# SMMK conditional proability and SDR

Chanwoo Lee, May 21, 2020

## 1 SMMK conditional probability

Main changes of SMMK algorithm are as follow.

1. Weighted hinge loss based algorithm is available for conditional probability estimation.
2. Symmetric adjustment procedure is added in nonlinear kernel cases. But nothing is changed in linear case.

I check that the new SMMK algorithm gives us reasonable conditional probability output from simple data simulation. Training data $\{(X_i, y_i)\}_{i=1}^{40}$ is generated with the following rule. $X_i \in \mathbb{R}^2$ is from i.i.d. multivariate normal distribution $N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 2\end{pmatrix}\right)$ and $y_i = \text{sign}\left(\phi(X_i) - \frac{1}{40}\sum_{i=1}^{40}\phi(X_i)\right)$ where $\phi(\cdot)$ is density function of $X_i$. Then, true boundary is an ellipse in the plane. First, I fit the each kernel types ("linear","polynomial","exponential") to training data. The following figure shows the boundary of classification.

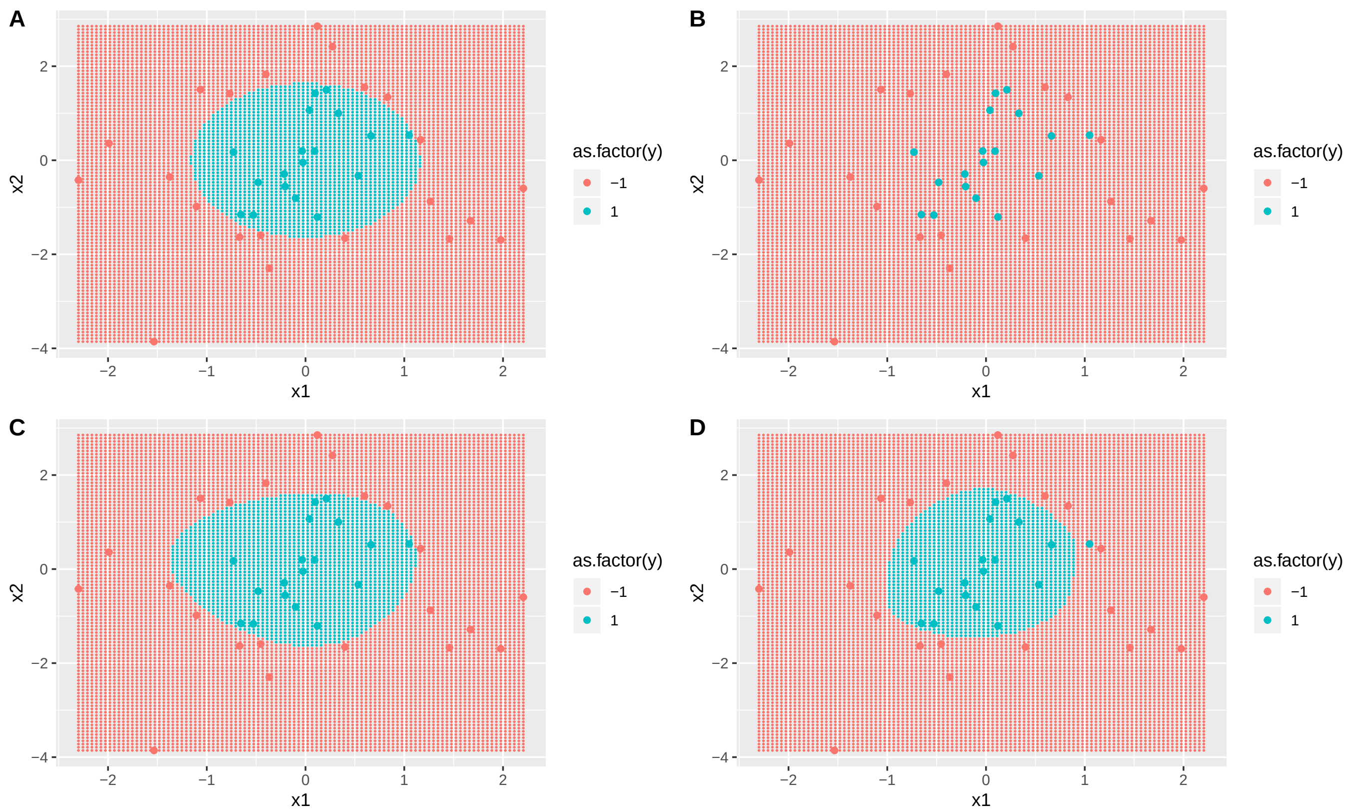


Figure 1: Subfigure A is true ellipsoid boundary. B is linear case boundary which assigns labels all 0. C and D show the boundary of polynomial and exponential kernel respectively.

Since there is no meaningful classification in linear case, I estimate conditional probability only in the case of polynomial and exponential kernel. Th2 shows the result of conditional probability estimation which looks reasonable.

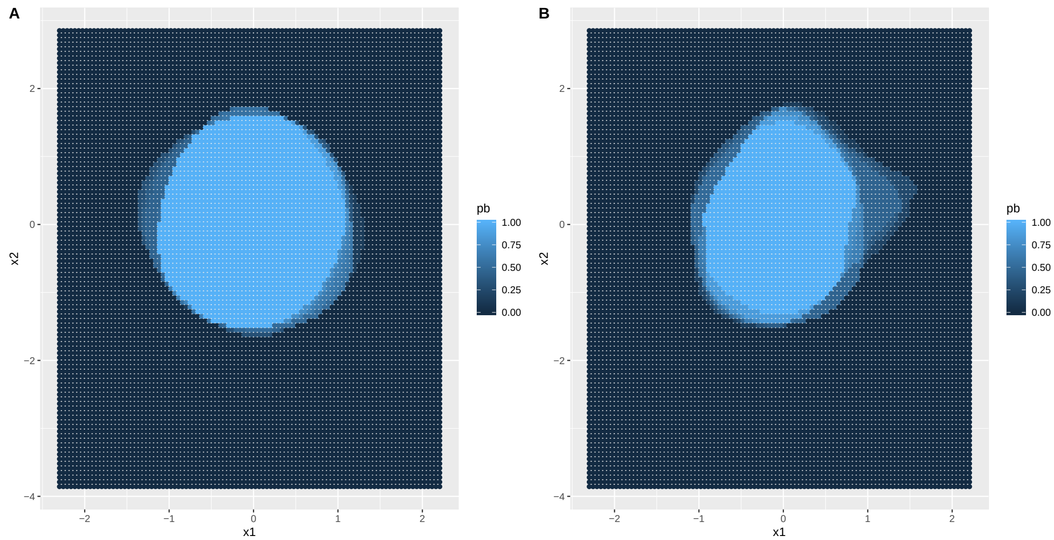


Figure 2: Figure A and B show the estimated probability with polynomial and exponential kernel respectively.

# 2 SDR summary and questions

## 2.1 SDR linear case

Sufficient dimension reduction here assumes that

$$Y \perp\!\!\!\perp X | B^T X.$$

We define the central subspace $S_{Y|X}$ as

$$S_{Y|X} = \bigcap_{\{B: Y \perp\!\!\!\perp X | B^T X\}} \text{span}(B).$$

Our goal is to estimate basis of $S_{Y|X}$ using $S_{Y|X} = S_{P(X|Y=1)|X}$.

### 2.1.1 Population level

For a pair of random variables $(X, Y) \in \mathbb{R}^p \times \{-1, +1\}$, the linear principal weighted support vector machine minimizes

$$\Lambda_\pi(\boldsymbol{\beta}, \alpha) = \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta} + \lambda \mathbb{E}\left[w_\pi(Y)(1 - Y f(X; \theta))_+\right], \tag{1}$$

where $\Sigma = \text{cov}(X)$, and $f(X; \boldsymbol{\beta}, \alpha) = \alpha + \langle \boldsymbol{\beta}, X - \mathbb{E}(X) \rangle$. Note (1) is reduced to SVM with weghted hinge loss with the population version when $\mathbb{E}(X) = 0$ and $\text{cov}(X) = I_p$. Fisher consistency of the weighted support vector machine ensures that a hyperplane $\{X : f(X; \boldsymbol{\beta}_{0,\pi}, \alpha_{0,\pi}) = 0\}$ optimally separates $S_\pi^+ = \{X : p(X) \geq \pi\}$ and $S_\pi^- = \{X : p(X) < \pi\}$ where $(\boldsymbol{\beta}_{0,\pi}, \alpha_{0,\pi}) = \arg\min \Lambda_\pi(\boldsymbol{\beta}, \alpha)$.

**Theorem 2.1.** *Assume that $\mathbb{E}(X|B^TX)$ is a linear function of $B^TX$. Then for any given weighted $\pi \in (0,1), \boldsymbol{\beta}_{0,\pi} \in S_{Y|X}$.*

The assumption in Theorem 2.1 is known as the linearity condition. It implies that $\mathbb{E}(\beta^TX|B^TX) = \beta^TP_B(\Sigma)X$ where $P_B(\Sigma) = B(B^T\Sigma B)^{-1}B^T\Sigma$. The condition holds when $X$ is elliptically symmetric and approximately holds when $p$ is large. We can assume that $\text{span}(\beta_{0.1}, \ldots, \beta_{0,H}) = S_{Y|X}$ whenever $H$ is large enough.

### 2.1.2 Finite sample estimation

For the finite sample case, $\Lambda_\pi$ in (1) changes to,

$$\hat{\Lambda}_{n,\pi}(\boldsymbol{\beta}, \alpha) = \boldsymbol{\beta}^T\hat{\Sigma}_n\boldsymbol{\beta} + \frac{\lambda}{n}\sum_{i=1}^{n} w_\pi(y_i)(1 - y_i\hat{f}_n(X_i; \boldsymbol{\beta}, \alpha))_+,$$

where $\hat{f}_n(X_i; \boldsymbol{\beta}, \alpha)) = \alpha + \langle \boldsymbol{\beta}, X_i - \bar{X}_n \rangle$, $\bar{X}_n$ is the sample mean, and $\hat{\Sigma}_n$ denotes the sample covariance matrix. Let $(\hat{\boldsymbol{\beta}}_{n,\pi}, \hat{\alpha}_{n,\pi}) = \arg\min_{\boldsymbol{\beta},\alpha} \Lambda_\pi(\boldsymbol{\beta}, \alpha)$. The candidate matrix of the linear principal weighted support vector machine is

$$\hat{M}_n = \sum_{i=1}^{H} \hat{\boldsymbol{\beta}}_{n,h}\hat{\boldsymbol{\beta}}_{n,h}^T.$$

The first $k$ eigenvectors of $\hat{M}_n$, denoted by $\hat{V}_n = (\hat{v}_1, \ldots, \hat{v}_k)$, estimate a basis of $S_{Y|X}$. To determine $k$, we consider

$$G_n(m; \rho, \hat{M}_n) = \sum_{j=1}^{m} \ell_j - \rho\frac{m\log n}{\sqrt{n}}\ell_1,$$

where $\ell_j$ is the $j$-th leading eigenvalues of $\hat{M}_n$ and $\rho$ is a tunning parameter. It is known that $\hat{k} = \arg\max_m G_n(m; \rho, \hat{M}_n)$ is a consistent estimator of $k$.

## 2.2 SDR nonlinear case

Sufficient dimension reduction here assumes that

$$Y \perp\!\!\!\perp X|\boldsymbol{\phi}(X), \tag{2}$$

where $\boldsymbol{\phi} : \mathbb{R}^p \to \mathbb{R}^k$ is an unknown vector valued function of $X$. We define unbiasedness notion in nonlinear case for the later use.

**Definition 1.** A function $\psi \in \mathcal{H}$ is unbiased for nonlinear sufficient dimension reduction (2) if it has a version that is measurable $\sigma\{\boldsymbol{\phi}(X)\}$

### 2.2.1 Population level

The principal weighted support vector machine minimizes

$$\Lambda_\pi(\psi, \alpha) = \text{var}(\psi(X)) + \lambda\mathbb{E}\left[\omega_\pi(Y)(1 - Yf(X; \psi, \alpha))_+\right], \tag{3}$$

where $f(X; \psi, \alpha) = \alpha + (\psi(X) - \mathbb{E}(\psi(X))$. We can equivalently express (3) as

$$\Lambda_\pi(\psi, \alpha) = \langle \psi, \Sigma\psi \rangle_{\mathcal{H}} + \lambda\mathbb{E}\left[\omega_\pi(Y)(1 - Yf(X; \psi, \alpha))_+\right], \tag{4}$$

where $\Sigma : \mathcal{H} \to \mathcal{H}$ is an operator such that $\langle f_1, \Sigma f_2 \rangle = \text{cov}[f_1(X), f_2(X)]$

**Theorem 2.2.** *Suppose the mapping*

$$\mathcal{H} \to L_2(P_X), \quad f \mapsto f$$

*is continuous and :*

1. *$\mathcal{H}$ is a dense subset of $L_2(P_X)$,*
2. *$Y \perp\!\!\!\perp X | \boldsymbol{\phi}(X)$.*

*If $(\psi^*, \alpha^*)$ minimizes (4) among all $(\psi, \alpha) \in \mathcal{H} \times \mathbb{R}$, then $\psi^*(X)$ is unbiased.*

### 2.2.2 Finite sample estimation

We define $K_n(\cdot) = (K(\cdot, X_1), \cdots, K(\cdot, X_n))^T$ from a given kernel $K$ and feature predictors $\{X_i\}_{i=1}^n$. Considering Hilbert space $\mathcal{H} = \{\boldsymbol{\beta}^T K_n(\cdot) = \sum_{i=1}^n \beta_i K(\cdot, X_i) : \boldsymbol{\beta} \in \mathbb{R}^n\}$ is too rich, so that the solution often overfits the data. So we consider new Hilbert space. Define basis of new $\mathcal{H}$ as

$$\begin{aligned} &\psi_j(X) = \tilde{K}_n(X)^T \boldsymbol{\omega}_j / \lambda_j \qquad (5) \\ \text{where } &\tilde{K}_n(X) = K_n(X) - \frac{1}{n}\sum_{i=1}^n K_n(X_i) \\ &\boldsymbol{\omega}_j, \lambda_j \text{ are j-th components s.t.} (\boldsymbol{I}_n - \boldsymbol{J}_n/n) K_n (\boldsymbol{I}_n - \boldsymbol{J}_n/n) \boldsymbol{\omega}_j = \lambda_j \boldsymbol{\omega}_j. \end{aligned}$$

Finally, we have new Hilbert space as $\mathcal{H} = \{\sum_{i=1}^k \beta_i \psi_i(\cdot) : \boldsymbol{\beta} \in \mathbb{R}^k\}$. Consider finite sample case of (4).

$$\Lambda_\pi(\boldsymbol{\beta}, \alpha) = \boldsymbol{\beta}^T \Psi^T \Psi \boldsymbol{\beta} + \frac{\lambda}{n}\sum_{i=1}^n \omega_\pi(y_i)\left(1 - y_i(\Psi_i^T \boldsymbol{\beta} + \alpha)\right)_+, \qquad (6)$$

$$\text{where } \Psi = \begin{pmatrix} \psi_1(X_1) & \cdots & \psi_k(X_1) \\ \vdots & & \vdots \\ \psi_1(X_n) & \cdots & \psi_k(X_n) \end{pmatrix} \text{ and } \Psi_i^T = (\psi_1(X_i), \ldots, \psi_k(X_i)).$$

Let $(\hat{\boldsymbol{\beta}}_{n,\pi}, \hat{\alpha}_{n,\pi}) = \arg\min_{\boldsymbol{\beta},\alpha} \Lambda_\pi(\boldsymbol{\beta}, \alpha)$, the minimizer of (6). It is shown that $\hat{\boldsymbol{\beta}}_{n,\pi} = \lambda \sum_{i=1}^n \hat{\gamma}_{i,\pi} y_i (\Psi^T\Psi)^{-1}\Psi/2$ where $\hat{\boldsymbol{\gamma}}_\pi = (\hat{\gamma}_{1,\pi}, \ldots, \hat{\gamma}_{n,\pi})^T$ solves

$$\max_{\boldsymbol{\gamma}} \sum_{i=1}^n \gamma_i - \frac{1}{4}\sum_{i=1}^n\sum_{j=1}^n \gamma_i\gamma_j y_i y_j [P_\Psi]_{i,j}$$

$$\text{subject to } 0 \le \gamma_i \le \lambda\omega_\pi(y_i), \quad \sum_{i=1}^n \gamma_i y_i = 0,$$

with $[P_\Psi]_{i,j}$ the $(i,j)$th element of $P_\Psi = \Psi(\Psi^T\Psi)^{-1}\Psi^T$. By the similar way in linear SDR case, The candidate matrix of the non linear principal weighted support vector machine is

$$\hat{M}_n = \sum_{i=1}^H \hat{\boldsymbol{\beta}}_{n,h}\hat{\boldsymbol{\beta}}_{n,h}^T.$$

Denote first $k$ eigenvectors of $\hat{M}_n$, by $\hat{V}_n = (\hat{v}_1, \ldots, \hat{v}_k)$. Then, the $s$th sufficient predictor evaluated at $X$ is $\sum_{i=1}^n v_{si}\psi_i(X)$.

### 2.3 Question

There are some questions I had while summarizing the papers. I will think about those questions until tomorrow meeting.

1. In linear case, we define the central subspace as

$$S_{Y|X} = \bigcap_{\{B: Y \perp\!\!\!\perp X | B^T X\}} \text{span}(B).$$

   If we find the basis $\{v_1, \ldots, v_k\}$ such that $\text{span}(\{v_1, \ldots, v_k\}) = S_{Y|X}$, is it true that $Y \perp\!\!\!\perp X | (v_1, \ldots, v_k)^T X$?

2. In linear case, obtained $\hat{\boldsymbol{\beta}}_{n,\pi}$ is normal vector of hyperplane that optimally separate $S_1 = \{X_i : P(X_i | y_i = 1) > \pi\}$ and $S_2 = \{X_i : P(X_i | y_i = 1) < \pi\}$. Intuitively, the normals of these hyperplanes are roughly aligned with the directions that forms the central subspace. We use the principal components of these $\hat{\boldsymbol{\beta}}_{n,\pi}$s to estimate the central subspace. However, I cannot get the intuition that how eigen-vectors of $\hat{M}_n = \sum_{i=1}^{H} \hat{\boldsymbol{\beta}}_{n,h} \hat{\boldsymbol{\beta}}_{n,h}^T$ give us estimation of the central subspace.

3. I understand overall procedures of nonlinear SDR. However, I actually, do not understand how we define new Hilbert space as in Equation (5) and under what geometric intuition.