4/30/25

Least Squares Regression and Interaction Matrices for a Finite Dimensional Data Set

APPM 3310

CHIARA PESCE, DANNY ALEMAYEHU, CALEN MICHEL

Abstract

Understanding how variables in air quality impact one another is essential for assessing pollutant behavior and developing long-term solutions to air quality challenges. This project employs least-squares linear regression in combination with the construction of an interaction matrix to classify variables (such as carbon monoxide (CO), carbon dioxide (CO2), and volatile organic compounds (VOCs)) with respect to their systemic roles. Specifically, each variable is categorized as more or less interactive, and as dominant or subordinate within the system interaction. The dataset used in this analysis was intended to investigate carbon monoxide concentrations from tailpipe emissions with respect to distance from a highway in high alpine environments, however any data set can be analyzed using the processes outlined in this paper.

Attribution

~

Introduction

Interaction matrices have been used in environmental science since the development of the Leopold Matrix in 1971 (Leopold et al., (1971)) primarily as a tool to identify and effectively weight different activities to assess the potential environmental impact of a system. The interaction matrix utilized in this paper is a modification on the Leopold matrix originally proposed by Mavroulidou et al. (2004) to assess air quality in urban environments along with GIS mapping, the latter of which will not be investigated in this paper. A simple interaction matrix follows the format below:

$$\begin{bmatrix} Parameter\ 1 & \cdots & I_{1,n} \\ \vdots & \ddots & \vdots \\ I_{n,1} & \cdots & Parameter\ n \end{bmatrix}$$

Where the parameters of interest are placed along the diagonal and interaction coefficients are placed in each of the off-diagonal entries. A clockwise convention is adopted such that $I_{1,n}$ is the influence of parameter 1 on parameter n and $I_{n,1}$ is the influence of parameter n on parameter 1 for an n dimensional system. The challenge then comes with assigning appropriate weighing factors for each interaction. For this paper, we will employ linear least-squares regression to numerically calculate each interaction value.

Linear regression will be performed between each variable of interest and the value of the least-squares line will be taken as the coefficient for the variable interaction. For example, in the matrix described above, the slope from least-squares computation with parameter 1 as the independent variable and parameter n as the dependent variable will be taken as the value for $I_{1,n}$. In this way, we construct an interaction matrix representative of the system.

$$\begin{bmatrix} & j=1 & j=2 & j=3 & Cause \sum I_j \\ i=1 & Parameter\ 1 & I_{1,2} & I_{1,3} & I_{1,2}+I_{1,3} \\ i=2 & I_{2,1} & Parameter\ 2 & I_{2,3} & I_{2,1}+I_{2,3} \\ i=3 & I_{3,1} & I_{3,2} & Parameter\ 3 & I_{3,1}+I_{3,2} \\ Effect \sum I_i & I_{2,1}+I_{3,1} & I_{1,2}+I_{3,2} & I_{1,3}+I_{2,3} & \sum I \end{bmatrix}$$

From here, each row and column can be summed and expressed as a percentage of the whole such that each variable can be classified as more or less interactive and subordinate or dominant with respect to the system (Mavroulidou et al. (2007)). A representative chart can then be created to visually classify each variable in the categories listed above.

In this paper, the parameters of interest are chosen to be relative humidity (%), temperature (°C), light VOCs (ADU), heavy VOCs (ADU), ozone (ADU), CO (ADU), CO_2 (ADU), elevation (ft) and distance from the neatest highway (miles).

Mathematical Formulation

The bulk of the relevant computation will be performing least squares between each respective variable in the data set. Let the basic equation for a linear line be:

$$y = \beta_0 + \beta_1 \cdot x$$

For an nth dimensional system of equations, we can form the system

$$\begin{bmatrix} y_1 = \beta_0 + \beta_1 \cdot x_1 \\ y_2 = \beta_0 + \beta_1 \cdot x_2 \\ \vdots \\ y_n = \beta_0 + \beta_1 \cdot x_n \end{bmatrix}$$

Which can be represented in matrix form by the following equations.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

This system follows the standard form $\vec{b} = A \cdot \vec{x}$ where $\vec{b} \in \mathbb{R}^n$, $A \in \mathbb{M}^{nx2}$, and $\vec{x} \in \mathbb{R}^2$. As you may note, this system is overdetermined and thus inconsistent. A is rectangular and noninvertible, however its columns are linearly independent $\forall\ x_1 \neq x_2 \cdots \neq x_n$. This means that $\vec{b} \notin img(A)$, and thus no solution exists. For an inconsistent system, the best way to approximate a solution is to project $\vec{b}$ onto the img(A) which yields the smallest possible deviation from the system solution (Olver, Chehrzad (2018)). Let the vector $\vec{b}$ be represented as

$$\vec{b} = \vec{b}^* + (\vec{b} - \vec{b}^*)$$

Where $\vec{b}^*$ represents the component of $\vec{b}^* \in img(A)$ and the quantity $(\vec{b} - \vec{b}^*)$ is component of the vector $\vec{b}$ perpendicular to the img(A). Since the img(A) and the ker (A^T) are complimentary subspaces, we can then conclude that the quantity $(\vec{b} - \vec{b}^*) \in$ ker (A^T). Using this idea, we can multiply the equation $\vec{b} = A \cdot \vec{x}$ by A^T to eliminate the component of $\vec{b} \in$ ker (A^T).

$$A^T \cdot \vec{b} = A^T \cdot A \cdot \vec{x}$$

Since the columns of A are linearly independent , the matrix formed by computing $A^T \cdot A$ is invertible (Olver, Chehrzad (2018)) and thus we can rearrange the above equation to get

$$A \cdot (A^T \cdot A)^{-1} \cdot A^T \cdot \vec{b} = A \cdot \vec{x} = \vec{b}^*$$

Where the sequence $A \cdot (A^T \cdot A)^{-1} \cdot A^T$ represents a projection matrix for the system. In this way, we can then solve the system for $\vec{x}$ and thus obtain our respective values of β_0 and β_1, taking β_1 as our interaction coefficient for the respective parameters in the interaction matrix.

Computationally, this method can induce large amounts of floating point error especially since the condition number of $A^T \cdot A$ will be the condition number of A squared. Because of this computational limit, we may choose to create an orthogonal basis for A using QR factorization such that Q forms an orthonormal basis on $\mathbb{M}^{nx2}$, and R is an upper triangular, invertible matrix on $\mathbb{M}^{2x2}$.

$$(Q \cdot R)^T \cdot \vec{b} = (Q \cdot R)^T \cdot Q \cdot R \cdot \vec{x}$$

And thus

$$Q \cdot R \cdot \left((Q \cdot R)^T \cdot (Q \cdot R)\right)^{-1} \cdot (Q \cdot R)^T \cdot \vec{b} = (Q \cdot R) \cdot \vec{x}$$

Which will simplify down to

$$Q^T \cdot \vec{b} = R \cdot \vec{x}$$

$$Q \cdot Q^T \cdot \vec{b} = \vec{b}^*$$

Where the sequence $Q \cdot Q^T$ represents a projection matrix for the system. Similarly, we can then solve the system for $\vec{x}$ and thus obtain our respective values of β_0 and β_1, taking β_1 as our interaction coefficient for the respective parameters in the interaction matrix. This will be more numerically stable than using the normal equations.

References

Maria Mavroulidou; Susan J. Hughes; Eemma E. Hellawell. (2004). “A qualititative tool combining an interaction matrix and a GIS to map vulnerability to traffic induced air pollution”. *Journal of Environmental Management*, Volume 70, Issue 4. p. 283-289. https://doi.org/10.1016/j.jenvman.2003.12.002.

Maria Mabroulidou; Susan J. Hughes; Emma E. Hellawell. (2007). “Developing the interaction atrix technique as a tool assessing the impact of traffic on air quality”. *Journal of Environmental Management*, Volume 84, Issue 4. p. 513-522. https://doi.org/10.1016/j.jenvman.2006.07.002.

Mazzoccola D. F; Hudson J. A. (1996). “A comprehensive method of rock mass characterization for indication natural slope instability”. *Quarterly Journal of Engineering Geology*, Volume 29, Issue 1. The Geological Society of London. doi:10.1144/GSL.QJEGH.1996.029.P1.03

Leopold, Luna Bergere; Clarke, Frank Eldridge; Hanshaw, Bruce B.; Balsley, James R. (1971). “A Procedure for Evaluating Environmental Impact”. *Circular.* Washington, D.C. p. 19. doi:10.3133/cir645 – via U.S. Geological Survey.

Olver, Peter J, and Chehrzad Shakiban. *Applied Linear Algebra*. 2nd ed., Springer International Publishing, 2018.

Watson, Geoffrey S. “Linear Least Squares Regression.” *The Annals of Mathematical Statistics*, vol. 38, no. 6, 1967, pp. 1679–99. *JSTOR*, http://www.jstor.org/stable/2238648.

Appendix

Data set used: https://o365coloradoedu-my.sharepoint.com/:x:/g/personal/chpe5809_colorado_edu/Ea6s5LFXd-BLqcmaAcQ0UIkBUFGun13I2MpWIhNEBjj3Ig?e=xTc8E1