Correspondence Analysis of Government Expenditure Patterns

Hsiang Hsu[1], Flavio P. Calmon[1], José Cândido Silveira Santos Filho[1], Andre P. Calmon[2], Salman Salamatian[3]

HARVARD School of Engineering and Applied Sciences

[1]School of Engineering and Applied Science, Harvard University
[2]Technology and Operations Management, INSEAD
[3]Research Laboratory of Electronics , Massachusetts Institute of Technology

Background

- Open data trend in executive and legislative branches of governments: transparency ↑, corruption ↓, and democratic institutions ↑
- Machine learning to quantify, model, and evaluate the performance of public administration
- An active area of research in social/ political science: **Public Expenditure Analysis (PEA)**
- Why Brazil?
  - Brazilian government has a large open data initiative: untouched by Advanced ML!
  - High-profile budget misuse problems: $>$ 30% of congress members under investigation!
- Goal: **Introducing a neural network-based method to analyze expenditure patterns**

Data Description

- **Operation Serenata de Amor**: Discretionary expenditure by Brazilian Congress members
- Made available by Brazil government
- 513 Congress members/ 26 states/ $>$ 30 parties in Brazil
- $>$ 7 million expenditure records from 2009 to 2018
- Monthly budgets: BRL$ 45k (≈USD$ 13k)
- Data pre-processing
  - Translation: Portuguese ⇒ English
  - Most recent term: 2015 - 2018
  - Dropped missing data/ eliminated categories appearing less than 500 times.
  - 1.1 million records/ 16 categories/ 595 congress members/ 26 parties/ 27 states

Congress Member ID	Start Term	State	Party	Category	vendor	Value
213	2015	RR	DEM	Maintenance of office	COMPANHIA DE AGUAS E ESGOTOS DE RORAIMA	BRL$ 165.65
452	2015	RR	DEM	Fuel and lubricants	CASCOL COMBUSTIVEIS PARA VEICULOS LTDA	BRL$ 40
97	2015	CE	PODE	Food for the congressperson	BOI ZANGADO_FRANCISCA DE SENA VASCONCELOS - ME	BRL$ 52.4

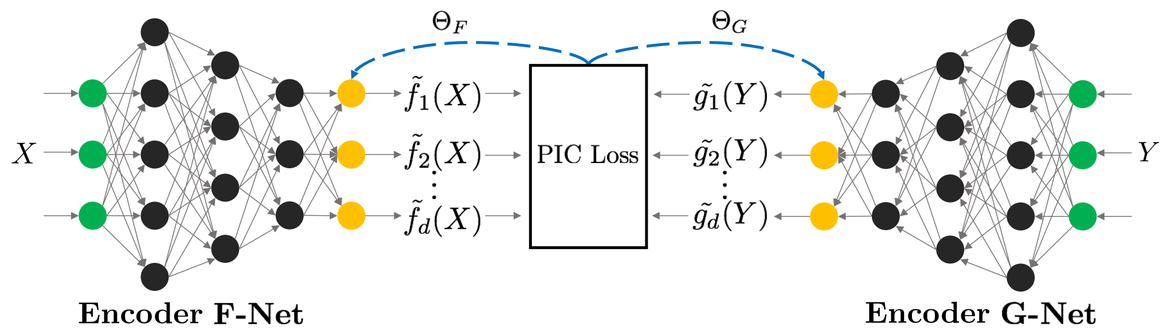
X — Data samples in Operation Serenata de Amor — Y

Methodology

- **Correspondence Analysis (CA)**
  - An exploratory multivariate statistical technique
    ⇒ used in genealogy, epidemiology, social and environmental sciences
  - Similar to PCA/CCA: Low-dimensional orthogonal representations
    ⇒ visualization/ interpretability
  - Consider two random variables X and Y of finite cardinality

$$\mathbf{Q} \triangleq \mathbf{D}_X^{-1/2}(\mathbf{P}_{X,Y} - \mathbf{p}_X\mathbf{p}_Y^T)\mathbf{D}_Y^{-1/2},\ \mathbf{D}_X \triangleq \mathrm{diag}(\mathbf{p}_X), \mathbf{D}_Y \triangleq \mathrm{diag}(\mathbf{p}_Y)$$
$$= \mathbf{U}\mathbf{Z}\mathbf{V}^T \quad \text{(Singular Value Decomposition)}$$

  - d = $\min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1$, $\{\sigma_i\}_{i=1}^{d}$: the singular values
    - Orthogonal factors of X: $\mathbf{L} \triangleq \mathbf{D}_X^{-1/2}\mathbf{U}$
    - Orthogonal factors of Y: $\mathbf{R} \triangleq \mathbf{D}_Y^{-1/2}\mathbf{V}$
    - Factor scores: $\lambda_i = \sigma_i^2, 1 \le i \le d$
    - Factor score ratios: $\frac{\lambda_i}{\sum_{i=1}\lambda_i}, 1 \le i \le d$
  - The first and second columns of **L** and **R** plotted on a 2-D plane: factoring plane
- **Limitations of SVD-based CA:** restricted to categorical data and requires estimating $\mathbf{P}_{X,Y}$
- **A novel neural network-based approach for CA:**
  - $\tilde{\mathbf{f}}(X) \triangleq [\tilde{f}_1(X), \cdots, \tilde{f}_d(X)]^\intercal \in \mathbb{R}^{d\times 1}$, and $\tilde{\mathbf{g}}(Y) \triangleq [\tilde{g}_1(Y), \cdots, \tilde{g}_d(Y)]^\intercal \in \mathbb{R}^{d\times 1}$
  - $\mathbf{C}_f = \mathbb{E}[\tilde{\mathbf{f}}(X)\tilde{\mathbf{f}}(X)^\intercal]$, $\mathbf{C}_{fg} = \mathbb{E}[\tilde{\mathbf{f}}(X)\tilde{\mathbf{g}}(Y)^\intercal]$, $\|\mathbf{Z}\|_d$: d-th Ky-Fan norm
  - Loss Function for back-propagation: $\min_{\tilde{\mathbf{f}},\tilde{\mathbf{g}}} -2\|\mathbf{C}_f^{-\frac{1}{2}}\mathbf{C}_{fg}\|_d + \mathbb{E}[\|\tilde{\mathbf{g}}(Y)\|_2^2]$
  - $\tilde{\mathbf{f}}(X)$ and $\tilde{\mathbf{g}}(Y)$: generalizations of **L** and **R**



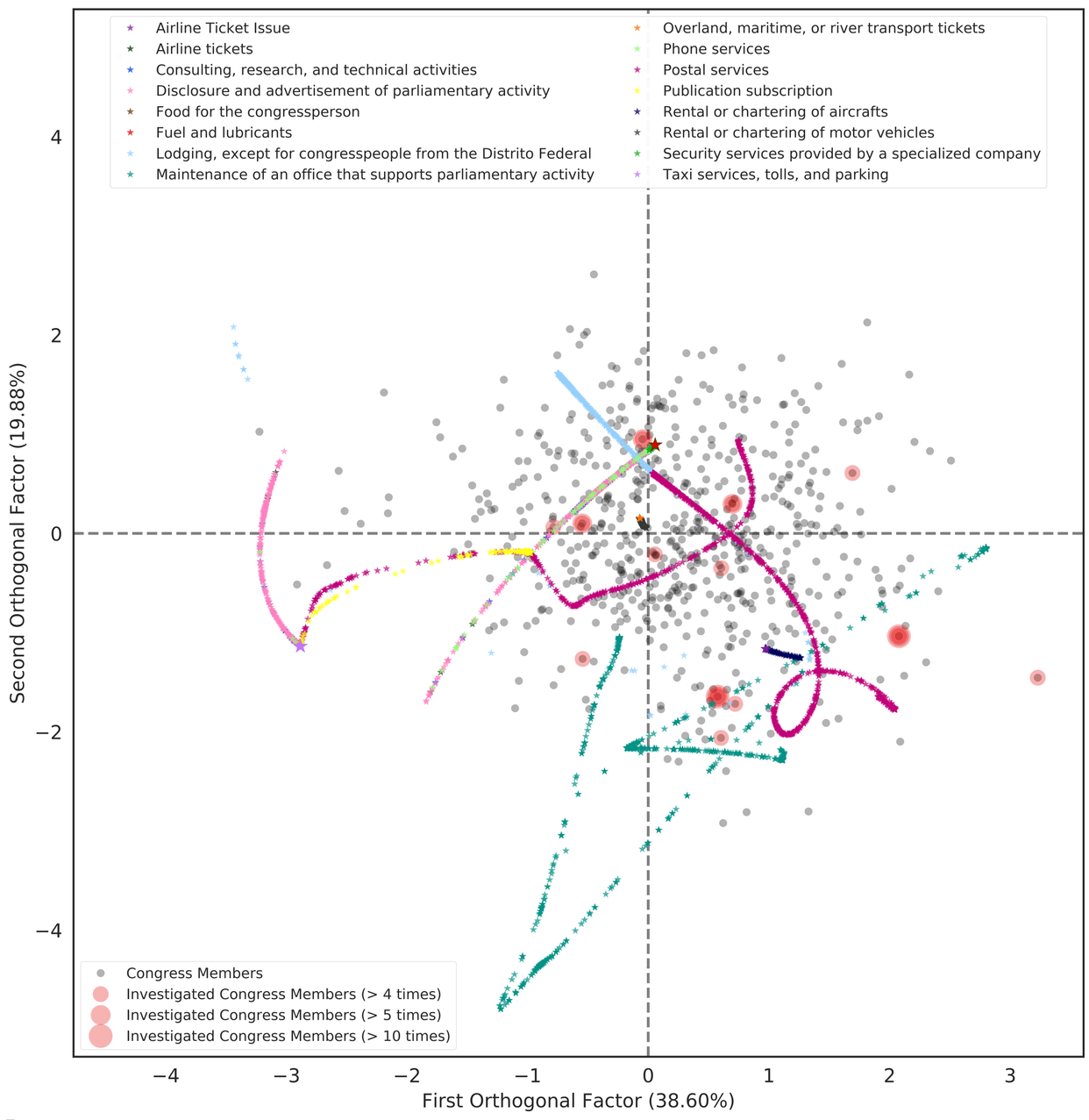
The architecture of the CA-NN, consisting of two encoders F-Net and G-Net for X and Y respectively to estimate the orthogonal factors.

Main Results and Discussions

- **Expenditure Pattern**
  - Automatically clusters related expenses together since they have close patterns.
    Aviation-related expenses: "Airline Ticket Issue", "Rental of aircrafts"
    Transportation-related expenses: "River transport tickets", "Rental of motor vehicles"
    Daily expenses: "Food", "Fuel", "Security services"
  - Certain categories not correlated with congress members: "Food", "Fuel and lubricants", "River transport tickets", "Rental of motor vehicles", "Security services", and "Taxi services and parking".
  - High-variation, overlapping traces of "Publication subscription" and "Postal service", and "Airline tickets", "Consulting, research, and technical activities" and "Disclosure and advertisement of parliamentary activity": mishandling of this category by certain congress members
  - "Maintenance of an office" and "Lodging": outlying patterns
- **Charged Congress members**
  - Investigated congress members near expenditure patterns with large variation
  - Ongoing work: Discretionary funding may be predictive of budget misuse problems

Potential Use Cases

- Anomalous expenditure discovery, interpretation and visualization
- Clustering of congress members in terms of their discretionary expenditure pattern
- Algorithmic watchdogs for predictive models of budget misuse: proactive reactions
- New methodological approaches transferable to other civic projects for government transparency



- First factoring plane of the expenditures of 595 congress members in Brazil from 2015 to 2018

$$(\tilde{f}_1(x_i), \tilde{f}_2(x_i)), \text{ and } (\tilde{g}_1(y_i), \tilde{g}_2(y_i)),\ \forall i$$

- Higher factor score ratio: more correlation captured by the orthogonal factor
- Colored traces: 16 expenditure patterns for all congress members
- Grey dots: congress members without investigations
- Red dots: congress members under investigations
- Points and lines close to the center (the origin): small correlation.

Selected Reference

- Bates, J. (2012). this is what modern deregulation looks like: co-optation and contestation in the shaping of the UK open government data initiative. The Journal of Community Informatics, 8(2).
- Shah, A. (2005). Public Expenditure Analysis. The World Bank.
- Winter, B. (2017). Brazils never-ending corruption crisis: Why radical transparency is the only fix. Foreign Aff., 96:87.
- Greenacre, M. J. (1984). Theory and applications of correspondence analysis. London Academic Press.
- Hsu, H., Salamatian, S., and Calmon, F. P. (2018). Deep orthogonal representations: Fundamental properties and applications. arXiv preprint arXiv:1806.08449.

Contact

Hsiang Hsu

Paper

GitHub and Data Release