

## Topic 15: Sparse Orthogonal Factor Regression

by Sai Zhang

**Key points:** Sparsity and dimensionality reduction for Multivariate Linear Regression models.

**Disclaimer:** The note is built on Prof. *Jinchi Lv*'s lectures of the course at USC, DSO 607, High-Dimensional Statistics and Big Data Problems.

### 15.1 Motivation

Consider a Multivariate Linear Regression (MLR) model

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{C} + \mathbf{E}$$

$n \times q \quad n \times p \quad p \times q \quad n \times q$

How to apply regularization methods to this model? There are several approaches to consider

- **Shrinkage**: ridge regression to overcome multicollinearity
- **sparsity**: variable selection in multivariate setting
- **Reduced-rank**
  - **Dimension reduction** via reducing rank of  $\mathbf{C}$
  - $\min \|\mathbf{Y} - \mathbf{XC}\|_F^2$  s.t.  $\text{rank}(\mathbf{C}) \leq r$
- **Combinations**
- **Low-rank** plus **sparse decomposition**: robust PCA, latent variable graphical models, covariance estimation
- **Regularized matrix** or **tensor regression**

Or, we can introduce a very attractive sparsity structure to achieve simultaneous dimension reduction and variable selection. This structure should be characterized by

- Having a few **distinct** channels/pathways relating responses and predictors
- Each of such associations may involve only a **smaller subset**, but not all of the responses and predictors

that is

$$\begin{aligned} \mathbf{Y} &= \mathbf{XC} + \mathbf{E} \\ &= \mathbf{X} \cdot \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1q} \\ c_{21} & c_{22} & \cdots & c_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pq} \end{pmatrix} + \mathbf{E} \\ &= \mathbf{X} \cdot \begin{pmatrix} 0 & u_{12} & \cdots & u_{1r} \\ u_{21} & 0 & \cdots & c_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pr} \end{pmatrix} \cdot \begin{pmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 & \cdots & v_{q1} \\ v_{12} & v_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ v_{1r} & v_{2r} & \cdots & v_{qr} \end{pmatrix} + \mathbf{E} \end{aligned}$$

This way, we can have

- **Sparsity**: selection of both latent and original variables
- **Low-rank SVD**: different subsets of responses allowed to be associated with different subsets of predictors

Consider an example:

**Example 15.1.1: Dimension Reduction and Variable Selection via Sparse SVD**

Consider the case where  $p = 1000, q = 100$ , then  $C$ , as a  $p \times q$  matrix, contains 100000 coefficients. Meanwhile, for a rank-3 SVD model:

$$C = d_1 \mathbf{u}_1 \mathbf{v}_1' + d_2 \mathbf{u}_2 \mathbf{v}_2' + d_3 \mathbf{u}_3 \mathbf{v}_3'$$

where  $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$  are all  $p \times 1$ ,  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$  are all  $q \times 1$ ,  $d_1, d_2, d_3$  are all scalars. Hence, there are only  $3 \times (1000 + 100 + 1) = 3303$  parameters to estimate. If further assume sparsity, the dimension would be even lower.

Now let's develop a scalable procedure for this idea.

## 15.2 Sparse Orthogonal Factor Regression

Consider the singular value decomposition of  $C$

$$C = \mathbf{U} \mathbf{D} \mathbf{V}' = \sum_{k=1}^r d_k \mathbf{u}_k \mathbf{v}_k'$$

where  $\mathbf{U}$  and  $\mathbf{V}$  are both **orthonormal**:  $\mathbf{U}\mathbf{U}' = \mathbf{V}\mathbf{V}' = \mathbf{I}$ . Then we can achieve dimension reduction via **low-dimensional latent model**

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}} \mathbf{D} + \tilde{\mathbf{E}}$$

where

- $\tilde{\mathbf{Y}} = \mathbf{Y}\mathbf{V}$ :  $\mathbf{V}$  sparsity leads to response variable selection
- $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{U}$ :  $\mathbf{U}$  sparsity leads to predictor variable selection

How consider

$$(\hat{\mathbf{D}}, \hat{\mathbf{U}}, \hat{\mathbf{V}}) = \arg \min_{\mathbf{D}, \mathbf{U}, \mathbf{V}} \left\{ \frac{1}{2} \|\mathbf{Y} - \mathbf{X} \mathbf{U} \mathbf{D} \mathbf{V}'\|_F^2 + \lambda_d \|\mathbf{D}\|_1 + \lambda_a \rho_a(\mathbf{U} \mathbf{D}) + \lambda_b \rho_b(\mathbf{V} \mathbf{D}) \right\} \quad \text{s.t. } \mathbf{U}' \mathbf{U} = \mathbf{V}' \mathbf{V} = \mathbf{I}_m \quad (15.1)$$

where

- $\rho_a(\cdot), \rho_b(\cdot)$  are penalty functions with regularization parameters  $\lambda_d, \lambda_a, \lambda_b \geq 0$ . These sparsity penalizations on  $\mathbf{U} \mathbf{D}$  and  $\mathbf{V} \mathbf{D}$  can be thought as **importance weighting**
- $\|\cdot\|_F$  is the nuclear norm, defined as the **sum** of its singular values  $\|\mathbf{A}\|_F = \sum_i \sigma_i(\mathbf{A})$ . It encourages sparsity among singular values and achieve **rank reduction**
- The orthogonality on  $\mathbf{U}, \mathbf{V}$  allow a flexible form of sparsity-inducing penalties

If we further enrich this model by introducing an **adaptive weighting  $\mathbf{W}$  matrices**

$$(\hat{\mathbf{\Theta}}, \hat{\mathbf{\Omega}}) = \arg \min_{\mathbf{\Theta}, \mathbf{\Omega}} \left\{ \frac{1}{2} \|\mathbf{Y} - \mathbf{X} \mathbf{U} \mathbf{D} \mathbf{V}'\|_F^2 + \lambda_d \|\mathbf{W}_d \circ \mathbf{D}\|_1 + \lambda_a \rho_a(\mathbf{W}_a \circ \mathbf{A}) + \lambda_b \rho_b(\mathbf{W}_b \circ \mathbf{B}) \right\}$$

s.t.  $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}_m$ ,  $\mathbf{UD} = \mathbf{A}$ ,  $\mathbf{VD} = \mathbf{B}$ . But why? Singular values and singular vectors of **larger magnitude** should be **less penalized** to reduce bias and improve efficiency.

Two applications are

- Biclustering with sparse SVD

$$(\hat{\mathbf{D}}, \hat{\mathbf{U}}, \hat{\mathbf{V}}) = \arg \min_{\mathbf{D}, \mathbf{U}, \mathbf{V}} \left\{ \frac{1}{2} \|\mathbf{X} - \mathbf{UDV}'\|_F^2 + \lambda_d \|\mathbf{D}\|_1 + \lambda_a \rho_a(\mathbf{UD}) + \lambda_b \rho_b(\mathbf{VD}) \right\} \quad \text{s.t. } \mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}_m$$

- Sparse PCA (sparsity in loadings of principal components)

$$(\hat{\mathbf{A}}, \hat{\mathbf{V}}) = \arg \min_{\mathbf{A}, \mathbf{V}} \left\{ \frac{1}{2} \|\mathbf{X} - \mathbf{XAV}'\|_F^2 + \lambda_a \rho_a(\mathbf{A}) \right\} \quad \text{s.t. } \mathbf{V}'\mathbf{V} = \mathbf{I}_m$$

### 15.3 Nonasymptotic Properties of SOFAR

First, define the robust spark for the regularity conditions

**Definition 15.3.1: The robust spark  $\kappa_c$**

The robust spark  $\kappa_c$  of the  $n \times p$  design matrix  $\mathbf{X}$  is defined as the smallest possible positive integer such that there exists an  $n \times \kappa_c$  submatrix of  $\frac{1}{\sqrt{n}}\mathbf{X}$  having a **singular value less than** a given positive constant  $c$

The robust spark  $\kappa_c$  here can be at least of order  $O\left(\frac{n}{\log p}\right)$  with large probability for Gaussian design with dependency.