Innovation State Space Models

The idea behind the ISSM is that a time series y_t can be described in terms of its unobserved components and that this can be used to predict its future behavior.

Forecasting Methods vs Models

A *forecasting method* is an algorithm that produces a point forecast, while a *forecasting model* is an stochastic data generating process that can be used to generate an entire probability distribution for a future time period. A point forecast can be obtained from a forecasting model by taking the mean or the median of the probability distribution. Furthermore, a forecasting model allows us to compute prediction intervals with a given level of confidence.

Structure of ISSM

Let y_t be the observation at time t and x_t a *state vector* containing unobserved components that describe the level, trend, and seasonality of the series. Then a linear innovations state space model is given by

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

$$x_t = F x_{t-1} + g \epsilon_t \tag{2}$$

Here w, F and g are coefficients, while $\{\epsilon_t\}$ is a white noise series.

It is usually assumed that ϵ_t are independent and identically distributed, following a normal distribution with mean 0 and variance σ^2.

Equation (1) is called the *measurement equation* and it describes the relationship between the observations and the unobserved stated. Equation (2) is called the *transition equation* and it describes the evolution of the unobserved states over time.

These two equations use identical errors, also known as *innovations* since they represent what is new and unpredictable. Models that used identical errors like this are known as *Single Source of Error*.

The general model has a vector of unobserved states given by $x_t = (l_t, b_t, s_t, s_{t-1}, \ldots s_{t-m+1})$

Holt’s linear model with additive errors

Holt’s linear model assumes that the data y_t has a trend. Hence, we need to consider two components: level (l_t) and trend (b_t).

Let $x_t = \begin{bmatrix} l_t \\ b_t \end{bmatrix}$ the vector of unobserved states. Then the ISSM for Holt’s linear model with additive errors is given by

$$\begin{aligned} y_t &= \begin{bmatrix} 1 & 1 \end{bmatrix} x_{t-1} + \epsilon_t \\ x_t &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} x_{t-1} + \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \epsilon_t \end{aligned}$$

The model is completely specified once the distribution of ϵ_t has been set (usually ϵ_t is assumed to be $NID(0, \sigma^2)$.

Initialization and Estimation

We need to estimate x_0 and the values of the parameters $\theta = (\alpha, \beta, \gamma, \phi)$. We can do this by minimizing

$$L^*(\theta, x_0) = n\, log\left(\sum_{t=1}^{n} \epsilon_t^2 \right)$$

which is equal to twice the negative of the likelihood function without constant terms. Here n is the number of observations.

Model Selection

For model selection, use the *Aikake Information Criterion*, given by

$$AIC = L^*(\hat{\theta}, \hat{x_0}) + 2K$$

with K the number of parameters in θ plus the number of free states in x_0, and $\hat{\theta}$ and $\hat{x_0}$ the estimates for θ and x_0.

The initial values θ and x_0 are selected heuristically. The initial x_0 is also known as the *seed vector*.

Automatic Forecasting

For a given series,

1. Generate all the models that are appropriate.
2. Select the best model using the AIC.
3. Produce point forecasts using the best model with optimized parameters.
4. Obtain the prediction intervals for the required confidence levels, either using analytical methods or with simulation.

The General ISSM

> 💡 **Theorem 1**: Under the ISSM framework, the forecast is a linear function of past observations and the seed vector.

Proof:

Substitute equation (1) in equation (2) to get

$$\begin{aligned} x_t &= Fx_{t-1} + g(y_t - w'x_{t-1}) \\ &= (F - gw')x_{t-1} + gy_t \\ &= Dx_{t-1} + gy_t \end{aligned}$$

with $D = F - gw'$. D is known as the *discount matrix*.

For $t = 1$, this equation becomes

$$x_1 = Dx_0 + gy_1$$

For $t = 2$,

$$\begin{aligned} x_2 &= Dx_1 + gy_2 \\ &= D(Dx_0 + gy_1) + gy2 \\ &= D^2 x_0 + Dgy_1 + gy_2 \end{aligned}$$

For $t = 3$,

$$\begin{aligned} x_3 &= Dx_2 + gy_3 \\ &= D(D^2 x_0 + Dgy_1 + gy_2) + gy_3 \\ &= D^3 x_0 + D^2 gy_1 + Dgy_2 + gy_3 \end{aligned}$$

Hence, for an arbitrary t,

$$x_t = D^t x_0 + \sum_{j=0}^{t-1} D^j gy_{t-j} \tag{3}$$

Hence, the forecast is a linear function of past observations and the seed vector. □

From equation (3), it can be seen that to minimize the effect of the seed states, D^t must converge to zero. This equivalent to ask that all eigenvalues of D lie inside the unit circle (i.e., their absolute values are less than 1).

Estimation of Innovations State Space Models

Assuming a Gaussian distribution, it can be shown that

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^{n} \epsilon_t^2 \tag{4}$$

Proof: The Gaussian log likelihood is

$$logL = \frac{-n}{2} log(2\pi\sigma^2) - \frac{1}{2}\sum_{t=1}^{n}\frac{\epsilon_t^2}{\sigma^2}$$

Taking the partial derivative with respecto to σ^2 and setting it to zero gives us the maximum likelihood estimate of the innovations variance σ^2.

$$\begin{aligned}\frac{\partial logL}{\partial\sigma^2} &= -\frac{n}{2}\frac{2\pi}{2\pi\sigma^2} - \frac{1}{2}\sum_{t=1}^{n}\epsilon_t^2\frac{-1}{(\sigma^2)^2} \\ &= \frac{-n}{2\sigma^2} + \frac{1}{2}\sum_{t=1}^{n}\frac{\epsilon_t^2}{(\sigma^2)^2}\end{aligned}$$

Setting it to zero gives us the maximum likelihood estimate of the innovations variance σ^2.

$$\begin{aligned}\frac{-n}{2\sigma^2} + \frac{1}{2}\sum_{t=1}^{n}\frac{\epsilon_t^2}{(\sigma^2)^2} = 0 \\ \frac{n}{\sigma^2} &= \frac{\sum_{t=1}^{n}\epsilon_t^2}{(\sigma^2)^2} \\ \sigma^2 &= \frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2\end{aligned}$$

which is equation (4).

Normalization of Seasonal Components

Normalization is only necessary when the seasonal component of a series is going to be analyzed separately or when a seasonal adjustment will be perform.

TBATS Model

The TBATS model uses the exponencial smoothing framework to forecast time series with multiple seasonal patterns.

Trigonometric exponential smoothing models

The following Fourier series representation can be used to capture a seasonal pattern s_t with frequency m.

$$s_t = \sum_{j=1}^{k} \alpha_j cos(\lambda_j t) + b_j sin(\lambda_j t)$$

Here $\lambda_j = \frac{2\pi j}{m}$ for $j = 1, 2, \ldots, k$ and a_j and b_j are the trigonometric coefficients. Notice that with this representation m doesn't have to be an integer.

The exact approximation is achieved when

$$k = \begin{cases} \frac{m}{2} & \text{if m is even} \\ \frac{m-1}{2} & \text{if m is odd} \end{cases}$$

If the series has a smooth pattern, then fewer terms can be used to approximate the seasonal pattern. The selection of k is an important decision since using more terms than necessary might result in overfitting, while using less terms might result in loss of information.

How to select k

There are two ways of selecting k:

1. Start with one value, say k_i^*, for the *i*th seasonal component. Fit a model to the data with $k_i = k_i^*$ and compute the AIC. Changing one seasonal component at a time, repeatedly fit the model to the data until the minimum AIC is achieved. This is only practical for series with low m (for example, quarterly or monthly data).

2. Use multiple linear regression

A new exponential smoothing framework

Uses:

1. A Box-Cox transformation for handling non-linearity.
2. An ARMA representation to capture autocorrelations in the residuals.
3. M seasonal patterns as follows.

Let y_t be the observation at time t.

$$y_t^{(w)} = \begin{cases} log(y_t) & w = 0 \\ \frac{y_t^w - 1}{w} & w \neq 0 \end{cases} \tag{5}$$

$$y_t^{(w)} = l_{t-1} + \phi b_{t-1} + \sum_{i=1}^{M} s_{t-m_i}^{(i)} + d_t \tag{6}$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha d_t \tag{7}$$

$$b_t = (1 - \phi)b + \phi b_{t-1} + \beta d_t \tag{8}$$

$$s_t^{(i)} = s_{t-m_i}^{(i)} + \gamma_i d_t \tag{9}$$

$$d_t = \sum_{i=1}^{p} \varphi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \tag{10}$$

Here

- $m_1, \ldots, m_M$ denote the seasonal periods.
- l_t represents the level at time t.
- b represents the long-run trend.
- b_t represents the short-run trend in period t.
- $s_t^{(i)}$ represents the *i*th seasonal component at time t.
- d_t denotes an ARMA(p,q) process.
- ϵ_t is a Gaussian white noise process with zero mean and constant variance σ^2.
- $\alpha, \beta, \gamma_i, i = 1, \ldots, M$ are the smoothing parameters.
- ϕ is the damping parameter.

Notice that the model uses a damped trend with damping parameter ϕ, but it supplements it with a long-run trend b.

Trigonometric representation

We'll replace equation (9) with a trigonometric representation of the seasonal components based on Fourier series.

$$s_t^{(i)} = \sum_{j=1}^{k_i} s_{j,t}^{(i)} \tag{11}$$

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} cos\lambda_j^{(i)} + s_{j,t-1}^{*(i)} sin\lambda_j^{(i)} + \gamma_1^{(i)} d_t \tag{12}$$

$$s_{j,t}^{*(i)} = -s_{j,t-1} sin\lambda_j^{(i)} + s_{j,t-1}^{*(i)} cos\lambda_j^{(i)} + \gamma_2^{(i)} d_t \tag{13}$$

Here

- γ_1^i and γ_2^i are the smoothing parameters.
- $\lambda_j^i = \frac{2\pi j}{m}$
- k_i denotes the number of harmonics for the *i*th seasonal component.
- A deterministic representation of the seasonal components can be obtained by setting the smoothing parameters equal to zero.

A TBATS model requieres the estimation of $2 * (k_1, \ldots, k_M)$ initial seasonal values.

ISSM for TBATS

The ISSM for TBATS is given by

$$y_t^{(w)} = w' x_{t-1} + \epsilon_t \tag{14}$$

$$x_t = F x_{t-1} + g\epsilon_t \tag{15}$$

Now, the state vector x_t can be defined as follows

$$x_t = (l_t, b_t, s_t^{(1)}, \dots, s_t^{(M)}, d_t, d_{t-1}, \dots, d_{t-p+1}, \epsilon_t, \epsilon_{t-1}, \dots, \epsilon_{t-q+1})' \quad (16)$$

Here $s_t^{(i)}$ is the vector

$$s_t^{(i)} = (s_{1,t}^{(i)}, s_{2,t}^{(i)}, \dots, s_{k_i,t}^{(i)}, s_{1,t}^{*(i)}, s_{2,t}^{*(i)}, \dots, s_{k_i,t}^{*(i)})$$

Let

$$1_r = (1, 1, \dots, 1)$$

and

$$0_r = (0, 0, \dots, 0)$$

be row vectors of length r.

Let

$$\gamma_1^{(i)} = \gamma_1^{(i)} 1_{k_i} \quad \text{and} \quad \gamma_2^{(i)} = \gamma_2^{(i)} 1_{k_i}$$

Then define

$$\begin{aligned} \gamma^{(i)} &= (\gamma_1^{(i)}, \gamma_2^{(i)}) \\ \gamma &= (\gamma^{(1)}, \dots, \gamma^{(M)}) \end{aligned}$$

Let $\varphi = (\varphi_1, \dots, \varphi_p)$ and $\theta = (\theta_1, \dots, \theta_q)$.

Define $O_{u,v}$ be a $u \times v$ matrix of zeros and $I_{u,v}$ a $u \times v$ rectangular diagonal matrix of 1s.

Let

$$\begin{aligned} a^{(i)} &= (1_{k_i}, 0_{k_i}) \\ a &= (a^{(1)}, \dots, a^{(M)}) \end{aligned}$$

Define matrices $B = \gamma$'φ and $C = \gamma$'θ.

Let

$$A_i = \begin{bmatrix} C^{(i)} & S^{(i)} \\ -S^{(i)} & C^i \end{bmatrix}$$

$$\tilde{A}_i = \begin{bmatrix} 0_{m_i-1} & 1 \\ I_{m_i-1} & 0'_{m_i-1} \end{bmatrix}$$

and

$$A = \bigoplus_{i=1}^{M} A_i$$

Here $C^{(i)}$ and $S^{(i)}$ are $k_i \times k_i$ diagonal matrices with elements $cos(\lambda_j^{(i)})$ and $sin(\lambda_j^{(i)})$ respectively with $j = 1, 2, \ldots, k_i$ and $i = 1, \ldots, M$.

The symbol $\bigoplus$ denotes the direct sum of matrices.

Finally, define

$$\tau = 2 \sum_{i=1}^{M} k_i$$

With this, we can define the matrices for the ISSM as follows:

$$w = (1, \phi, a, \varphi, \theta)' \tag{17}$$

$$g = (\alpha, \beta, \gamma, 1, 0_{p-1}, 1, 0_{q-1})' \tag{18}$$

$$F = \begin{bmatrix} 1 & \phi & 0_\tau & \alpha\varphi & \alpha\theta \\ 0 & \phi & 0_\tau & \beta\varphi & \beta\theta \\ 0'_\tau & 0'_\tau & A & B & C \\ 0 & 0 & 0_\tau & \varphi & \theta \\ 0'_{p-1} & 0'_{p-1} & O_{p-1,\tau} & I_{p-1,p} & O_{p-1,q} \\ 0 & 0 & 0_\tau & 0_p & 0_q \\ 0'_{q-1} & 0'_{q-1} & O_{q-1,\tau} & O_{q-1,p} & I_{q-1,q} \end{bmatrix} \tag{19}$$

Notice that 0_τ denotes a vector of zeros of size τ, while $O_{p-1,q}$ is a matrix of zeros of shape $(p-1, q)$.

Estimation

We need to estimate:

- the smoothing parameters $\alpha, \beta, \gamma_i, \ldots \gamma_M$
- the damping parameter ϕ.
- the Box-Cox transformation parameter w.
- The ARMA coefficients, $\varphi_i, i = 1, \ldots, p$ and $\theta_i, i = 1, \ldots, q$.

Let ϑ denote a vector containing the Box-Cox parameter, the smoothing parameters, and the ARMA coefficients.

We also need to estimate the seed states x_0.

> 💡 The output of the optimization process contains the optimal values for ϑ

Initialization

Starting with $w = 0$ seems to lead to better post sample results for forecasting than other initial values of for w.

The initial values of the seed states can be found using conventional linear least squares methods since ϵ_t is a linear function of the seed vector x_0 (see theorem 1).

Optimization

Both the unknown parameters and the seed states are chosen to maximize the *conditional* likelihood function.

The log-likelihood of this problem is given by

For integer period seasonality, the seasonal values can also be constrained when optimizing, so that each seasonal component sums to zero.

The smoothing parameters are restricted to the *forecastibility* region, w and ϕ are restricted to lie between 0 and 1, and the ARMA coefficients are restricted to the stationary region.

References

De Livera, A. M. (2010). Forecasting time series with complex seasonal patterns using exponential smoothing, PhD thesis, Monash University.

Hyndman, R., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media.

The forecasts from the TBATS model depend on the number of harmonics k_i used for the seasonal component i. It is impractical to consider all possible combinations in the quest for the best combination. After much experimentation we found that the following approach leads to good models and that further improvement can rarely be achieved.

Using the first few seasons, multiple linear regression is applied to the approximation of the initial seasonal component obtained from the following procedure.

Assume that the seasonal components are arranged so that the seasonal periods m_i, $i = 1, \dots, M$ are of increasing order and m_i, $i = 1, \dots, M$ are factors of m_M.

1. Compute a 2 × m_M moving average through the first few seasons of the data. Denote this by f_t for $t = \frac{m_M}{2} + 1, \frac{m_M}{2} + 2, \dots$ to estimate the trend of the first few seasons of the data by removing the seasonality.
2. Obtain an approximation of the overall seasonal component using $z_t = y_t - f_t$. The ith seasonal component can then be approximated by
   $z_t^{(i)} \approx \sum_{i=1}^{M} \sum_{j=1}^{k_i} a_j^{(i)} cos(\lambda_j^{(i)} t) + b_j^{(i)} sin(\lambda_j^{(i)} t)$
   Here $a_j^{(i)}$ and $b_j^{(i)}$ are estimated by regressing z_t against the trigonometric terms, and $\lambda_j^i = \frac{2\pi j}{m_i}$.

Starting with a single harmonic, gradually add harmonics, testing the significance of each one using F-tests. Let k_i^* number of highly significant harmonics (with p < 0.001) for the ith seasonal component. The use of multiple linear regression gives a rough indication as to which harmonics are likely to contribute the most.