State-space model for interesting term(set)s

Younos Aboulnaga[1]

David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada
yaboulna@uwaterloo.ca

1 Introduction

We have chosen the state space model because of its ability to represent hidden component models with minimum assumptions. Alternatives we considered are the Box-Jenxins ARIMA models, and Hidden Markov Models. In the Box–Jenkins approach, trend and seasonal effects are treated as nuisance parameters. These effects are removed from the series before any analysis can begin. In Hidden Markov Models, the discretization of the level component into states would require imposing unnecessary assumptions on the behaviour of the system, let alone specifying an arbitrary number of states. "State space methods provide an explicit structural framework for the decomposition of time series in order to diagnose all the dynamics in the time series data simultaneously." [1]

2 Univariate model

We model the count of appearances of a term or termset, denoted term(set) from now on, at any point of time t as a local level model; each *observation* y_t is equal to the current level plus an *irregular component* ε_t. We assume the irregular components are drawn independently from a normal distribution with zero mean and a fixed variance; that is, it is homoscedastic *white noise*. The level component μ_t is assumed to be stochastic, with stochastic slope ν_t to account for the increasing number of users of social media. Sochasticity of both the level and the slope components is modelled as irregular components, ξ_t and ζ_t respectively. We assume ξ_t and ζ_t are also independent and normally distributed random variables with zero means and a fixed variances. The assumption of homoscedasticity might not be well justified for those two random variables, but it is an accepted simplification in the time series analysis of most domains except those with high volatility such as financial markets. We actually believe that there is evidence of volatility in the use of terms in social media, shown by the high churn in the 1000 most frequent terms hour after hour, and the high number of out of vocabulary terms. However, we chose to reduce the number of stochastic variables in the timeseries model of each term(set), and we can compensate for this by using wider confidence intervals (lower confidence levels) based on the deterministic variance.

We also add 2 seasonal components to model the natural (uninteresting) change in the use of certain terms. The first seasonal $\gamma_t^{(h)}$ models the seasonality according to the hour of day, since different terms are used in different hours of the day both within the same language and across languages. Within the same language, the frequency of some

terms such as "night, morning and lunch" increase as one region passes through different periods of the day. The frequency of all terms of a certain language also change as people in different locations on earth wake up, sleep and go through the various phases of the common 9-5 daily routine. The second seasonal $\gamma_t^{(d)}$ models the change of use of terms according to the day of the week, depending on how far the day is from the weekend. Therefore, there are $23 + 6 = 29$ seasonal equations all in all, and 27 of them are just for adjusting the lag of the seasonal effect. For simplicity we assume that the seasonal components are both deterministic; that is, with no corresponding irregular components. This assumption is supported by how there is always people going through a pattern change in their daily lives at each hour, and how the weekend days are different in different regions (Thursday and Friday in the Arab peninsula, for example).

Equations (1) represents the described model. The notation assumes that the time index t increases with each time step; whether continuous or discrete. The equations calculate the values of the next state with respect to the current state, because time series models are normally used to predict the next state then adjust the model according to the deviation from the actual observation. As explained in [1], "unlike classical regression, therefore, in state space methods the (hyper)parameter estimates are obtained by minimising the prediction errors and their variances, not by minimising the observation errors or disturbances and their variance." A popular method for this end is the use of a Kalman filter, but in its batch version it uses observations from the future to smooth the state and disturbances. It also does a second pass on the data for smoothing.

$$
\begin{aligned}
y_t &= \mu_t + \gamma_{1,t}^{(h)} + \gamma_{1,t}^{(d)} + \varepsilon_t, && \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \\
\mu_{t+1} &= \mu_t + \nu_t + \xi_t, && \xi_t \sim NID(0, \sigma_\xi^2) \\
\nu_{t+1} &= \nu_t + \zeta_t, && \zeta_t \sim NID(0, \sigma_\zeta^2) \\
\gamma_{1,t+1}^{(h)} &= -\textstyle\sum_{h=1}^{23} \gamma_{h,t}^{(h)} && \\
\gamma_{l,t+1}^{(h)} &= \gamma_{l-1,t}^{(h)}, && \forall l \in \{2 \dots 23\} \\
\gamma_{1,t+1}^{(d)} &= -\textstyle\sum_{d=1}^{6} \gamma_{d,t}^{(d)} && \\
\gamma_{l,t+1}^{(d)} &= \gamma_{l-1,t}^{(d)}, && \forall l \in \{2 \dots 6\}
\end{aligned}
\tag{1}
$$

3 Ranking according to user interest

To use this model to rank term(set)s according to users interest we differentiate between three cases, then we use rank fusion to rank term(set)s of high interest. The three cases are as follows:

1. To detect spikes of interest, the instantaneous residual of the slope ζ_t can be used. If its value falls over the upper bound of the confidence band this indicates a rise in the slope higher than expected, thus this difference can be used to rank term(set)s according to unexpected "acceleration" of interest. If CI(x) is one side of the confidence interval and E[x] is the expected value (assumed to be zero) acceleration is given by:

$$\mathcal{A} = \zeta_t - (E[\zeta_t] + CI(\zeta_t)) \propto Z = \frac{\zeta_t}{\sigma_{\zeta_t}} \tag{2}$$

2. To detect gradually increasing interest, we have to compare the current level with historic levels. However, to avoid keeping a history of levels per term(set)s, we note that our model is equivalent to an exponential moving average. Exponential moving average gives data an exponentially decreasing weight according to recency, therefore it will be higher than simple or cumulative moving average at times when the recent counts are higher than older ones. Therefore, it is enough to keep a moving average for each term(set), and use its difference form the level component to rank term(set)s according to gradual increase in interest. Since unweighted moving averages don't take trend into account when making predictions, we make our model's prediction closer to that of the moving average by subtracting the trend component from the level prediction. Thus the deviation from historic level (moving average) is given by:

$$\mathcal{H} = \mu_t - \nu_{t-1} - \bar{\mu}_t \tag{3}$$

3. The moving average is also a good indicator for term(set)s with sustained high popularity. The simple moving average is more responsive than the cumulative moving average since data points that are *too old* are subtracted, but calculating it over a history of n time steps requires keeping a history of n observations. Therefore we use the Cumulative Moving Average as an indicator of sustained interest:

$$\bar{\mu}_t = \frac{(t-1)\bar{\mu}_{t-1} + y_t}{t} \tag{4}$$

4 The normality assumption

The model in section 2 assumes that disturbances are white noise. On the other hand, the most exact explanation of the observations is that they are drawn from a Poisson distribution with mean μ_t. However, the Poisson distribution can be approximated by a Gaussian distribution given that we will consider term(set)s with high enough support; $minsupp \geq 30$. This support is not prohibitively high in the context of social networks, and it is possible to devise many well known methods of estimation under the assumptions of normality, independence and homoscedasticity. This is particularly important for extending the model to the multivariate case where the multivariate normal distribution is one of a few well studied distributions.

4.1 Alternative views

It is possible to view the observations as coming from a Binomial distribution, where each Tweet is a Bernoulli trial for the occurrence of each term(set). However, this view requires

fixing a number of Tweets for which the number of successes is estimated. With a varying rate of arrival of Tweets, this would mean that the epochs of the series of observations would correspond to different time intervals, making the model hard to interpret. One possible solution is to fix the time interval of the epoch, and explain the observations as coming from a Multinomial distribution on different *levels* of occurrence. Since the Dirichlet distribution is the conjugate prior of the Multinomial distribution, then it is possible to use estimation methods from Bayesian inference. In this case, the relative frequencies of the term(set)s must be used to determine their level of occurrence, as the counts are meaningless given that the number of Tweets per epoch would be different and the model doesn't account for the changing rate. The model of section 2 accounts for the changing rate of arrival of Tweets by incorporating trend and seasonal variables.

To work around the problem of fixing an arbitrary time step for all term(set)s, the observations can be viewed as drawn from a Negative Binomial distribution. In this case, the likelihood function for the parameter representing the number of failures until the experiment is stopped would be the central equation. Using this likelihood function, we would estimate the number of failures until the number of term(set) occurrences exceeds the minimum support. This estimate would then be divided by the current rate of arrival of Tweets to get the "support lag", which is anti-monotonic and suitable for spiky interest as the effect of the spike naturally fades with time.

5 Multivariate model

The model in section 2 is a simplified univariate model, presented to explain the intuition behind it. It is possible to use this univariate model for individual terms, assuming independence. However, the assumption of independence is unacceptable if the model would be used for term sets, where the occurrences of term sets cannot be independent of their subsets. One solution is to mine maximal term sets, but then it is impossible to mine term sets that evolve over time. The natural solution is thus to refuse the assumption of independence and use a multivariate model. The most important difference between the univariate and the multivariate model is not in the model equations, but rather in the distributions of the irregular components. The equations simply use matrix notation to represent M different observations, levels and trends (with all state equations augmented in one block matrix), where M is the number of distinct term(set)s. On the other hand, the variances of disturbances become covariance matrices; one for covariances of the measurements disturbances, and another for the covariances of all state disturbances (with covariance between different disturbances set to zero). Equation (5) shows the multivariate model, where bold symbols indicate vectors, capital symbols indicate matrices, $\mathbf{I_{RxC}}$ is the Identity matrix and $\mathbf{0_{RxC}}$ is a matrix of zeros of R row and C columns.

$$\begin{aligned} \mathbf{y_t} &= Z_t\boldsymbol{\alpha_t} + \boldsymbol{\varepsilon_t}, & \boldsymbol{\varepsilon_t} &\sim NID(0, H) \\ \boldsymbol{\alpha_{t+1}} &= T_t\boldsymbol{\alpha_t} + R_t\boldsymbol{\eta_t}, & \boldsymbol{\eta_t} &\sim NID(0, Q) \end{aligned} \tag{5}$$

where

$$Z = \left[\mathbf{I_{MxM}}\ \mathbf{0_{MxM}}\ \mathbf{0_{MxM}}\ \mathbf{0_{MxM}}\right],$$

$$T = \begin{bmatrix} \mathbf{I_{MxM}} & \mathbf{I_{MxM}} & \mathbf{0_{Mx23M}} & \mathbf{0_{Mx6M}} \\ \mathbf{0_{MxM}} & \mathbf{I_{MxM}} & \mathbf{0_{Mx23M}} & \mathbf{0_{Mx6M}} \\ \mathbf{0_{23MxM}} & \mathbf{0_{23MxM}} & \mathbf{S_{23}} & \mathbf{0_{23Mx6M}} \\ \mathbf{0_{6MxM}} & \mathbf{0_{6MxM}} & \mathbf{0_{6Mx23M}} & \mathbf{S_6} \end{bmatrix},\ R = \begin{bmatrix} \mathbf{I_{2Mx2M}} \\ \mathbf{0_{29Mx2M}} \end{bmatrix},$$

$$\boldsymbol{\alpha'} = \left(\mu^{(1)} \dots \mu^{(M)}\ \nu^{(1)} \dots \nu^{(M)}\ \gamma_1^{(h,1)} \dots \gamma_1^{(h,M)} \dots \gamma_{23}^{(h,M)}\ \gamma_1^{(d,1)} \dots \gamma_1^{(d,M)} \dots \gamma_6^{(d,M)}\right),$$

$$\boldsymbol{\eta} = \begin{pmatrix} \xi^{(1)} \\ \vdots \\ \xi^{(M)} \\ \zeta^{(1)} \\ \vdots \\ \zeta^{(M)} \end{pmatrix},\ \mathbf{S}_l = \begin{bmatrix} -\mathbf{I_{MxM}} & -\mathbf{I_{MxM}} & \dots & -\mathbf{I_{MxM}} \\ \mathbf{I_{(l-1)Mx(l-1)M}} & & \mathbf{0_{MxM}} & \end{bmatrix},$$

$$H = \mathbf{COV}(\varepsilon),\ Q = \begin{bmatrix} \mathbf{COV}(\xi) & \mathbf{0_{MxM}} \\ \mathbf{0_{MxM}} & \mathbf{COV}(\zeta) \end{bmatrix},$$

$$\mathbf{COV}(x) = \begin{bmatrix} \sigma^2_{x^{(1)}} & cov(x^{(1)}, x^{(2)}) & \dots & cov(x^{(1)}, x^{(M)}) \\ \vdots & \ddots & & \vdots \\ cov(x^{(M)}, x^{(1)}) & cov(x^{(M)}, x^{(2)}) & \dots & \sigma^2_{x^{(M)}} \end{bmatrix}$$

References

[1] J. J. F. Commandeur. *An introduction to state space time series analysis CN - Guelph/Laurier/Waterloo users -(QA280) CN - Guelph McLaughlin Book Stacks(QA280 .C65 2007) CN - UW Davis. Book Stacks. Main Floor(QA280 .C65 2007)*. Oxford University Press, Oxford, 2007.