Spherical Regression under Mismatch Corruption with Application to Automated Knowledge Translation

Xu Shi, Xiaoou Li, Tianxi Cai
(STAT 8913 presentation Sungmin Park)

November 3, 2022

Spherical regression

Multivariate linear regression

- $X_i \in \mathbb{R}^p$ predictor,
- $Y_i \in \mathbb{R}^p$ response,
- $\mathbb{W} \in \mathbb{R}^{p \times p}$ regression parameter.

$$\mathbf{E}(Y_i | X_i) = \mathbb{W}^T X_i.$$

Spherical regression:

- $X_i, Y_i \in \mathcal{S}^{p-1}$ (unit sphere in $\mathbb{R}^p$),
- $\mathbb{W} \in \mathbb{R}^{p \times p}$ and $\mathbb{W}^T \mathbb{W} = I$ (orthogonal)
- That is, $\|X_i\| = \|\mathbb{W} X_i\| = \|Y_i\| = 1$.

von Mises-Fisher distribution

- The density of von Mises-Fisher (vMF) distribution us

$$f_p(x;\mu,\kappa) = C_p(\kappa)\exp(\kappa\mu^T x), \quad C_p(\kappa) = \frac{\kappa^{\frac{p}{2}-1}}{(2\pi)^{\frac{p}{2}} I_{\frac{p}{2}-1}(\kappa)}$$

  where $\kappa \geq 0, \|\mu\| = 1$, and $I_v(\cdot)$ denotes the modified Bessel function of the first kind at order v.
- For $Y \sim N(\mu, \frac{1}{\kappa}I)$, conditional distribution of Y given $\|Y\| = 1$ is a vMF distribution.

Examples of spherical regression?

- Directional data, word embedding
- For example, words are translated into vectors in natural language processing (NLP) the similarities in word embedding are often measured with cosine similarity:

$$\text{sim}(A, B) = \cos(\theta) = \frac{\langle A, B \rangle}{\|A\|\|B\|}.$$

Mismatch

The implicit assumption of a regression problem is that we observe the predictor and response in pairs (X_i, Y_i). In some situations, we may have data mismatch

Y_1	X_1
Y_2	X_4
Y_3	X_3
Y_4	X_2
Y_5	X_5

Table: Mismatch example

and observe $(X_{\pi(i)}, Y_i)$ where π is a permutation. In the above example, $\pi = (1, 4, 3, 2, 5)$.

Notation

Regression with mismatch

$$\mathbf{E}(Y_i|X) = \mathbb{W}^T X_{\pi(i)}$$

In matrix notation

$$\mathbf{E}(Y|X) = \Pi X \mathbb{W}$$

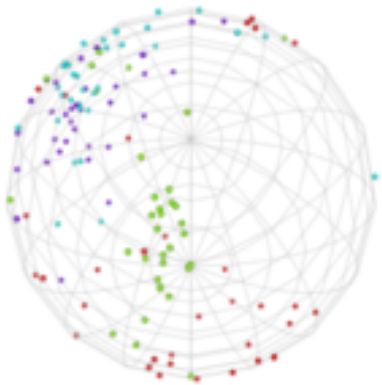
- $Y = [Y_1, \ldots, Y_n]^T$,
- $X = [X_1, \ldots, X_n]^T$,
- $\Pi \in \mathbb{R}^{n \times n}$ permutation matrix corresponding to π.

Automated Knowledge translation of medical codes

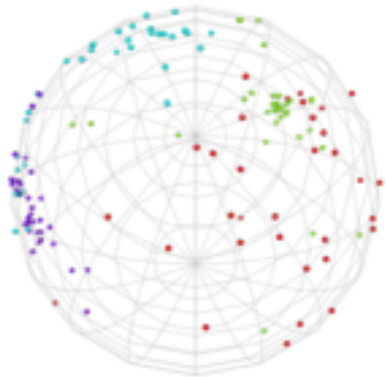
The goal is to combine different sources of electronic health records (EHR) to improve the quality of health information for the providers, clinicians, and patients.

- (spherical) Medical codes are translated into word embedding.
- (regression) Map medical codes from two different EHR systems.
- (mismatch) Medical codes from different EHR systems may not have exact match.

Exmaple: EHR embedding vectors



(a) Veterans Health Administration (VHA)



(b) Partners HealthCare Systems (PHS)

Figure: Embedding vectors in VHA and PHS

One-to-many mapping mismatch

- For permutation matrix Π, every response Y_i can correspond to exactly one predictor $X_{\pi(i)}$: one-to-one mapping.
- However, a medical code Y_i from one system may correspond to multiple medical codes $X_{j_1}, \ldots, X_{j_k}$ from another system in EHR data: one-to-many mapping.

ICD-10	Description		ICD-9	Description
A040	ENTEROPATHOGENIC ESCHERICHIA COLI INFECTION	→	00801	INTESTINAL INFECTION DUE TO ENTEROPATHOGENIC E. COLI
A041	ENTEROTOXIGENIC ESCHERICHIA COLI INFECTION	→	00802	INTESTINAL INFECTION DUE TO ENTEROTOXIGENIC E. COLI
A042	ENTEROINVASIVE ESCHERICHIA COLI INFECTION	→	00803	INTESTINAL INFECTION DUE TO ENTEROINVASIVE E. COLI
A043	ENTEROHEMORRHAGIC ESCHERICHIA COLI INFECTION	→	00804	INTESTINAL INFECTION DUE TO ENTEROHEMORRHAGIC E. COLI
A044	OTHER INTESTINAL ESCHERICHIA COLI INFECTIONS	→	00800	INTESTINAL INFECTION DUE TO E. COLI UNSPECIFIED
		→	00809	INTESTINAL INFECTION DUE TO OTHER INTESTINAL E. COLI INFECTIONS

Figure: Some examples of medical code mapping

One-to-many mapping mismatch

For example,

EHR 1	EHR 2
Y_1	X_1
Y_2	X_3, X_4
Y_3	X_3
Y_4	X_2
Y_5	X_5

Table: Mapping between different EHR embedding

the mapping Π is

$$\Pi = \begin{pmatrix} 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & * & * & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix}.$$

Problem statement

For spherical data $X = [X_1, \ldots, X_n]^T$, $Y = [Y_1, \ldots, Y_n]^T$ with $X_i, Y_j \in \mathcal{S}^{p-1}$, find Π and $\mathbb{W}$ that minimizes the squared error loss of the model $Y = \Pi X \mathbb{W} + E$:

$$\min_{\Pi, \mathbb{W}} \|Y - \Pi X \mathbb{W}\|_F^2$$

where Π is a mismatch matrix allowing both one-to-one and one-to-many mapping, and $\mathbb{W}$ is an orthogonal matrix.

Additional structure on mismatch

Moreover, assume that the mismatch only occurs within known groups indexed by $\{G_1, \ldots, G_K\}$, i.e. Π has a block diagonal structure. It is a reasonable assumption in EHR data where medical codes are often categorized within larger groups e.g. mental, eye, nutrition, blood, etc.

$$\Pi = \begin{pmatrix} \Pi_1 & \cdot & \cdot & \cdot \\ \cdot & \Pi_2 & \cdot & \cdot \\ & & \ddots & \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \Pi_K \end{pmatrix}, \quad \Pi_k : \text{ mismatch within group } k$$

An additional assumption is that the number of mismatch is small $n_{\text{mis}} = o(n)$. That is, the rows of Π and the identity matrix are the same except for n_{mis} cases.

Iterative spherical regression mapping (iSphereMAP)

Joint estimation of Π and $\mathbb{W}$ is a difficult problem.

$$\min_{\Pi,\mathbb{W}} \|Y - \Pi X \mathbb{W}\|_F^2 \quad \text{s.t.} \quad \mathbb{W}^T\mathbb{W} = I$$

To bypass this problem,

1. ignore mismatch ($\Pi = \hat{\Pi}^{[1]} = I$) and estimate $\hat{\mathbb{W}}^{[1]}$,
2. update mismatch $\hat{\Pi}^{[2]}$ given $\mathbb{W} = \hat{\mathbb{W}}^{[1]}$,
3. update $\hat{\mathbb{W}}^{[2]}$ given $\Pi = \hat{\Pi}^{[2]}$.

Step 1 (iSphereMAP)

Ignore Π, that is $\hat{\Pi}^{[1]} = I$ (identity) and find

$$\hat{\mathbb{W}}^{[1]} = \underset{\mathbb{W}:\mathbb{W}^T\mathbb{W}=I}{\operatorname{argmin}} \|Y - X\mathbb{W}\|_F^2.$$

This is called an *orthogonal Procrustes problem* where the solution is

$$\hat{\mathbb{W}}^{[1]} = UV^T$$

for singular value decomposition $X^TY = UDV^T$.

Step 2 (iSphereMAP)

Update Π given $\hat{\mathbb{W}}^{[1]}$ under group structure and sparsity:

1. group structure: obtain $\tilde{\Pi} = \mathrm{diag}\{\tilde{\Pi}^1, \ldots, \tilde{\Pi}^K\}$ where

$$\begin{aligned}\tilde{\Pi}^k &= \underset{\Pi^k}{\mathrm{argmin}} \left\| Y_{[G_k,:]} - \Pi^k X_{[G_k,:]} \hat{W}^{[1]} \right\|_F^2 \\ &= Y_{[G_k,:]} (X_{[G_k,:]} \hat{W}^{[1]})^T (X_{[G_k,:]} X_{[G_k,:]})^{-1}\end{aligned}$$

   and $X_{[G_k,:]}, Y_{[G_k,:]}$ are rows of X, Y with indexes G_k.

2. sparsity: define dissimilarity criterion

$$\tilde{\beta}_i = 1 - \max_j \cos(\tilde{\Pi}_{i\cdot}, I_{j\cdot}), \quad \tilde{j}_i = \underset{j}{\mathrm{argmax}} \cos(\tilde{\Pi}_{i\cdot}, I_{j\cdot})$$

   where $\tilde{\Pi}_{i\cdot}, I_{j\cdot}$ is the ith and jth row of $\tilde{\Pi}$ and I, and apply hard-thresholding

$$\hat{\Pi}_{i\cdot}^{[2]} = I_{\tilde{j}_i\cdot} \mathbf{1}(\tilde{\beta}_j \le \lambda_n) + \frac{\tilde{\Pi}_{i\cdot}}{\|(\tilde{\Pi}_{i\cdot} X)^T\|_2} \mathbf{1}(\tilde{\beta}_i > \lambda_n)$$

   for some threshold λ_n.

Step 3 (iSphereMAP)

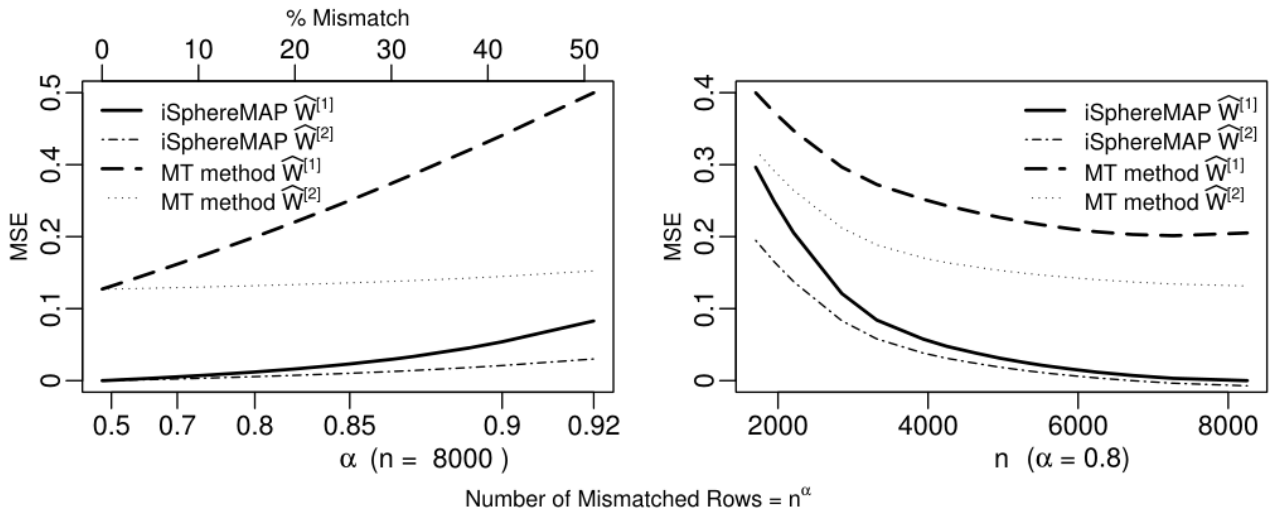
Refine estimate of $\mathbb{W}$, only using the indexes $\mathcal{S}(\hat{\Pi}^{[2]}) = \{i : \hat{\Pi}^{[2]}_{i\cdot} = I_{i\cdot}\}$ (estimated indexes of correct match):

$$\hat{\mathbb{W}}^{[2]} = \underset{\mathbb{W}:\mathbb{W}^T\mathbb{W}=I}{\operatorname{argmin}} \left\| Y_{[\mathcal{S}(\hat{\Pi}^{[2]}),:]} - X_{[\mathcal{S}(\hat{\Pi}^{[2]}),:]}\mathbb{W} \right\|_F^2 = U_2 V_2^T$$

for singular value decomposition $X^T_{[\mathcal{S}(\hat{\Pi}^{[2]}),:]} Y_{[\mathcal{S}(\hat{\Pi}^{[2]}),:]} = U_2 D_2 V_2^T$.

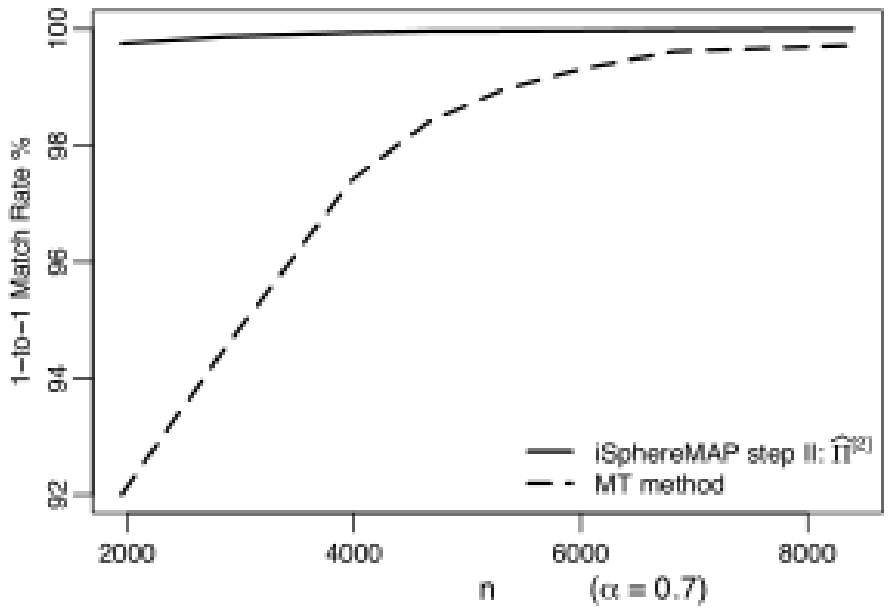
Numerical results for estimating $\mathbb{W}$

- Generated from vMF distribution with $p = 300$, $\kappa = 300$.
- Among n observations, there are n^α mismatch, half of which are one-to-one and the others are one-to-many.
- MSE $= \frac{1}{p}\|\mathbb{W} - \hat{\mathbb{W}}\|_F^2$ and averaged over 100 simulations.
- MT method: $\hat{\mathbb{W}}$ is OLS and $\hat{\pi}(i) = \text{argmax}_j \cos(Y_i, \hat{\mathbb{W}} X_j)$.
- $\hat{\mathbb{W}}^{[1]}$ is obtained from entire data and $\hat{\mathbb{W}}^{[2]}$ is obtained from refined data (observations estimated to be correct matches)



Numerical results for estimating Π

- The data are generated from the same vMF distribution
- Π is a permutation, i.e. here are no one-to-many mapping. (MT method does not allow one-to-many mapping)
- Compare the proportion of correctly matched rows for iSphereMAP and MT method.



Application to Electronic Health Records (EHR)

- ICD-10 code has larger and more complex set of codes compared to ICD-9 (one-to-many approximate mappings are frequent).
- Benchmark: General Equivalence Mappings (GEM) mapping (from National Bureau of Economic Research 2013)
- iSphereMAP correctly identified 49% of one-to-one and 54% of one-to-many mapping.
- MT method correctly identified 20% of the one-to-one mapping. (MT method cannot map one-to-many-mapping)

(c) GEM ICD-9-to-10 mapping for suicide and self-inflicted injuries (SSI)

(d) iSphereMAP estimated ICD-9-to-10 mapping for suicide and self-inflicted injuries (SSI)

Reference

- Shi, X., Li, X., & Cai, T. (2020). Spherical regression under mismatch corruption with application to automated knowledge translation. *Journal of the American Statistical Association*, 1-12.
- Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.