# Nonparametric learning with matrix-valued predictors in high dimensions

Chanwoo Lee, Miaoyan Wang

Department of Statistics, University of Wisconsin - Madison

## Problems & Existing methods

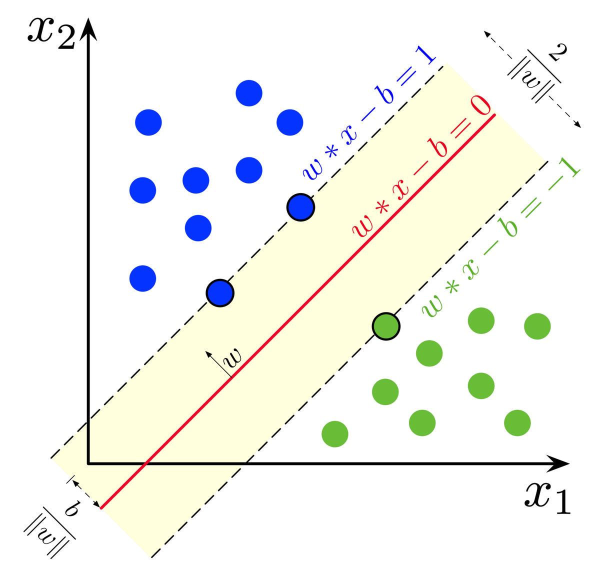


Fig. 1: Large margin classification with vector-valued predictors

**Problems** : Let $\{(\boldsymbol{X}_i, y_i) \in \mathbb{R}^{d_1\times d_2} \times \{-1,1\} : i = 1,\ldots,n\}$ denote an i.i.d. sample from unknown distribution $\mathcal{X}\times\mathcal{Y}$.

- Classification: Find a decision function $f : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}$ that has small error $\mathbb{P}_{\boldsymbol{X},y}(y \neq \text{sign}(f(\boldsymbol{X})))$.
- Regression: Estimate the regression function $\mathbb{E}(y|\boldsymbol{X})$. In binary label case, estimating regression is equivalent to estimating label probability $\mathbb{P}(y = 1|\boldsymbol{X})$.

**Existing methods** :

- Classification: Decision tree, Nearest neighbor, Neural network, and Support vector machine. However, most of methods have focused on vector valued features.
- Regression: Logistic regression and Linear discriminant analysis. However, it is often difficult to justify the assumptions made when features are matrices because of high-dimensionality.

**Goal** : We propose nonparametric learning approach with matrix-valued predictors, which is robust and preserves structural information of data matrices.

## Methods: 1. Classification

Nonparametric approach: 1. Classification ⇒ 2. Regression.

1. **Classification**:
   - We develop a large-margin classifiers for matrix predictors in high dimensions.

$$\hat f = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n L(y_i f(\boldsymbol{X}_i)) + \lambda J(f), \tag{1}$$

where $\mathcal{F}$ is considered decision function class, $J(f)$ is a regularization term for model complexity, and $L(z)$ is a margin loss.

- When we consider linear classifiers, we can set $\mathcal{F} = \{f : f(\cdot) = \langle \boldsymbol{B}, \cdot\rangle$ where $\text{rank}(\boldsymbol{B}) \le r\}$, $J(f) = \|\boldsymbol{B}\|_F^2$, and, $L(x) = (1-x)_+$.
- We consider linear classifiers in this poster but we can extend to nonlinear classifiers with a new definition of matrix feature mapping.

## Methods: 2. Regression

2. **Regression:**
   - We propose weighted loss function from (1),

$$\hat f_\pi = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\omega_\pi(y_i)L(y_i f(\boldsymbol{X}_i)) + \lambda J(f), \tag{2}$$

where $\omega_\pi(y) = 1-\pi$ if $y = 1$ and $\pi$ if $y = -1$.

- Consider two steps of approximations to the target probability function $p(\boldsymbol{X}) \stackrel{\text{def}}{=} \mathbb{P}(Y = 1|\boldsymbol{X})$:

$$p(\boldsymbol{X}) \approx \frac{1}{H}\sum_{h\in[H]} \mathbb{1}\left\{\boldsymbol{X} : p(\boldsymbol{X}) \le \frac{h}{H}\right\} \text{ (Discretization)}$$
$$\approx \frac{1}{H}\sum_{h\in[H]} \mathbb{1}\left\{\boldsymbol{X} : \text{sign}\left[\hat f_{\frac{h}{H}}(\boldsymbol{X})\right] = -1\right\},$$

where $H \in \mathbb{N}_+ \to \infty$ is the smoothing parameter.

- Last approximation used

$$\mathbb{1}\left\{\boldsymbol{X} : \underbrace{\text{sign}\left[\hat f_\pi(\boldsymbol{X})\right] = -1}_{\text{decision region from classification}}\right\} \xrightarrow{\text{in } p} \mathbb{1}\left\{\boldsymbol{X} : \underbrace{\mathbb{P}(Y = 1|\boldsymbol{X}) \le \pi}_{\text{target sublevel set}}\right\},$$

, which is verified in [1].

## Algorithms

- We focus on linear decision function case here

$$f(\boldsymbol{X}) = \langle \boldsymbol{B}, \boldsymbol{X}\rangle + b, \text{ where } \boldsymbol{B} = \boldsymbol{U}\boldsymbol{V}^T, \boldsymbol{U} \in \mathbb{R}^{d_1\times r}, \boldsymbol{V} \in \mathbb{R}^{d_2\times r}.$$

but we can extend to non linear case with kernel method.

- Optimization problem (2) can be solved by a little modification from (1).
- We take alternating optimization approach to solve non-convex problem (1).

**Algorithm 1: Classification algorithm**

**Input:** $(\boldsymbol{X}_1, y_1), \cdots, (\boldsymbol{X}_n, y_m)$,and prespecified rank $r$

**Initizlize:** $(\boldsymbol{U}^{(0)}, \boldsymbol{V}^{(0)}) \in \mathbb{R}^{d_1\times r} \times \mathbb{R}^{d_2\times r}$

**Do until converges**

**Update $\boldsymbol{U}$ fixing $\boldsymbol{V}$** :

Solve $\max_{\boldsymbol{\alpha}} -\sum_{i=1}^n \alpha_i + \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j \langle \boldsymbol{X}_i, \boldsymbol{X}_j \boldsymbol{V}(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{V}^T\rangle$

$\boldsymbol{U} = \sum_{i=1}^n \alpha_i y_i \boldsymbol{X}_i \boldsymbol{V}(\boldsymbol{V}^T\boldsymbol{V})^{-1}$.

**Update $\boldsymbol{V}$ fixing $\boldsymbol{U}$** :

Solve $\max_{\boldsymbol{\alpha}} -\sum_{i=1}^n \alpha_i + \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j \langle \boldsymbol{X}_i, \boldsymbol{U}(\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{X}_j\rangle$.

$\boldsymbol{V}^T = \sum_{i=1}^n \alpha_i y_i (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{X}_i$.

**Update $\boldsymbol{b}$ fixing $\boldsymbol{U}, \boldsymbol{V}$.**

**Output:** $\hat f(\boldsymbol{X}) = \langle \boldsymbol{U}\boldsymbol{V}^T, \boldsymbol{X}\rangle + b$

## Theoretical results: 1. Generalized bound of classification error

**Theorem 1.** *Let $\mathcal{F} = \{f : \boldsymbol{X} \mapsto \langle \boldsymbol{B}, \boldsymbol{X}\rangle |$ rank$(\boldsymbol{B}) \le r, \|\boldsymbol{B}\|_F \le C\}$. Assume that $\{\boldsymbol{X}_i\}_{i=1}^n$ be set of i.i.d. Gaussian distribution with bounded variation. Then with high probability,*

$$\mathbb{P}[Y \neq \text{sign}(f^*(\boldsymbol{X}))] - \mathbb{P}[Y \neq \text{sign}(\hat f(\boldsymbol{X}))] \le \frac{4C\sqrt{r(d_1+d_2)}}{\sqrt{n}},$$

*where $f^*$ is the best predictor in $\mathcal{F}$.*

## Theoretical results: 2. Consistency of regression estimation

**Theorem 2.** *Denote $\bar f_\pi =$ sign$(f_\pi)$ as Bayes rule with $f_\pi = \mathbb{P}(y = 1|\boldsymbol{X}) - \pi$. Let $e_V(f, \bar f_\pi) = \mathbb{E}\left[\omega_\pi(y)L(y(f(\boldsymbol{X})) - \omega_\pi(y)L(y(\bar f_\pi(\boldsymbol{X}))\right]$ be evaluation of distance between classifiers with respect to weighted loss. Assume that*

**A.1.** *For some positive sequence such that $s_n \to 0$ as $n \to \infty$, there exists $f_\pi^* \in \mathcal{F}$ such that $e_V(f_\pi^*, \bar f_\pi) \le s_n$. (The function class is enough to have elements close to Bayes rule.)*

**A.2.** *There exist constant $0 \le \alpha < \infty$, $a_1 > 0$ such that, for any sufficiently small $\delta > 0$.*

$$\sup_{\{f\in\mathcal{F}: e_V(f,\bar f_\pi)\le\delta\}} \|\text{sign}(f) - \text{sign}(\bar f_\pi)\|_1 \le a_1\delta^\alpha.$$

*(Any functions that have close distance to Bayes rule in the sense of $e_V$ are close to Bayes rule in L1 sense.)*

**A.3.** *Considered feature space is uniformly bounded such that there exists $0 < G < \infty$ satisfying $\sqrt{\mathbb{E}\|\boldsymbol{X}\|_F^2} \le G$*

*Then, for the estimator $\hat p$ obtained from our algorithm with function class $\mathcal{F}$,*

$$\mathbb{E}\|\hat p - p\|_1 = \mathcal{O}\left(\frac{1}{H} + a_1(H+1)\left(\frac{\log(n/r(d_1+d_2))}{(n/r(d_1+d_2))}\right)^{2/(2-\alpha\wedge 1)}\right)$$
$$= \mathcal{O}\left(\left(\frac{\log(n/r(d_1+d_2))}{(n/r(d_1+d_2))}\right)^{1/(2-\alpha\wedge 1)}\right),$$

*with a choice of $H = \mathcal{O}\left(\left(\frac{\log(n/r(d_1+d_2))}{(n/r(d_1+d_2))}\right)^{1/(2-\alpha\wedge 1)}\right)$.*

## References

[1] Junhui Wang, Xiaotong Shen, and Yufeng Liu. "Probability estimation for large-margin classifiers". In: *Biometrika* 95.1 (2008), pp. 149–167.