Sparse Sufficient Dimension Reduction via Penalized Principal Machines

true

2025-06-02

1 Introduction

Sufficient dimension reduction (SDR) aims to reduce the dimensionality of predictors $\mathbf{X}$ while preserving their relationship with the response Y. SDR methods estimate a projection matrix $\mathbf{B}$ spanning the **central subspace** $\mathcal{S}_{Y|\mathbf{X}}$ defined by

$$Y \perp \mathbf{X} \mid \mathbf{B}^\top \mathbf{X}.$$

In high-dimensional settings, **sparse SDR** improves interpretability and accuracy by encouraging solutions where many entries of $\mathbf{B}$ are zero. This is typically achieved by adding a sparsity-inducing penalty to the SDR optimization problem.

The **ppmSDR** package implements a unified framework for sparse SDR based on the penalized principal machine (P^2M) approach. It efficiently solves a broad family of SDR problems using group penalties and scalable optimization algorithms, supporting both regression and classification tasks. The package includes the following methods:

- **Penalized Principal Least Squares SVM** (P^2LSM)
- **Penalized Principal SVM** (P^2SVM)
- **Penalized Principal L2-hinge SVM** (P^2L2M)
- **Penalized Principal Quantile Regression** (P^2QR)
- **Penalized Principal Asymmetric Least Squares** (P^2AR)
- **Penalized Principal Logistic Regression** (P^2LR)
- **Weighted variants** for binary classification

2 Penalized Estimation for the Sparse SDR

2.1 Principal Machine

Suppose we are given a set of data $(y_i, \mathbf{x}_i) \in \mathbb{R} \times \mathbb{R}^p$, consisting of n i.i.d. samples of $(Y, \mathbf{X})$.
We assume that $\mathbf{x}_i$ are centered, i.e., $\sum_i \mathbf{x}_i = \mathbf{0}$, without loss of generality.

For a given sequence of cutoff values $r_1 < \cdots < r_k < \cdots < r_h$, the sample version of the principal machine estimator is:

$$(\beta_{0k}, \boldsymbol{\beta}_k) = \underset{\beta_0, \boldsymbol{\beta}}{\arg\min} \, \boldsymbol{\beta}^\top \hat{\Sigma} \boldsymbol{\beta} + \frac{c}{n} \sum_{i=1}^{n} L_k \left(\tilde{y}_{ik}, \, \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i \right), \quad k = 1, \ldots, h$$

where $\hat{\Sigma} = \sum_i \mathbf{x}_i \mathbf{x}_i^\top / n$ denotes the sample covariance matrix.
The basis of the central subspace $\mathcal{S}$, $\hat{\mathbf{B}}$, is then estimated by the first d leading eigenvectors of $\sum_{k=1}^{h} \hat{\boldsymbol{\beta}}_k \hat{\boldsymbol{\beta}}_k^\top$.

- For the **RPM** (response principal machine), $\tilde{y}_{ik} = \mathbb{I}\{y_i \geq r_k\} - \mathbb{I}\{y_i < r_k\}$, with the loss function L_k fixed as L for all r_k.
- For the **LPM** (loss principal machine), the loss function L_k varies with r_k, while $\tilde{y}_{ik}$ is fixed as y_i.

2.2 Penalized Sparse Estimation

Under the sparsity assumption, let $\boldsymbol{\beta}_0^\top = (\boldsymbol{\beta}_{0,+}^\top, \boldsymbol{\beta}_{0,-}^\top)$, where $\boldsymbol{\beta}_{0,+}^\top$ and $\boldsymbol{\beta}_{0,-}^\top$ denote solutions corresponding to $\mathbf{X}_+$ and $\mathbf{X}_-$, respectively.
Due to the sparsity structure, we have $\boldsymbol{\beta}_{0,-} = \mathbf{0}$ for any r_k.
Thus, the solutions $\boldsymbol{\beta}_k$ should share a common sparsity pattern across $k = 1, \ldots, h$, leading to the following penalized principal machine ($\mathrm{P}^2\mathrm{M}$) estimator:

$$\sum_{k=1}^h \left[\boldsymbol{\beta}_k^\top \hat{\Sigma} \boldsymbol{\beta}_k + \frac{c}{n} \sum_{i=1}^n L_k(\tilde{y}_{ik}, \beta_{0k} + \boldsymbol{\beta}_k^\top \mathbf{x}_i) \right] + \sum_{j=1}^p p_\lambda \left(\|\boldsymbol{\beta}_{(j)}\|_2 \right),$$

where $\boldsymbol{\beta}_{(j)} = (\beta_{1j}, \ldots, \beta_{hj})^\top$ and $p_\lambda(\cdot)$ is a sparsity-inducing penalty such as LASSO, SCAD, or MCP.

Remark: The penalty is applied group-wise to all coefficients associated with the j-th predictor, ensuring group sparsity across slices.

2.3 Supported Losses and Algorithms

The package implements multiple $\mathrm{P}^2\mathrm{M}$ variants, using different loss functions—some previously proposed and some new in this work.
Efficient computational algorithms such as group coordinate descent (GCD) and majorization-minimization (MM-GCD) are provided for each.

Below is a summary table of available models, loss functions, and algorithms.

Machine	Response	Type	Loss	Algorithm
$\mathrm{P}^2\mathrm{LSM}$	Continuous	RPM	$(1 - \tilde{y}_k f)^2$	GCD
$\mathrm{P}^2\mathrm{WLSM}$	Binary	LPM	$w_k(1 - yf)^2$	GCD
$\mathrm{P}^2\mathrm{LR}$	Continuous	RPM	$\log(1 + e^{-\tilde{y}_k f})$	Iterative GCD
$\mathrm{P}^2\mathrm{WLR}$	Binary	LPM	$w_k \log(1 + e^{-yf})$	Iterative GCD
$\mathrm{P}^2\mathrm{AR}$	Both	LPM	$(y - f)^2(\rho_k \mathbb{I}\{y \geq f\} + (1 - \rho_k)\mathbb{I}\{y < f\})$	Iterative GCD
$\mathrm{P}^2\mathrm{L2M}$	Continuous	RPM	$[\max\{0, 1 - \tilde{y}_k f\}]^2$	Iterative GCD
$\mathrm{P}^2\mathrm{WL2M}$	Binary	LPM	$w_k[\max\{0, 1 - yf\}]^2$	Iterative GCD
$\mathrm{P}^2\mathrm{SVM}$	Continuous	RPM	$\max\{0, 1 - \tilde{y}_k f\}$	MM-GCD
$\mathrm{P}^2\mathrm{WSVM}$	Binary	LPM	$w_k \max\{0, 1 - yf\}$	MM-GCD
$\mathrm{P}^2\mathrm{QR}$	Both	LPM	$(y - f)(\rho_k - \mathbb{I}\{y < f\})$	MM-GCD

Table 1: Table: List of $\mathrm{P}^2\mathrm{M}$ methods with loss functions and computational algorithms.

2.4 Group Coordinate Descent (GCD) for $\mathrm{P}^2\mathrm{LSM}$ and $\mathrm{P}^2\mathrm{WLSM}$

The penalized principal least squares SVM ($\mathrm{P}^2\mathrm{LSM}$) solves the following quadratic optimization problem:

$$\min_{\boldsymbol{\theta}} \sum_{k=1}^h \left[\boldsymbol{\theta}_k^\top \tilde{\Sigma} \boldsymbol{\theta}_k + \frac{c}{n} \sum_{i=1}^n (1 - \tilde{y}_{ik} \cdot \boldsymbol{\theta}_k^\top \tilde{\mathbf{x}}_i)^2 \right] + \sum_{j=1}^p p_\lambda(\|\boldsymbol{\theta}_{(j)}\|).$$

For binary classification, $\mathrm{P}^2\mathrm{WLSM}$ introduces sample weights w_{ik} to adjust for class imbalance:

$$\min_{\boldsymbol{\theta}} \sum_{k=1}^{h} \left[\boldsymbol{\theta}_k^\top \tilde{\Sigma} \boldsymbol{\theta}_k + \frac{c}{n} \sum_{i=1}^{n} w_{ik}(1 - \tilde{y}_{ik} \cdot \boldsymbol{\theta}_k^\top \tilde{\mathbf{x}}_i)^2 \right] + \sum_{j=1}^{p} p_\lambda(\|\boldsymbol{\theta}_{(j)}\|).$$

Both problems can be written in matrix form as a penalized least squares problem and efficiently solved via group coordinate descent (GCD). GCD requires only $\mathcal{O}(np^2)$ operations per iteration and can scale to large p and h due to the block-diagonal matrix structure.

3 Iterative GCD for Smooth Loss Functions

The GCD algorithm can be extended to smooth loss functions such as logistic loss or asymmetric squared loss by iteratively reweighting the quadratic approximation at each step.

- For penalized principal logistic regression (P^{2}LR/P^{2}WLR):

$$\sum_{k=1}^{h} \left[\boldsymbol{\theta}_k^\top \tilde{\Sigma} \boldsymbol{\theta}_k + \frac{c}{n} \sum_{i=1}^{n} w_{ik} \log(1 + e^{-\tilde{y}_{ik} \cdot \boldsymbol{\theta}_k^\top \tilde{\mathbf{x}}_i}) \right] + \sum_{j=1}^{p} p_\lambda(\|\boldsymbol{\theta}_{(j)}\|)$$

  The Newton–Raphson method constructs a local quadratic approximation at each iteration, which can be minimized using GCD.

- For penalized principal asymmetric least squares (P^{2}AR), the squared check loss

$$L(y, f) = \tau(y - f)^2 \mathbb{I}\{y \geq f\} + (1 - \tau)(y - f)^2 \mathbb{I}\{y < f\}$$

  for quantile regression can be decomposed into positive and negative parts, leading to a weighted least squares formulation that alternates updating the active sets.

- For penalized principal L$_2$-SVM (P^{2}L2M/P^{2}WL2M), the squared hinge loss

$$L(y, f) = (1 - yf)^2 \mathbb{I}\{yf < 1\}$$

  enables iterative updates using only the observations in the margin (the "active set").

These iterative schemes reduce smooth, non-quadratic losses to a sequence of penalized least squares problems, all solved by GCD.

4 MM-GCD for Non-differentiable Losses

For non-differentiable loss functions, the Majorization–Minimization (MM) algorithm constructs a quadratic majorizer at each step, enabling the use of GCD. The general form is:

$$\min_{\boldsymbol{\theta}} \sum_{k=1}^{h} \left[\boldsymbol{\theta}_k^\top \tilde{\Sigma} \boldsymbol{\theta}_k + \frac{c}{n} \sum_{i=1}^{n} w_{ik} \bar{L}_k(y_{ik}, \boldsymbol{\theta}_k^\top \tilde{\mathbf{x}}_i \mid \boldsymbol{\theta}_k^{(t)}) \right] + \sum_{j=1}^{p} p_\lambda(\|\boldsymbol{\theta}_{(j)}\|),$$

where $\bar{L}_k(\cdot)$ is a **quadratic surrogate** for the non-smooth loss at the current iterate $\boldsymbol{\theta}_k^{(t)}$. Examples include:
- For the hinge loss (SVM):

$$\bar{L}(y, f \mid f_0) = \frac{1}{4c_0} \left\{ (1 + c_0)y - f \right\}^2, \quad c_0 = |1 - yf_0|$$

- For the quantile check loss:

$$\bar{L}_\tau(y, f \mid f_0) = \frac{1}{4} \left[\frac{1}{\epsilon + |r_0|}(y - f)^2 + (4\tau - 2)(y - f) + |r_0| \right], \quad r_0 = y - f_0$$

After majorization, the resulting penalized quadratic problem is solved by GCD at each MM iteration.

5 List of Available Methods

The `ppmSDR` package supports the following penalized principal machine methods for sufficient dimension reduction:

- **Penalized Principal Least Squares SVM** (P^2LSM)
- **Penalized Principal SVM** (P^2SVM)
- **Penalized Principal L2-hinge SVM** (P^2L2M)
- **Penalized Principal Quantile Regression** (P^2QR)
- **Penalized Principal Asymmetric Least Squares** (P^2AR)
- **Penalized Principal Logistic Regression** (P^2LR)
- **Weighted versions** of the above for binary classification

Each method supports group-structured sparsity via penalties such as group SCAD, group MCP, and group Lasso.

6 Key Features

- Single wrapper function ppm() to access all penalized PM-based SDR estimators
- Support for group SCAD, group Lasso, and group MCP penalties
- Fast computation for large-scale problems
- Outputs include estimated SDR directions, eigenvalues, and more

7 Example: SDR for Regression

```r
set.seed(1)
n <- 1000
p <- 10
B <- matrix(0, p, 2)
B[1,1] <- B[2,2] <- 1
x <- MASS::mvrnorm(n, rep(0, p), diag(1,p))
y <- (x %*% B[,1] / (0.5 + (x %*% B[,2] + 1)^2)) + 0.2 * rnorm(n)

# Penalized principal least squares SVM (P^2LSM)
fit <- ppm(x, y, H = 10, C = 1, loss = "lssvm", penalty = "grSCAD", lambda = 0.01)
round(fit$evectors[,1:2], 3)
#>          [,1]   [,2]
#>  [1,]  1.000 0.005
#>  [2,] -0.005 1.000
#>  [3,]  0.000 0.000
#>  [4,]  0.000 0.000
#>  [5,]  0.000 0.000
#>  [6,]  0.000 0.000
#>  [7,]  0.000 0.000
#>  [8,]  0.000 0.000
#>  [9,]  0.000 0.000
#> [10,]  0.000 0.000
```

8 Example: SDR for Classification

```r
set.seed(1)
y.binary <- sign(y)
```

```
# Penalized principal weighted least squares SVM (P^2WLSM)
fit2 <- ppm(x, y.binary, H = 10, C = 1, loss = "wlssvm", penalty = "grSCAD", lambda = 0.0005)
round(fit2$evectors[,1:2], 3)
#>          [,1]    [,2]
#>  [1,]  0.999  0.024
#>  [2,]  0.024 -0.999
#>  [3,]  0.000  0.000
#>  [4,]  0.012 -0.033
#>  [5,] -0.024 -0.035
#>  [6,]  0.000  0.000
#>  [7,]  0.000  0.000
#>  [8,]  0.000  0.000
#>  [9,]  0.000  0.000
#> [10,]  0.000  0.000
```

9 References

- Artemiou, A. and Dong, Y. (2016). Sufficient dimension reduction via principal lq support vector machine, *Electronic Journal of Statistics*, **10**: 783–805.
- Artemiou, A., Dong, Y. and Shin, S. J. (2021). Real-time sufficient dimension reduction through principal least squares support vector machines, *Pattern Recognition*, **112**: 107768.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**: 1348–1360.
- Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms, *The American Statistician*, **58**(1): 30–37.
- Jang, H. J., Shin, S. J. and Artemiou, A. (2023). Principal weighted least square support vector machine: An online dimension-reduction tool for binary classification, *Computational Statistics & Data Analysis*, **187**: 107818.
- Kim, B. and Shin, S. J. (2019). Principal weighted logistic regression for sufficient dimension reduction in binary classification, *Journal of the Korean Statistical Society*, **48**(2): 194–206.
- Li, B., Artemiou, A. and Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction, *Annals of Statistics*, **39**(6): 3182–3210.
- Shin, J. and Shin, S. J. (2024). A concise overview of principal support vector machines and its generalization, *Communications for Statistical Applications and Methods*, **31**(2): 235–246.
- Shin, J., Shin, S. J. and Artemiou, A. (2024). The R package psvmsdr: A unified algorithm for sufficient dimension reduction via principal machines, *arXiv preprint* arXiv:2409.01547.
- Shin, S. J. and Artemiou, A. (2017). Penalized principal logistic regression for sparse sufficient dimension reduction, *Computational Statistics & Data Analysis*, **111**: 48–58.