A Forecasting Procedure involving Bayesian Exponential Smoothing and LASSO

Avner Abrami

November 15, 2014

Contents

1 Settings and Notations

One has p different time series $Y^1, Y^2, ...Y^p$ that we observe for $t \in [0, n]$.

They constitute the columns of Y

$$Y = \begin{bmatrix} Y_1^1 & Y_1^2 & \cdots & Y_1^p \\ Y_1^2 & Y_2^2 & \cdots & Y_2^p \\ \vdots & & \ddots & \vdots \\ Y_n^1 & Y_n^2 & \cdots & Y_n^p \end{bmatrix}$$

1.1 Assumption: High Dimensional Data Analysis

$$p >> n$$

2 Model Formulation

$$\forall i \in [1, p], \forall t \in [1, n], Y_{t+1}^i = a_t^i + b_t^i - \sum_{j \neq i} \omega_j^i Y_{t+1}^j + \epsilon_{t+1}^i$$

Where:

$$\Omega = \begin{bmatrix} 1 & \omega_1^1 & \cdots & \omega_1^p \\ \omega_2^1 & 1 & \cdots & \omega_2^p \\ \vdots & & \ddots & \vdots \\ \omega_p^1 & \omega_p^2 & \cdots & 1 \end{bmatrix}$$

Where the (a^i, b^i)[1] are the smoothing coefficients associated with the i^{th} time series.
It is time to see how we construct this model. But first a few words on exponential smoothing.

3 Exponential Smoothing in brief

This section has been extracted from *Forecasting with Exponential Smoothing* by Rob J.Hyndlan, Anne B.Koehler, J.Jeith Ord, Ralph D.Snyder.

3.1 General Decomposition of Time Series

It is common in business and economics to think of a time series as a combi-nation of various components such as the trend (T), cycle (C), seasonal (S), and irregular or error (E) components. These can be defined as follows:

- **Trend (T)**: The long-term direction of the series
- **Seasonal (S)**: A pattern that repeats with a known periodicity (e.g., 12 months per year, or 7 days per week)
- **Cycle (C)**: A pattern that repeats with some regularity but with unknown and changing periodicity (e.g., a business cycle)
- **Irregular or error (E)**: The unpredictable component of the series

[1] to be precisely defined in the next paragraphs

A *purely additive model* can be expressed as :

$$y = T + S + E \tag{1}$$

where the three components are added together to form the observed series.

A *purely multiplicative* model is written as

$$y = T * S * E \tag{2}$$

where the data are formed as the product of the three components

3.2 Short history of Exponential Smoothing

Historically, exponential smoothing describes a class of forecasting methods. In fact, some of the most successful forecasting methods are based on the concept of exponential smoothing. There are a variety of methods that fall into the exponential smoothing family, each having the property that forecasts are weighted combinations of past observations, with recent observations given relatively more weight than older observations. The name 'exponential smoothing' reflects the fact that the weights decrease exponentially as the observations get older.

The idea seems to have originated with Robert G. Brown in about 1944 while he was working for the US Navy as an Operations Research analyst. Independently, Charles Holt was also working on an exponential smoothing method for the US Office of Naval Research (ONR)

3.3 State Space Models

State space models allow considerable flexibility in the specification of the parametric structure. In this book, we will use the innovations formulation of the model (e.g., Anderson and Moore 1979; Aoki 1987; Hannan and Deistler 1988). Let y_t denote the observation at time t, and let x_t be a 'state vector' containing unobserved components that describe the level, trend and seasonality of the series. Then a linear innovations state space model can be written as:

$$y_t = w' x_{t-1} + \epsilon_t \tag{3}$$

It is known as the measurement (or observation) equation; it describes the relationship between the unobserved states x_{t-1} and the observation y_t .

$$x_t = F x_{t-1} + g\epsilon_t \tag{4}$$

where ϵ_t is a white noise series and F , g and w are coefficients.

It is known as the transition (or state) equation; it describes the evolution of the states over time.

4 Holt-Winters Model

In this section, we will be talking about any of our time series Y^i but we will omit its index i for more convenience.

Reminder: Y^i are the columns of Y.

$$Y = \begin{bmatrix} Y_1^1 & Y_1^2 & \cdots & Y_1^p \\ Y_1^2 & Y_2^2 & \cdots & Y_2^p \\ \vdots & & \ddots & \vdots \\ Y_n^1 & Y_n^2 & \cdots & Y_n^p \end{bmatrix}$$

4.1 Transition Equations

The additive Holt-Winters decomposition can be defined through the following transition equations which define the Holt-Winters filter :

Mean Equation	$a_t = \alpha(y_t - c_{t-p}) + (1-\alpha)(a_{t-1} + b_{t-1})$
Trend equation	$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}$
Seasonality equation	$c_i = \gamma(y_t - a_{t-1} - b_{t-1}) + (1-\gamma)c_{t-p}$

where $\mathbf{y} = (y_1, ..., y_n)'$ is the vector of observed data, $\theta = (\alpha, \beta, \gamma)'$ is the smoothing parameters vector, and p is the seasonal period.

4.2 Statistical Model

A stochastic component is added to the above filter in order to obtain a statistical model :

$$Y_t = (a_{t-1} + b_{t-1} + c_{t-p}) + \epsilon_t$$

$\mu(t) = a_{t-1} + b_{t-1} + c_{t-p}$ is the one-step forecast made at time $t-1$ by the Holt-Winters filter, and ϵ_t is the forecast error at time t.

The combination of the above equation and transitions equations give a state representation of the additive Holt-Winters model that will be modified in the next section to remove the seasonality.

5 Holt-Winters Model without seasonality : Holt's linear Model

5.1 Model Formulation

We are interested in developing the Holt-Winters method for data without the seasonality component. The corresponding filter is obtained from the initial one through the suppression of the seasonality equation, the γ smoothing parameter, the period p, and the $(c_0, ..., c_{p-1})$ factors.

The combination of the obtained filter with the revised equation,

$$Y_t = a_{t-1} + b_{t-1} + \epsilon_t \tag{5}$$

gives the following state representation for the additive non-seasonal Holt-Winters model :

$$Y = (Y_1, Y_2, ..., Y_n)^T = M\psi + L\varepsilon \tag{6}$$

where :

$$M = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & n \end{pmatrix} \in \mathcal{M}_{n\times 2}(\mathbb{R}) \text{ a design matrix}$$

$\psi = (a_0, b_0)' \in \mathbb{R}^2$ the initial conditions vector

$$L = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & \vdots \\ l_{i,j|i>j} & & & \ddots & 0 \\ = (1+(i-j)\beta)\alpha & & & & 1 \end{pmatrix} \in \mathcal{M}_{n\times n}(\mathbb{R}) \text{ the smoothing matrix}$$

$\varepsilon = (\epsilon_1, ..., \epsilon_n)' \in \mathbb{R}^n$ the forecast errors vector

5.2 Sum up

$$Y_t^i = a_{t-1}^i + b_{t-1}^i + \epsilon_t^i$$

$$Y^i = (Y_1^i, Y_2^i, ..., Y_n^i)^T = M^i\psi^i + L^i\varepsilon^i \tag{7}$$

6 Why is our new model compatible with Holt's Linear Model?

Our model is:

$$\forall i \in [1,p], \forall t \in [1,n], Y_{t+1}^i = a_t^i + b_t^i - \sum_{j\neq i} \omega_j^i Y_{t+1}^j + \epsilon_{t+1}^i$$

Which can easily be rewritten as:

$$Y_{t+1}^i + \sum_{j\neq i} \omega_j^i Y_{t+1}^j = \sum_j \omega_j^i Y_{t+1}^j = (\omega^i)^T Y_{t+1}) = a_t^i + b_t^i + \epsilon_{t+1}^i$$

$$\omega_i^i = 1$$

According to section 5:

$$(\omega^i)^T Y_{t+1} = M\psi^i + L^i\epsilon^i$$

But more generally:

$$Y\Omega = \mathcal{M}\Psi + L\epsilon$$

$$Y = \begin{bmatrix} Y_1^1 & Y_1^2 & \cdots & Y_1^p \\ Y_1^2 & Y_2^2 & \cdots & Y_2^p \\ \vdots & & \ddots & \vdots \\ Y_n^1 & Y_n^2 & \cdots & Y_n^p \end{bmatrix}$$

$$\Psi = \left[\begin{array}{c|c|c|c} \psi^1 & \psi^2 & \cdots & \psi^p \end{array} \right]$$

$$\mathcal{M} = \begin{bmatrix} M & 0 & \cdots & 0 \\ 0 & M & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & M \end{bmatrix}$$

$$L = \begin{bmatrix} L^1 & 0 & \cdots & 0 \\ 0 & L^2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & L^p \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} \epsilon_1^1 & \epsilon_2^1 & \cdots & \epsilon_p^1 \\ \epsilon_1^2 & \epsilon_2^2 & \cdots & \epsilon_p^2 \\ \vdots & & \ddots & \vdots \\ \epsilon_1^n & \epsilon_2^n & \cdots & \epsilon_p^n \end{bmatrix}$$

7 Optimization Program and Algorithm

7.1 Optimization Program

In order to assess our statistical model, we would like to solve:

$$min_{\theta \in [0,1]} \{ min_{\omega} \{ ||L^{-1} Y \Omega - L^{-1} \mathcal{M} \Psi||_F + \lambda ||\Omega|| \} \}$$

Where:

$$\theta = \left[\begin{array}{c|c|c|c} (\alpha^1, \beta^1) & (\alpha^2, \beta^2) & \cdots & (\alpha^p, \beta^p) \end{array} \right]$$

If $Y_\theta = (L^{-1} Y \Omega)$, $W_\theta = L^{-1} \mathcal{M} \Psi$, the program becomes:

$$min_{\theta \in [0,1]} \{ min_{\Omega} \{ ||Y_\theta \Omega - W_\theta||_F + \lambda ||\Omega|| \} \}$$

Which is a problem of the form

$$min_{(\theta \in [0,1]^2, \Omega)} g(\theta, \Omega) \tag{8}$$

It is gonna be a two step algorithm. The first one (apparently the hardest?) is the θ optimization. Indeed Y_θ, W_θ are polynomial fractions of θ (cf L^{-1}).

But then obtaining Ω via a Lasso based algorithm would be a quite easy task.

7.2 Variable Projection : General Formulation

This section has been extracted from *Estimating Nuisance Parameters in Inverse Problems* by Aleksandr Y. Aravkin and Tristan van Leeuwen.

We consider problems of the form (8). As explained previously, for any given θ one can easily find (via a Lasso based Algorithm):

$$\hat{\Omega}(\theta) \in argmin_{\Omega} g(\theta, \Omega)$$

This condition can be relaxed, and θ can be considered a local minimum. Rather than working to solve (8), we can instead focus on the reduced objective:

$$\hat{g}(\theta) = g(\theta, \hat{\Omega}(\theta))$$

This approach is justified by a theorem (Theorem 2.1 page 3 in Prof. Aravkin's paper)

7.3 First

We are in a simple case as $\theta \in [0,1]^2$ is in a closed and bounded set.

To solve our problem, Prof Aravkin suggests a modified projected gradient method in his paper:

$$\theta^{k+1} = P_{[0,1]^2}[\theta^k - \gamma \nabla_x \hat{g}(\theta^k)]] \tag{9}$$

Where P is the projection on $[0,1]^2$.

7.3.1 Second step of the algorithm : Determining Ω

If θ is fixed, solving the program for Ω is the same as resolving a Lasso Problem.

7.3.2 Sumarizing the Algorithms

Initialization: $(\Omega)_0 = I_p$

Compute θ_0 = ProjectionVariable

while not converged **do**:

$(\Omega)_{k+1}$ = Lasso

θ_{k+1} = ProjectionVariable

end while

Output :Ω,θ

8 Forecasting Procedure

Let Y_{obs} be the $n \times 1$ vector of observations and Y_{new} the $h \times 1$ vector of predictions. We suppose that the joint vector $Y = (Y_{obs}, Y_{new})'$ still follows the model described in 5.1 with the same assumptions :

$$Y\Omega = \begin{pmatrix} Y_{\text{obs}}\Omega \\ Y_{\text{new}}\Omega \end{pmatrix} = \begin{pmatrix} \mathcal{M}_1 \\ \mathcal{M}_2 \end{pmatrix} \psi + \begin{pmatrix} L_1 & 0 \\ L_{21} & L_2 \end{pmatrix} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

As a consequence:

$$\forall i \in [1,n], Y_{\text{new}}\Omega = \mathcal{M}_2\tilde{\psi} + L_{21}L_1^{-1}(Y_{\text{obs}} - \mathcal{M}_1\tilde{\psi}) + L_2(\epsilon_2)$$

$$\forall i \in [1, n], E[Y_{\text{new}}] = (\mathcal{M}_2\tilde{\psi} + L_{21}L_1^{-1}(Y_{\text{obs}} - \mathcal{M}_1\tilde{\psi}))\Omega^{-1}$$

Indeed, by definition $E[\epsilon] = 0$.

References

[1] HYNDAMN, KOEHLER, ORD, SNYDER (2008) *Forecasting with Exponential Smoothing*

[2] BERMUDEZ, JD., SEGURA, JV., VERCHER, E. (2010) *Bayesian forecasting with the Holt–Winters model*

[3] Aleksandr Y. Aravkin and Tristan van Leeuwen (2012) *Estimating Nuisance Parameters in Inverse Problems*