Modified SDR for matrix predictors

Chanwoo Lee, May 28, 2020

1 Definition of SDR

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

$$Y \perp\!\!\!\perp X | \boldsymbol{B}^T X,$$

where $\boldsymbol{B} \in \mathbb{R}^{d \times k}$. For a matrix predictor $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$, sufficient dimension reduction assumes that

$$Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X} \times_1 \boldsymbol{U} \times_2 \boldsymbol{V}, \tag{1}$$

where $\boldsymbol{U} \in \mathbb{R}^{d_1 \times k_1}, \boldsymbol{V} \in \mathbb{R}^{d_2 \times k_2}$. We can equivalently express (1) as

$$\begin{aligned} &Y \perp\!\!\!\perp \boldsymbol{X} | \left\{ \boldsymbol{u}_i^T \boldsymbol{X} \boldsymbol{v}_j \right\}_{i \in [k_1], j \in [k_2]}, \text{ or equivalently} \\ &Y \perp\!\!\!\perp \boldsymbol{X} | \left\{ \left\langle \boldsymbol{u}_i \boldsymbol{v}_j^T, \boldsymbol{X} \right\rangle \right\}_{i \in [k_1], j \in [k_2]} \end{aligned} \tag{2}$$

where $\boldsymbol{u}_i$ is i-th column of $\boldsymbol{U}$ and $\boldsymbol{v}_j$ is j-th column of $\boldsymbol{V}$. The central subspace in matrix case is defined as

$$S_{Y|\boldsymbol{X}} = \bigcap_{\{(\boldsymbol{U},\boldsymbol{V}): Y \perp\!\!\!\perp X | \boldsymbol{X} \times_1 \boldsymbol{U} \times_2 \boldsymbol{V}\}} \text{span}(\boldsymbol{U}) \times \text{span}(\boldsymbol{V}),$$

Remark 1. The assumption (1) is different from the previous note (under structural conditions, they might be the same but different in general). The main difference is the dimension reduction part (1) has $k_1 k_2$ linear equations but k in previous notes. There are some reasons I choose the current one. For previous equation, we have to assume all $\boldsymbol{B}_i$ has the same rank r structure, which does not make sense well to me. In addition, we have to choose the rank k_1 and k_2 later from previous definition, which does not have clear connection with the number of equations k.

2 Estimation of SDR

We use rank 1 SMM method motivated by formula in (2). The weighted SMM finds a matrix $\mu \boldsymbol{v}^T$ that optimizes the following problem.

$$\min_{\boldsymbol{u} \in \mathbb{R}^{d_1}, \boldsymbol{v} \in \mathbb{R}^{d_2}} \|\boldsymbol{u}\boldsymbol{v}^T\|^2 + \frac{\lambda}{N} \sum_{i=1}^{N} \omega_\pi(Y_i) \left(1 - Y_i f(\boldsymbol{X}_i; \boldsymbol{u}, \boldsymbol{v}, \alpha)\right)_+,$$

where $f(\boldsymbol{X}_i; \boldsymbol{u}, \boldsymbol{v}, \alpha) = \alpha + \langle \boldsymbol{u}\boldsymbol{v}^T, \boldsymbol{X}_i \rangle$.

By the similar way, we can extend the linear principal weighted vector machine to the matrix case with a pair of random variables $(\boldsymbol{X}, Y) \in \mathbb{R}^{d_1 \times d_2} \times \{+1, -1\}$. We look for optimizer that minimizes

$$\Lambda_\pi(\boldsymbol{u}\boldsymbol{v}^T, \alpha) = \text{Vec}\left(\boldsymbol{u}\boldsymbol{v}^T\right)^T \text{cov}\left(\text{Vec}(\boldsymbol{X})\right) \text{Vec}\left((\boldsymbol{u}\boldsymbol{v}^T\right) + \lambda \mathbb{E}\left[\omega_\pi(Y)\left(1 - Y f(\boldsymbol{X}; \boldsymbol{u}, \boldsymbol{v}, \alpha)\right)_+\right] \tag{3}$$

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

$$\hat{\Lambda}_{N,\pi} = \mathrm{Vec}\,(\boldsymbol{u}\boldsymbol{v}^T)^T \hat{\boldsymbol{\Sigma}}_{\mathbf{N}}\, \mathrm{Vec}\,((\boldsymbol{u}\boldsymbol{v}^T) + \frac{\lambda}{N}\sum_{i=1}^{N}\left[\omega_\pi(Y_i)\left(1 - Y_i\hat{f}_N(\boldsymbol{X}_i; \boldsymbol{u}, \boldsymbol{v}, \alpha)\right)_+\right], \tag{4}$$

where $\hat{f}_N(\boldsymbol{X}_i, \boldsymbol{u}, \boldsymbol{v}, \alpha) = \alpha + \langle \boldsymbol{X}_i - \bar{\boldsymbol{X}}_N, \boldsymbol{u}\boldsymbol{v}^T\rangle$, $\bar{\boldsymbol{X}}_N$ is the sample mean, and $\boldsymbol{\Sigma}_{\mathbf{N}}$ denotes the sample covariance matrix of $\{\mathrm{Vec}(\boldsymbol{X}_i)\}_{i=1}^N$. With transformations $\mathrm{Vec}(\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}^T) = \hat{\boldsymbol{\Sigma}}_{\mathbf{N}}^{\frac{1}{2}}\,\mathrm{Vec}(\boldsymbol{u}\boldsymbol{v}^T)$ and $\boldsymbol{Z}_i = \hat{\boldsymbol{\Sigma}}_{\mathbf{N}}^{-\frac{1}{2}}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_n)$, (4) becomes

$$\tilde{\Lambda}_{N,\pi} = \|\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}^T\|^2 + \frac{\lambda}{N}\sum_{i=1}^{N}\left[\omega_\pi(Y_i)\left(1 - Y_i\hat{f}_N(\boldsymbol{Z}_i; \tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}, \alpha)\right)_+\right] \tag{5}$$

Denote the optimizer of (5) as $(\tilde{\boldsymbol{u}}_{n,\pi}, \tilde{\boldsymbol{v}}_{n,\pi})$, then we can obtain the optimizer of (3) from $\mathrm{Vec}(\hat{\boldsymbol{u}}_{n,\pi}\hat{\boldsymbol{v}}_{n,\pi}^T) = \hat{\boldsymbol{\Sigma}}_{\mathbf{N}}^{-\frac{1}{2}}\,\mathrm{Vec}(\tilde{\boldsymbol{u}}_{n,\pi}\tilde{\boldsymbol{v}}_{n,\pi}^T)$

Given a grid $0 < \pi_1 < \cdots < \pi_H < 1$, we obtained H-candidates $\{\hat{\boldsymbol{u}}_{n,\pi_h}\hat{\boldsymbol{v}}_{n,\pi_h}^T\}_{h=1}^H$ for the central subspace.

We can perform principal component analysis to get the k_1 column basis elements and k_2 row basis elements in $S_{Y|\boldsymbol{X}}$ with the following procedure.

1. Obtain the candidate tensor $\hat{\mathcal{M}} \in \mathbb{R}^{H\times d_1\times d_2}$ such that $\hat{\mathcal{M}}_{h\cdot\cdot} = \hat{\boldsymbol{u}}_{n,\pi_h}\hat{\boldsymbol{v}}_{n,\pi_h}^T$
2. From N-th order SVD (Thm 2 in Lathauwer et al. [2000]),
$$\hat{\mathcal{M}} = \hat{\mathcal{C}} \times_1 \hat{\boldsymbol{U}}_1 \times_2 \hat{\boldsymbol{U}}_2 \times_3 \hat{\boldsymbol{U}}_3,$$
where $\mathcal{C} \in \mathbb{R}^{H\times d_1\times d_2}$, $\boldsymbol{U}_1 \in \mathbb{R}^{H\times H}$, $\boldsymbol{U}_2 \in \mathbb{R}^{d_2\times d_2}$, and $\boldsymbol{U}_3 \in \mathbb{R}^{d_3\times d_3}$.
3. Estimate the central subspace as
$$\hat{S}_{Y|\boldsymbol{X}} = \{[\hat{\boldsymbol{U}}_2]_i\}_{i=1}^{k_1} \times \{[\hat{\boldsymbol{U}}_3]_i\}_{i=1}^{k_2}.$$

We can reduce the dimension of $\boldsymbol{X}$ as $\{\boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v} \;\text{ where }\; (\boldsymbol{u}, \boldsymbol{v}) \in \hat{S}_{Y|\boldsymbol{X}}\}$

Remark 2. We consider criterion to estimate k_1 and k_2

$$G_{n,1}(m; \rho, \hat{\mathcal{M}}_n) = \sum_{j=1}^{m}\|\mathcal{C}_{i_2=j}\| - \rho\frac{m\log n}{\sqrt{n}}\|\mathcal{C}_{i_2=1}\|,$$
$$G_{n,2}(m; \rho, \hat{\mathcal{M}}_n) = \sum_{j=1}^{m}\|\mathcal{C}_{i_3=j}\| - \rho\frac{m\log n}{\sqrt{n}}\|\mathcal{C}_{i_3=1}\|,$$

where ρ is a tuning parameter. We have the consistent estimators

$$\hat{k}_1 = \arg\min_m G_{n,1}(m; \rho, \hat{\mathcal{M}}_n),$$
$$\hat{k}_2 = \arg\min_m G_{n,2}(m; \rho, \hat{\mathcal{M}}_n).$$

Remark 3. With some structural assumption on $\boldsymbol{X}$, we can simplify (4).

References

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Analysis Applications*, 21:1253–1278, 2000.