

SDR for matrix predictors

Chanwoo Lee, May 24, 2020

1 SDR for matrix

For a vector predictor $X \in \mathbb{R}^d$, sufficient dimension reduction assumes that

$$Y \perp\!\!\!\perp X | \mathbf{B}^T X, \quad (1)$$

where $\mathbf{B} \in \mathbb{R}^{d \times k}$. We define with $\mathcal{X} \bar{\times}_N \mathcal{Y}$ a sequence of contracted products between the $(K + N)$ -order tensor $\mathcal{X} \in \mathbb{R}^{J_1 \times \dots \times J_K \times I_1 \times \dots \times I_N}$ and the $(N + M)$ -order tensor $\mathcal{Y} \in \mathbb{R}^{I_1 \times \dots \times I_N \times H_1 \times \dots \times H_M}$. Entry-wise, it is defined as

$$(\mathcal{X} \bar{\times}_N \mathcal{Y})_{j_1, \dots, j_K, h_1, \dots, h_M} = \sum_{i_1=1}^{I_1} \dots \sum_{i_N=1}^{I_N} \mathcal{X}_{j_1, \dots, j_K, i_1, \dots, i_N} \mathcal{Y}_{i_1, \dots, i_N, h_1, \dots, h_M}.$$

For a matrix predictor $\mathbf{X} \in \mathbb{R}^{m \times n}$, sufficient dimension reduction assumes that

$$Y \perp\!\!\!\perp \mathbf{X} | \mathcal{B} \bar{\times}_2 \mathbf{X}, \quad (2)$$

where $\mathcal{B} \in \mathbb{R}^{k \times m \times n}$. If we define $\mathcal{B}_{i..} = \mathbf{B}_i$, we have

$$\mathcal{B} \bar{\times}_2 \mathbf{X} = (\langle \mathbf{B}_1, \mathbf{X} \rangle, \dots, \langle \mathbf{B}_k, \mathbf{X} \rangle)^T.$$

Remark 1. The predictor matrix \mathbf{X} is a vector where $n = 1$, (2) is reduced down to (1) ($\mathbf{B} \bar{\times}_1 X$). In addition, we can extend to tensor case with order d as

$$Y \perp\!\!\!\perp \mathcal{X} | \mathcal{B} \bar{\times}_d \mathcal{X},$$

where $\mathcal{X} \in \mathbb{R}^{I_1 \times \dots \times I_d}$ and $\mathcal{B} \in \mathbb{R}^{k \times I_1 \times \dots \times I_d}$.

Remark 2. If we do not assume low rank matrix structure on \mathbf{B}_i , (2) is equivalent to (1) with predictor \mathbf{X} replaced by $\text{Vec}(\mathbf{X})$.

Remark 3. My guess of defining the central subspace in matrix case as follows. First, define span of tensor \mathcal{B} as

$$\text{span}(\mathcal{B}) = \left\{ \mathbf{U} \mathbf{V}^T : \mathbf{U} = \sum_{i=1}^k \alpha_i \mathbf{U}_i, \quad \mathbf{V} = \sum_{i=1}^k \beta_i \mathbf{V}_i \quad \text{where } \alpha, \beta \in \mathbb{R}^k \text{ and } \mathcal{B}_{i..} = \mathbf{U}_i \mathbf{V}_i^T \right\}$$

From this span of the tensor, the central subspace in matrix case is defined as

$$S_{Y|\mathbf{X}} = \bigcap_{\{\mathcal{B}: Y \perp\!\!\!\perp X | \mathcal{B} \bar{\times}_2 \mathbf{X}\}} \text{span}(\mathcal{B}),$$

We extended weighted SVM to SMM to find the best hyperplane that separate $S_\pi = \{\mathbf{X} : \mathbb{P}(\mathbf{X}|y = 1) > \pi\}$ and $S_{-\pi} = \{\mathbf{X} : \mathbb{P}(\mathbf{X}|y = 1) < \pi\}$. The weighted SMM finds a matrix $\mathbf{B} \in \mathbb{R}^{m \times n}$ that optimizes the following problem.

$$\min_{\mathbf{B} \in \mathbb{R}^{m \times n}} \|\mathbf{B}\|^2 + \frac{\lambda}{N} \sum_{i=1}^N \omega_\pi(Y_i) (1 - Y_i f(\mathbf{X}_i; \mathbf{B}, \alpha))_+,$$

where $f(\mathbf{X}_i; \mathbf{B}, \alpha) = \alpha + \langle \mathbf{B}, \mathbf{X}_i \rangle$. We make distinction from SVM assuming low rank structure to $\mathbf{B} = \mathbf{U}\mathbf{V}^T$ where $\mathbf{U} \in \mathbb{R}^{m \times r}$, $\mathbf{V} \in \mathbb{R}^{n \times r}$.

By the similar way, we can extend the linear principal weighted vector machine to the matrix case with a pair of random variables $(\mathbf{X}, Y) \in \mathbb{R}^{m \times n} \times \{+1, -1\}$. We look for optimizer that minimizes

$$\Lambda_\pi(\mathbf{B}, \alpha) = \text{Vec}(\mathbf{B})^T \text{cov}(\text{Vec}(\mathbf{X})) \text{Vec}((\mathbf{B}) + \lambda \mathbb{E} [\omega_\pi(Y) (1 - Y f(\mathbf{X}; \mathbf{B}, \alpha))_+]) \quad (3)$$

Denote the observed data by $\{(\mathbf{X}_i, Y_i) : \mathbf{X}_i \in \mathbb{R}^{m \times n}, Y_i \in \{+1, -1\}, i = 1, \dots, N\}$. The sampled version of Λ_π in (3) is,

$$\hat{\Lambda}_{n,\pi} = \text{Vec}(\mathbf{B})^T \hat{\Sigma}_{\mathbf{N}} \text{Vec}((\mathbf{B}) + \frac{\lambda}{n} \sum_{i=1}^N \left[\omega_\pi(Y_i) (1 - Y_i \hat{f}_n(\mathbf{X}_i; \mathbf{B}, \alpha))_+ \right]), \quad (4)$$

where $\hat{f}_n(\mathbf{X}_i, \mathbf{B}, \alpha) = \alpha + \langle \mathbf{X}_i - \bar{\mathbf{X}}_n, \mathbf{B} \rangle$, $\bar{\mathbf{X}}_n$ is the sample mean, and $\Sigma_{\mathbf{n}}$ denotes the sample covariance matrix of $\{\text{Vec}(\mathbf{X}_i)\}_{i=1}^N$. With transformations $\text{Vec}(\mathbf{D}) = \hat{\Sigma}_{\mathbf{N}}^{\frac{1}{2}} \mathbf{B}$ and $\mathbf{Z}_i = \hat{\Sigma}_{\mathbf{N}}^{-\frac{1}{2}} (\mathbf{X}_i - \bar{\mathbf{X}}_n)$, (4) becomes

$$\hat{\Lambda}'_{n,\pi} = \|\mathbf{D}\|^2 + \frac{\lambda}{n} \sum_{i=1}^N \left[\omega_\pi(Y_i) (1 - Y_i \hat{f}_n(\mathbf{Z}_i; \mathbf{D}, \alpha))_+ \right] \quad (5)$$

Denote the optimizer of (5) as $\hat{\mathbf{D}}_{n,\pi}$, then the optimizer of (3) is $\hat{\mathbf{B}}_{n,\pi} = \hat{\Sigma}_{\mathbf{N}}^{-\frac{1}{2}} \hat{\mathbf{D}}_{n,\pi}$

Remark 4. If we assumes \mathbf{B} as full rank, then all the procedures are reduced down to the linear principal weighted vector machine with sample $\{\text{Vec}(\mathbf{X}_i)\}_{i=1}^N$

Remark 5. Since the transformation $\text{Vec}(\mathbf{D}) = \hat{\Sigma}_{\mathbf{N}}^{\frac{1}{2}} \mathbf{B}$ does not change the rank. we can assume the low rank structure as $\mathbf{D} = \mathbf{U}\mathbf{V}^T$ and solve the weighted SMM problem.

Given a grid $0 < \pi_1 < \dots < \pi_H < 1$, we obtained H -candidates $\{\hat{\mathbf{B}}_{n,\pi_h}\}_{h=1}^H$ of the central subspace. We can perform principal component analysis to get the k basis elements of $S_{Y|\mathbf{X}}$ with the following procedure.

1. Obtain column part matrices $\{\hat{\mathbf{U}}_h\}_{h=1}^H$ and row part matrices $\{\hat{\mathbf{V}}_h\}_{h=1}^H$ through SVD such that

$$\hat{\mathbf{B}}_{n,\pi_h} = \hat{\mathbf{U}}_h \hat{\Sigma}_h \hat{\mathbf{V}}_h^T \quad h = 1, \dots, H.$$

2. Calculate row-candidate matrix $\hat{\mathbf{M}}_n^r$ and column-candidate matrix $\hat{\mathbf{M}}_n^c$ as

$$\begin{aligned} \hat{\mathbf{M}}_n^r &= \sum_{h=1}^H \text{Vec}(\hat{\mathbf{V}}_h) \text{Vec}(\hat{\mathbf{V}}_h)^T \\ \hat{\mathbf{M}}_n^c &= \sum_{h=1}^H \text{Vec}(\hat{\mathbf{U}}_h) \text{Vec}(\hat{\mathbf{U}}_h)^T. \end{aligned}$$

3. The first k eigenmatrices (folded from eigenvectors) of $\hat{\mathbf{M}}_n^r$, denoted by $\{\tilde{\mathbf{V}}_1, \dots, \tilde{\mathbf{V}}_k\}$, estimates the row-part basis. By the same way, The first k eigenmatrices $\{\tilde{\mathbf{U}}_1, \dots, \tilde{\mathbf{U}}_k\}$ of $\hat{\mathbf{M}}_n^c$ estimates the column-part basis

4. Estimate the central subspace as

$$S_{Y|\mathbf{X}} = \left\{ \left(\sum_{i=1}^k \alpha_i \tilde{\mathbf{U}}_i \right) \left(\sum_{i=1}^k \beta_i \tilde{\mathbf{V}}_i \right)^T : \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^k \right\}.$$

Remark 6. These principal component procedures can be reduced down to the vector case if we standardize the estimated normal vectors as $\{\boldsymbol{\beta}_h / \|\boldsymbol{\beta}_h\|\}_{h=1}^H$.

2 Generating matrix valued training data for SDR

We can consider simple model that can show matrix valued SDR performance. First, generate matrix valued $\{\mathbf{X}_i\}_{i=1}^N \in \mathbb{R}^{m \times n}$ whose entries are from i.i.d. $N(0, 1)$. Next, we generate $\mathcal{B} \in \mathbb{R}^{2 \times m \times n}$ such that

$$\mathcal{B}_{1..} = \mathbf{u}_1 \mathbf{v}_1^T, \quad \mathcal{B}_{2..} = \mathbf{u}_2 \mathbf{v}_2^T,$$

where $\mathbf{u}_i \in \mathbb{R}^{m \times r}$ and $\mathbf{v}_i \in \mathbb{R}^{n \times r}$, $i = 1, 2$. Denote $\mathbf{Z}_{1i} = \langle \mathcal{B}_{1..}, \mathbf{X}_i \rangle$ and $\mathbf{Z}_{2i} = \langle \mathcal{B}_{2..}, \mathbf{X}_i \rangle$. We assign the label $Y_i \in \{+1, -1\}$ as

$$Y_i = \text{sign}(2\mathbf{Z}_{1i} + \mathbf{Z}_{2i} + 0.2\epsilon) \quad \text{where } \epsilon \sim N(0, 1).$$

In this way, we can generate the training data $\{(\mathbf{X}_i, Y_i)\}_{i=1}^N$ and check whether estimated the central subspace is close to true one.

If we set the rule of labeling Y_i as

$$Y_i = \text{sign}(\mathbf{Z}_{1i}^2 + \mathbf{Z}_{2i}^2 - 1)$$

We can check weather the kernel method works well with good visualization which we considered in the last meeting.