Technical details of the linear model calculations for the article Prediction of employment and unemployment rates from Twitter daily rhythms in the US

Eszter Bokányi, bokanyi@complex.elte.hu*,

Zoltán Lábszki, labszkizoltan@gmail.com,

Gábor Vattay, vattay@complex.elte.hu

Department of Physics of Complex Systems

Pázmány Péter sétány 1/A, Eötvös Loránd University, Budapest H-1117, Hungary

* corresponding author

Technical details for the Methods section

We define a daily activity pattern with hourly resolution for each county that are enumerated by $k = 1 \dots M$. Thus, each county (k) is represented by a 24-dimensional vector ($\mathbf{y}^{(k)}$), where the elements of $\mathbf{y}^{(k)}$ are aggregated normalized hourly tweeting activities.

We assume that the tweeting pattern of a county can be represented by the linear combination of only two universal patterns ($\mathbf{A}$ and $\mathbf{B}$) that are mixed for each county k with a proportion of $\alpha^{(k)}$ and $1 - \alpha^{(k)}$, respectively. We have no further restriction on these $\alpha^{(k)}$ values, they can be any arbitrary real numbers. $\mathbf{A}$ and $\mathbf{B}$ are both 24-dimensional vectors normalized to 1, the 24 dimensions representing the 24 hours of the day.

Then the predicted activity $x_i^{(k)}$ of a county k in hour i would be

$$x_i^{(k)} = \alpha^{(k)} \cdot A_i + (1 - \alpha^{(k)}) \cdot B_i = \alpha^{(k)}(A_i - B_i) + B_i. \tag{1}$$

Let us denote the weight of each county by $w^{(k)}$, which is proportional to its population $p^{(k)}$, such that $w^{(k)} = p^{(k)} / \sum_{k=1}^{M} p^{(k)}$. We then define the squared error of our model as

$$E = \sum_{i,k} w^{(k)} \left(y_i^{(k)} - \underbrace{\left(\alpha^{(k)}(A_i - B_i) + B_i \right)}_{x_i^{(k)}} \right)^2 .$$

We would like to minimize this error with subject to the two conditions $\sum_i A_i = 1, \sum_i B_i = 1$, which leads to the following expression to minimize with Lagrange multipliers λ_a and λ_b:

$$E + \lambda_a \left(\sum_i A_i - 1 \right) + \lambda_b \left(\sum_i B_i - 1 \right) = \min. \tag{2}$$

The derivatives yield the following linear equation system:

$$\frac{\partial}{\partial A_j}: \quad \sum_k 2w^{(k)} \left(y_j^{(k)} - \alpha^{(k)}(A_j - B_j) - B_j \right) \left(-\alpha^{(k)} \right) + \lambda_a = 0 \tag{3}$$

$$\frac{\partial}{\partial B_j}: \quad \sum_k 2w^{(k)} \left(y_j^{(k)} - \alpha^{(k)}(A_j - B_j) - B_j \right) \left(-(1 - \alpha^{(k)}) \right) + \lambda_b = 0 \tag{4}$$

$$\frac{\partial}{\partial \alpha^{(m)}}: \quad \sum_i 2w^{(m)} \left(y_i^{(m)} - \alpha^{(m)}(A_i - B_i) - B_i \right) (-(A_i - B_i)) = 0 \tag{5}$$

Summing Eq 3 and Eq 4 for j yield 0 for the Lagrange multipliers λ_a and λ_b. Thus, the problem reduces to minimizing E, which actually measures the sum of squared distances from the line parametrized by $\mathbf{A} - \mathbf{B}$, $\mathbf{B}$ and $\alpha^{(k)}$ for a county k.

Since

$$\sum_j [(3) + (4)] \cdot (A_j - B_j) = \sum_m (5), \tag{6}$$

the equation system is not linearly independent. Thus, we cannot obtain all exact values for A_j, B_j and $\alpha^{(k)}$, they will be dependent on each other.

Expressing $\alpha^{(k)}$ from our equation system yields:

$$\alpha^{(m)} = \frac{\sum_i \left(y_i^{(m)} - B_i \right) (A_i - B_i)}{\sum_i (A_i - B_i)^2} = \frac{(\mathbf{y}^{(m)} - \mathbf{B})(\mathbf{A} - \mathbf{B})}{(\mathbf{A} - \mathbf{B})^2}. \tag{7}$$

The line from which the summed distance of the datapoints is minimal is the line whose direction is parallel to the eigenvector ($\mathbf{m}$) corresponding to the largest eigenvalue of the covariance matrix $\mathbf{C}$, where

$$C_{ij} = \langle y_i y_j \rangle - \langle y_i \rangle \langle y_j \rangle, \tag{8}$$

if $\langle\rangle$ denotes the weighted mean ($\sum_k w^{(k)} = 1$, $w^{(k)} \geq 0 \, \forall k = 1 \ldots M$)

$$\langle y_j \rangle = \sum_k w^{(k)} y_j^{(k)}. \tag{9}$$

By substituting the expression for $\alpha^{(k)}$ into Eq 3)+Eq 4, and averaging over k we get that the point $\langle \mathbf{y} \rangle$ should fit onto our line.

Thus, we get a valid solution of our error minimization problem, if we choose

$$\sigma(\alpha) = \sqrt{\frac{\sum_{k=1}^{M} \left(\alpha^{(k)}\right)^2}{M}},$$

$$\mathbf{A} = \langle \mathbf{y} \rangle + 2 \cdot \mathbf{m} \cdot \sigma(\alpha), \tag{10}$$

$$\mathbf{B} = \langle \mathbf{y} \rangle - 2 \cdot \mathbf{m} \cdot \sigma(\alpha), \tag{11}$$

and calulate $\alpha^{(k)}$ values according to Eq 7.