.

Proof.
...

□

Properties of π_i and π_{ij}

Suggested readings: Valliant et al. (2018)
Pfeffermann (1993)

Def. (Effective sample size)
We define $\nu(s) :=$ "# of different units in the sample" as the ***effective sample size*** of the sample. In general, $\nu(s) \neq n$ if the sampling is done with replacement.

Let $p(s)$ be a design, $n(s)$ the sample size, $\nu(s)$ be the *effective sample size* and $\nu = \mathbb{E}[\nu(s)]$, then we have that

$$\sum_{i=1}^{N} \pi_i = \nu,$$

$$\sum_{i=1}^{N} \sum_{j \neq i} \pi_{ij} = \mathbb{V}[\nu(s)] + \nu(\nu - 1).$$

Proof.
Slides

□

If the sample is ***without replacement*** (WOR), then we have

$$n(s) = n = \nu(s) = \nu,$$

hence the previous results become

$$\sum_{i=1}^{N} \pi_i = n,$$

$$\sum_{i=1}^{N} \sum_{j \neq i} \pi_{ij} = n(n-1)$$

moreover,

$$\sum_{j \neq i} \pi_{ij} = \pi_i (n-1).$$

8.2 Selection procedures

In general, simple random samples where every unit has the same inclusion probability are not used when designing sample surveys.

Example (Shoplifting)

Suppose we are interested in the total number of theft and total value of theft from the shops in a city. We have different strategies

a) Assign the same probability to each shop.

b) Give to the shop a different probability, since we think that the larger are the shop then the larger will be the number of thefts. Then we need to find a *proxy variable* that is correlated to the number of thefts.

In this example, we use

$$\pi_i \propto A, \quad A := \text{“floor area of the shop”},$$

hence the sample will contain a lot of large shops and will not be representative of the population.

Statistical analysis provides us the tools to deal with this situation by *weighting* the estimator so that the estimator is unbiased for the total.

- Sample units proportional to a measure of their size
- The size $\mathcal{X}$ is strongly correlated with $\mathcal{Y}$ (*probability proportional to sample size*)
- X_i must be known in advance for all $i \in U$.
- $P_i = X_i / X$, where $X = \sum_{i=1}^{N} X_i$
- We hope to have a more accurate estimator than by using simple random sampling
- The final estimator $\widehat{\vartheta}$ will be a weighted average so that the bias is removed.

8.2.1 PPS with replacement (PPSWR)

A PPS with replacement (PPSWR) can be drawn using different methods.

Cumulative total method Suppose X_i are general numbers, then we can use the following steps to sample with replacement proportionally to size:

1. Write down the cumulative totals for the size $Q_i = \sum_{j \leq i} P_j$
2. Choose a random number $r \sim \text{Unif}(0, 1)$
3. Select the i^{th} population unit if $Q_{i-1} < r \leq Q_i$.
4. Repeat until the size of the sample equals n.

Lahiri's method To avoid the calculation of the cumulative sum (not a problem anymore), we can use

1. Select i such that $1 \leq i \leq N$.
2. Select a j such that $1 \leq j \leq \max\{X_1, \dots, X_N\}$.
3. If $j \leq X_i$, the i^{th} unit is selected, otherwise the pair (i, j) is rejected and another pair is chosen by repeating steps (1) and (2).

8.2.2 PPS without replacement (PPSWOR)

Samples without replacement is more complicated than sampling with replacement, and more than a hundred methods have been proposed. The probability π_{ij} is very difficult to express in a closed form when the sample size is greater than 2 or 3.
Methods are useful if they satisfy some properties, i.e.

- $\pi_i = nP_i$, calculations are simplified;
- $\pi_{ij} > 0$ yields unbiased estimator of the variance;
- $\pi_i\pi_j - \pi_{ij} \geq 0$ yields nonnegative variances of the estimators.

Yates-Grundy ($n = 2$)

1. Select first unit with probability $P_i = X_i/X$
2. Modify the total size X and select the second unit after removing the i^{th} unit.

The inclusion probabilities become

$$\pi_i = P_i\Big(1 + \sum_{j \neq i} \frac{P_j}{1 - P_j}\Big)$$

$$\pi_{ij} = P_iP_j\Big(\frac{1}{1 - P_i} + \frac{1}{1 - P_j}\Big)$$

Proof.

$$
\begin{aligned}
\pi_i &= \mathbb{P}\big((i_1 = i) \cup (i_1 \neq i, i_2 = i)\big) \\
&= \mathbb{P}(i_1 = i) + \sum_{j \neq i} \mathbb{P}(i_1 = j)\mathbb{P}(i_2 = i | i_1 = j) \\
&= P_i + \sum_{j \neq i} P_j \frac{X_i}{X - X_j} \\
&= P_i\Big(1 + \sum_{j \neq i} \frac{P_j}{1 - P_j}\Big)
\end{aligned}
$$

Exercise Try to calculate the probabilities for a sample of size $n = 3$.

□

Hartley-Rao-Cochran Allows us to select a sample of size n without heavy computations.

1. Partition the population in n groups such that the sizes are $N_1, \dots, N_n$ and $\sum_{i=1}^{n} N_i = N$.
2. From each group select a single unit using the CTM or Lahiri's method.

In general the method depends on the way we split the population.

Midzuno-Sen Widely used in practice

1. Select the first unit using PPSWR
2. The remaining units are selected with simple random sampling, $\mathbb{P}(i_k = i) = 1/(N-1)$.

The inclusion probabilities are given by

$$
\left.
\begin{aligned}
\pi_i &= P_i \frac{N-n}{N-1} + \frac{n-1}{N-1} \\
\pi_{ij} &= \frac{n-1}{N-1} \dots
\end{aligned}
\right.
\implies P_i = \frac{(N-1)\pi_i - (n-1)}{N-n}
$$

8.2.3 PPSWR and Hansen-Hurwitz estimator

Consider a population $U = \{1, \dots, N\}$ and we have Y_i variable and X_i auxiliary variable for each population. Let (y_i, p_i) be the values of the study variable and selection probability for each unit; then, an estimator $\widehat{Y}$ can be constructed as

$$
\widehat{Y} = \sum_{i=1}^{n} d_i y_i = \sum_{i=1}^{N} d_i Y_i \gamma_i,
$$

where

- d_i is a ***design-based weight*** to make the estimator unbiased.
- Y_i is the constant value of each unit i.

- γ_i is the number of times the i^{th} population unit appears in the sample,

$$\gamma_i \sim \text{Bin}(n, P_i).$$

Hence, the estimator is such that

$$\mathbb{E}[\widehat{Y}] = \sum_{i=1}^{N} d_i Y_i \mathbb{E}[\gamma_i] = n \sum_{i=1}^{N} d_i Y_i P_i = \sum_{i=1}^{N} Y_i,$$

which is unbiased for Y if $d_i = \frac{1}{nP_i}$, and thus we obtain the ***Hansen-Hurwitz estimator*** (Hansen and Hurwitz, 1943)

$$\widehat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i}.$$

We have that

$$\mathbb{V}[\widehat{Y}_{HH}] = \frac{1}{n} \sum_{i=1}^{N} P_i \left(\frac{Y_i}{P_i} - Y \right)^2,$$

which can be unbiasedly estimated by

$$\widehat{\mathbb{V}}[\widehat{Y}_{HH}] = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \widehat{Y}_{HH} \right)^2.$$

Remarks

1. $\mathbb{V}[\widehat{Y}_{HH}] = 0 \iff P_i = Y_i/Y$ for all i.
2. If $Y_i \propto X_i$, then $\mathbb{V}[\widehat{Y}_{HH}]$ is low, and $\mathbb{V}[\widehat{Y}_{HH}] = 0$ when $Y_i = \beta X_i$.
3. If $Y_i \propto \alpha + \beta X_i$, then $\mathbb{V}[\widehat{Y}_{HH}] \propto \alpha$.

Confidence intervals We have that $\widehat{Y}_{HH} \to \mathcal{N}(Y, \mathbb{V}[\widehat{Y}_{HH}])$, hence an *approximate* confidence interval for Y is

$$\dots$$

SRSWR If $P_i = 1/N$ we obtain the simple random sampling with replacement, and this is called the *expansion estimator.*

Efficiency In general, $\widehat{Y}_{HH}$ is more efficient than the simple $\widehat{Y}$ if

$$\text{Cov}\left(\frac{\mathcal{Y}^2}{\mathcal{X}}, \mathcal{X} \right) > 0.$$

8.2.4 PPSWOR and Horvitz-Thompson estimator

With the same reasoning we can build the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) using

$$\widehat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{N} \frac{Y_i}{\pi_i} \delta_i,$$

where the design weights are $d_i = 1/\pi_i$.

We have that $\delta_i \sim \text{Ber}(\pi_i)$, hence $\mathbb{E}[\delta_i] = \pi_i$, $\mathbb{V}[\delta_i] = \pi_i(1-\pi_i)$, and Cov(...)

$$\begin{aligned}\mathbb{E}[\widehat{Y}_{HT}] &= Y \\ \mathbb{V}[\widehat{Y}_{HT}] &= \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i^2}\pi_i(1-\pi_i) + \sum_{i=1}^{N}\sum_{j\neq i} \frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}(\pi_{ij} - \pi_i\pi_j).\end{aligned}$$

We can extend the sampling version of the estimator in order to show that the estimated variance $\widehat{\mathbb{V}}[Y_{HT}]$ is unbiased.

The Yates-Grundy variance can be applied when the sample size n is fixed in advance, and

$$\widehat{\mathbb{V}}[\widehat{Y}_{HT}] = \sum_{i=1}^{n}\sum_{j>i} \left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

Remark 1 $\pi_{ij} > 0$ for all i, j is required for unbiasedness of $\widehat{\mathbb{V}}[\widehat{Y}_{HT}]$.

Remark 2 $\pi_i\pi_j - .. \geq 0$ is required so that... Hence, the usefulness of the Midzuno-Sen procedure.

In this case, we also see that a high value of intercept increases the value of the variance of the Horvitz-Thompson estimator.

Problem A big problem in the HT estimator is calculating the second-order probabilities, unless we make some approximations. We usually approximate using

$$\pi_{ij} = \pi_i\pi_j\frac{c_i + c_j}{2},$$

where c_i and c_j are appropriately chosen by different authors. In general, it has been proven (Raj, 1968) that if for all i, j

$$\pi_{ij} > \frac{n-1}{n}\pi_i\pi_j \implies \mathbb{V}[\widehat{Y}_{HH}] > \mathbb{V}[\widehat{Y}_{HT}].$$

Sample size calculation The sample size calculation is usually dependent on various considerations

1. the budget of the survey;
2. the individual cost of surveying one unit;
3. the accuracy required for statistical inference from the survey data.

Assuming that the sample is being selected by SRSWOR and the parameter to estimate is the population mean $\overline{Y}$. Under SRSWOR, the *expansion estimator* is unbiased

$$\bar{y} = \sum_{i=1}^{n} y_i/n,$$

which has variance $\mathbb{V}[\bar{y}] = \frac{1-f}{n} S_y^2$, with $f = n/N$. Assume that V_0 is a desirable value for the standard error of $\bar{y}$, then we constrain

$$\mathbb{V}[\bar{y}] = \frac{1-f}{n} S_y^2 = V_0^2,$$

and solving for n we obtain

$$n_0 = \frac{S_y^2}{V_0^2 + S_y^2/N}.$$

The problem is that S_y^2 is usually unknown, unless we can produce a guess for the variance driven by e.g. census data, previous surveys, etc... Otherwise we can conduct a small pilot sample survey only to estimate S_y^2.

Another way to proceed is to fix the *margin of error* using the confidence interval,

$$\mathbb{P}(|\bar{y} - \overline{Y}| \leq e) = 1 - \alpha,$$

hence solving for n using the normal distribution confidence intervals,

$$n = \frac{z_{1-\frac{\alpha}{2}} S_y^2}{e^2 + \frac{z_{1-\frac{\alpha}{2}} S_y^2}{N}}.$$

In practice We usually have multiple variables to survey, and each variable has to have different criteria to satisfy. In this case, we might focus on one or two variables that are of primary interest.

8.3 Stratified sampling

Applying a stratified sampling approach might yield more precise estimates of population parameters, when the survey variable takes different mean values in different subgroups of the population.

Def. (Stratification)
Stratification means dividing the population units into H subpopulations called strata according to one or more variables.

Remark Strata are formed so that each unit within the stratum is as much similar as possible w.r. to the target variable $\mathcal{Y}$. We do so by stratifying using other auxiliary variables $\mathcal{X}_1, \ldots, \mathcal{X}_k$, which are correlated with $\mathcal{Y}$.

Pooling From each stratum an independent sample of size n_h is selected with sampling designs which might differ across strata. Estimates from each strata are *pooled* to obtain overall population estimates.

1. This reduces the possibility of obtaining a "bad" sample, hence it might produce a better estimate of the parameters.
2. We can get estimates within each subgroup, hence we can find confidence levels for the mean of each strata.

3. More convenient to administer and might result in a lower cost for the survey.

Stratification is a complex topic since we need to prune the number of stratification variables, select how many strata we want to use, which sampling design to adopt in each stratum, and how many units to sample from each strata.

An unbiased estimator for the population total Y is

$$\widehat{Y}_{\text{STR}} = \sum_{h=1}^{H} N_h \bar{y}_h = N \sum_{h=1}^{H} W_h \bar{y}_h,$$

provided that $n_h \geq 1$ for each h. The variance of the estimator is given by the sum of the variances, since the estimators are independent across the strata,

$$\mathbb{V}[\widehat{Y}_{\text{STR}}] = \sum_{h=1}^{H} N_h^2 \mathbb{V}[\bar{y}_h] = N^2 \sum_{h=1}^{H} W_H^2 \frac{1-f_h}{n_h} S_h^2.$$

$$\widehat{\mathbb{V}}[\widehat{Y}_{\text{STR}}] = \sum_{h=1}^{H} N_h^2 \mathbb{V}[\bar{y}_h] = N^2 \sum_{h=1}^{H} W_H^2 \frac{1-f_h}{n_h} s_h^2.$$

In general, the lower S_h^2, the higher will the precision of the estimate be. Hence, we are looking to construct strata which are *highly homogeneous* within themselves.

We want to determine the allocation of sample sizes n_h in order to minimize the variance of the estimator,

$$\mathbb{V}[\widehat{Y}_{\text{STR}}] = \underbrace{N^2 \sum_{h=1}^{H} \frac{W_h^2 S_h^2}{n_h}}_{\text{minimize}} - \sum_{h=1}^{H} N_h S_h^2,$$

under the constraint given by the total cost $C_t = C_0 + C_v$, where

$$C_0 = \text{fixed}$$

$$C_v = \sum_{h=1}^{H} c_h n_h,$$

where c_h is the cost of surveying one unit in stratum h.

$$\min_{\boldsymbol{n}} \sum_{h=1}^{H} \frac{W_h^2 S_h^2}{n_h}$$

$$\text{s.t.} \sum_{h=1}^{H} c_h n_h = C_v,$$

and by using the Lagrange multiplier solution we obtain the optimal sample allocation as

$$n_h = n \frac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{H} W_h S_h / \sqrt{c_h}}.$$

In general, the higher the variability (S_h) within the strata and the larger the strata is (W_h), the more units we need to sample from it. Higher cost (c_h) means a lower number of units.
In many cases, $c_h = c$ and the optimal allocation reduces to the ***Neyman allocation***,

$$n_h = n \frac{W_h S_h}{\sum_{h=1}^{H} W_h S_h},$$

which yields an expression for the variance of the estimator,

$$\mathbb{V}_{\text{opt}}[\widehat{Y}_{\text{STR}}] = N^2 \left[\frac{\cdots^2}{\cdots} - \frac{\cdots}{\cdots} \right].$$

Finally, if $S_h = S$ in all strata, the Neyman allocation reduces to the ***proportional allocation***,

$$n_h = n \cdot \frac{N_h}{N}.$$

In general,

$$\mathbb{V}[\widehat{Y}_{\text{STR}}] \geq \mathbb{V}_{\text{prop}}[\widehat{Y}_{\text{STR}}] \geq \mathbb{V}_{\text{opt}}[\widehat{Y}_{\text{STR}}].$$

Lecture 9: Auxiliary information in estimation

2022-02-10

Introduce and discuss the idea of including auxiliary information in the estimation stage. Previous lecture was auxiliary information in the design stage, today in estimation. Information to include information in the estimation stage is less restrictive than before: we only need to know the value of the mean in the whole population, instead than the value of auxiliary variables for each unit in the population.

9.1 Ratio method

Suppose $\mathcal{X}$ is auxiliary whose total is known in advance, $\mathcal{Y}$ the variable of interest such that $\mathrm{Cov}(\mathcal{X}, \mathcal{Y}) > 0$. Let $\widehat{Y}, \widehat{X}$ be unbiased estimators under a generic sampling design $p(s)$ and define the estimated ratio as $\widehat{R} = \widehat{Y}/\widehat{X}$.

Def. (Ratio estimator)

We define the ***ratio estimator*** between the two variables $\mathcal{X}$ and $\mathcal{Y}$ as

$$\widehat{Y}_R = \widehat{Y}\frac{X}{\widehat{X}} = \widehat{R}X,$$

Remark The quantity $X/\widehat{X}$ is called ***adjustment factor*** which tends to reduce the variability of $\widehat{Y}$ around Y.

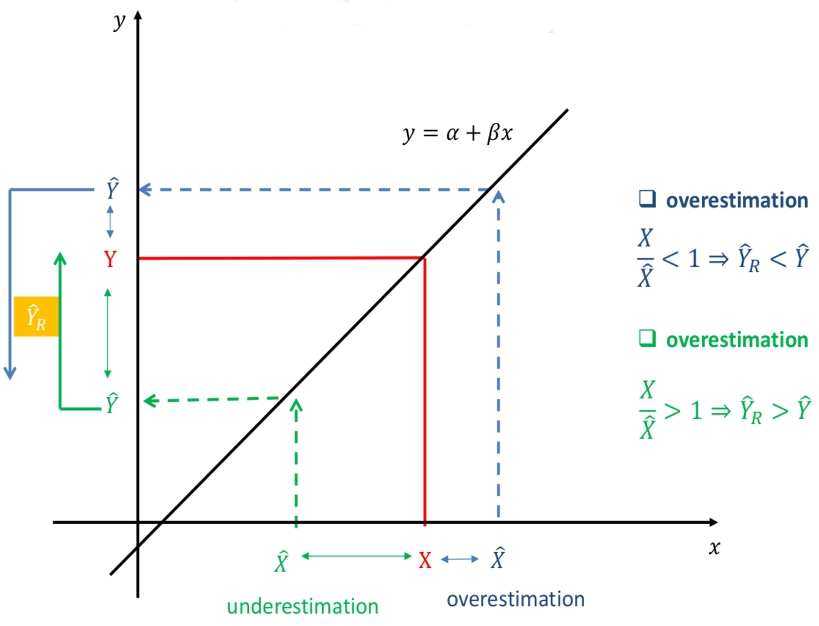

Figure 5: Intuition behind the ratio estimator $\widehat{Y}_R$: since $\mathrm{Cov}(X, Y) > 0$, and we know X in advance, $\widehat{X}$ gives information whether we also overestimated/underestimated Y. Hence, $\widehat{Y}_R$ pulls $\widehat{Y}$ towards the regression line.

Bias In general, $\widehat{Y}_R$ is biased for Y and its bias is

$$\mathbb{E}[\widehat{Y}_R - Y] = -\operatorname{Cov}(\widehat{R}, \widehat{X}).$$

Proof.

$$\begin{aligned}\operatorname{Cov}(\widehat{R}, \widehat{X}) &= \mathbb{E}[\widehat{R}\widehat{X}] - \mathbb{E}[\widehat{R}]\mathbb{E}[\widehat{X}] \\ &= \mathbb{E}[\widehat{Y}] - \mathbb{E}[\widehat{R}]X \\ &= Y - \mathbb{E}[\widehat{Y}_R] \\ &= -B(\widehat{Y}_R)\end{aligned}$$

In general, it can also be shown that the bias is negligible when the sample size is large enough.

□

Consider now the mean squared error of $\widehat{Y}_R$, $\operatorname{MSE}(\widehat{Y}_R)$, then with the first-order approximation of the MSE we have

$$\operatorname{MSE}(\widehat{Y}_R) \approx \mathbb{V}[\widehat{Y}_R] \overset{\text{Taylor}}{=} \mathbb{V}[\widehat{Y}] - 2R\operatorname{Cov}(\widehat{X}, \widehat{Y}) + R^2\mathbb{V}[\widehat{X}].$$

Proof.
Using the *Delta method* we have that if $\delta_y = (\widehat{Y} - Y)/Y$ and $\delta_x = (\widehat{X} - X)/X$, then

$$\widehat{Y}_R = \widehat{Y}\frac{X}{\widehat{X}} = Y(1 + \delta_y)(1 + \delta_x)^{-1}$$

Considering the Taylor expansion of $1/(1 + \delta_x)^{-1}$ in $\delta_x = 0$, we have

$$\begin{aligned}\widehat{Y}_R &= Y(1 + \delta_y)(1 - \delta_x + \delta_x^2 - \delta_x^3 + \ldots) \\ &= Y(1 + \delta_y - \delta_x + \delta_x^2 - \delta_y\delta_x + \ldots) \\ &\approx Y(1 + \delta_y - \delta_x).\end{aligned}$$

Hence, by calculating the MSE we obtain

$$\operatorname{MSE}(\widehat{Y}_R) \approx Y^2\mathbb{E}[(\delta_y - \delta_x)]^2 = \mathbb{V}[\widehat{Y}] + R^2\mathbb{V}[\widehat{X}] - 2R\operatorname{Cov}(\widehat{X}, \widehat{Y}).$$

□

Improvement Considering the expression of the MSE, then we obtain an improvement over the estimation with $\widehat{Y}$ if

$$\mathbb{V}[\widehat{Y}] \geq \mathbb{V}[\widehat{Y}_R] \iff \rho(\widehat{X}, \widehat{Y}) > \frac{1}{2}\frac{C(\widehat{X})}{C(\widehat{Y})},$$

where $C(\cdot)$ is the ***coefficient of variation***, $C(Z) = \sigma_z/\mu_z$.
Since $\rho(\widehat{X}, \widehat{Y}) \in [0, 1]$, this condition is never satisfied if

$$C(\widehat{X}) > 2C(\widehat{Y}).$$

Hence, if we choose $\mathcal{X}$ which is characterized by a high variability, then we expect our estimator to be less efficient than $\widehat{Y}$.

The above expression can be generalized to different sampling schemes, hence with different estimators $\widehat{Y}$ for which we know the expression of $\mathbb{V}[\widehat{Y}]$ and $\mathbb{V}[\widehat{X}]$. Now, we can write the covariance between the HH and HT estimators,

$$\operatorname{Cov}(\widehat{X}, \widehat{Y}) = \begin{cases} \dots & \text{for HH} \\ \dots & \text{for HT} \end{cases}$$

which can be particularized to simple random sampling using the expansion estimator.
The ratio estimator works well when

$$Y_i = \beta X_i + \varepsilon, \quad \rho_{xy} > 0,$$

whereas if $\rho_{xy} < 0$ then the ***product estimator*** can be used,

$$\widehat{Y}_P = \widehat{Y}\widehat{X}/X.$$

Note that the regression *does not have an intercept*: in general, having an intercept $\alpha \neq 0$ is a bad property in sampling schemes.

Note
The regression estimator $Y = \alpha + \beta X + \varepsilon$ is the best estimator we can use in terms of estimation of Y. The ratio estimator is still used today, and the justifications for its use are not very convincing apart from the need to publish.

9.2 Regression estimator

The idea is to improve an estimator for $\mathcal{Y}$ using the linear regression on X, with the idea that usually X is easier to measure than Y.

1. X_i's measured in the sample
2. Y_i's measured in the sample
3. X total in the population

Consider a linear combination

$$\widehat{Y}^* = a\widehat{Y} + b\widehat{X} + cX,$$

and determine the coefficients a, b, c such that the estimator is unbiased for Y. Imposing unbiasedness yields the following construction,

$$\mathbb{E}[\widehat{Y}^*] = Y \implies a = 1, c = -b \implies \widehat{Y}_d = \widehat{Y} + c(X - \widehat{X}).$$

Interpretation We adjust the estimate of $\widehat{Y}$ through a quantity which is proportional to the known estimation error $X - \widehat{X}$.

In order to determine c, we consider the variance of the difference estimator,

$$\mathbb{V}[\widehat{Y}_d] = \mathbb{V}[\widehat{Y}] + c^2\mathbb{V}[\widehat{X}] - 2c\,\mathrm{Cov}(\widehat{Y}, \widehat{X}),$$

which is similar to the ratio estimator as before. The above quantity is minimized by setting

$$\widehat{c} = \underset{c}{\operatorname{argmin}}\,\mathbb{V}[\widehat{Y}_d] = \frac{\mathrm{Cov}(\widehat{Y}, \widehat{X})}{\mathbb{V}[\widehat{X}]},$$

which is the regression coefficient between Y and X.

Def. (Regression estimator)
We define the ***regression estimator*** for Y as

$$\widehat{Y}_{\mathrm{reg}} = \widehat{Y} + \beta(X - \widehat{X}),$$

which has variance given by

$$\mathbb{V}[\widehat{Y}_{\mathrm{reg}}] = \mathbb{V}[\widehat{Y}]\big(1 - \rho^2(\widehat{Y}, \widehat{X})\big).$$

Improvement In general we observe that $\mathbb{V}[\widehat{Y}] \geq \mathbb{V}[\widehat{Y}_{\mathrm{reg}}]$, hence this is *always* better or equal than the estimator without regression.

Ratio estimator This is a better estimator than the ratio estimator, which can be worse than $\widehat{Y}$ for some values of $C(X)$. The two estimator can be equal in terms of efficiency if

$$\mathrm{MSE}(\widehat{Y}_R) - \mathrm{MSE}(\widehat{Y}_{\mathrm{reg}}) = \mathbb{V}[\widehat{X}](R - \beta)^2 \geq 0,$$

and this holds when the intercept is null, $Y_i = \beta X_i$.

9.3 Optimal use of multi-auxiliary estimation

We want to consider the best possible estimator for Y under auxiliary information.

Notation $\boldsymbol{X} = (\mathcal{X}_1, \dots, \mathcal{X}_p)^\top$ auxiliary variable, for which we both know $\overline{\boldsymbol{X}}$ and $\boldsymbol{S}$. We also define $\boldsymbol{v} = (\bar{\boldsymbol{x}}, \boldsymbol{s})^\top$.

Consider a wide class of estimators

$$\bar{y}_g = g(\bar{y}, \boldsymbol{v}),$$

which satisfies some regularity conditions (diana perri 2007). Expanding $\bar{y}_g$ sing a

$$\bar{y}_g \approx \bar{y} + (\boldsymbol{v} - \boldsymbol{V})^\top \boldsymbol{g}_v,$$

and show that the MSE is

$$\mathrm{MSE}(\bar{y}_g) \approx \frac{1-f}{n}(S_y^2 + \boldsymbol{g}_v^\top \Sigma_{xx} \boldsymbol{g}_v + 2\boldsymbol{g}_v^\top \Omega_{y\boldsymbol{z}}),$$

which is minimized over $\boldsymbol{g}_v$, yielding

$$\boldsymbol{g}_v^* = \operatorname*{argmin}_{\boldsymbol{g}_v} \mathrm{MSE}(\bar{y}_g) = \Sigma_{zz}^{-1}\Omega_y \boldsymbol{z},$$

with $\beta_{y\cdot}$ are the ***partial regression coefficients***.

The best estimator in the class is the unbiased ***multivariate regression estimator***, given by

$$\bar{y}_{\text{reg}} = \bar{y} + (\overline{\boldsymbol{X}} - \bar{\boldsymbol{x}})^\top \beta_{y\boldsymbol{x}} + (\boldsymbol{S} - \boldsymbol{s})^\top \beta_{y\boldsymbol{\delta}},$$

using the partial regression coefficients of each X_i, and not the total estimator of the linear regression.

Estimators for instance of the form

$$\bar{\bar{y}} = \bar{y} + b_1(\overline{X}_1 - \bar{x}_1) + b_2(\overline{X}_2 - \bar{x}_2),$$

where the b_i's are the ***total*** regression coefficients of Y on X, and not the ***partial*** coefficients, are *not optimal* for estimating Y.

9.4 Regression estimator in stratified sampling

We now try to generalize the regression estimator by assuming that the total X_h is known for each stratum $h = 1, \dots, H$. The regression of $\mathcal{Y}$ on $\mathcal{X}$ produces different regression coefficients across the strata, and we can obtain the regression estimator within each strata

$$\widehat{Y}_{\text{reg},h} = \widehat{Y}_h + \beta_h(X_h - \widehat{X}_h).$$

Def. (Separate regression estimator)
We define the ***separate regression estimator*** as

$$\widehat{Y}_{\text{reg}}^{(s)} = \sum_{h=1}^{H} \widehat{Y}_{\text{reg},h}.$$

We can calculate the mean squared error in terms of the covariances and variances inside each stratum.

Estimation When we do not know the value of the variables X_h inside each strata, we can construct another estimator based on the estimation

$$\widehat{Y}_{\text{str}} = \sum_{h=1}^{H} \widehat{Y}_h, \quad \widehat{X}_{\text{str}} = \sum_{h=1}^{H} \widehat{X}_h,$$

and calculate the regression coefficient

$$\beta_c = \mathrm{Cov}(\widehat{Y}_{\mathrm{str}}, \widehat{X}_{\mathrm{str}})/\mathbb{V}[\widehat{X}_{\mathrm{str}}].$$

The combined estimator is the regression estimator when applied to Y_{str},

$$\widehat{Y}_{\mathrm{reg}}^{(c)} = \widehat{Y}_{\mathrm{str}} + \beta_c(X - \widehat{X}_{\mathrm{str\ fs}}),$$

and the calculable first-order MSE approximation is valid if n_h is large enough in each strata. In general, we observe that the separate estimator is always equal or better than the combined estimator,

$$\mathbb{V}[\widehat{Y}_{\mathrm{reg}}^{(c)}] - \mathbb{V}[\widehat{Y}_{\mathrm{reg}}^{(s)}] = N^2 \sum_{h=1}^{H} a_h (b_h - b_c)^2 \geq 0,$$

and it is equivalent whenever b_h is constant in all strata.

9.5 Regression-type estimators and missing data

Another contribution is to study the regression-type estimators whenever we are in the context of missing data. Let s be a SRSWOR of size n drawn from U to estimate $\overline{Y}$. Assume now that the y_i's can be observed only on a subset of $s_R \subset s$ of size $n_r < n$, and that auxiliary information $\mathcal{X}$ is available, for which $\overline{X}$ might be known or not.

Define $\bar{x}_n$ to be the mean on all the sample, $\bar{y}_r$ and $\bar{x}_r$ to be the mean values on the observable units, and $b = s_{xy}/s_x^2$ computed on s_R.

The best estimators for $\overline{Y}$ are (Diana perri 2010) are the following:

$$\bar{y}_1 = \bar{y}_r + b(\overline{X} - \bar{x}_n)$$

$$\bar{y}_2 = \bar{y}_r + b(\overline{X} - \bar{x}_r)$$

$$\bar{y}_3 = \bar{y}_r + b(\bar{x}_n - \bar{x}_n)$$

Lecture 10: More technical topics

2022-02-10

Randomized response theory (RRT) is a technique for tackling sensitive and confidential topics such as gambling, alcoholism, sexual abuse, drug addiction, tax evasion, ...
The idea is

a) Randomized device is used to hide the answers in

b) Since the question is kept random, both the researcher and the interviewer are unaware of which question has been answered.

to sample indirect questions about the topic from a bigger pool of observation, and an estimator is constructed such that the quantities of interest are appropriately estimated.

10.1 Randomized response theory

Example (Drug use)
Question is "have you ever used drugs?", and the interview proceeds as follows:

1. A deck of cards is considered as a randomizer
2. On each card is written one of two statements "I used drugs" or "I did not use drugs" with proportions p and $1-p$.
3. Respondents are asked to select a card and report `yes` or `no` to match the question, *without revealing which question* they have drawn.

An unbiased estimator of π_A is

$$\widehat{\pi}_W = \frac{\widehat{\lambda} - (1-p)}{2p-1}, \quad \widehat{\lambda} = \frac{\#\text{ "yes" response in } s}{n}.$$

In the above example, we have an unbiased estimator which is however less efficient than the estimators with the direct method. However, this method is useful to *improve the quality* of the data with respect to the direct method.

Problem In the above example, the answer no to the negative question is itself a

Unrelated question models We associate to the sensitive question a completely unrelated question, for instance

a) "Have you ever used drugs?"

b) "Do you like soccer?"

With this method we have an easy estimator for the second question, using a second sample, and we can estimate the proportion as

...

Perturb the response on Y and an auxiliary variable X using W, U, T, H ***scrambling variables*** with known distributions, and the response becomes a scrambled version of the true values,

$$S = \varphi(Y; W, U), \quad R = \phi(X; T, H).$$

Mechanism n individuals run a Bernoulli trial with probability p: if success, the respondent provides the true X and Y, whereas if there isa failure we obtain R and S. The distribution of the response is

$$(Z, V) = \begin{cases} (Y, X) & \text{with probability } p \\ (S, R) & \text{with probability } 1 - p \end{cases}$$

We can obtain an efficient estimator for Z by augmenting with V, and the procedures based on auxiliary information are at least as efficient as the analogous procedures defined without them.

10.2 Calibration

A more general approach to incorporate auxiliary information into the estimates is based on the idea of ***calibration***, introduced by Deville and Särndal 1992. With this method we can eliminate many restrictions of the other estimators. We replace the estimator using

$$\widehat{Y}_c = \sum_{i \in S} w_i y_i,$$

where the w_i's are determined to minimize the distance to $d_i = 1/\pi_i$, while respecting the ***calibration equation***

$$\sum_{i \in S} w_i \boldsymbol{x}_i = \boldsymbol{T}_X.$$

There are several distance functions, and among them the most used is the $\boldsymbol{\chi^2}$ ***distance***

$$G_S(w, d) = \sum_{i \in S} \frac{(w_i - d_i)^2}{2 d_i g_i},$$

where the g_i's are known positive weights unrelated to d_i chosen by the researcher. This choice yields the ***calibration estimator*** of the total of $\mathcal{Y}$,

$$\widehat{Y}_c = \widehat{Y}_{HT} + (\boldsymbol{T} - \widehat{\boldsymbol{X}}_{HY})^\top \widehat{\boldsymbol{B}},$$

where

$$\widehat{\boldsymbol{B}} = \left(\sum_{i \in S} d_i g_i \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \sum_{i \in S} d_i g_i \boldsymbol{x}_i y_i,$$

which is called GREG estimator. Other choices of distance are possible, but asymptotically they are equal to the χ^2distance functions.

When the relationship between X's and Y is not linear, the method can be generalized using a ***model calibration*** approach.

Def. (Superpopulation)
We assume that the population under study is generated by a stochastic mechanism, where $Y_1, \dots, Y_N$ are a realization of a random variable $\boldsymbol{Y} = (Y_1, \dots, Y_N)^\top$ whose distribution defines a ***superpopulation***.

A superpopulation model specifies certain features of the generating process, i.e. we can specify

a) the first moments of the marginals;

b) the correlation between each Y_i and Y_j;

c) the functional form of the superpopulation, $\boldsymbol{Y} \to f(\boldsymbol{Y}|\boldsymbol{\vartheta})$.

We assume that there is a relationship between X and Y of the form

$$\mathbb{E}_\xi[Y_i|\boldsymbol{x}_i] = \mu(\boldsymbol{x}_i, \boldsymbol{\vartheta}), \quad \mathbb{V}_\xi[Y_i|\boldsymbol{x}_i] = \sigma^2 v_i^2,$$

and this model includes both LM's and GLM's.

The idea is not to calibrate w.r. to the value of X, but w.r. to the *fitted* model.
We want an estimator

$$\widehat{Y}_{\text{mc}} = \sum_{i \in S} w_i y_i,$$

where w_i are the weights obtained as a solution of the minimization problem

$$\begin{aligned} \min \sum_{i \in S} & \frac{(w_i - d_i)^2}{2 d_i g_i} \\ \text{s.t.} \sum_{i \in s} & w_i = N \\ \sum_{i \in S} & w_i \mu(\boldsymbol{x}_i, \widehat{\boldsymbol{\vartheta}}) = \sum_{i \in U} \mu(\boldsymbol{x}_i, \widehat{\boldsymbol{\vartheta}}). \end{aligned}$$

Lecture 11: More technical topics (ii)

2022-02-11

cluster sampling is a complex design, and the complexity increases as the hierarchy becomes deeper. We will mostly focus on two-stage cluster design, which is the simplest possible cluster design.

11.1 Cluster sampling

The PPS sampling framework requires a list of elementary units from the population that we want to sample. In many cases, however, the list is either not available or unfeasible/time consuming to produce. Hence, we prefer to apply a ***cluster sampling*** procedure.

1. Split the population in groups and compile a list of *primary sampling units.*
2. A sample of clusters is selected from the list, according to a sampling design (usually PPS).

If after sampling a cluster

a) all elementary units from the cluster are included, we have *single-stage cluster sampling*;

b) a sample of elementary units from the cluster is selected, we have *two-stage sampling*;

c) other more complicated are *multi-stage sampling.*

Notation The population $U = \{1, \dots, N\}$ is made by the set of clusters (N PSUs). Within each cluster i, we have M_i SSUs.

We define the cluster inclusion probabilities of first and second order, respectively, by π_i and π_{ik}. Moreover, we define the inclusion probabilities for the j^{th} and l^{th} second-order units as $\pi_{j|i}$ and $\pi_{jl|i}$, conditionally on the fact that the i^{th} cluster has been selected. The first-order inclusion probability of the j^{th} SSU is therefore

$$\tilde{\pi}_j = \pi_i \pi_{j|i},$$

and the second order inclusion probability is

$$\tilde{\pi}_{jl} = \begin{cases} \pi_i \pi_{jl|i} & \text{if } j, l \in i^{\text{th}} PSU \\ \pi_{ik} \pi_{j|i} \pi_{l|k} & \text{if } j \in i^{\text{th}} PSU \text{ and } k \in l^{\text{th}} PSU \end{cases}$$

The HT estimator for the total Y under cluster sampling is

$$\widehat{Y}_{HT} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{y_{ij}}{\tilde{\pi}_j} = \sum_{i=1}^{n} \frac{1}{\pi_i} \sum_{j=1}^{m_i} \frac{y_{ij}}{\pi_{j|i}} = \underbrace{\sum_{i=1}^{n} \frac{\widehat{Y}_{i\cdot}}{\pi_i}}_{\text{cluster } \widehat{Y}_{HT}\text{'s}} .$$

The HT estimator can be shown to be unbiased with the tower rule of the expectation,

$$\begin{aligned} \mathbb{E}[\widehat{Y}_{HT}] &= \mathbb{E}_2\big[\mathbb{E}_1[\widehat{Y}_{HT}]\big] \\ &= \mathbb{E}_2\Big[\sum_{i=1}^{n} \frac{\mathbb{E}_1[\widehat{Y}_{i\cdot}]}{\pi_i}\Big] \\ &= Y_i. \end{aligned}$$

As for the variance of the HT estimator, we need to consider the total variance,

$$\begin{aligned}\mathbb{V}[\widehat{Y}_{HT}] &= \mathbb{V}_1[\mathbb{E}_2[\widehat{Y}_{HT}]] + \mathbb{V}_2[\mathbb{E}_1[\widehat{Y}_{HT}]] \\ &= \mathbb{V}_1\Big[\sum_{i=1}^{n} \frac{Y_{i\cdot}}{\pi_i}\Big] + \mathbb{E}_1\Big[\sum_{i=1}^{n} \frac{\mathbb{V}_2[\widehat{Y}_{i\cdot}]}{\pi_i^2}\Big]\end{aligned}$$

We can see that the quantities involved are:

$$\mathbb{V}_2[\widehat{Y}_{i\cdot}] = \sum_{j=1}^{M_i} \sum_{l\neq j}^{M_i} (\tilde{\pi}_j\tilde{\pi}_l - \tilde{\pi}_{jl}) \left(\frac{Y_{ij}}{\tilde{\pi}_j} - \frac{Y_{il}}{\tilde{\pi}_l}\right)^2 .$$

11.2 Nonresponse

In a perfect world, we expect that all units that have been selected and are subject to an interview will provide answers without measurement error, i.e. they are completely honest in their responses. Hence, all statistical variability is derived from the *sampling error* induced by the sampling design.

Def. (Nonresponse)
Without loss of generality, we define ***nonresponse*** as any kind of lack of information from the surveyed unit.

Problems When nonresponse is present in a significant way, the researcher loses control on the sampling mechanism.

We can specialize nonresponse as

1. *Unit nonresponse*: the survey unit does not provide any information at all, meaning that the questionnaire form remains completely empty
2. *Item nonresponse*: the participant to the survey answers only some questions, leaving out more personal and sensitive topics.

The main problem of nonresponse is that estimates may be affected by a non-negligible bias: some groups are overrepresented whereas other are underrepresented $\implies$ ***selective nonresponse***.

In order to investigate the impact of non-response, we need to incorporate it in the sampling theory. We can follow two approaches:

1. ***Fixed response model***: we partition the population into two groups U_R of respondents and $U_{\bar{r}}$
2. ***Random response model***: every unit $i \in U$ has an unknown probability to participate if selected in the sample. We define the ***propensity score*** of i as

   $$q_{i|s} = q_i,$$

   and if R_i is the event of response we have

   $$\mathbb{P}(R_i = 1|s) = q_i.$$

The bias in estimating the mean of the population is

$$B(\widehat{Y}_R) = \mathbb{E}[\overline{Y}_R] - Y = \frac{N_{\overline{R}}}{N}(\overline{Y}_R - \overline{Y}_{\overline{R}}),$$

and we can observe that this quantity is driven by two factors:

- difference $\overline{Y}_R - \overline{Y}_{\overline{R}}$ between the two groups;
- the relative size $N_{\overline{R}}/N$ of the non-response group.

Hence, we have identified two ways of reducing the bias. Since in general we have no data for non-respondents, there is no way to evaluate $\overline{Y}_R - \overline{Y}_{\overline{R}}$, and the only way to combat nonresponse is to employ *ad hoc* survey procedures.

Sampling Note that the bias does not depend on the size of s_R.

11.2.1 Horvitz-Thompson and underestimation

In general, the HT estimator produces underestimation in presence of nonresponse, since

$$\widehat{Y}_{HT} = \sum_{i=1}^{N} \frac{Y_i}{\pi_i}\delta_i R_i,$$

and its expected value is

$$\mathbb{E}[\widehat{Y}_{HT}] = \mathbb{E}_p\big[\mathbb{E}_q[\widehat{Y}_{HT}]\big] = \mathbb{E}_p\Big[\sum_{i=1}^{N} \frac{Y_i}{\pi_i}\delta_i \underbrace{\mathbb{E}_q[R_i]}_{q_i}\Big] = \sum_{i=1}^{N} \frac{Y_i}{\pi_i} \underbrace{\mathbb{E}_p[\delta_i]}_{\pi_i} q_i = \sum_{i=1}^{N} Y_i q_i \leq Y.$$

From the above expression, we can try to remove the bias by introducing weights w_i,

$$\widehat{Y}^*_{HT} = \sum_{i=1}^{N} w_i \frac{Y_i}{\pi_i}\delta_i R_i,$$

and the necessary weights are

$$w_i = \frac{1}{\pi_i q_i}.$$

In general, we can find that we have a price to pay in terms of accuracy when accounting for non-response:

$$\mathbb{V}[\widehat{Y}^*_{HT}] = \mathbb{V}[\widehat{Y}_{HT}] + \sum_{i=1}^{N} \frac{Y_i^2}{\varphi_i}(1 - q_i) \geq \mathbb{V}[\widehat{Y}_{HT}].$$

The problem of the adjusted HT estimator is that q_i is unknown in practice.

Propensity score regression The solution is to replace q_i by some estimated response propensity, for instance using a logistic regression with auxiliary variables

$$q_i = \mathbb{P}(R_i = 1|\boldsymbol{X}_i).$$

Weighting-class adjustment Another approach is to apply the *weighting-class adjustment*: we partition unit in C classes, such that within each class we assume respondents and non-respondents to be similar, and we are able to estimate

$$\widehat{q}_c = \frac{\sum_{j \in s_{R,c}} d_j}{\sum_{j \in s_c} d_j}$$

11.2.2 Selective nonresponse

Def. (Selective nonresponse)
Nonresponse is ***selective*** if some groups in the population are under- or over-represented in the sample, and if these groups behave differently from the characteristics being sampled.

Problem Available data is useless for this aim, since they are only available for the respondents and for non-respondents.

Solution We can employ auxiliary variables assuming that these variables are known for each unit of the population.

Assumptions We need assumptions on the mechanisms of non-response and the relationship between variables involved. In particular, we can assume

1. *Missing completely at random* (MCAR): $q_i \perp\!\!\!\perp (Y_i, \boldsymbol{X}_i)$ and the respondents are representative of the population, hence the estimates are unbiased. In this case we can analyze only the units with complete data, and the resulting estimators will be unbiased for the population quantities.
2. *Missing at random* (MAR): $q_i \perp\!\!\!\perp y_i | \boldsymbol{X}_i$, hence we can apply weighting techniques based on $\boldsymbol{x}_i$ to improve the estimate of the quantity of interest.

   Example (MAR)
   The response probability depends on age, sex, . . . which we know for each respondent without knowing the value of y_i. Hence, we can apply a GLM to estimate the probability of inclusion $\widehat{q}_i$ for each unit.

3. *Missing not at random* (MNAR): the response probability only depends on y_i and cannot be explained by the $\boldsymbol{X}_i$'s.

   Example (MNAR)
   Individuals may be more reluctant to disclose the true value of their income depending on whether it is high or low.

The problem is that we cannot distinguish between MNAR and MAR, however we can distinguish between MCAR and MAR by fitting a logistic model to predict the observed probabilities.

References

Hansen, M. H. and Hurwitz, W. N. (1943). «On the Theory of Sampling from Finite Populations». In: *The Annals of Mathematical Statistics* 14.4, 333–362.

Hedayat, A. S. and Sinha, B. K. (1991). *Design and Inference in Finite Population Sampling.* Wiley-Interscience.

Horvitz, D. G. and Thompson, D. J. (1952). «A Generalization of Sampling Without Replacement From a Finite Universe». In: *Journal of the American Statistical Association* 47.260, 663–685.

Pfeffermann, D. (1993). «The Role of Sampling Weights When Modeling Survey Data». In: *International Statistical Review* 61.2, 317–337.

Raj, D. (1968). *Sampling Theory.* McGraw-Hill.

Särndal, C.-E. et al. (2003). *Model Assisted Survey Sampling.* New York Berlin Heidelberg: Springer.

Tille, Y. (2020). *Sampling and Estimation from Finite Populations.* Trans. by I. Hekimi. Hoboken, NJ: John Wiley & Sons Inc.

Valliant, R. et al. (2018). *Practical Tools for Designing and Weighting Survey Samples.* Second. Cham: Springer Nature.

Bayesian Data Analysis and Computation

Lecture 12: Bayesian inference

2022-05-23

12.1 Introduction

In Bayesian statistics, the prior distribution introduces extra-experimental information in the process of statistical inference for the parameters of interest. For this reason there is an ongoing and vivid debate among "subjective" and "objective" Bayesian statisticians, although we will not discuss it at the moment.

Example
Let $\boldsymbol{y} = (y_1, \dots, y_n) \overset{\text{iid}}{\sim} N(\vartheta, 1)$ and two people have different priors:

a) Ann has prior $\vartheta \sim N(\mu, \tau^2)$.

b) Bob has a more diffuse prior $\vartheta \sim N(\mu, 1)$.

Ann. Ann can compute her posterior distribution as

$$\vartheta|\boldsymbol{y} \sim N(\mu^*, \tau^{2*}),$$

where

$$\mu^* = W\mu + (1-W)\bar{y}, \quad \tau^{2*} = \frac{\tau^2}{n\tau^2+1},$$

and $W = 1/(n\tau^2 + 1)$.

Bob. On the other hand, Bob cannot have a closed-form expression for his posterior, since

$$\pi^B(\vartheta|\boldsymbol{y}) \propto \frac{1}{1+(\vartheta-\mu)^2/\tau^2} \exp\left\{-\frac{n}{2}(\bar{y}-\vartheta)^2\right\},$$

for which he cannot compute the normalizing constant.

From a historical perspective, before MCMC methods were available the statistical practice was skewed towards frequentist statistics.

› Frequentist methods are based on optimization of functions.

› Bayesian methods are based on optimization of decision rules, and it is performed through expectation with respect to posterior probability density functions.

› Until the 80's of the last century, optimization was a much easier mathematical task than integration.

› The vast majority of applied statistics was performed in a frequentist way, more suitable for standardized statistical packages.

› Bayesian thinking was merely a philosophical disturbance.

› The usual answer was "We would like to be Bayesian but we have no prior!"

The appearance of simulation methods completely turned around the story. Statisticians realized the huge potential of Markov Chain Monte Carlo methods in 1990 with the Gelfand & Smith JASA paper. In particular, applied Bayesian nonparametric methods, historically too difficult to perform, became often easier than standard parametric methods.

Consequence 1. A first consequence of this shift in statistical practice was that many applied researchers have started to use Bayesian methods, as an alternative and convenient tool to provide estimators, standard errors, etc…

Consequence 2. The choice of the prior is often based on convenience and manageability rather than an important piece of the statistical model.

12.1.1 Starting point

From a purely Bayesian perspective, the role of a prior distribution in a statistical model has been clarified by the celebrated de Finetti's theorem, at least under the exchangeability assumption. In particular, de Finetti (1931) shows that all exchangeable binary sequences are mixtures of Bernoulli sequences:

Def. (Exchangeable sequence)
A binary sequence $X_1, \dots, X_n, \dots$ is exchangeable if and only if there exists a cumulative distribution function F on $[0,1]$ such that for all n

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \int_0^1 \vartheta^{\sum_{i=1}^n x_i}(1-\vartheta)^{n-\sum_{i=1}^n x_i} dF(\vartheta).$$

Theorem 4 (de Finetti)
It further holds that F is the distribution function of the limiting frequency of the exchangeable sequence,

$$F(y) = \mathbb{P}(Y \leq y),$$

where

$$Y = X_\infty = \lim_{n\to\infty} \sum_{i=1}^n X_i/n.$$

Remark. The Bernoulli distribution is obtained by conditioning to the event $Y = \vartheta$,

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n | Y = \vartheta) = \vartheta^{\sum_{i=1}^n x_i}(1-\vartheta)^{n-\sum_{i=1}^n x_i}.$$

Remark. In general, we replace the unknown limiting distribution $F(y)$ with the prior information over the parameter ϑ.

Although this result is valid for Bernoulli random variables, Hewitt and Savage (1955) generalized the result to all random variables over an abstract space $\mathcal{X}$.

Theorem 5 (Generalized de Finetti)
Let $X_1, \dots, X_n, \dots$ be an exchangeable sequence of random variables with values in $\mathcal{X}$. Then there exists a probability measure Q on the set of probability measures $\mathscr{F}(\mathcal{X})$ such that

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \int_{\mathscr{F}} \prod_{i=1}^{n} F(x_i) dQ(F),$$

Remark. It further holds that Q is the limiting distribution function of the empirical distribution function process.

Remark. This theorem is the basis for the application of Bayesian nonparametric models.

As a comparison, we can write the simplest models under the Bayesian and frequentist paradigms:

- Simplest frequentist model: There is an unknown distribution Q so that $X_1, \dots, X_n$ are independent and identically distributed with distribution Q, $Q(A)$ being defined as the limiting proportion of X's in A.
- Simplest Bayesian model: Subjective exchangeable probability distribution Q representing your expectations for the behaviour of $X_1, \dots, X_n$.

12.2 Objective priors

For many years objective priors were called noninformative, which are the simplest ones.

Location. For a location parameter, we use $\pi(\mu) \propto 1$.

Scale. For a scale parameter, we use $\pi(\sigma) \propto \sigma^{-1}$.

Although $\pi(\mu)$ is improper, there are some reasons to justify these results under some conditions. Specifically, the only way to obtain perfect frequentist properties is to use the improper prior.

12.2.1 Jeffreys' prior

This is by far the simplest and most immediate way of obtaining a prior distribution in absence of information. Jeffreys' prior is defined as

$$\pi(\vartheta) \propto \det \mathcal{I}(\vartheta)^{1/2},$$

where $\mathcal{I}(\vartheta)$ is the Fisher information matrix, whose elements are

$$\mathcal{I}_{ij}(\vartheta) = -\mathbb{E}_{\vartheta}\left[\frac{\partial^2}{\partial\vartheta_i \partial\vartheta_j} \log f(Y|\vartheta)\right].$$

Property. The prior is invariant to the choice of parametrization due to the Jacobian, hence it is a 1st order matching prior. Moreover, it is optimal in many senses if the entire vector ϑ is of interest.

Example (Poisson problem)
Let $X_1, \dots, X_n | \vartheta \overset{\text{iid}}{\sim} \text{Pois}(\vartheta)$, then the likelihood is

$$L(\vartheta; x) \propto \exp\{-n\vartheta\} \vartheta^{\sum_{i=1}^n x_i}.$$

For $n = 1$,

$$L(\vartheta; x) \propto \exp\{-\vartheta\} \vartheta^x,$$

hence the Fisher information is

$$-\frac{\partial^2 \ell(\vartheta)}{\partial \vartheta^2} = \frac{x}{\vartheta^2},$$

and then $\mathcal{I}(\vartheta) = \mathbb{E}_\vartheta[X]/\vartheta^2 = 1/\vartheta$. Therefore,

$$\pi(\vartheta) \propto \frac{1}{\sqrt{\vartheta}},$$

and the posterior distribution is

$$\pi(\vartheta | x) \propto \vartheta^{t-1/2} \exp\{-n\vartheta\}.$$

that is,

$$\vartheta | X \sim \text{Gamma}(n, t + 1/2).$$

Probability matching. The Jeffreys' prior yields a coverage probability of the resulting Bayesian one-sided credible interval which matches asymptotically the coverage probability of the corresponding frequentist confidence interval.

Problems. It has problems in the multidimensional case. Loosely speaking, if there is a parameter of interest then the best prior depends on which of them is the nuisance one.

Example (Problems)
Suppose $X_i \sim N(\mu_i, 1)$ for $i = 1, \dots, p$ and the parameter of interest is

$$\vartheta = \frac{1}{p} \sum_{i=1}^n \mu_i^2 = \frac{1}{p} \|\mu\|_2^2.$$

The Fisher matrix is diagonal and the diagonal elements do not depend on the ϑ's, thus

$$\pi(\mu_1, \dots, \mu_p) \propto 1.$$

Therefore, we obtain

$$(\mu_1, \dots, \mu_p) | X \sim N_p(x, 1),$$

which in turn implies that

$$p \cdot \vartheta \sim \chi_p^2 \left(\sum_{i=1}^p x_i^2 \right),$$

where $\sum_{i=1}^p x_i^2$ is the non-centrality parameter.

Problem. This is NOT a "good" posterior distribution for ϑ due to Stein's Paradox, since

$$\lim_{p\to\infty} \mathbb{E}_\pi[\vartheta|x] - \vartheta = 2.$$

The same happens for the posterior mode or median.

Most important issue. The Jeffreys' method seeks for the noninformative prior for the entire vector of the parameters. If the parameter of interest is just a function of it, say ϑ, this introduces a "bias" into the procedure.

12.2.2 Matching prior

Another way of defining the "noninformativeness" of a prior distribution is to do so is in terms of frequentist coverage.

Idea. A noninformative prior should provide inferences similar to the classical frequentist methods.

Def. (Frequentist coverage probability (FCP))
We define the frequentist coverage probability (FCP) as

$$\pi(\cdot) \to C_\pi(X, 1-\alpha),$$

where C is the one-sided credible interval. The FCP is defined as

$$\mathbb{P}(\vartheta \in C_\pi(X, 1-\alpha)|\vartheta).$$

Def. (Matching prior)
A prior π is called a matching prior of order γ if

$$\mathbb{P}(\vartheta \in C_\pi(X, 1-\alpha)|\vartheta) = 1-\alpha + O(n^{-\gamma/2})$$

Remark. The idea is that a matching prior produces inference procedures that substantially agree with frequentist methods.

12.2.3 Reference prior

The reference prior is the one which maximizes the expected (with respect to the sampling distribution) Kullback-Leibler divergence between the prior and the posterior distribution. Specifically, if the KL divergence is defined as

$$D_{\mathrm{KL}}(p, q) = \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx,$$

which is zero if and only if $q \overset{\text{a.s.}}{=} p$. then we want to find

$$\operatorname*{argmin}_\pi \mathbb{E}_X \left[D_{\mathrm{KL}}\big(\pi(\vartheta|\boldsymbol{x}), \pi(\vartheta)\big)\right].$$

This is a variational problem, for which we obtain a solution using various approximations.

Idea. We use a sequence of Jeffreys' prior one component at a time to minimize the KL divergence.

Def. (Entropy)
The entropy of a probability measure π defined on a probability space Ω is defined as

$$\mathcal{E} = -\int_\Omega \pi(\omega) \log \pi(\omega) d\omega.$$

Given the experiment $E_k = (\mathcal{X}_k, \Omega, \mathscr{P})$, one can define the "information contained in E_k" with respect to a prior π as

$$I_{E_k}(\pi) = \int_{\mathcal{X}_k} \int_\Omega m(x_k) \pi(\omega|x_k) \log \frac{\pi(\omega|x_k)}{\omega(\pi)} d\omega dx_k.$$

Here, $m(x)$ is the marginal distribution of X. The reference prior makes $k \to \infty$ and tries to find the $\pi(\omega)$ that maximizes such information.

Example (Trinomial model)
Consider a trinomial model with probabilities $(\omega_1, \omega_2, 1 - \omega_1 - \omega_2)$. Then, imagine that ω_1 is the parameter of interest but we have to deal with $\lambda = \omega_2$. Then,

$$\mathcal{I}(\vartheta, \lambda) = \frac{1}{1 - \vartheta - \lambda} \cdot \begin{pmatrix} \frac{1-\lambda}{\vartheta} & 1 \\ 1 & \frac{1-\vartheta}{\lambda} \end{pmatrix},$$

and thus the Jeffreys' prior is the Dirichlet distribution with $\boldsymbol{\alpha} = (1/2, 1/2, 1/2)$:

$$\pi(\vartheta, \lambda) \propto \frac{1}{\sqrt{\vartheta \lambda (1 - \vartheta - \lambda)}}.$$

The reference prior is instead the one that we would get by considering the counts in the second and third cells together,

$$\pi^r(\vartheta, \lambda) \propto \frac{1}{\sqrt{\vartheta \lambda (1 - \vartheta)(1 - \vartheta - \lambda)}}.$$

Remark. The problem is the dilemma between exchangeability (Jeffreys' prior) and marginalization (reference prior).

12.2.4 Discrete parameter space

Many problems are discrete by nature, for instance:

› choice of the number of mixture components;

› choosing between candidate models;

› ...

For Θ discrete, methods as Jeffreys' prior or reference priors do not work or give non-sensible results.

Villa and Walker approach We assign a worth to each $\vartheta \in \Theta$ by objectively measuring what is lost if ϑ is removed from the parameter space and it is the true value. This loss may be formalized as (Merhav and Feder, 1998)

$$\text{Loss}(\vartheta) = -\log \pi(\vartheta) \iff \text{Utility}(\vartheta) = \log \pi(\vartheta).$$

The formal definition for the prior can be obtained as

$$\pi(\vartheta) \propto \exp\left\{ \min_{\vartheta' \neq \vartheta} D_{\text{KL}}\big(f(\cdot|\vartheta)||f(\cdot|\vartheta')\big) \right\} - 1.$$

12.2.5 Current state of practical objective Bayesian analysis

Ad hoc objective Bayesian analysis can be successful, especially if validated by experience or extensive sensitivity studies. However, Gelman is a strong supporter of the so-called weakly informative priors since a noninformative prior may give strong information to unlikely values of the parameter.
A constant prior can yield improper posteriors or can swamp the data in high dimensions:

› the posterior may be improper;

› the posterior may be dominated by the prior;

› the prior is not vanishing.

12.3 Modern statistical practice

Modern statistical practice usually faces problems for which objective priors are hard to compute or even virtually impossible to obtain. Nevertheless a large body of research has been devoted over the recent years to develop default priors for high dimensional models, typically relying on the notion of sparsity:

› spike-and-slab priors (Ishwaran and Rao, 2005; Ročková and George, 2014), such as

$$\pi(\vartheta_i) = \gamma\delta_{\vartheta^*}(\vartheta) + (1-\gamma)g(\vartheta),$$

where $\delta_z(\cdot)$ is the Dirac delta centered in z and $g(\cdot)$ is a diffuse proper prior. Practical use of those prior is not easy from a computational perspective, due to the presence of discrete components and a huge model space.

› horseshoe priors (Carvalho et al., 2010), which approximate the spike using a mixture of distributions:

$$\begin{aligned} \vartheta_i|\lambda_i,\tau &\overset{\text{iid}}{\sim} N(0,\lambda_i^2) && i=1,\dots,p \\ \lambda_i|\tau &\overset{\text{iid}}{\sim} \text{Cauchy}(0,\tau), && i=1,\dots,p \\ \tau &\overset{\text{iid}}{\sim} \text{Cauchy}^+(0,1) \end{aligned}$$

By reparametrizing using $\kappa_i = (1+\tau^2\lambda_i^2)^{-1}$, marginally we have

$$\vartheta_i|\kappa_i,\tau \sim N\left(0, \frac{1-\kappa_i}{\kappa_i}\right).$$

Hence, $\kappa_i \in [0,1]$ is a local shrinkage factor for the i^{th} component of the model. Polson and Scott (2012) called this class of priors the global-local shrinkage priors.

- in the horseshoe prior, the dependence between of λ_i's cannot be too large. Therefore, **bhattacharya2014** proposed the Dirichlet-Laplace prior

$$\vartheta_j \overset{\text{ind}}{\sim}_t DE(\tau, \phi_j), \phi \sim \text{Dir}(a, \dots, a), \tau \sim g(\tau),$$

which yields a much stronger dependence.

- Another approach is when we want use a model $f(x|\vartheta)$ where $f(x|\vartheta_0)$ gives a simpler version of the model. Hence, we consider a function of the KL distance of $f(x|\vartheta)$ from $f(x|\vartheta_0)$,

$$d(\vartheta) = \sqrt{2\text{KL}\big(f(x|\vartheta)|f(x|\vartheta_0)|\big)}.$$

Thus one might construct a prior for $d(\vartheta)$ using an exponential prior

$$\pi(d(\vartheta)) = \lambda e^{-\lambda d(\vartheta)},$$

where λ determines the thickness of the tail.

Remark. The performance of high-dimensional default priors is typically measured in terms of frequentist and asymptotic properties of the posterior estimators. For instance, one could estimate posterior concentration on the true value of the parameter as the sample size increases.

Lecture 13: Monte Carlo methods

2022-05-25

13.1 Introduction

Bayesian methods require the computation of integrals both to normalize posterior distributions and to evaluate posterior summaries. When integrals do not have a closed form expression, numerical integration and asymptotic approximations suffer from the curse of dimensionality. Markov Chain Monte Carlo methods sample dependent draws from a Markov chain whose limiting distribution is the posterior distribution.

Suppose that we wish to calculate

$$I = \int_{\mathcal{X}} h(x) f(x) dx,$$

where $f(x)$ is a density and $h(x)$ is a function of x. For instance,

› $h(x) = x \implies \mathbb{E}[X]$

› $h(x) = x^2 \implies \mathbb{E}[X^2]$

› $h(x) = \mathbb{1}_A(x) \implies \mathbb{P}[X \in A]$

If $|I| < \infty$ and $X_1, \ldots, X_T \overset{\text{iid}}{\sim} f$, then by the strong law of large numbers the empirical mean is consistent for I,

$$\widehat{I} = \frac{1}{T}\sum_{i=1}^{T} h(x_i) \xrightarrow{\text{a.s.}} \mathbb{E}_f[h(x)] \quad \text{as } T \to \infty.$$

Moreover, the variance of $\widehat{I}$ can be estimated by

$$\widehat{V} = \frac{1}{T^2}\sum_{i=1}^{T} \left\{h(x_i) - \widehat{I}\right\}^2,$$

and when T is large it is approximately true that

$$\frac{\widehat{I} - \mathbb{E}_f[h(X)]}{\sqrt{\widehat{V}}} \mathrel{\dot\sim} \mathcal{N}(0,1),$$

hence we can obtain a confidence interval for I.

Remark. We need to sample from f in order to obtain the numerical approximation.

13.2 Monte Carlo sampling

13.2.1 Accept-reject methods

Alternative technique for computing integrals when it is impossible to directly sample from the target density. Suppose that we need to evaluate

$$I = \int g(\vartheta)\pi(\vartheta) d\vartheta,$$

and we cannot sample from $\pi(\vartheta)$ which is continuous and possibly unnormalized,

$$\pi(\vartheta) = f(\vartheta)/K.$$

However, we can sample from another function $h(\vartheta)$ and it holds that

$$f(\vartheta) < c \cdot h(\vartheta).$$

Then, the accept-reject algorithm is able to generate values from $\pi(\vartheta)$ by alternating the following steps:

1. Draw a candidate $w \sim h(\vartheta)$ and a value $u \sim \text{Unif}(0, 1)$.
2. If
$$u \leq \frac{f(w)}{c \cdot h(w)},$$
then set $\vartheta = w$, otherwise reject the candidate and go back to step 1.

Theorem 6 (Accept-reject)
It holds that

a) he distribution of the accepted values generated from the previous algorithm is exactly the target density $\pi(\vartheta)$;

b) the marginal probability that a single candidate w is accepted is K/c.

Remark. When K is unknown, we must choose c such that $f(\vartheta) < c \cdot h(\vartheta)$ for all ϑ, then

$$c = \sup_{\vartheta} \frac{f(\vartheta)}{h(\vartheta)}.$$

13.2.2 Importance sampling

Importance sampling is based on the following representation for the integral of interest,

$$\begin{aligned} I &= \int_{\mathcal{X}} h(x)f(x)dx = \mathbb{E}_f[h(X)] \\ &= \int_{\mathcal{X}} h(x)\frac{f(x)}{g(x)}g(x)dx = \mathbb{E}_g[h(X)\frac{f(X)}{g(X)}], \end{aligned}$$

where g is an arbitrary density whose support contains the support of f. Then, the integral can be estimated using

$$\tilde{I} = \frac{1}{T}\sum_{i=1}^{T} h(x_i)\frac{f(x_i)}{g(x_i)} = \sum_{i=1}^{T} h(x_i)w(x_i),$$

where $w(x) = f(x)/g(x)$ is called the importance function.

$$\mathbb{V}[\widehat{I}] = T^{-1} \int \{h(x) - I\}^2 f(x)dx$$

$$\mathbb{V}[\tilde{I}] = T^{-1} \int \left\{h(x)\frac{f(x)}{g(x)} - I\right\}^2 g(x)dx$$

and one can work on g to minimize the variance of $\tilde{I}$.

Remark. Importance Sampling variance is finite only when

$$\mathbb{E}_g[h(X)^2\frac{f(X)^2}{g(X)^2}] = \int h(X)^2\frac{f(X)^2}{g(X)}dx < \infty, \tag{15}$$

hence densities g with lighter tails than f are not good proposals.

Remark. Note that since (15) can be rewritten as

$$\int h(X)^2\frac{f(X)}{g(X)}f(x)dx,$$

the ratio $f(x)/g(x)$ must be bounded when $f(x)$ is non negligible. This means that the modes of $f(x)$ and $g(x)$ should be close to each other.

Both the estimators

$$\widehat{\mu} = \frac{\sum_i w_i h(X_i)}{\sum_i w_i}$$

$$\tilde{\mu} = \frac{\sum_i w_i h(X_i)}{n}$$

and $\widehat{\mu}$ is called self-normalized estimator.

Theorem 7 (Bias and variance of IS)
For the estimators $\widehat{\mu}$ and $\tilde{\mu}$ we have that, respectively,

$$\mathbb{E}[\tilde{\mu}] = \mu$$

$$\mathbb{V}[\tilde{\mu}] = \frac{1}{n}\,\mathbb{V}[w(X)h(X)]$$

and

$$\mathbb{E}[\widehat{\mu}] = \mu + \frac{1}{n}\big(\mu\,\mathbb{V}[w(X)] - \mathrm{Cov}\,\big(w(X), w(X)h(X)\big)\big) + O(n^{-2})$$

$$\mathbb{V}[\tilde{\mu}] = \frac{1}{n}\,\mathbb{V}[w(X)h(X)] - 2\mu\,\mathrm{Cov}\,\big(w(X), w(X)h(X)\big) + \mu^2\,\mathbb{V}\,\big(w(X)\big) + O(n^{-2}),$$

where all expectations are taken with respect to g.

13.2.3 Sampling Importance Resampling

The idea is to start from a proposal distribution $g(\vartheta)$ and to convert it into a sample from $\pi(\vartheta|y)$. In order to do so, we can apply the following algorithm:

1. For each $j = 1, \ldots, J$ we compute weights

$$\psi_j = \frac{\pi(\vartheta_j|y)}{g(\vartheta_j)},$$

   and normalize them using

$$w_j = \frac{\psi_j}{\sum_h \psi_h}.$$

2. We draw a new sample $\vartheta_1^*, \ldots, \vartheta_J^*$ without replacement from the discrete distribution $\vartheta_1, \ldots, \vartheta_J$ using weights $w_1, \ldots, w_J$.

Theorem 8 (Sampling importance Resampling)
The sample of the resampled ϑ's is approximately a sample from $\pi(\vartheta|y)$.

Proof.
In the univariate case,

$$\begin{aligned}
\mathbb{P}(\vartheta^* \leq a) &= \sum_{i=1}^{J} w_i \mathbb{1}_{(-\infty,a]}(\vartheta_i) \\
&= \frac{n^{-1}\sum_i \psi_i \mathbb{1}_{(-\infty,a]}(\vartheta_i)}{n^{-1}\sum_i \psi_i} \\
&\to \frac{\mathbb{E}_g[\pi(\vartheta|y)]/g(\vartheta)\mathbb{1}_{(-\infty,a]}(\vartheta)}{\mathbb{E}_g[\pi(\vartheta|y)]/g(\vartheta)} \\
&= \int_{-\infty}^{a} \pi(\vartheta|y)d\vartheta.
\end{aligned}$$

□

Remark. The size J of the resampled values can be as large as desired.

Remark. The more g resembles $\pi(\vartheta|y)$, the smaller the sample size is needed to approximate the target distribution.

Prop. 1 (Estimator for the normalizing constant)
A consistent estimator for the normalizing constant of the density is

$$\widehat{I} = \frac{1}{J}\sum_{j=1}^{J} \psi_j.$$

Proof.

$$m(y) = \int \pi(\vartheta|y)d\vartheta = \int \frac{\pi(\vartheta|y)}{g(\vartheta)}g(\vartheta)d\vartheta \approx \frac{1}{J}\sum_{j=1}^{J}\frac{\pi(\vartheta_j|y)}{g(\vartheta_j)} = \frac{1}{J}\sum_{j=1}^{J}\psi_j.$$

□

Sequential Monte Carlo (SMC). The SIR methodology can be extended to sample from the posterior distribution when the latter evolves over time, such as time series models and online monitoring.

13.3 Markov Chain Monte Carlo

Define an invariant distribution π to be some distribution such that $\pi = \pi P$. In our case, the target distribution is $\pi(\vartheta|y)$ and we want to devise a Markov chain so that the posterior distribution is the stationary distribution.

Theorem 9 (Ergodic)
If $\{\xi_i, i \in \mathbb{N}\}$ is an irreducible, recurrent, with state space E, $\mathbb{R}^d$-valued Markov chain which admits a stationary distributionπ, then for any integrable function $f : E \to \mathbb{R}$ it holds that

$$\lim_{t\to\infty} \frac{1}{T}\sum_{i=1}^{T} f(\xi_i) \xrightarrow{a.s.} \int_{\mathbb{R}} f(x)\pi(x)dx,$$

for almost every starting value x of the chain.

Remark. The Ergodic Theorem is the Markov-Chain analogue to the SLLN. It allows one to ignore the dependence between draws of the Markov chain when we calculate quantities of interest from the posterior draws.

Moreover, the asymptotic variance of a MCMC estimator is approximately

$$\mathbb{V}[f(\widehat{X})] = \frac{\sigma^2}{n}\left(1 + 2\sum_i \rho_i\right),$$

where ρ_i is the i^{th} lag autocorrelation of the sequence of the $f(\xi_i)$'s.

13.3.1 Gibbs sampling

We can use the Gibbs sampler to sample from the joint posterior distribution if we know the full conditional distributions for each parameter. For each parameter, the full conditional is the distribution of each single component of the parameter vector, conditional on the known information and all the other parameters,

$$p(\vartheta_j|\vartheta_{-j}, y) = p(\vartheta_j|\vartheta_1, \dots, \vartheta_{j-1}, \vartheta_{j+1}, \dots, \vartheta_J, y).$$

Theorem 10 (Hammersley-Clifford in 2 dimensions)
Suppose we have a joint density $f(x, y)$, then we can write the joint density in terms of the conditional densities $f(x|y)$ and $f(y|x)$ as

$$f(x, y) = \frac{f(y|x)}{\int \frac{f(y|x)}{f(x|y)} dx}.$$

Proof.
The denominator can be written as

$$\int \frac{f(y|x)}{f(x|y)} dx = \frac{\int \frac{f(x,y)}{f(x)} dx}{\int \frac{f(x,y)}{f(y)} dx} = \int \frac{f(y)}{f(x)} dy = \frac{1}{f(x)}.$$

□

Def. (Positivity condition)
A distribution with density $f(x_1, ..., x_p)$ and marginal densities $f_{X_i}(x_i)$ is said to satisfy the positivity condition if

$$f(x_1, \dots, x_p) > 0$$

for all $x_1, \dots, x_p$ such that $f_{X_i}(x_i) > 0$.

Theorem 11 (Hammersley-Clifford)
Let $(X_1, \dots, X_p)$ satisfy the postiivity condition and have joint density $f(x_1, \dots, x_p)$. Then for all $(fx_1, \dots, \xi_p) \in \operatorname{supp}(f)$ we have

$$f(x_1, \dots, x_p) = \prod_{j=1}^{p} \frac{f_{X_j|X_{-j}}(x_j|x_1, \dots, x_{j-1}, \xi_{j+1}, \dots, \xi_p)}{f_{\xi_j|X_{-j}}(\xi_j|x_1, \dots, x_{j-1}, \xi_{j+1}, \dots, \xi_p)}$$

Remark. Note that the theorem does not guarantee the existence of a joint distribution for every set of full conditionals. Therefore, a Gibbs sampler can be used only when the existence of a joint distribution has already been established.

13.3.2 Metropolis-Hastings

If all else fails, we can use the Metropolis-Hastings algorithm, which is guaranteed to always work. The problem, however, is that we need to implement a carefully-designed proposal distribution in order to obtain the posterior distribution.

Algorithm 1 Metropolis-Hastings

```
Choose a starting value $\vartheta^{(0)}$
At iteration $t$, draw a candidate $\vartheta^*$ from a proposal distribution $q_t(\vartheta^*|\vartheta^{(t-1)})$.
Compute an acceptance ratio
    $$\alpha = \min\left\{1, \frac{\pi(\vartheta^*|y)q_t\big(\vartheta^{(t-1)}\vartheta^*\big)}{\pi(\vartheta^{(t-1)}|y)q_t(\vartheta^*|\vartheta^{(t-1)})}\right\}.$$
Accept $\vartheta^{(t)} \longleftarrow \vartheta^*$ with probability $\alpha$, otherwise $\vartheta^{(t)} \longleftarrow \vartheta^{(t-1)}$.
```

Remark. The proposal distribution $q_t(\vartheta^*|\vartheta^{(t-1)})$ determines where we move to in the next iteration of the Markov chain (analogous to the transition kernel). The correction term in the calculation of α is needed when the distribution q is not symmetric. Moreover, the support of the proposal distribution must contain the support of the posterior.

Proportionality. Since α is a ratio, $m(y)$ cancels out in both the numerator and denominator and thus we only need $\pi(\vartheta|y)$ up to a constant of proportionality.

Acceptance. Using theoretical arguments, the optimal acceptance rate α using a normal random walk proposal $\mathcal{N}(\vartheta^{(t-1)}, \sigma_q^2)$ can be shown to be in $[0.25, 0.45]$.

13.3.3 Rao-Blackwellization

Suppose that $(\vartheta_1, \dots, \vartheta_n)$ is the parameter and that $h(\vartheta_1)$ is our function of interest. Then, the naive approach for estimating $h(\vartheta_1)$ would be to use the first element of the output from a Gibbs sampler and use

$$\delta_0 = \frac{1}{T}\sum_{t=1}^{T} h(\vartheta_1^{(t)}) \xrightarrow{T\to\infty} \int h(\vartheta_1)\pi(\vartheta_1)d\vartheta_1,$$

which is unbiased for $h(\vartheta_1)$.

The Rao-Blackwellization procedure replaces δ_0 with its conditional expectation

$$\delta_{\text{RB}} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[h(\vartheta_1)|\vartheta_2^{(t)}, \dots, \vartheta_p^{(t)}\right].$$

Corollary 1 (Rao-Blackwellization)
Due to Rao-Blackwell theorem, we have that

- *Both* δ_0 *and* δ_{RB} *converge to* $\mathbb{E}[h(\vartheta_1)]$.
- *Both estimators are unbiased.*
- $\mathbb{V}[\delta_{RB}] \leq \mathbb{V}[\delta_0]$.

Remark. This implies that δ_{RB} is uniformly better than δ_0.
Another substantial benefit of RB is in the approximation of densities of different components of ϑ without using nonparametric density estimation methods.

Lemma 1 (Conditional convergence)
The estimator

$$\frac{1}{T}\sum_{t=1}^{T} f_i(\vartheta_i|\vartheta_j^{(t)}, j \neq i) \xrightarrow{T\to\infty} f_i(\vartheta_i),$$

and it is unbiased.

13.4 Convergence diagnostics

Recall that MCMC is an iterative procedure, such that it converges to the target distribution as $t \longrightarrow \infty$ for any starting value $\vartheta^{(0)}$. However, the early samples are strongly influenced by the distribution $\vartheta^{(0)}$, which is presumably not drawn from $\pi(\vartheta|y)$.

13.4.1 Geweke diagnostic

Geweke (1992) proposed a convergence test based on a time-series analysis approach. Informally, if the chain has reached convergence then statistics calculated over different portions of the chain should be close to each other.

By default, we select the first 10% and last 50% of the chain, and calculate the mean over these two sets. If the samples are drawn from the stationary distribution of the chain, the two means are equal and Geweke's test statistic has an asymptotically standard normal distribution.

13.4.2 Gelman & Rubin test

Gelman and Rubin (1992) proposed a convergence test based on output from two or more multiple runs of the MCMC simulation, where the chains are started from different initial over-dispersed values relative to the posterior distribution.

The method compares the within and between chain variances for each variable. When the chains have "mixed" (converged) then the variance within each sequence and the variance between sequences for each variable will be roughly equal.

$$B = \frac{n}{J-1}\sum_j (\bar{\eta}_j - \bar{\eta})^2,$$

where

$$\bar{\eta}_j = \frac{1}{n}\sum_i \eta_{ij}, \quad \bar{\eta} = \frac{1}{J}\sum_j \bar{\eta}_j,$$

and the within variance is

$$W = \frac{1}{J(n-1)}\sum_i \sum_j (\eta_{ij} - \bar{\eta}_j)^2.$$

An unbiased estimator of $\mathbb{V}[\eta|y]$ is the weighted average

$$\widehat{\mathbb{V}}[\eta|y] = \frac{n-1}{n}W + \frac{1}{n}B,$$

from which we can obtain a factor

$$\widehat{R} = \sqrt{\frac{\widehat{\mathbb{V}}[\eta|y]}{W}} \approx \left(1 + \frac{B}{nW}\right).$$

Remark. When $\widehat{R}$ is high (perhaps greater than 1.1 or 1.2), then we should run our chains out longer to improve convergence to the stationary distribution.

Remark. There is an improved $\widehat{R}$ version which has been introduced by Vehtari et al. (2020).

13.4.3 Effective sample size

Since the valued are autocorrelated, we can reduce the sample size to what we would observe compared to i.i.d samples from the posterior. The effective sample size (ESS) is defined as

$$\text{ESS} = \frac{T}{\left(1 + 2\sum_{j=1}^{k}\rho(j)\right)},$$

and the closer it is to T the better it is.

Lecture 14: Bayesian linear model

2022-06-06

Bayesian inference implies the use of a prior distribution for (β, σ^2). Because of practical considerations, we can use a conjugate prior to use a conjugate prior of the form

$$\pi(\beta, \sigma^2) = \pi(\beta|\sigma^2)\pi(\sigma^2),$$

where

$$\begin{aligned}\beta|\sigma^2 &\sim N(\beta_0, \sigma^2 V_0),\\ \sigma^2 &\sim \mathrm{IGamma}(c_0/2, d_0/2).\end{aligned}$$

Since the likelihood can be written as

$$\begin{aligned}L(\beta, \sigma^2; \boldsymbol{y}) &\propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}(y-\widehat{y})^\top(y-\widehat{y}) + (\beta-\widehat{\beta})^\top X^\top X(\beta-\widehat{\beta})\right\}\\ &\propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(nS^2 + (\beta-\widehat{\beta})^\top X^\top X(\beta-\widehat{\beta})\right)\right\},\end{aligned}$$

where $nS^2 = (\boldsymbol{y}-\widehat{\boldsymbol{y}})^\top(\boldsymbol{y}-\widehat{\boldsymbol{y}})$, we can apply the prior distribution in order to show that it is conjugate.

Theorem 12 (Quadratic form equivalence)
Let $\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^k$ and let $A, B \in \mathbb{R}^{k\times k}$ be symmetric matrices such that $(A+B)^{-1}$ exists. Then,

$$(\boldsymbol{x}-\boldsymbol{a})^\top A(\boldsymbol{x}-\boldsymbol{a}) + (\boldsymbol{x}-\boldsymbol{b})B(\boldsymbol{x}-\boldsymbol{b}) = (\boldsymbol{x}-\boldsymbol{c})^\top(A+B)(\boldsymbol{x}-\boldsymbol{c}) + (\boldsymbol{a}-\boldsymbol{b})^\top A(A+B)^{-1}B(\boldsymbol{a}-\boldsymbol{b}),$$

where $\boldsymbol{c} = (A+B)^{-1}(A\boldsymbol{a} + B\boldsymbol{b})$.

Using the above theorem, we have that the posterior distribution is

$$\pi(\beta, \sigma^2|\boldsymbol{y}) \propto \frac{1}{(\sigma^2)^{n/2+c_0/2+p/2+1}} \exp\left\{-\frac{1}{2\sigma^2}\left(nS^2 + d_0 + Q(\beta)\right)\right\},$$

where

$$Q(\beta) = (\beta-\widehat{\beta})X^\top X(\beta-\widehat{\beta}) + (\beta-\beta_0)^\top V_0^{-1}(\beta-\beta_0).$$

Theorem 13 (Dickey)

Let $\boldsymbol{X}$ be a k-dimensional random vector and Y be a scalar random variable such that

$$\boldsymbol{X}|Y \sim N(\mu, Y\Psi), \quad Y \sim IGamma(a, b),$$

then ht marginal distribution of $\boldsymbol{X}$ is multivariate Student

$$\boldsymbol{X} \sim St_k\left(2a, \mu, \frac{b}{a}\Psi\right).$$

Hence, we can write the posterior distribution of the β vector as

$$\beta \sim \mathrm{St}_p\left(c^*, \tilde{\beta}, \frac{d^*}{c^*}\tilde{V}\right).$$

Finally, the full conditional distribution of $\pi(\sigma^2|\beta, y)$ is

$$\sigma^2|\beta, y \sim \mathrm{IGamma}\left(\frac{n + c_0 + p}{2}, \frac{d_0 + k\tilde{S}^2 + Q(\beta)}{2}\right).$$

Def. (Zellner's *g*-prior)

We define Zellner's *g*-prior as the prior distribution that places the prior distribution of β as

$$\beta|\sigma^2 \sim N(\beta_0, g\sigma^2 \cdot (X^\top X)^{-1}),$$

where g is the only hyperparameter of choice.

References

Carvalho, C. M. et al. (2010). «The Horseshoe Estimator for Sparse Signals». In: *Biometrika* 97.2, 465–480.

de Finetti, B. (1931). «Sul significato soggettivo della probabilità». In: *Fundamenta Mathematicae* 17.1, 298–329.

Gelman, A. and Rubin, D. B. (1992). «Inference from Iterative Simulation Using Multiple Sequences». In: *Statistical Science* 7.4, 457–472.

Hewitt, E. and Savage, L. J. (1955). «Symmetric Measures on Cartesian Products». In: *Transactions of the American Mathematical Society* 80.2, 470–501.

Ishwaran, H. and Rao, J. S. (2005). «Spike and Slab Variable Selection: Frequentist and Bayesian Strategies». In: *The Annals of Statistics* 33.2, 730–773.

Merhav, N. and Feder, M. (1998). «Universal Prediction». In: *IEEE Transactions on Information Theory* 44.6, 2124–2147.

Polson, N. G. and Scott, J. G. (2012). «Local Shrinkage Rules, Lévy Processes and Regularized Regression». In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.2, 287–311.

Ročková, V. and George, E. I. (2014). «EMVS: The EM Approach to Bayesian Variable Selection». In: *Journal of the American Statistical Association* 109.506, 828–846.

Vehtari, A. et al. (2020). «Rank-Normalization, Folding, and Localization: An Improved $\widehat{R$ for Assessing Convergence of MCMC». In: *arXiv:1903.08008 [stat]*. arXiv: 1903.08008 [stat].