Time Series Modelling *with* Unobserved Components

Matteo M. Pelagatti

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Time Series Modelling
with
Unobserved Components

Time Series Modelling
with
Unobserved Components

Matteo M. Pelagatti
University of Milano-Bicocca, Italy


CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20150602

International Standard Book Number-13: 978-1-4822-2501-3 (eBook - PDF)

To Antje and Julian

Contents

List of figures **xi**

List of symbols **xiii**

Preface **xv**

I Statistical prediction and time series **1**

1 Statistical Prediction **3**
1.1 Optimal predictor 4
1.2 Optimal linear predictor 6
1.3 Linear models and joint normality 12

2 Time Series Concepts **15**
2.1 Definitions 15
2.2 Stationary processes 17
2.3 Integrated processes 29
2.4 ARIMA models 35
2.5 Multivariate extensions 44

II Unobserved components **49**

3 Unobserved Components Model **51**
3.1 The unobserved components model 51
3.2 Trend 53
3.2.1 Genesis 53
3.2.2 Properties 54
3.2.3 Relation with Hodrick–Prescott filtering and cubic splines 56
3.3 Cycle 59
3.3.1 Genesis 59
3.3.2 Properties 61
3.3.3 A smooth extension 64
3.4 Seasonality 67

3.4.1 Genesis 68
3.4.2 Properties 69

4 Regressors and Interventions 73
4.1 Static regression 73
4.2 Regressors in components and dynamic regression 79
4.3 Regression with time-varying coefficients 86

5 Estimation 91
5.1 The state space form 91
5.2 Models in state space form 93
5.2.1 ARIMA processes in state space form 93
5.2.2 UCM in state space form 99
5.2.2.1 Local linear trend 99
5.2.2.2 Stochastic cycle 100
5.2.2.3 Trigonometric seasonal component 101
5.2.2.4 Stochastic dummy seasonal component 101
5.2.2.5 Building UCM in state space form 102
5.2.3 Regression models in state space form 105
5.2.4 Putting the pieces together 109
5.3 Inference for the unobserved components 111
5.3.1 The Kalman filter 112
5.3.2 Smoothing 118
5.3.3 Forecasting 124
5.3.4 Initialisation of the state space recursion 125
5.3.5 Adding the Gaussianity assumption 127
5.4 Inference for the unknown parameters 127
5.4.1 Building and maximising the likelihood function 128
5.4.2 Large sample properties of the maximum likelihood estimator 129
5.4.3 Coding the maximum likelihood estimator in a model in state space form 135

6 Modelling 137
6.1 Transforms 137
6.2 Choosing the components 142
6.3 State space form and estimation 148
6.4 Diagnostics checks, outliers and structural breaks 154
6.5 Model selection 158

7 Multivariate Models 165
7.1 Trends 165
7.2 Cycles 170
7.3 Seasonalities 173
7.4 State space form and parametrisation 175

III Applications 177

8 Business Cycle Analysis with UCM 179
8.1 Introduction to the spectral analysis of time series 179
8.2 Extracting the business cycle from one time series 188
8.3 Extracting the business cycle from a pool of time series 194

9 Case Studies 199
9.1 Impact of the point system on road injuries in Italy 199
9.2 An example of benchmarking: Building monthly GDP data 203
9.3 Hourly electricity demand 208

10 Software for UCM 215
10.1 Software with ready-to-use UCM procedures 215
10.1.1 SAS/ETS 216
10.1.2 STAMP 219
10.1.3 Stata 226
10.2 Software for generic models in state space form 230
10.2.1 EViews 230
10.2.2 Gretl 233
10.2.3 Ox/SsfPack 236
10.2.4 R 241
10.2.5 Stata 244

Bibliography 247

Index 255

List of figures

2.1 Monthly time series of international airline passengers. 30
2.2 Sample paths of the processes Y_t and W_t. 34
2.3 Box–Jenkins strategy for ARIMA model identification. 43
2.4 Sample paths of cointegrated time series. 47

3.1 Log-airline time series and deterministic components. 52
3.2 Sample path of a local linear trend and its embedded trends. 55
3.3 European Union real GDP and HP-filter-based trend. 57
3.4 Deterministic function plus noise and UCM-based spline. 58
3.5 Sinusoids at different frequencies and phases. 59
3.6 Geometric visualisation of a sinusoid at frequency λ. 60
3.7 Geometric representations of one stochastic cycle step. 62
3.8 Sample path of a stochastic cycle. 64
3.9 Sample paths of standardised higher-order stochastic cycles. 65
3.10 Higher-order cycles in U.S. investments. 66

4.1 Sample path of real trend component and UCM estimate. 75
4.2 Variables used to fit abrupt changes in time series. 76
4.3 Flow of the Nile River and trend with and without break. 77
4.4 Effect of the Euro currency on Italian CPI for restaurants. 78
4.5 Approximations of a smooth seasonal component by sinusoids. 80
4.6 Number of weddings in England and Wales per quarter. 83
4.7 Seasonal component of the number of weddings in England and Wales. 84
4.8 Impulse response functions of two transfer function models. 85
4.9 Effect of 9/11 and Iraq war on Australian air passengers. 86
4.10 Impulse response function of 9/11 and Iraq war. 87
4.11 Smooth component of time varying Okun's coefficients. 88

5.1 Example of Kalman filter: conditional mean and MSE. 117
5.2 Comparison of filter and smoother: conditional mean and MSE. 122
5.3 Actual vs. asymptotic distribution of variance estimators. 132
5.4 Variance estimator distribution when the variance is zero. 133

6.1 Time series with multiplicative components and its log. 138
6.2 Box–Cox transforms of annual U.S. airline passengers. 141

6.3 Means vs. standard deviations scatter plots. 142
6.4 Seasonal component sinusoids for the NYC temperature. 144
6.5 Smoothed components: diffuse vs. informative initialisation. 150
6.6 Density of one point with respect to different variances. 151
6.7 Innovation plots of the LLT plus noise model (Nile data). 157
6.8 Auxiliary residuals of the LLT plus noise model (Nile data). 158
6.9 Approximated auxiliary residuals of the LLT plus noise model. 159

7.1 Common cycle extracted from the Conference Board series. 173

8.1 Periodograms of time series already encountered in this book. 182
8.2 Spectra of various ARMA processes. 186
8.3 Squared gain functions of the filters difference and seasonal difference. 188
8.4 Gain functions of four band-pass filters. 190
8.5 Cycles extracted by four band-pass filters. 193
8.6 U.S. cycles extracted by HT filters of different order. 194
8.7 Common cycles extracted from the Conference Board series. 197

9.1 Monthly number of motor vehicle injuries in Italy. 200
9.2 Auxiliary residual t-tests for a level shift. 200
9.3 UCM plots for the Italian motor vehicle injuries series. 201
9.4 Smoothed trend-cycle of the monthly GDP. 205
9.5 Comparison of monthly GDP with Conference Board Index. 206
9.6 Sample ACF of model standardised residuals. 209
9.7 One week of Italian hourly electricity demand observations. 210
9.8 Low-pass filtered daily demand of electricity in Italy. 211
9.9 Yearly seasonal pattern with sinusoid-based approximation. 212
9.10 Actual values and out-of-sample forecast of the hourly load. 213

10.1 Residual diagnostics plot of the SAS/ETS PROC UCM. 219
10.2 Model window in which the components are chosen. 221
10.3 Multivariate settings window. 222
10.4 Component graphs produced by STAMP by default. 224
10.5 Prediction graphs produced by STAMP. 225
10.6 Stata window to define an UCM. 227
10.7 Stata window to define an UCM. 229

List of symbols

x	Scalar deterministic quantity.
$\boldsymbol{x}$	Deterministic (column) vector.
$\mathbf{0}$	Column vector of zeros.
$\mathbf{1}$	Column vector of ones.
$\mathbf{X}$	Deterministic matrix.
$\mathbf{X}^\top$	Transposition of $\mathbf{X}$.
$\mathbb{T}\mathrm{r}(\mathbf{X})$	Trace of the square matrix $\mathbf{X}$.
$\lvert\mathbf{X}\rvert$ or $\det(\mathbf{X})$	Determinant of the square matrix $\mathbf{X}$.
$\mathrm{diag}(\mathbf{X})$	Vector containing the elements on the main diagonal of the square matrix $\mathbf{X}$.
$\mathrm{vec}(\mathbf{X})$	Vector obtained by staking the columns of the matrix $\mathbf{X}$.
$\mathbf{I}_p$	Identity matrix of dimensions $p \times p$.
$\mathbf{O}_p$	Matrix of zeros of dimensions $p \times p$.
X	Scalar random quantity.
$\boldsymbol{X}$	Random (column) vector.
$\{X_t\}$	Random scalar sequence (also stochastic process in discrete time).
$\{X(t)\}$	Scalar stochastic process in continuous time (also random function).
$\{\boldsymbol{X}_t\}$	Random vector-valued sequence (also vector-valued stochastic process in discrete time).
$\{\boldsymbol{X}(t)\}$	Stochastic vector-valued process in continuous time (also vector-valued random function).
$(\Omega, \mathcal{F}, P)$	Probability space, where Ω is the sample space (a set whose generic element is indicated with ω), $\mathcal{F}$ a σ-algebra of events (subsets of Ω), P a probability measure.
$\mathbb{E}[X]$ or $\mathbb{E}X$	Expectation of the random variable X.

$\mathbb{P}[Y\|X_1,\ldots,X_m]$ or $\mathbb{P}[Y\|\boldsymbol{X}]$	Best linear predictor of the random variable Y based on the random variables $\{X_1,\ldots,X_m\}$ (projection of Y onto the linear span of $\{1,X_1,\ldots,X_m\}$). The second notation is to be interpreted in the same way provided the elements of $\boldsymbol{X}$ are $\{X_1,\ldots,X_m\}$.
$\mathbb{V}\mathrm{ar}(X)$	Variance of the random variable X.
$\mathbb{V}\mathrm{ar}(\boldsymbol{X})$	Covariance matrix of the random vector $\boldsymbol{X}$.
$\mathbb{C}\mathrm{ov}(X,Y)$	Covariance between the random variables X and Y.
$\mathbb{C}\mathrm{ov}(\boldsymbol{X},\boldsymbol{Y})$	Matrix of the covariances of the elements of $\boldsymbol{X}$ with the elements of $\boldsymbol{Y}$: $\mathbb{E}[(\boldsymbol{X}-\mathbb{E}\boldsymbol{X})(\boldsymbol{Y}-\mathbb{E}\boldsymbol{Y})^\top]$.
$\mathbb{C}\mathrm{or}(X,Y)$	Correlation between the random variables X and Y.
$\mathbb{C}\mathrm{or}(\boldsymbol{X},\boldsymbol{Y})$	Matrix of the correlations of the elements of $\boldsymbol{X}$ with the elements of $\boldsymbol{Y}$.
$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$	Multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
χ^2_m	Chi-square distribution with m degrees of freedom.
$\boldsymbol{X}_n \xrightarrow{d} \mathcal{D}$	Convergence in distribution of $\boldsymbol{X}_n$ to $\mathcal{D}$ (which is a place-holder for a distribution such as $\mathcal{N}$ or χ^2_m).
$\mathrm{WN}(0,\sigma^2)$	White noise sequence with mean 0 and variance σ^2. (White noise sequences are zero-mean by definition; here we keep the mean parameter for conformity with the next two notational entries.)
$\mathrm{IID}(\mu,\sigma^2)$	Independent identically distributed random sequence with mean μ and variance σ^2.
$\mathrm{NID}(\mu,\sigma^2)$	Normally independently distributed random sequence with mean μ and variance σ^2.
$\mathbb{B}$	Backward shift (also lag) operator: $\mathbb{B}X_t = X_{t-1}$.
$\mathbb{I}(\text{condition})$	Indicator: it takes the value 1 when the condition is true and 0 otherwise.
$\mathbb{N}$	Set of natural numbers.
$\mathbb{Z}$	Set of integer numbers.
$\mathbb{R}$	Set of real numbers.
$\mathbb{C}$	Set of complex numbers.
$\lfloor x \rfloor$	Floor of x: the largest integer equal or smaller than x.
$x \wedge y$	$\min(x,y)$.
$x \vee y$	$\max(x,y)$.

Preface

On the 25th birthday of the path-breaking book *Forecasting, Structural Time Series Models and the Kalman Filter* written by Professor Andrew Harvey, I was reflecting on the relatively scarce diffusion of unobserved component models (UCM) among practitioners outside the academic community.

From (not only) my own experience, I know that UCM have many advantages over more popular forecasting techniques based on regression analysis, exponential smoothing and ARIMA. Indeed, being based on how humans naturally conceive time series, UCM are simple to specify, the results are easy to visualise and communicate to non-specialists (for example to the boss) and their forecasting performance is generally very competitive. Moreover, various types of outliers can easily be identified, missing values are effortlessly managed and working contemporaneously with time series observed at different frequencies presents no problem.

I concluded that the limited spread of UCM among practitioners could be attributed to one or more of the following causes:

1. Lack of mainstream software implementing UCM,
2. Few books on UCM and mostly academic rather than practical,
3. Limited number of university classes in which UCM are taught.

For a long time the only ready-to-use software package for UCM was STAMP, whose first version is contemporaneous with Harvey's book. STAMP is really an excellent software, but I am afraid its use outside academia is rather limited. However, in the last few years UCM procedures have started to appear in software systems such as SAS (since version 8.9) and Stata (since version 12) with a larger audience also outside the academic community. Thus, Point 1 seems to be at least partially resolved and in the future it is likely that more and more software packages will offer UCM procedures.

As for Point 2, for more than ten years the only two books on UCM were Harvey (1989) and West and Harrison (1989, from a Bayesian perpective), with the only exception being the volume by Kitagawa and Gersch (1996) which deals with similar topics but with a different approach. Again, these books are appreciated by academic scholars but are not that accessible to practitioners. The first decade of the new millennium witnessed the introduction of two new volumes on state space modelling: Durbin and Koopman (2001) and Commandeur and Koopman (2007). The first one is a great book, but quite technical, with an emphasis on state space methods rather than on

modelling. The second one is introductory and, although very clear, it lacks some topics needed in economic time series modelling.

Finally, Point 3 is not easy to verify, but it is certainly linked to the first two points: if the first two causes are solved, then it is likely that the number of courses covering UCM will increase both in and outside universities.

Now, if the product (the UCM) is excellent, the supporting technology (the software) is available, then probably the product has to be popularised in a different way. So, I reflected on how a book should be designed to achieve this goal and concluded that such a book should

- Focus on the UCM approach rather than on general state space modelling
- Be oriented toward the applications
- Review the available software
- Provide enough theory to let the reader understand what's under the hood
- Keep the rigour to a level that is appropriate for academic teaching

This book has been written with those aims in mind but, of course, I am not the one who can judge if they were achieved.

Prerequisites. It is assumed that the reader has a basic knowledge of calculus, matrix algebra, probability and statistical inference at the level commonly met in the first year of undergraduate programmes in statistics, economics, mathematics, computer science and engineering.

Structure. The book is organised in three parts.

The first one covers propaedeutic time series and prediction theory, which the reader acquainted with time series analysis can skip. Unlike many other books on time series, I put the chapter on prediction at the beginning, because the problem of predicting is not limited to the field of time series analysis.

The second part introduces the UCM, presents the state space form and the related algorithms, and provides practical modelling strategies to build and select the UCM which best fits the needs of the time series analyst.

The third part presents some real-world applications with a chapter that focusses on business cycle analysis. Despite the seemingly economic-centric scope of the business cycle chapter, its content centres on the construction of band-pass filters using UCM, and this has obvious applications in many other fields. The last chapter reviews software packages that offer ready-to-use procedures for UCM and systems that are popular among statisticians and econometricians and that allow general estimation of models in state space form.

Website. Information, corrections, data and code are available at the book's website

`http://www.ucmbook.info`

I am grateful to all those readers who want to share their comments and signal errors in the book so that corrections can be placed on the site.

Acknowledgements. My gratitude goes to Professors Andrew Harvey, Siem Jan Koopman and Tommaso Proietti, whose ideas inspired this book. I thank my colleagues at Bicocca, Lisa Crosato, Gianna Monti, Alessia Paccagnini and Biancamaria Zavanella who read and corrected parts of the manuscript. Finally, many thanks go to my family who tolerated my numerous nights, weekends and holidays working on the manuscript.

Matteo M. Pelagatti
Milan, Italy

Part I

Statistical prediction and time series

Chapter 1

Statistical Prediction

A *statistical prediction* is a guess about the value of a random variable Y based on the outcome of other random variables $X_1, \ldots, X_m$. Thus, a *predictor*[1] is a (measurable) function, say $p(\cdot)$, of the random variables $X_1, \ldots, X_m$. In order to select an optimal predictor, we need a *loss function*, say $\ell(\cdot)$, which maps the prediction error to its cost. In principle, the loss function has to be determined case by case, but we can harmlessly assume that if the prediction error is zero also the loss is zero and that $\ell(\cdot)$ is non-decreasing in the absolute value of the prediction error. Indeed, it is reasonable to assume that an exact guess of the outcome of Y will induce no losses, while the greater the prediction error, the higher the cost.

A loss function can be symmetric about zero (i.e., $\ell(-x) = \ell(x)$) or asymmetric. In the former case positive and negative errors of equal modulus produce the same loss, while in the latter case a different weight is given to positive and negative prediction errors. While there can be many reasons for the loss function to be asymmetric (cf. Example 1.1), the most used loss functions are generally symmetric. In particular the quadratic loss function $\ell_2(x) = x^2$ is for practical reasons the most frequently used in time series analysis and statistics in general.

Example 1.1 (Asymmetric loss function)**.**
Suppose that for a firm that produces sunglasses, the variable costs of producing one pair of glasses is 1 Euro, and its wholesale value is 20 Euro. If one pair of sunglasses is produced but not sold the firm will have to pay 1 Euro per piece for storage and recycling costs.

Every year the firm has to decide how many pairs of glasses to produce and in order to do this it needs a prediction of sunglasses sales for that year. The cost of the prediction error will be higher if the predicted sales are lower than the actual, in fact for each produced pair of glasses the cost of not selling them is just 2 Euro (production cost plus storage/recycling) while the cost of not producing them when they would

[1] Notice that the term *predictor* is also commonly used for each of the variables $X_1, \ldots, X_m$ on which the prediction is based. We will avoid this second meaning of the term to prevent misunderstandings.

sell is 19 Euro of lost profits (20 Euro of lost sales revenues minus 1 Euro of production cost).

In formulas, let us call Y the unknown future value of the sunglasses demand, $\hat{Y}$ its prediction and $E = Y - \hat{Y}$ the prediction error. Then, the loss function for this problem is given by

$$\ell(E) = \begin{cases} 19E & \text{for } E \geq 0 \\ 2E & \text{for } E < 0. \end{cases}$$

Intuitively, by observing this cost function one can expect that it is convenient to build predictions that tend to be positively biased so that negative errors (less costly) are more frequent than positive ones (more costly).

A predictor is optimal if it minimises the *expected loss* (i.e., the expectation of the loss function) among the class of measurable functions.

Definition 1.1 (Optimal prediction)**.** Let $Y, X_1, \ldots, X_m$ be random variables defined on the same probability space, $\mathcal{M}$ be the class of $(X_1, \ldots, X_m)$-measurable functions and $\ell(\cdot)$ be a non-negative loss function; then the predictor $\hat{Y} = \hat{p}(X_1, \ldots, X_m)$, with $\hat{p} \in \mathcal{M}$, is optimal for Y with respect to the loss ℓ if

$$\mathbb{E}\,\ell\big(Y - \hat{Y}\big) = \min_{p \in \mathcal{M}} \mathbb{E}\,\ell\big(Y - p(X_1, \ldots, X_m)\big).$$

In particular the quadratic loss function $\ell_2(x) = x^2$ is for practical reasons the most frequently used in time series analysis and statistics in general. By using this loss function one assumes that the loss grows quadratically with the prediction error, and positive and negative errors of the same entity correspond to equal losses. As it will become clear from the next lines, the quadratic loss function has many mathematical advantages that make it a good choice if no better reasons suggest to the use of different loss curves.

In the rest of the book, predictions will always be made with respect to the quadratic loss function unless otherwise specified.

1.1 Optimal predictor

We are now in the condition to derive the optimal predictor under the quadratic loss function.

Theorem 1.1 (Optimal predictor under quadratic loss)**.** *Let $Y, X_1, \ldots, X_m$ be random variables with finite variance, then the optimal predictor for Y based on $X_1, \ldots, X_m$ with respect to the quadratic loss function, $\ell_2(x) = x^2$, is the conditional expectation $\mathbb{E}[Y|X_1, \ldots, X_m]$.*

Proof. We have to show that there is no other measurable function $p(X_1, \ldots, X_m)$ that has smaller expected loss than $\mathbb{E}[Y|X_1, \ldots, X_m]$. The expected (quadratic) loss of the optimal predictor (*mean squared error*) is

$$MSE_{opt} = \mathbb{E}\big\{Y - \mathbb{E}[Y|X_1, \ldots, X_m]\big\}^2.$$

If we compute the expected loss of the generic predictor and subtract and add the optimal predictor $\mathbb{E}[Y|X_1, \ldots, X_m]$ we can write the mean squared error of $p(\cdot)$ as

$$\begin{aligned} MSE_{p(\cdot)} &= \mathbb{E}\big\{Y - \mathbb{E}[Y|X_1, \ldots, X_m] + \mathbb{E}[Y|X_1, \ldots, X_m] - p(X_1, \ldots, X_m)\big\}^2 \\ &= MSE_{opt} + \mathbb{E}\big\{\mathbb{E}[Y|X_1, \ldots, X_m] - p(X_1, \ldots, X_m)\big\}^2, \end{aligned}$$

since, by Property 2. of Theorem 1.2 below, we have

$$\mathbb{E}\Big\{\Big(Y - \mathbb{E}[Y|X_1, \ldots, X_m]\Big)\Big(\mathbb{E}[Y|X_1, \ldots, X_m] - p(X_1, \ldots, X_m)\Big)\Big\} = 0.$$

Thus, $MSE_{p(\cdot)}$ is the sum of a fixed number and a non-negative term which is zero if and only if $p(X_1, \ldots, X_m) = \mathbb{E}[Y|X_1, \ldots, X_m]$ with probability one. □

The following properties of the conditional expectation will be useful throughout the entire book.

Theorem 1.2 (Properties of the conditional expectation)**.** *Let Y, X and Z be random variables (or vectors) with finite expectation and $g(\cdot)$ a function (measurable with respect to X) such that $\mathbb{E}g(X)$ exists, then*

1. *(Linearity) $\mathbb{E}[aY + bZ + c|X] = a\mathbb{E}[Y|X] + b\mathbb{E}[Z|X] + c$, with a, b, c constants;*
2. *(Functions of conditioning variables) $\mathbb{E}[Yg(X)|X] = \mathbb{E}[Y|X]g(X)$;*
3. *(Independence with the conditioning variables) $\mathbb{E}[Y|X] = \mathbb{E}[Y]$ when Y is independent from X;*
4. *(Law of iterated expectations) $\mathbb{E}[Y] = \mathbb{E}\{\mathbb{E}[Y|X]\}$;*
5. *(Orthogonality of the prediction error) $\mathbb{E}\big\{(Y - \mathbb{E}[Y|X])\, g(X)\big\} = 0$;*
6. *(Law of total variance) $\mathbb{V}\mathrm{ar}[Y] = \mathbb{E}[\mathbb{V}\mathrm{ar}(Y|X)] + \mathbb{V}\mathrm{ar}[\mathbb{E}[Y|X]]$.*

Proof. We prove the theorem point by point.

Linearity. Being the conditional expectation, an integral, linearity of the expectation is just a consequence of the linearity of the integral.

Functions of conditioning variables. For any value x that the random variable X can take, the expectation $\mathbb{E}[Y g(X)|X = x]$ is equal to $g(x)\mathbb{E}[Y|X = x]$ as, given $X = x$, $g(X)$ becomes the constant $g(x)$. Since this holds for all the values x in the range of X, the result follows.

Independence with the conditioning variable. Under independence of Y and X, the joint distribution of (X, Y) is the product of the two marginal distributions and, thus, the (conditional) distribution of $(Y|X)$ is equal to the marginal distribution of Y. Therefore, the expectation of $(Y|X)$ and Y is equal.

Law of iterated expectations. For a general proof of Property 1, refer to any measure-theoretic book on probability (for instance Shorack, 2000, Chapter 8), we provide a proof only for the case of absolutely continuous random variables using elementary probability notions:

$$\begin{aligned}\mathbb{E}[Y] &= \int y \int f(y, x)\,\mathrm{d}x\,\mathrm{d}y = \int y \int f(y|x) f(x)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int f(x) \int y f(y|x)\,\mathrm{d}y\,\mathrm{d}x = \int f(x)\,\mathbb{E}[Y|X = x]\,\mathrm{d}x \\ &= \mathbb{E}\big\{\mathbb{E}[Y|X]\big\}.\end{aligned}$$

The change in the order of integration is allowed by the assumption of finiteness of the expectations of X and Y. The reader should try to replicate this proof for X and Y discrete random variables.

Orthogonality of the prediction error. Using Properties 1, 2, and 4 we have

$$\begin{aligned}\mathbb{E}\big\{(Y - \mathbb{E}[Y|X])\, g(X)\big\} &= \mathbb{E}\left[\mathbb{E}\big\{(Y - \mathbb{E}[Y|X])\, g(X)|X\big\}\right] \\ &= \mathbb{E}\left[\mathbb{E}\big\{Y - \mathbb{E}[Y|X]|X\big\} g(X)\right] \\ &= \mathbb{E}\big[\{\mathbb{E}[Y|X] - \mathbb{E}[Y|X]\} g(X)\big] = 0\end{aligned}$$

Law of total variance. Using Property 5, we can write

$$\mathbb{V}\mathrm{ar}[Y] = \mathbb{V}\mathrm{ar}\left[Y - \mathbb{E}(Y|X) + \mathbb{E}(Y|X)\right] = \mathbb{V}\mathrm{ar}[Y - \mathbb{E}[Y|X]] + \mathbb{V}\mathrm{ar}[\mathbb{E}(Y|X)].$$

Using Property 4, the first addend after the last equal sign can be written as

$$\mathbb{V}\mathrm{ar}[Y - \mathbb{E}[Y|X]] = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] = \mathbb{E}[\mathbb{E}(Y|X - \mathbb{E}[Y|X])^2] = \mathbb{V}\mathrm{ar}[Y|X].$$

□

1.2 Optimal linear predictor

Sometimes, instead of looking for an optimal predictor among the class of all measurable functions, it can be easier to limit the search to a smaller class of

functions, such as linear combinations. As will be clear from the next lines, the advantage of basing the prediction of Y on the class of linear combinations of the random variables $X_1, \dots, X_m$ (plus a constant) is that (under quadratic loss) the covariance structure of the random variables $Y, X_1, \dots, X_m$ is all that is needed to compute the prediction.

Let $\boldsymbol{X} = (X_1, \dots, X_m)^\top$, $\mu_Y = \mathbb{E}[Y]$, $\boldsymbol{\mu}_X = \mathbb{E}[\boldsymbol{X}]$, $\boldsymbol{\Sigma}_{XX} = \mathbb{E}[(\boldsymbol{X} - \boldsymbol{\mu}_x)(\boldsymbol{X} - \boldsymbol{\mu}_X)^\top]$, $\boldsymbol{\Sigma}_{YX} = \mathbb{E}[(Y - \mu_Y)(\boldsymbol{X} - \boldsymbol{\mu}_X)]$ and $\boldsymbol{\Sigma}_{XY} = \boldsymbol{\Sigma}_{YX}^\top$. As the next theorem states, this information is sufficient and necessary to compute the optimal linear predictor.

Theorem 1.3 (Optimal linear predictor)**.** *Let $Y, X_1, \dots, X_m$ be random variables with finite variance, and let $\mathcal{L}$ be the class of linear functions $\{\beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m : (\beta_0, \beta_1, \dots, \beta_m) \in \mathbb{R}^{m+1}\}$; then the optimal predictor in the class $\mathcal{L}$ with respect to the quadratic loss function, $\ell_2(x) = x^2$, is the linear predictor*

$$\mathbb{P}[Y|X_1, \dots, X_m] = \mu_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_X),$$

where, if $\boldsymbol{\Sigma}_{XX}$ is singular, $\boldsymbol{\Sigma}_{XX}^{-1}$ is to be substituted with a generalised inverse. The optimal linear predictor is unique.

In time series, there is no standard symbol for the linear prediction (or linear projection), thus, we will use $\mathbb{P}[Y|X]$ that recalls the conditional expectation notation.

Notice that if $\boldsymbol{\Sigma_{XX}}$ is singular, its generalised inverse[2] is not unique, but, as the theorem states, the projection $\mathbb{P}[Y|\boldsymbol{X}]$ will be unique. This means that there are more choices of the vector $(\beta_0, \dots, \beta_m)$ that yield the identical prediction.

Proof. We need to minimise the following MSE with respect to $\boldsymbol{\beta}$

$$MSE_{\boldsymbol{\beta}} = \mathbb{E}[(Y - \boldsymbol{\beta}^\top \tilde{\boldsymbol{X}})(Y - \boldsymbol{\beta}^\top \tilde{\boldsymbol{X}})^\top],$$

where we have set $\tilde{\boldsymbol{X}} = (1, X_1, \dots, X_m)^\top$. If we define $\boldsymbol{\Omega}_{XX} = \mathbb{E}[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top]$, $\boldsymbol{\Omega}_{YX} = \mathbb{E}[Y\tilde{\boldsymbol{X}}^\top] = \boldsymbol{\Omega}_{XY}^\top$, we can write

$$MSE_{\boldsymbol{\beta}} = \mathbb{E}(Y^2) + \boldsymbol{\beta}^\top \boldsymbol{\Omega}_{XX}\boldsymbol{\beta} - 2\boldsymbol{\Omega}_{YX}\boldsymbol{\beta}.$$

By setting equal to zero the derivative with respect to $\boldsymbol{\beta}$, we obtain the system

[2]If $\boldsymbol{A}$ is a matrix, its generalised inverse is a matrix $\boldsymbol{B}$ such that $\boldsymbol{ABA} = \boldsymbol{A}$. Every matrix has at least one generalised inverse.

of *normal equations*

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\beta}} MSE_{\boldsymbol{\beta}} = \mathbf{0}^\top \\
&2\boldsymbol{\beta}^\top \boldsymbol{\Omega}_{XX} - 2\boldsymbol{\Omega}_{YX} = \mathbf{0}^\top \\
&\boldsymbol{\beta}^\top \boldsymbol{\Omega}_{XX} = \boldsymbol{\Omega}_{YX},
\end{aligned}$$

which is a system of $m+1$ linear equations in $n+1$ unknowns. Thus, if $\boldsymbol{\Omega}_{XX}$ is invertible there is only the solution $\boldsymbol{\beta} = \boldsymbol{\Omega}_{XX}^{-1}\boldsymbol{\Omega}_{\boldsymbol{XY}}$; otherwise there are infinitely many choices of $\boldsymbol{\beta}$ that solve the system.

To prove the uniqueness of the optimal linear prediction also when the matrix $\boldsymbol{\Omega}_{XX}$ is non-invertible, consider two arbitrary solutions of the linear system, say $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$, and the distance between the predictions $\hat{\boldsymbol{\beta}}^\top \tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{\beta}}^\top \tilde{\boldsymbol{X}}$:

$$\begin{aligned}
\mathbb{E}\left[(\hat{\boldsymbol{\beta}}^\top \tilde{\boldsymbol{X}} - \tilde{\boldsymbol{\beta}}^\top \tilde{\boldsymbol{X}})(\hat{\boldsymbol{\beta}}^\top \tilde{\boldsymbol{X}} - \tilde{\boldsymbol{\beta}}^\top \tilde{\boldsymbol{X}})^\top\right] &= (\hat{\boldsymbol{\beta}}^\top - \tilde{\boldsymbol{\beta}}^\top)\boldsymbol{\Omega}_{XX}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) \\
&= (\boldsymbol{\Omega}_{YX} - \boldsymbol{\Omega}_{YX})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) = 0,
\end{aligned}$$

where we used the fact that both coefficient vectors satisfy the above normal equations. This zero mean-square distance implies that $\hat{\boldsymbol{\beta}}^\top \tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{\beta}}^\top \tilde{\boldsymbol{X}}$ are equal with probability one.

Notice that the optimal linear predictor in the theorem is expressed in a slightly different form. There, we have $\boldsymbol{\beta}$-coefficients that solve $\boldsymbol{\Sigma}_{XX}\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_X Y$ and $\beta_0 + \boldsymbol{\mu}_X^\top \boldsymbol{\beta}_1 = \mu_Y$. We can see the equivalence of the two systems of normal equations if we write $\boldsymbol{\Omega}_{XX}\boldsymbol{\beta} = \boldsymbol{\Omega}_{XY}$ in blocks as follows:

$$\begin{bmatrix} 1 & \boldsymbol{\mu}_X^\top \\ \boldsymbol{\mu}_X & \boldsymbol{\Sigma}_{XX} + \boldsymbol{\mu}_X\boldsymbol{\mu}_X^\top \end{bmatrix} \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{bmatrix} = \begin{bmatrix} \mu_Y \\ \boldsymbol{\Sigma}_{XY} + \boldsymbol{\mu}_X\mu_Y \end{bmatrix}.$$

From the first line we obtain $\beta_0 = \mu_Y - \boldsymbol{\mu}_X^\top \boldsymbol{\beta}_1$ and, substituting in the second block we get,

$$\boldsymbol{\mu}_X(\mu_Y - \boldsymbol{\mu}_X^\top \boldsymbol{\beta}_1) + \boldsymbol{\Sigma}_{XX}\boldsymbol{\beta}_1 + \boldsymbol{\mu}_X\boldsymbol{\mu}_X^\top \boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{XY} + \boldsymbol{\mu}_X\mu_Y,$$

which simplifies to $\boldsymbol{\Sigma}_{XX}\boldsymbol{\beta} = \boldsymbol{\Sigma}_{XY}$. □

Notice that if instead of predicting the scalar Y we need to predict the vector $\boldsymbol{Y} = (Y_1, \ldots, Y_k)^\top$, we have a set of k independent optimisations, and the prediction formula in Theorem 1.3 simply generalises to

$$\boldsymbol{Y} = \boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}(\boldsymbol{X} - \boldsymbol{\mu}_X).$$

We provide a list of properties that the optimal linear predictor enjoys in the more general case in which $\boldsymbol{Y}$ is a vector.

Theorem 1.4 (Properties of the optimal linear predictor)**.** *Let all the conditions and the notation of Theorem 1.3 hold, a, b, c be constants, $\boldsymbol{Z}$ be a random vector with finite variances; then the optimal linear predictor satisfies the following properties*

1. *(Unbiasedness)* $\mathbb{E}\big[\boldsymbol{Y} - \mathbb{P}[\boldsymbol{Y}|\boldsymbol{X}]\big] = 0$*;*
2. *(Orthogonality of the prediction error)* $\mathbb{E}\big[(\boldsymbol{Y} - \mathbb{P}[\boldsymbol{Y}|\boldsymbol{X}])\,\boldsymbol{X}^\top\big] = \mathbf{0}$*;*
3. *(Mean square error of the prediction)* $MSE_{lin} = \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}$*;*
4. *(Linearity)* $\mathbb{P}[a\boldsymbol{Y} + b\boldsymbol{Z} + c|\boldsymbol{X}] = a\mathbb{P}[\boldsymbol{Y}|\boldsymbol{X}] + b\mathbb{P}[\boldsymbol{Z}|\boldsymbol{X}] + c$*;*
5. *(Law of iterated projections)* $\mathbb{P}[\boldsymbol{Y}|\boldsymbol{X}] = \mathbb{P}\big\{\mathbb{P}[\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{X}]\big|\boldsymbol{X}\big\}$*;*
6. *(Projection on orthogonal variables) if* $\mathbb{E}(\boldsymbol{X} - \boldsymbol{\mu}_X)(\boldsymbol{Z} - \boldsymbol{\mu}_Z)^\top = \mathbf{0}$ *then*

$$\mathbb{P}[\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}] = \boldsymbol{\mu}_Y + \mathbb{P}[\boldsymbol{Y} - \boldsymbol{\mu}_Y|\boldsymbol{X}] + \mathbb{P}[\boldsymbol{Y} - \boldsymbol{\mu}_Y|\boldsymbol{Z}];$$

7. *(Updating)*

$$\begin{aligned}\mathbb{P}[\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{X}] &= \mathbb{P}[\boldsymbol{Y}|\boldsymbol{X}] + \mathbb{P}\big[\boldsymbol{Y} - \mathbb{P}[\boldsymbol{Y}|\boldsymbol{X}]\big|\boldsymbol{Z} - \mathbb{P}[\boldsymbol{Z}|\boldsymbol{X}]\big] \\ &= \mathbb{P}[\boldsymbol{Y}|\boldsymbol{X}] + \boldsymbol{\Sigma}_{YZ|X}\boldsymbol{\Sigma}_{ZZ|X}^{-1}\big(\boldsymbol{Z} - P[\boldsymbol{Z}|\boldsymbol{X}]\big)\end{aligned}$$

with

$$\begin{aligned}\boldsymbol{\Sigma}_{YZ|X} &= \mathbb{E}\left[(\boldsymbol{Y} - \mathbb{P}[\boldsymbol{Y}|\boldsymbol{X}])(\boldsymbol{Z} - \mathbb{P}[\boldsymbol{Z}|\boldsymbol{X}])^\top\right], \\ \boldsymbol{\Sigma}_{ZZ|X} &= \mathbb{E}\left[(\boldsymbol{Z} - \mathbb{P}[\boldsymbol{Z}|\boldsymbol{X}])(\boldsymbol{Z} - \mathbb{P}[\boldsymbol{Z}|\boldsymbol{X}])^\top\right].\end{aligned}$$

If we call $MSE_{Y|X}$ the mean square error of $\mathbb{P}[\boldsymbol{Y}|\boldsymbol{X}]$, then the MSE of $\mathbb{P}[\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{X}]$ is given by

$$MSE_{Y|Z,X} = MSE_{Y|X} - \boldsymbol{\Sigma}_{YZ|X}\boldsymbol{\Sigma}_{ZZ|X}^{-1}\boldsymbol{\Sigma}_{ZY|X}.$$

Proof. For notational compactness, let $\mathbf{B}_{YX} = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}$.
Unbiasedness.

$$\mathbb{E}\left\{\boldsymbol{Y} - \boldsymbol{\mu}_Y - \mathbf{B}_{YX}(\boldsymbol{X} - \boldsymbol{\mu}_X)\right\} = \boldsymbol{\mu}_Y - \boldsymbol{\mu}_Y - \mathbf{B}_{YX}(\boldsymbol{\mu}_X - \boldsymbol{\mu}_X) = \mathbf{0}.$$

Orthogonality.

$$\begin{aligned}&\mathbb{E}\left\{[\boldsymbol{Y} - \boldsymbol{\mu}_Y - \mathbf{B}_{YX}(\boldsymbol{X} - \boldsymbol{\mu}_X)]\boldsymbol{X}^\top\right\} = \\ &\mathbb{E}\left\{[(\boldsymbol{Y} - \boldsymbol{\mu}_Y) - \mathbf{B}_{YX}(\boldsymbol{X} - \boldsymbol{\mu}_X)][(\boldsymbol{X} - \boldsymbol{\mu}_X) + \boldsymbol{\mu}_X]^\top\right\} = \\ &\boldsymbol{\Sigma}_{YX} - \mathbf{B}_{YX}\boldsymbol{\Sigma}_{XX} + \mathbb{E}\big\{\boldsymbol{Y} - \boldsymbol{\mu}_Y\big\}\boldsymbol{\mu}_X^\top + \mathbb{E}\big\{\boldsymbol{Y} - \mathbb{P}[\boldsymbol{Y}|\boldsymbol{X}]\big\}\boldsymbol{\mu}_X^\top = 0\end{aligned}$$