Principal Differences Analysis

Interpretable Characterization of Differences between Distributions

Jonas Mueller

jonasmueller@csail.mit.edu

Objectives

Consider high-dimensional random variables $X, Y \in \mathbb{R}^d$

X = measurements of various variables under condition 1
Y = measurements of same variables under condition 2

Objectives

Consider high-dimensional random variables $X, Y \in \mathbb{R}^d$

X = measurements of various variables under condition 1
Y = measurements of same variables under condition 2

Given unpaired samples $x_1, \ldots, x_n \overset{iid}{\sim} \mathbb{P}_X$, $y_1, \ldots, y_m \overset{iid}{\sim} \mathbb{P}_Y$:

Objectives

Consider high-dimensional random variables $X, Y \in \mathbb{R}^d$

X = measurements of various variables under condition 1
Y = measurements of same variables under condition 2

Given unpaired samples $x_1, \ldots, x_n \overset{iid}{\sim} \mathbb{P}_X$, $y_1, \ldots, y_m \overset{iid}{\sim} \mathbb{P}_Y$:

(Q1) Is $\mathbb{P}_X = \mathbb{P}_Y$? (Two-sample testing)

Objectives

Consider high-dimensional random variables $X, Y \in \mathbb{R}^d$

X = measurements of various variables under condition 1
Y = measurements of same variables under condition 2

Given unpaired samples $x_1, \ldots, x_n \overset{iid}{\sim} \mathbb{P}_X$, $y_1, \ldots, y_m \overset{iid}{\sim} \mathbb{P}_Y$:

(Q1) Is $\mathbb{P}_X = \mathbb{P}_Y$? (Two-sample testing)

(Q2) If not, what is minimal feature subset $S \subseteq \{1, \ldots, d\}$ such that marginal distributions $\mathbb{P}_{X_S} \neq \mathbb{P}_{Y_S}$ while $\mathbb{P}_{X_{S^C}} \approx \mathbb{P}_{Y_{S^C}}$?

Objectives

Consider high-dimensional random variables $X, Y \in \mathbb{R}^d$

X = measurements of various variables under condition 1
Y = measurements of same variables under condition 2

Given unpaired samples $x_1, \ldots, x_n \overset{iid}{\sim} \mathbb{P}_X$, $y_1, \ldots, y_m \overset{iid}{\sim} \mathbb{P}_Y$:

(Q1) Is $\mathbb{P}_X = \mathbb{P}_Y$? (Two-sample testing)

(Q2) If not, what is minimal feature subset $S \subseteq \{1, \ldots, d\}$ such that marginal distributions $\mathbb{P}_{X_S} \neq \mathbb{P}_{Y_S}$ while $\mathbb{P}_{X_{S^C}} \approx \mathbb{P}_{Y_{S^C}}$?

(Q3) How much does each feature contribute to the overall difference?

Motivation

- Understanding differences between populations
  = fundamental scientific problem

Motivation

- Understanding differences between populations
  = fundamental scientific problem
- General differences beyond mean shifts are of interest
  (e.g. variance/covariance)

Motivation

- Understanding differences between populations
  = fundamental scientific problem
- General differences beyond mean shifts are of interest
  (e.g. variance/covariance)
- Undesirable to restrict the analysis to specific parametric differences

Motivation

- Understanding differences between populations
  = fundamental scientific problem
- General differences beyond mean shifts are of interest
  (e.g. variance/covariance)
- Undesirable to restrict the analysis to specific parametric differences
- Often many variables are measured (high-dimensional data), but only a small subset expected to exhibit differences between populations

Motivation

- Understanding differences between populations
  = fundamental scientific problem
- General differences beyond mean shifts are of interest
  (e.g. variance/covariance)
- Undesirable to restrict the analysis to specific parametric differences
- Often many variables are measured (high-dimensional data), but only a small subset expected to exhibit differences between populations
- Two-sample testing is easy in univariate case: can use any statistical divergence D that measures difference between univariate distributions (e.g. Kullback-Leibler, Kolmogorov-Smirnov)

Related Work

Related Work

1. Marginal per-variable analysis
   (ignores potentially important interactions between variables)

Related Work

1. Marginal per-variable analysis
   (ignores potentially important interactions between variables)
2. Logistic regression with lasso (Tibshirani, 1996)
   (requires (log)linear relationships, only models expectation)

Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*.

Related Work

1. Marginal per-variable analysis
   (ignores potentially important interactions between variables)
2. Logistic regression with lasso (Tibshirani, 1996)
   (requires (log)linear relationships, only models expectation)
3. Sparse linear discriminants analysis (Clemmensen, 2011)
   (requires multivariate Gaussianity)

Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*.
Clemmensen L, Hastie T, Witten D, Ersbo Il B (2011). Sparse Discriminant Analysis. *Technometrics*.

Related Work

1. Marginal per-variable analysis
   (ignores potentially important interactions between variables)
2. Logistic regression with lasso (Tibshirani, 1996)
   (requires (log)linear relationships, only models expectation)
3. Sparse linear discriminants analysis (Clemmensen, 2011)
   (requires multivariate Gaussianity)
4. Random projection (Lopes, 2011)
   Direction-projection-permutation (Wei, 2015)
   (only suited for specific types of differences)

Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*.

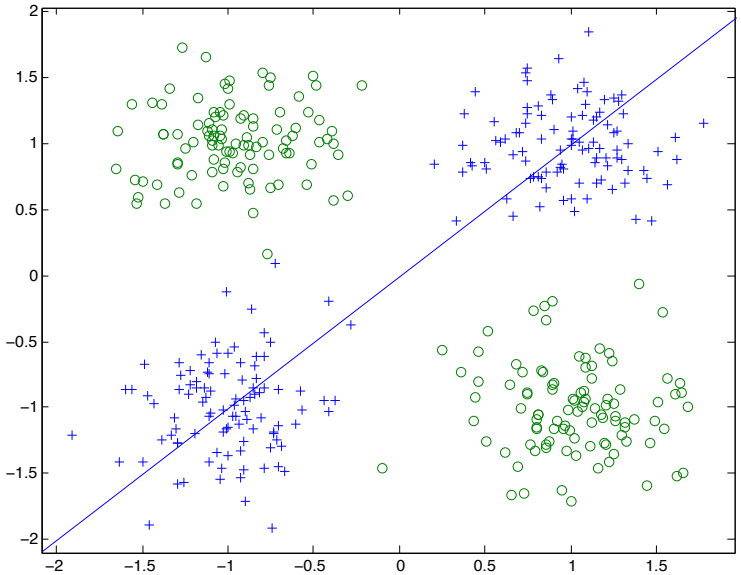
Clemmensen L, Hastie T, Witten D, Ersbo Il B (2011). Sparse Discriminant Analysis. *Technometrics*.

Lopes M, Jacob L, Wainwright M (2011). A More Powerful Two-Sample Test in High Dimensions using Random Projection. *NIPS*.

Wei S, Lee C, Wichers L, Marron JS (2015). Direction-Projection-Permutation for High Dimensional Hypothesis Tests. *Journal of Computational and Graphical Statistics*.

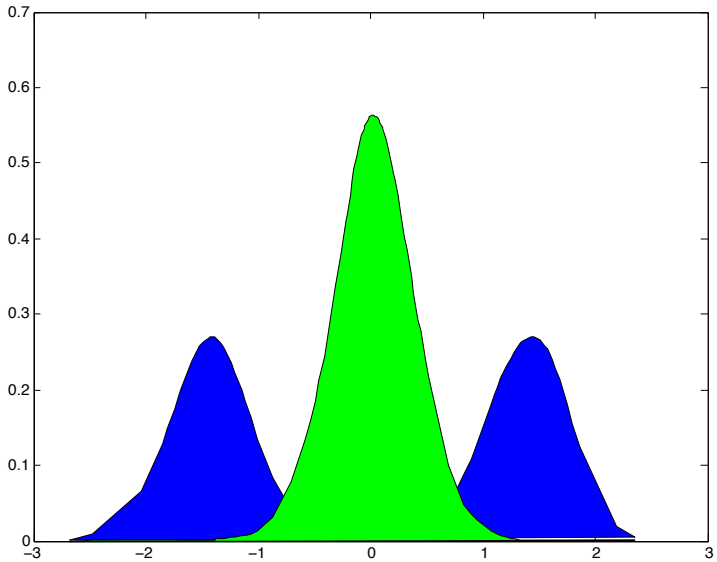
Principal Differences Analysis (PDA)

- User chooses statistical divergence D
- Goal: Find (unit-norm) projection β which maximizes $D(\beta^T X, \beta^T Y)$
- Transforms hard high-dimensional statistical problem into simple 1-D measure



PDA

$\Rightarrow$



Cramer-Wold Device

Theorem (Cramer & Wold, 1936)

Multivariate $X \overset{d}{=} Y$ *if and only if* $\beta^T X \overset{d}{=} \beta^T Y$ *for all* $\beta \in \mathbb{R}^d$

- If $\mathbb{P}_X \neq \mathbb{P}_Y$ and D is *positive definite* divergence, then PDA-projection β^* is guaranteed to ensure $D(\beta^{*T} X, \beta^{*T} Y) > 0$
- PDA can capture any type of difference between populations, using a single linear projection

Cramer H, Wold H (1936). Some Theorems on Distribution Functions. *Journal of the London Mathematical Society*.

Sparse Differences Analysis (SPARDA)

- Additional Goal: Select features over which populations differ

Sparse Differences Analysis (SPARDA)

- Additional Goal: Select features over which populations differ
- Method: Impose sparsity on β and examine features with nonzero weight in resulting projection-vector

Sparse Differences Analysis (SPARDA)

- Additional Goal: Select features over which populations differ
- Method: Impose sparsity on β and examine features with nonzero weight in resulting projection-vector

SPARDA

Find projection $\widehat{\beta}$ *that solves:* $$\max_{\beta \in \mathcal{B}, ||\beta||_0 \leq k} \left\{ D(\beta^T \widehat{X}^{(n)}, \beta^T \widehat{Y}^{(m)}) \right\}$$

$\mathcal{B} := \{\beta \in \mathbb{R}^d : ||\beta||_2 \leq 1, \beta_1 \geq 0\}$, $\beta^T \widehat{X}^{(n)}$ = *projected empirical distribution*

Sparse Differences Analysis (SPARDA)

- Additional Goal: Select features over which populations differ
- Method: Impose sparsity on β and examine features with nonzero weight in resulting projection-vector

SPARDA

Find projection $\widehat{\beta}$ *that solves:* $\max_{\beta \in \mathcal{B}, ||\beta||_0 \leq k} \left\{ D(\beta^T \widehat{X}^{(n)}, \beta^T \widehat{Y}^{(m)}) \right\}$

$\mathcal{B} := \{\beta \in \mathbb{R}^d : ||\beta||_2 \leq 1, \beta_1 \geq 0\}$, $\beta^T \widehat{X}^{(n)}$ = *projected empirical distribution*

- Cardinality constraint may be relaxed by adding $\lambda||\beta||_1$ penalty

Sparse Differences Analysis (SPARDA)

- Additional Goal: Select features over which populations differ
- Method: Impose sparsity on β and examine features with nonzero weight in resulting projection-vector

SPARDA

Find projection $\widehat{\beta}$ *that solves:* $$\max_{\beta \in \mathcal{B}, ||\beta||_0 \leq k} \left\{ D(\beta^T \widehat{X}^{(n)}, \beta^T \widehat{Y}^{(m)}) \right\}$$

$\mathcal{B} := \{\beta \in \mathbb{R}^d : ||\beta||_2 \leq 1, \beta_1 \geq 0\}$, $\beta^T \widehat{X}^{(n)}$ = *projected empirical distribution*

- Cardinality constraint may be relaxed by adding $\lambda||\beta||_1$ penalty
- In practice: choose λ or k by maximizing projected divergence between held-out samples

Choice of divergence D

- PDA enables application of rank-based measures (eg. Mann-Whitney) to high-dimensional data

Choice of divergence D

- PDA enables application of rank-based measures (eg. Mann-Whitney) to high-dimensional data
- Support Vector Machine $=$ special case of PDA where D measures margin between projected distributions

Choice of divergence D

- PDA enables application of rank-based measures (eg. Mann-Whitney) to high-dimensional data
- Support Vector Machine $=$ special case of PDA where D measures margin between projected distributions
- Fisher's Discriminant Analysis $=$ special case of PDA where D is ratio of within-vs-between-class variance (Bhattacharyya distance for Gaussian X, Y with identical covariance)

Choice of divergence D

- PDA enables application of rank-based measures (eg. Mann-Whitney) to high-dimensional data
- Support Vector Machine $=$ special case of PDA where D measures margin between projected distributions
- Fisher's Discriminant Analysis $=$ special case of PDA where D is ratio of within-vs-between-class variance (Bhattacharyya distance for Gaussian X, Y with identical covariance)
- If D defined over densities (eg. f-divergence), can use kernel density estimation. For smooth kernel (eg. Gaussian), locally optimal projection can be found via projected gradient methods.

Choice of divergence D

- PDA enables application of rank-based measures (eg. Mann-Whitney) to high-dimensional data
- Support Vector Machine $=$ special case of PDA where D measures margin between projected distributions
- Fisher's Discriminant Analysis $=$ special case of PDA where D is ratio of within-vs-between-class variance (Bhattacharyya distance for Gaussian X, Y with identical covariance)
- If D defined over densities (eg. f-divergence), can use kernel density estimation. For smooth kernel (eg. Gaussian), locally optimal projection can be found via projected gradient methods.
- Our focus is $D =$ Wasserstein distance; natural choice when variables are measured on common scale (eg. expression of various genes).

Wasserstein Distance

Definition (Squared Wasserstein Distance)

$$D(X,Y) = \min_{\mathbb{P}_{XY}} \mathbb{E}_{\mathbb{P}_{XY}} ||X - Y||^2$$

where $(X,Y) \sim \mathbb{P}_{XY}$ *and* $X \sim \mathbb{P}_X,\ Y \sim \mathbb{P}_Y$

Mallows, C (1972). A note on asymptotic joint normality. *The Annals of Mathematical Statistics*.

Wasserstein Distance

Definition (Squared Wasserstein Distance)

$$D(X,Y) = \min_{\mathbb{P}_{XY}} \mathbb{E}_{\mathbb{P}_{XY}} ||X - Y||^2$$

where $(X,Y) \sim \mathbb{P}_{XY}$ *and* $X \sim \mathbb{P}_X,\ Y \sim \mathbb{P}_Y$

- Canonical divergence between distributions on metric space, successfully used in many applications (eg. shape/image data)

Mallows, C (1972). A note on asymptotic joint normality. *The Annals of Mathematical Statistics*.

Wasserstein Distance

Definition (Squared Wasserstein Distance)

$$D(X,Y) = \min_{\mathbb{P}_{XY}} \mathbb{E}_{\mathbb{P}_{XY}} ||X - Y||^2$$

where $(X,Y) \sim \mathbb{P}_{XY}$ *and* $X \sim \mathbb{P}_X,\ Y \sim \mathbb{P}_Y$

- Canonical divergence between distributions on metric space, successfully used in many applications (eg. shape/image data)
- Intuitively: minimal amount of work to transform $\mathbb{P}_X$ into $\mathbb{P}_Y$ where work = probability mass moved $\times$ distance transported

Mallows, C (1972). A note on asymptotic joint normality. *The Annals of Mathematical Statistics*.

Wasserstein Distance

Definition (Squared Wasserstein Distance)

$$D(X,Y) = \min_{\mathbb{P}_{XY}} \mathbb{E}_{\mathbb{P}_{XY}} ||X - Y||^2$$

where $(X,Y) \sim \mathbb{P}_{XY}$ *and* $X \sim \mathbb{P}_X,\ Y \sim \mathbb{P}_Y$

- Canonical divergence between distributions on metric space, successfully used in many applications (eg. shape/image data)
- Intuitively: minimal amount of work to transform $\mathbb{P}_X$ into $\mathbb{P}_Y$ where work = probability mass moved $\times$ distance transported
- Natural dissimilarity measure between populations: integrates both fraction of individuals which are different & magnitude of differences

Mallows, C (1972). A note on asymptotic joint normality. *The Annals of Mathematical Statistics*.

Wasserstein Distance

Definition (Squared Wasserstein Distance)

$$D(X,Y) = \min_{\mathbb{P}_{XY}} \mathbb{E}_{\mathbb{P}_{XY}} ||X - Y||^2$$

where $(X,Y) \sim \mathbb{P}_{XY}$ *and* $X \sim \mathbb{P}_X,\ Y \sim \mathbb{P}_Y$

- Canonical divergence between distributions on metric space, successfully used in many applications (eg. shape/image data)
- Intuitively: minimal amount of work to transform $\mathbb{P}_X$ into $\mathbb{P}_Y$ where work = probability mass moved $\times$ distance transported
- Natural dissimilarity measure between populations: integrates both fraction of individuals which are different & magnitude of differences
- Statistically & computationally inefficient in high dimensions

Mallows, C (1972). A note on asymptotic joint normality. *The Annals of Mathematical Statistics*.

SPARDA with Wasserstein distance

Objective

$$\textit{Find}\quad \widehat{\beta} = \underset{\substack{\beta \in \mathcal{B} \\ ||\beta||_0 \leq k}}{\operatorname{argmax}} \left\{ \min_{M \in \mathcal{M}} \beta^T W_M \beta \right\}$$

$\mathcal{M}$:= *set of* $n \times m$ *matching matrices (entries* ≥ 0, *row sums* $= \frac{1}{n}$, *column sums* $= \frac{1}{m}$*)*

$W_M := \sum_{i,j} [Z_{ij} \otimes Z_{ij}] M_{ij}$ $\qquad$ $Z_{ij} := x^{(i)} - y^{(j)}$

Since: $D\left(\beta^T \widehat{X}^{(n)}, \beta^T \widehat{Y}^{(m)}\right) = \min_{M \in \mathcal{M}} \sum_{i,j} M_{ij} \left(\beta^T x^{(i)} - \beta^T y^{(j)}\right)^2 = \min_{M \in \mathcal{M}} \beta^T W_M \beta$

SPARDA with Wasserstein distance

Objective

$$\textit{Find} \quad \widehat{\beta} = \underset{\substack{\beta \in \mathcal{B} \\ ||\beta||_0 \leq k}}{\operatorname{argmax}} \left\{ \min_{M \in \mathcal{M}} \beta^T W_M \beta \right\}$$

$\mathcal{M}$:= *set of* $n \times m$ *matching matrices (entries* ≥ 0, *row sums* $= \frac{1}{n}$, *column sums* $= \frac{1}{m}$*)*

$W_M := \sum_{i,j} [Z_{ij} \otimes Z_{ij}] M_{ij}$ $\qquad$ $Z_{ij} := x^{(i)} - y^{(j)}$

Since: $D\left(\beta^T \widehat{X}^{(n)}, \beta^T \widehat{Y}^{(m)}\right) = \min_{M \in \mathcal{M}} \sum_{i,j} M_{ij} \left(\beta^T x^{(i)} - \beta^T y^{(j)}\right)^2 = \min_{M \in \mathcal{M}} \beta^T W_M \beta$

- Non-concave max-min optimization

SPARDA with Wasserstein distance

Objective

$$\textit{Find}\ \ \widehat{\beta} = \underset{\substack{\beta \in \mathcal{B} \\ ||\beta||_0 \leq k}}{\operatorname{argmax}} \left\{ \min_{M \in \mathcal{M}} \beta^T W_M \beta \right\}$$

$\mathcal{M}$:= *set of* $n \times m$ *matching matrices (entries* ≥ 0, *row sums* $= \frac{1}{n}$, *column sums* $= \frac{1}{m}$*)*

$W_M := \sum_{i,j} [Z_{ij} \otimes Z_{ij}] M_{ij}$ $\qquad$ $Z_{ij} := x^{(i)} - y^{(j)}$

Since: $D\left(\beta^T \widehat{X}^{(n)}, \beta^T \widehat{Y}^{(m)}\right) = \min_{M \in \mathcal{M}} \sum_{i,j} M_{ij} \left(\beta^T x^{(i)} - \beta^T y^{(j)}\right)^2 = \min_{M \in \mathcal{M}} \beta^T W_M \beta$

- Non-concave max-min optimization

Two-step relax-tighten procedure:

1. Solve convex relaxation (semidefinite program).
2. Run steepest ascent method to greedily improve the current projection with respect to the original nonconvex objective (if relaxation is not tight).

Wang Z, Lu H, Liu H (2014). Tighten after Relax: Minimax-Optimal Sparse PCA in Polynomial Time. *NIPS*.

Semidefinite Relaxation

- Can rewrite SPARDA objective:

$$\max_{B} \min_{M \in \mathcal{M}} \text{tr}\left(W_M B\right) \text{ subject to } B \in \mathcal{B}_r,\ ||B||_0 \leq k^2,\ \text{rank}(B) = 1$$

where $B = \beta \otimes \beta$ and $\mathcal{B}_r = \{B \in \mathbb{R}^{d \times d} : \text{tr}(B) = 1,\ B \succeq 0\}$

Semidefinite Relaxation

- Can rewrite SPARDA objective:

$$\max_{B} \min_{M \in \mathcal{M}} \text{tr}\left(W_M B\right) \text{ subject to } B \in \mathcal{B}_r,\ ||B||_0 \leq k^2,\ \text{rank}(B) = 1$$

where $B = \beta \otimes \beta$ and $\mathcal{B}_r = \{B \in \mathbb{R}^{d \times d} : \text{tr}(B) = 1,\ B \succeq 0\}$

- Relaxation: $\max_{B \in \mathcal{B}_r} \left\{ \min_{M \in \mathcal{M}} \text{tr}\left(W_M B\right) - \lambda ||B||_1 \right\}$

Semidefinite Relaxation

- Can rewrite SPARDA objective:

$$\max_{B} \min_{M \in \mathcal{M}} \text{tr}\left(W_M B\right) \text{ subject to } B \in \mathcal{B}_r,\ ||B||_0 \leq k^2,\ \text{rank}(B) = 1$$

where $B = \beta \otimes \beta$ and $\mathcal{B}_r = \{B \in \mathbb{R}^{d \times d} : \text{tr}(B) = 1,\ B \succeq 0\}$

- Relaxation: $\max_{B \in \mathcal{B}_r} \left\{ \min_{M \in \mathcal{M}} \text{tr}\left(W_M B\right) - \lambda ||B||_1 \right\}$

- While concave, max-min relaxation remains computationally demanding. Instead, turn to the dual:

$$\max_{\substack{B \in \mathcal{B}_r \\ u \in \mathbb{R}^n, v \in \mathbb{R}^m}} \frac{1}{m} \sum_{i,j} \min\{0, \text{tr}([Z_{ij} \otimes Z_{ij}]\, B) - u_i - v_j\} + \frac{1}{n} \sum_{i=1}^{n} u_i + \frac{1}{m} \sum_{j=1}^{m} v_j - \lambda ||B||_1$$

Semidefinite Relaxation

- Can rewrite SPARDA objective:

$$\max_{B} \min_{M \in \mathcal{M}} \text{tr}\,(W_M B) \text{ subject to } B \in \mathcal{B}_r,\ ||B||_0 \leq k^2,\ \text{rank}(B) = 1$$

where $B = \beta \otimes \beta$ and $\mathcal{B}_r = \{B \in \mathbb{R}^{d \times d} : \text{tr}(B) = 1,\ B \succeq 0\}$

- Relaxation: $\max_{B \in \mathcal{B}_r} \left\{ \min_{M \in \mathcal{M}} \text{tr}\,(W_M B) - \lambda ||B||_1 \right\}$

- While concave, max-min relaxation remains computationally demanding. Instead, turn to the dual:

$$\max_{\substack{B \in \mathcal{B}_r \\ u \in \mathbb{R}^n, v \in \mathbb{R}^m}} \frac{1}{m} \sum_{i,j} \min\{0, \text{tr}([Z_{ij} \otimes Z_{ij}]\, B) - u_i - v_j\} + \frac{1}{n} \sum_{i=1}^{n} u_i + \frac{1}{m} \sum_{j=1}^{m} v_j - \lambda ||B||_1$$

- Find optimal B^* via projected subgradient method, take largest eigenvalue of B^* as best projection vector $\widehat{\beta}_{\text{relax}}$

Subgradient Algorithm for Semidefinite Relaxation

RELAX Algorithm: Solves the dualized semidefinite relaxation of SPARDA. Returns the largest eigenvector of the solution to (2) as the desired projection direction for SPARDA.

Input: d-dimensional data $x^{(1)}, \ldots, x^{(n)}$ and $y^{(1)}, \ldots, y^{(m)}$ (with $n \geqslant m$)
Parameters: $\lambda \geqslant 0$ controls the amount of regularization, $\gamma > 0$ is the step-size used for B updates, $\eta > 0$ is the step-size used for updates of dual variables u and v, T is the maximum number of iterations without improvement in cost after which algorithm terminates.

Initialize $\beta^{(0)} \leftarrow \left[\frac{\sqrt{d}}{d}, \ldots, \frac{\sqrt{d}}{d}\right]$, $B^{(0)} \leftarrow \beta^{(0)} \otimes \beta^{(0)} \in \mathcal{B}_r$, $u^{(0)} \leftarrow \mathbf{0}_{n\times 1}$, $v^{(0)} \leftarrow \mathbf{0}_{m\times 1}$
While the number of iterations since last improvement in objective function is less than T:
  $\partial u \leftarrow [1/n, \ldots, 1/n] \in \mathbb{R}^n$, $\partial v \leftarrow [1/m, \ldots, 1/m] \in \mathbb{R}^m$, $\partial B \leftarrow \mathbf{0}_{d\times d}$
  For $i, j \in \{1, \ldots, n\} \times \{1, \ldots, m\}$:
    $Z_{ij} \leftarrow x^{(i)} - y^{(j)}$
    If $\text{tr}([Z_{ij} \otimes Z_{ij}]B^{(t)}) - u_i^{(t)} - v_j^{(t)} < 0$:
      $\partial u_i \leftarrow \partial u_i - 1/m$, $\partial v_j \leftarrow \partial v_j - 1/m$, $\partial B \leftarrow \partial B + Z_{ij} \otimes Z_{ij}/m$
  End For
  $u^{(t+1)} \leftarrow u^{(t)} + \eta \cdot \partial u$ and $v^{(t+1)} \leftarrow v^{(t)} + \eta \cdot \partial v$
  $B^{(t+1)} \leftarrow$ **Projection**$\left(B^{(t)} + \frac{\gamma}{||\partial B||_F} \cdot \partial B \,;\, \lambda,\, \gamma/||\partial B||_F\right)$

Output: $\widehat{\beta}_{\text{relax}} \in \mathbb{R}^d$ defined as the largest eigenvector (based on corresponding eigenvalue's magnitude) of the matrix $B^{(t^*)}$ which attained the best objective value over all iterations.

Projection Algorithm: Projects matrix onto positive semidefinite cone of unit-trace matrices $\mathcal{B}_r$ (the feasible set in our relaxation). Step 4 applies soft-thresholding proximal operator for sparsity.

Input: $B \in \mathbb{R}^{d\times d}$
Parameters: $\lambda \geqslant 0$ controls the amount of regularization, $\delta = \gamma/||\partial B||_F \geqslant 0$ is the actual step-size used in the B-update.

$Q\Lambda Q^T \leftarrow$ eigendecomposition of B
$w^* \leftarrow \arg\min\left\{||w - \text{diag}(\Lambda)||_2^2 : w \in [0,1]^d, ||w||_1 = 1\right\}$ (Quadratic program)
$\widetilde{B} \leftarrow Q \cdot \text{diag}\{w_1^*, \ldots, w_d^*\} \cdot Q^T$
If $\lambda > 0$: **For** $r, s \in \{1, \ldots, d\}^2$: $\widetilde{B}_{r,s} \leftarrow \text{sign}(\widetilde{B}_{r,s}) \cdot \max\{0, |\widetilde{B}_{r,s}| - \delta\lambda\}$

Output: $\widetilde{B} \in \mathcal{B}_r$

Subgradient Algorithm for Semidefinite Relaxation

RELAX Algorithm: Solves the dualized semidefinite relaxation of SPARDA. Returns the largest eigenvector of the solution to (2) as the desired projection direction for SPARDA.

Input: d-dimensional data $x^{(1)}, \ldots, x^{(n)}$ and $y^{(1)}, \ldots, y^{(m)}$ (with $n \geqslant m$)
Parameters: $\lambda \geqslant 0$ controls the amount of regularization, $\gamma > 0$ is the step-size used for B updates, $\eta > 0$ is the step-size used for updates of dual variables u and v, T is the maximum number of iterations without improvement in cost after which algorithm terminates.

1: Initialize $\beta^{(0)} \leftarrow \left[\frac{\sqrt{d}}{d}, \ldots, \frac{\sqrt{d}}{d}\right]$, $B^{(0)} \leftarrow \beta^{(0)} \otimes \beta^{(0)} \in \mathcal{B}_r$, $u^{(0)} \leftarrow \mathbf{0}_{n\times 1}$, $v^{(0)} \leftarrow \mathbf{0}_{m\times 1}$
2: **While** the number of iterations since last improvement in objective function is less than T:
3: $\quad \partial u \leftarrow [1/n, \ldots, 1/n] \in \mathbb{R}^n$, $\partial v \leftarrow [1/m, \ldots, 1/m] \in \mathbb{R}^m$, $\partial B \leftarrow \mathbf{0}_{d\times d}$
4: $\quad$ **For** $i, j \in \{1, \ldots, n\} \times \{1, \ldots, m\}$:
5: $\quad\quad Z_{ij} \leftarrow x^{(i)} - y^{(j)}$
6: $\quad\quad$ **If** $\text{tr}([Z_{ij} \otimes Z_{ij}]B^{(t)}) - u_i^{(t)} - v_j^{(t)} < 0$:
7: $\quad\quad\quad \partial u_i \leftarrow \partial u_i - 1/m$, $\partial v_j \leftarrow \partial v_j - 1/m$, $\partial B \leftarrow \partial B + Z_{ij} \otimes Z_{ij}/m$
8: $\quad$ **End For**
9: $\quad u^{(t+1)} \leftarrow u^{(t)} + \eta \cdot \partial u$ and $v^{(t+1)} \leftarrow v^{(t)} + \eta \cdot \partial v$
10: $\quad B^{(t+1)} \leftarrow$ **Projection**$\left(B^{(t)} + \frac{\gamma}{||\partial B||_F} \cdot \partial B \,;\, \lambda,\, \gamma/||\partial B||_F\right)$

Output: $\widehat{\beta}_{\text{relax}} \in \mathbb{R}^d$ defined as the largest eigenvector (based on corresponding eigenvalue's magnitude) of the matrix $B^{(t^*)}$ which attained the best objective value over all iterations.

Projection Algorithm: Projects matrix onto positive semidefinite cone of unit-trace matrices $\mathcal{B}_r$ (the feasible set in our relaxation). Step 4 applies soft-thresholding proximal operator for sparsity.

Input: $B \in \mathbb{R}^{d\times d}$
Parameters: $\lambda \geqslant 0$ controls the amount of regularization, $\delta = \gamma/||\partial B||_F \geqslant 0$ is the actual step-size used in the B-update.

1: $Q\Lambda Q^T \leftarrow$ eigendecomposition of B
2: $w^* \leftarrow \arg\min\left\{||w - \text{diag}(\Lambda)||_2^2 : w \in [0,1]^d, ||w||_1 = 1\right\}$ (Quadratic program)
3: $\widetilde{B} \leftarrow Q \cdot \text{diag}\{w_1^*, \ldots, w_d^*\} \cdot Q^T$
4: **If** $\lambda > 0$: **For** $r, s \in \{1, \ldots, d\}^2$: $\quad \widetilde{B}_{r,s} \leftarrow \text{sign}(\widetilde{B}_{r,s}) \cdot \max\{0, |\widetilde{B}_{r,s}| - \delta\lambda\}$

Output: $\widetilde{B} \in \mathcal{B}_r$

- Guaranteed to converge to global optimum (Bertsekas, 2011)

Bertsekas DP (2011). Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*.

Subgradient Algorithm for Semidefinite Relaxation

RELAX Algorithm: Solves the dualized semidefinite relaxation of SPARDA. Returns the largest eigenvector of the solution to (2) as the desired projection direction for SPARDA.

Input: d-dimensional data $x^{(1)}, \ldots, x^{(n)}$ and $y^{(1)}, \ldots, y^{(m)}$ (with $n \geqslant m$)
Parameters: $\lambda \geqslant 0$ controls the amount of regularization, $\gamma > 0$ is the step-size used for B updates, $\eta > 0$ is the step-size used for updates of dual variables u and v, T is the maximum number of iterations without improvement in cost after which algorithm terminates.

Initialize $\beta^{(0)} \leftarrow \left[\frac{\sqrt{d}}{d}, \ldots, \frac{\sqrt{d}}{d}\right]$, $B^{(0)} \leftarrow \beta^{(0)} \otimes \beta^{(0)} \in \mathcal{B}_r$, $u^{(0)} \leftarrow \mathbf{0}_{n\times 1}$, $v^{(0)} \leftarrow \mathbf{0}_{m\times 1}$
While the number of iterations since last improvement in objective function is less than T:
  $\partial u \leftarrow [1/n, \ldots, 1/n] \in \mathbb{R}^n$, $\partial v \leftarrow [1/m, \ldots, 1/m] \in \mathbb{R}^m$, $\partial B \leftarrow \mathbf{0}_{d\times d}$
  For $i, j \in \{1, \ldots, n\} \times \{1, \ldots, m\}$:
    $Z_{ij} \leftarrow x^{(i)} - y^{(j)}$
    If $\text{tr}([Z_{ij} \otimes Z_{ij}]B^{(t)}) - u_i^{(t)} - v_j^{(t)} < 0$:
      $\partial u_i \leftarrow \partial u_i - 1/m$, $\partial v_j \leftarrow \partial v_j - 1/m$, $\partial B \leftarrow \partial B + Z_{ij} \otimes Z_{ij}/m$
  End For
  $u^{(t+1)} \leftarrow u^{(t)} + \eta \cdot \partial u$ and $v^{(t+1)} \leftarrow v^{(t)} + \eta \cdot \partial v$
  $B^{(t+1)} \leftarrow$ **Projection**$\left(B^{(t)} + \frac{\gamma}{||\partial B||_F} \cdot \partial B \,;\, \lambda,\, \gamma/||\partial B||_F\right)$

Output: $\widehat{\beta}_{\text{relax}} \in \mathbb{R}^d$ defined as the largest eigenvector (based on corresponding eigenvalue's magnitude) of the matrix $B^{(t^*)}$ which attained the best objective value over all iterations.

Projection Algorithm: Projects matrix onto positive semidefinite cone of unit-trace matrices $\mathcal{B}_r$ (the feasible set in our relaxation). Step 4 applies soft-thresholding proximal operator for sparsity.

Input: $B \in \mathbb{R}^{d\times d}$
Parameters: $\lambda \geqslant 0$ controls the amount of regularization, $\delta = \gamma/||\partial B||_F \geqslant 0$ is the actual step-size used in the B-update.

$Q\Lambda Q^T \leftarrow$ eigendecomposition of B
$w^* \leftarrow \arg\min\left\{||w - \text{diag}(\Lambda)||_2^2 : w \in [0,1]^d, ||w||_1 = 1\right\}$ (Quadratic program)
$\widetilde{B} \leftarrow Q \cdot \text{diag}\{w_1^*, \ldots, w_d^*\} \cdot Q^T$
If $\lambda > 0$: **For** $r, s \in \{1, \ldots, d\}^2$: $\widetilde{B}_{r,s} \leftarrow \text{sign}(\widetilde{B}_{r,s}) \cdot \max\{0, |\widetilde{B}_{r,s}| - \delta\lambda\}$

Output: $\widetilde{B} \in \mathcal{B}_r$

- Guaranteed to converge to global optimum (Bertsekas, 2011)
- To scale to large datasets, can employ incremental subgradients by drawing random (i, j) pairs

Bertsekas DP (2011). Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*.

Subgradient Algorithm for Semidefinite Relaxation

RELAX Algorithm: Solves the dualized semidefinite relaxation of SPARDA. Returns the largest eigenvector of the solution to (2) as the desired projection direction for SPARDA.

Input: d-dimensional data $x^{(1)}, \ldots, x^{(n)}$ and $y^{(1)}, \ldots, y^{(m)}$ (with $n \geqslant m$)
Parameters: $\lambda \geqslant 0$ controls the amount of regularization, $\gamma > 0$ is the step-size used for B updates, $\eta > 0$ is the step-size used for updates of dual variables u and v, T is the maximum number of iterations without improvement in cost after which algorithm terminates.

Initialize $\beta^{(0)} \leftarrow \left[\frac{\sqrt{d}}{d}, \ldots, \frac{\sqrt{d}}{d}\right]$, $B^{(0)} \leftarrow \beta^{(0)} \otimes \beta^{(0)} \in \mathcal{B}_r$, $u^{(0)} \leftarrow \mathbf{0}_{n\times 1}$, $v^{(0)} \leftarrow \mathbf{0}_{m\times 1}$
While the number of iterations since last improvement in objective function is less than T:
  $\partial u \leftarrow [1/n, \ldots, 1/n] \in \mathbb{R}^n$, $\partial v \leftarrow [1/m, \ldots, 1/m] \in \mathbb{R}^m$, $\partial B \leftarrow \mathbf{0}_{d\times d}$
  For $i, j \in \{1, \ldots, n\} \times \{1, \ldots, m\}$:
    $Z_{ij} \leftarrow x^{(i)} - y^{(j)}$
    If $\text{tr}([Z_{ij} \otimes Z_{ij}]B^{(t)}) - u_i^{(t)} - v_j^{(t)} < 0$:
      $\partial u_i \leftarrow \partial u_i - 1/m$, $\partial v_j \leftarrow \partial v_j - 1/m$, $\partial B \leftarrow \partial B + Z_{ij} \otimes Z_{ij}/m$
  End For
  $u^{(t+1)} \leftarrow u^{(t)} + \eta \cdot \partial u$ and $v^{(t+1)} \leftarrow v^{(t)} + \eta \cdot \partial v$
  $B^{(t+1)} \leftarrow$ **Projection**$\left(B^{(t)} + \frac{\gamma}{||\partial B||_F} \cdot \partial B \,;\, \lambda,\, \gamma/||\partial B||_F\right)$
Output: $\widehat{\beta}_{\text{relax}} \in \mathbb{R}^d$ defined as the largest eigenvector (based on corresponding eigenvalue's magnitude) of the matrix $B^{(t^*)}$ which attained the best objective value over all iterations.

Projection Algorithm: Projects matrix onto positive semidefinite cone of unit-trace matrices $\mathcal{B}_r$ (the feasible set in our relaxation). Step 4 applies soft-thresholding proximal operator for sparsity.

Input: $B \in \mathbb{R}^{d\times d}$
Parameters: $\lambda \geqslant 0$ controls the amount of regularization, $\delta = \gamma/||\partial B||_F \geqslant 0$ is the actual step-size used in the B-update.

$Q\Lambda Q^T \leftarrow$ eigendecomposition of B
$w^* \leftarrow \arg\min\left\{||w - \text{diag}(\Lambda)||_2^2 : w \in [0,1]^d, ||w||_1 = 1\right\}$ (Quadratic program)
$\widetilde{B} \leftarrow Q \cdot \text{diag}\{w_1^*, \ldots, w_d^*\} \cdot Q^T$
If $\lambda > 0$: **For** $r, s \in \{1, \ldots, d\}^2$: $\widetilde{B}_{r,s} \leftarrow \text{sign}(\widetilde{B}_{r,s}) \cdot \max\{0, |\widetilde{B}_{r,s}| - \delta\lambda\}$
Output: $\widetilde{B} \in \mathcal{B}_r$

- Guaranteed to converge to global optimum (Bertsekas, 2011)
- To scale to large datasets, can employ incremental subgradients by drawing random (i, j) pairs
- Use different learning rates for B, u, and v (eg. Adagrad).

Bertsekas DP (2011). Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*.

Tightening Step

- Projected subgradient method to greedily improve original nonconvex objective $J(\beta) = \min_{M \in \mathcal{M}} \beta^T W_M \beta$ s.t. $\beta \in \mathcal{B}, ||\beta||_0 \leq k$

Tightening Step

- Projected subgradient method to greedily improve original nonconvex objective $J(\beta) = \min_{M \in \mathcal{M}} \beta^T W_M \beta$ s.t. $\beta \in \mathcal{B}, ||\beta||_0 \leq k$
- Sparsity level $k := ||\widehat{\beta}_{\text{relax}}||_0$

Tightening Step

- Projected subgradient method to greedily improve original nonconvex objective $J(\beta) = \min_{M \in \mathcal{M}} \beta^T W_M \beta$ s.t. $\beta \in \mathcal{B}, ||\beta||_0 \leq k$
- Sparsity level $k := ||\widehat{\beta}_{\text{relax}}||_0$
- For any β: matching-minimization (and subgradients of J) computed by sorting scalars $\beta^T x^{(1)}, \ldots, \beta^T x^{(n)}, \beta^T y^{(1)}, \ldots, \beta^T y^{(m)}$ (matching matrices not needed)

Fact

In 1-D, Wasserstein distance = L_2 norm between quantile functions.

$$D(X, Y) = \int_0^1 \left[F_Y^{-1}(p) - F_X^{-1}(p)\right]^2 \mathrm{d}p$$

Tightening Step

- Projected subgradient method to greedily improve original nonconvex objective $J(\beta) = \min_{M \in \mathcal{M}} \beta^T W_M \beta$ s.t. $\beta \in \mathcal{B}, ||\beta||_0 \leq k$
- Sparsity level $k := ||\widehat{\beta}_{\text{relax}}||_0$
- For any β: matching-minimization (and subgradients of J) computed by sorting scalars $\beta^T x^{(1)}, \ldots, \beta^T x^{(n)}, \beta^T y^{(1)}, \ldots, \beta^T y^{(m)}$ (matching matrices not needed)

Fact

In 1-D, Wasserstein distance = L_2 norm between quantile functions.

$$D(X, Y) = \int_0^1 \left[F_Y^{-1}(p) - F_X^{-1}(p)\right]^2 \mathrm{d}p$$

- Time-complexity (per iteration):
  Tightening procedure $= O(dn \log n)$, RELAX algorithm $= O(d^3 n^2)$

Some cases where relaxation is tight

1. The *projected* Wasserstein distance between X and Y in some direction is nearly as large as overall Wasserstein distance in $\mathbb{R}^d$.

   Ex: if $||\mathbb{E}[X] - \mathbb{E}[Y]||_2 \gg \max\{||\text{Cov}(X)||_F, ||\text{Cov}(Y)||_F\}$

Some cases where relaxation is tight

1. The *projected* Wasserstein distance between X and Y in some direction is nearly as large as overall Wasserstein distance in $\mathbb{R}^d$.

   Ex: if $||\mathbb{E}[X] - \mathbb{E}[Y]||_2 \gg \max\{||\text{Cov}(X)||_F, ||\text{Cov}(Y)||_F\}$

2. $X \sim N(\mu_X, \Sigma_X)$ and $Y \sim N(\mu_Y, \Sigma_Y)$ with $\mu_X \neq \mu_Y$ and $\Sigma_X \approx \Sigma_Y$

Some cases where relaxation is tight

1. The *projected* Wasserstein distance between X and Y in some direction is nearly as large as overall Wasserstein distance in $\mathbb{R}^d$.

   Ex: if $||\mathbb{E}[X] - \mathbb{E}[Y]||_2 \gg \max\{||\text{Cov}(X)||_F, ||\text{Cov}(Y)||_F\}$

2. $X \sim N(\mu_X, \Sigma_X)$ and $Y \sim N(\mu_Y, \Sigma_Y)$ with $\mu_X \neq \mu_Y$ and $\Sigma_X \approx \Sigma_Y$

3. $X \sim N(\mu_X, \Sigma_X)$ and $Y \sim N(\mu_Y, \Sigma_Y)$ where $\mu_X \approx \mu_Y$ and $\underset{B \in \mathcal{B}_r}{\text{argmax}} ||(B^{1/2}\Sigma_X B^{1/2})^{1/2} - (B^{1/2}\Sigma_Y B^{1/2})^{1/2}||_F^2$ is nearly rank 1.

   Ex: $\Sigma_Y \approx V \cdot \Sigma_X$ (where V is a diagonal matrix)

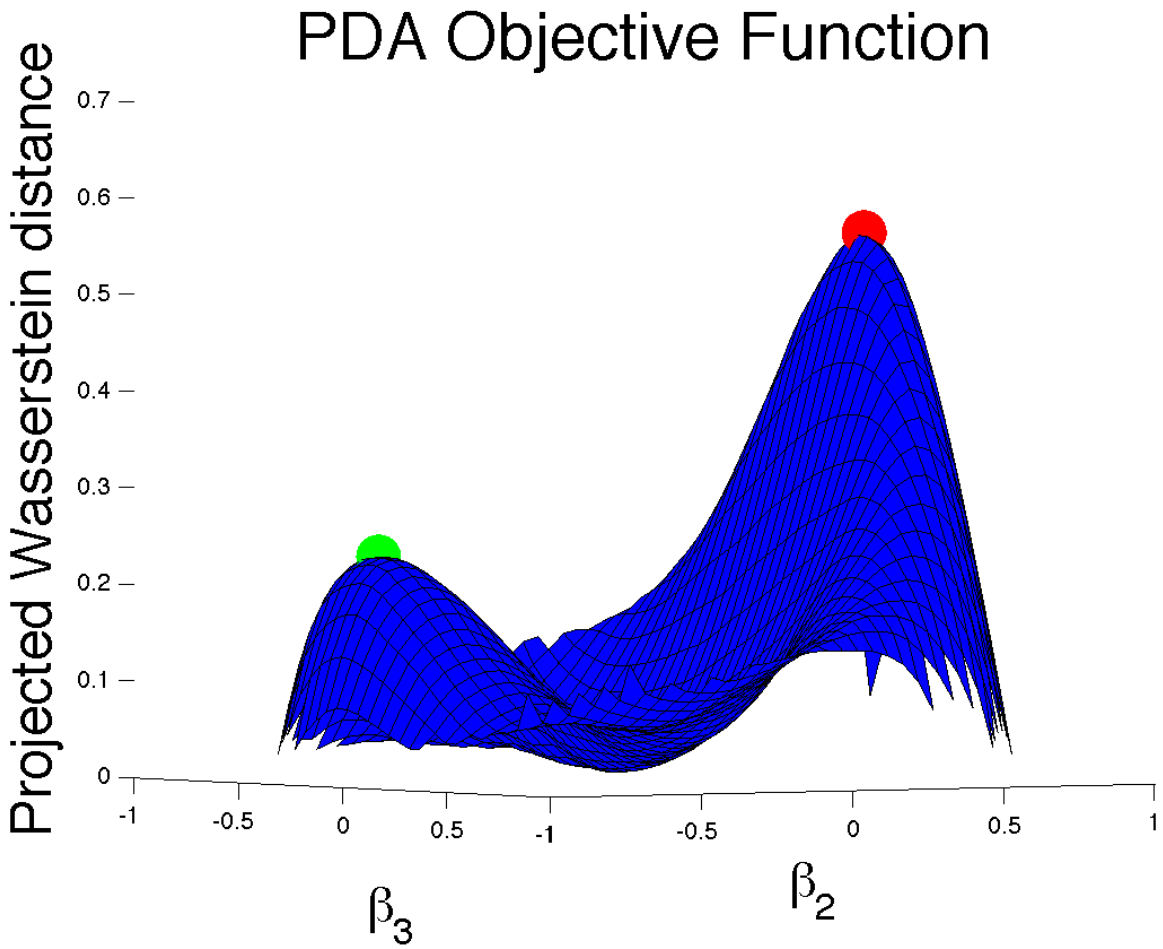


Figure: PDA objective for two 3-dimensional Gaussian distributions.
Green = solution found by tightening procedure.
Red = solution found by RELAX algorithm.

Statistical Properties

Simplifying assumptions: (A1) $n = m$ (A2) X, Y admit continuous density functions (A3) X, Y compactly supported with nonzero density in Euclidean ball of radius R (A4) $\widehat{\beta} = \arg\max_{\beta \in \mathcal{B}} D(\beta^T \widehat{X}^{(n)}, \beta^T \widehat{Y}^{(n)})$

Statistical Properties

Simplifying assumptions: (A1) $n = m$ (A2) X, Y admit continuous density functions (A3) X, Y compactly supported with nonzero density in Euclidean ball of radius R (A4) $\widehat{\beta} = \arg\max_{\beta \in \mathcal{B}} D(\beta^T \widehat{X}^{(n)}, \beta^T \widehat{Y}^{(n)})$

Theorem 1

If there exists direction $\beta^* \in \mathcal{B}$ such that $D(\beta^{*T} X, \beta^{*T} Y) \geq \Delta$, then:

$$D(\widehat{\beta}^T \widehat{X}^{(n)}, \widehat{\beta}^T \widehat{Y}^{(n)}) > \Delta - \epsilon \qquad \textit{with probability } \geq 1 - 4\exp\left(-\frac{n\epsilon^2}{16R^4}\right)$$

Statistical Properties

Simplifying assumptions: (A1) $n = m$ (A2) X, Y admit continuous density functions (A3) X, Y compactly supported with nonzero density in Euclidean ball of radius R (A4) $\widehat{\beta} = \arg\max_{\beta \in \mathcal{B}} D(\beta^T \widehat{X}^{(n)}, \beta^T \widehat{Y}^{(n)})$

Theorem 1

If there exists direction $\beta^* \in \mathcal{B}$ such that $D(\beta^{*T} X, \beta^{*T} Y) \geq \Delta$, then:

$$D(\widehat{\beta}^T \widehat{X}^{(n)}, \widehat{\beta}^T \widehat{Y}^{(n)}) > \Delta - \epsilon \qquad \textit{with probability } \geq 1 - 4\exp\left(-\frac{n\epsilon^2}{16R^4}\right)$$

Theorem 2

If X and Y are identically distributed in $\mathbb{R}^d$, then:
$D(\widehat{\beta}^T \widehat{X}^{(n)}, \widehat{\beta}^T \widehat{Y}^{(n)}) < \epsilon$

with probability $\geq 1 - C_1 \left(1 + \frac{R^2}{\epsilon}\right)^d \exp\left(-\frac{C_2}{R^4} n\epsilon^2\right)$

Statistical Properties

Extra assumptions: (A4) Y sub-Gaussian, (A5) $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, (A6) Var$[X_\ell] = 1$

Jirak M (2011). On the maximum of covariance estimators. *Journal of Multivariate Analysis*.

Statistical Properties

Extra assumptions: (A4) Y sub-Gaussian, (A5) $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, (A6) Var$[X_\ell] = 1$

Let $T_a(X, Y) = |\Pr(|X_1| \leq a, \ldots, |X_d| \leq a) - \Pr(|Y_1| \leq a, \ldots, |Y_d| \leq a)|$
(measures difference between $X, Y \in \mathbb{R}^d$, parameterized by $a \geq 0$)

Jirak M (2011). On the maximum of covariance estimators. *Journal of Multivariate Analysis*.

Statistical Properties

Extra assumptions: (A4) Y sub-Gaussian, (A5) $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, (A6) $\text{Var}[X_\ell] = 1$

Let $T_a(X, Y) = |\Pr(|X_1| \leq a, \ldots, |X_d| \leq a) - \Pr(|Y_1| \leq a, \ldots, |Y_d| \leq a)|$
(measures difference between $X, Y \in \mathbb{R}^d$, parameterized by $a \geq 0$)

Define $h\left(g(\Delta)\right) := \min\{\Delta_1, \Delta_2\}$

$$\Delta_1 = (a+d)^d(g(\Delta)+d) + \exp(-a^2/2) + \psi \exp\left(-1/(\sqrt{2}\psi)\right)$$

$$\Delta_2 = \left(g(\Delta) + \exp(-a^2/2)\right) \cdot d$$

$\psi = ||\text{Cov}(X)||_1$, $g(\Delta) = \Delta^4 \cdot (1+\Phi)^{-4}$, and $\Phi = \sup_{\alpha \in \mathcal{B}} \left\{ \sup_y |f_{\alpha^T Y}(y)| \right\}$
with $f_{\alpha^T Y}(y)$ defined as the density of the projection of Y in the α direction.

Jirak M (2011). On the maximum of covariance estimators. *Journal of Multivariate Analysis*.

Statistical Properties

Extra assumptions: (A4) Y sub-Gaussian, (A5) $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, (A6) $\text{Var}[X_\ell] = 1$

Let $T_a(X, Y) = |\Pr(|X_1| \leq a, \ldots, |X_d| \leq a) - \Pr(|Y_1| \leq a, \ldots, |Y_d| \leq a)|$
(measures difference between $X, Y \in \mathbb{R}^d$, parameterized by $a \geq 0$)

Define $h\left(g(\Delta)\right) := \min\{\Delta_1, \Delta_2\}$

$$\Delta_1 = (a+d)^d (g(\Delta) + d) + \exp(-a^2/2) + \psi \exp\left(-1/(\sqrt{2}\psi)\right)$$

$$\Delta_2 = \left(g(\Delta) + \exp(-a^2/2)\right) \cdot d$$

$\psi = ||\text{Cov}(X)||_1$, $g(\Delta) = \Delta^4 \cdot (1+\Phi)^{-4}$, and $\Phi = \sup_{\alpha \in \mathcal{B}} \left\{ \sup_y |f_{\alpha^T Y}(y)| \right\}$
with $f_{\alpha^T Y}(y)$ defined as the density of the projection of Y in the α direction.

Theorem 3

If $\exists\ a \geq 0$ *s.t.* $T_a(X, Y) > h\left(g(\Delta)\right)$, *then:* $D(\widehat{\beta}^T \widehat{X}^{(n)}, \widehat{\beta}^T \widehat{Y}^{(n)}) > C\Delta - \epsilon$
with probability $\geq\ 1 - C_1 \exp\left(-\frac{C_2}{R^4} n\epsilon^2\right)$

Jirak M (2011). On the maximum of covariance estimators. *Journal of Multivariate Analysis*.

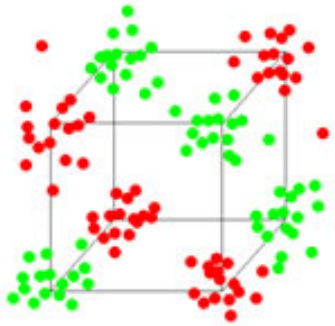
Statistical Properties

Define C as in Theorem 3 and $\widehat{\beta}^{(k)} := \underset{\beta \in \mathcal{B}, ||\beta||_0 \leq k}{\operatorname{argmax}} \{D(\beta^T \widehat{X}^{(n)}, \beta^T \widehat{Y}^{(n)})\}$

Theorem 4 (Sparsistency)

Suppose there exists feature subset $S \subset \{1, \ldots, d\}$ *s.t.* $|S| = k$, $T(X_S, Y_S) \geq h\left(g\left(\epsilon(d+1)/C\right)\right)$, *and remaining marginal distributions* X_{S^C} , Y_{S^C} *are identical. Then:* $\widehat{\beta}_i^{(k)} \neq 0$, $\widehat{\beta}_j^{(k)} = 0 \quad \forall\, i \in S, j \in S^C$ *with probability* $\geq 1 - C_1 \left(1 + \frac{R^2}{\epsilon}\right)^{d-k} \exp\left(-\frac{C_2}{R^4} n\epsilon^2\right)$

Feature Selection



- Two-class MADELON dataset from NIPS 2003 feature selection challenge
- $n = m = 1000, d = 500$
- 20 features with differences between groups
- 480 noise features with no difference
- Only differences present are in interactions between features (resembles parity problem)

Guyon I, Gunn S, Nikravesh M, Zadeh LA (2006). Feature Extraction: Foundations and Applications. *NIPS 2003 Feature Selection Challenge*

Feature Selection: Results

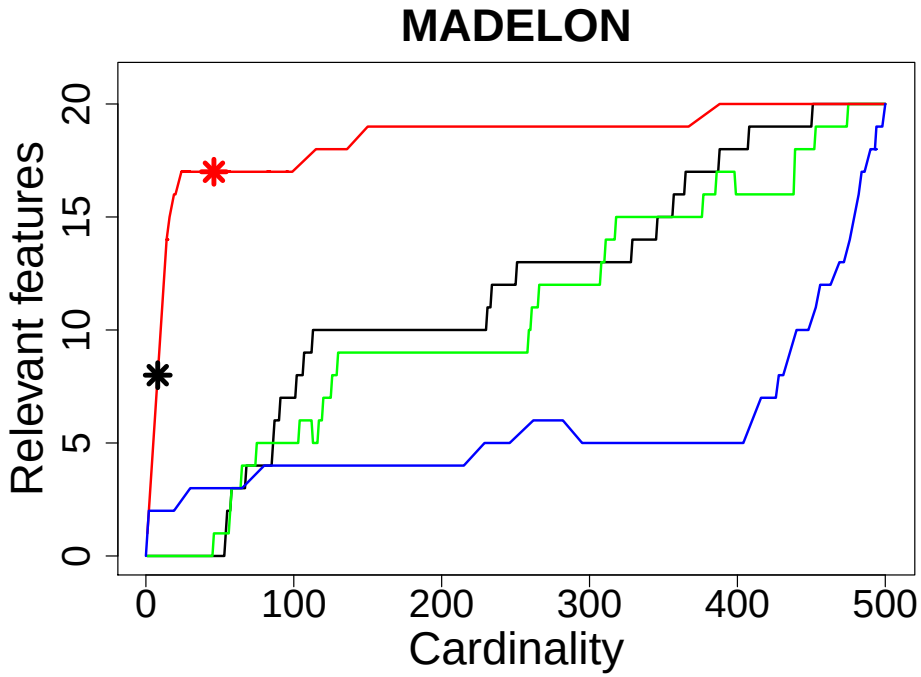


Figure: How well SPARDA (red), top sparse principal component (black), sparse linear discriminants analysis (green), and logistic lasso (blue) identify the 20 relevant features over different regularization settings.

$*$ = SPARDA result with λ chosen via cross-validation (46 features selected with 17 relevant). Many λ-settings return 14 relevant features with no false positives.

$*$ = best result (ARD kernel SVM) in the NIPS challenge (8 of the 20 relevant features selected).

Two-sample Testing

- Generated 20 datasets of varying dimensionality, only first 3 features in each dataset differ
- Centered multivariate Gaussian with covariances $\sim$ Wishart $(n = m = 1000)$

Two-sample Testing

- Generated 20 datasets of varying dimensionality, only first 3 features in each dataset differ
- Centered multivariate Gaussian with covariances $\sim$ Wishart $(n = m = 1000)$
- For PDA/SPARDA, test statistic $= D(\widehat{\beta}^T \widehat{X}^{(n)}, \widehat{\beta}^T \widehat{Y}^{(m)})$

Two-sample Testing

- Generated 20 datasets of varying dimensionality, only first 3 features in each dataset differ
- Centered multivariate Gaussian with covariances $\sim$ Wishart ($n = m = 1000$)
- For PDA/SPARDA, test statistic $= D(\widehat{\beta}^T \widehat{X}^{(n)}, \widehat{\beta}^T \widehat{Y}^{(m)})$
- Evaluate significance of all test statistics via permutation testing (exact Type I error control)

Two-sample Testing: Results

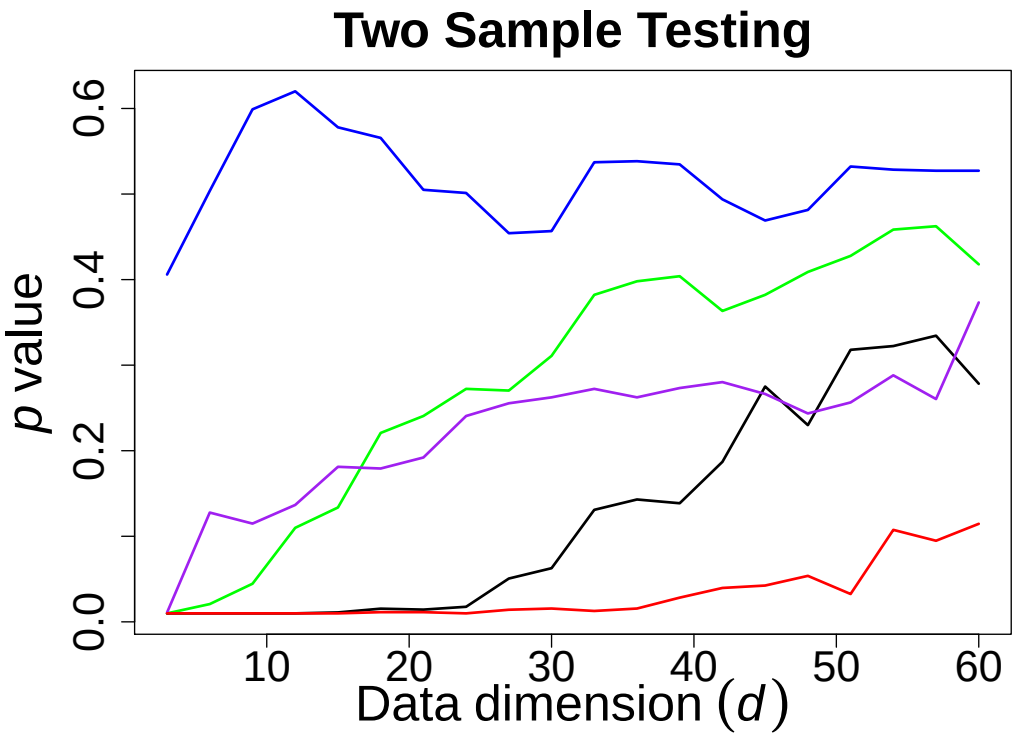


Figure: (10-repetition averaged) p-values produced by SPARDA (red), PDA (purple), overall Wasserstein distance in $\mathbb{R}^d$ (black), Maximum Mean Discrepancy[1] (green), and DiProPerm[2] (blue).

[1] Gretton A, Borgwardt KM, Rasch MJ, Scholkopf B, Smola A (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research*.

[2] Wei S, Lee C, Wichers L, Marron JS (2015). Direction-Projection-Permutation for High Dimensional Hypothesis Tests. *Journal of Computational and Graphical Statistics*.

Cellular gene expression in cortex vs. hippocampus

- From juvenile mice: 1,691 cells sampled from somatosensory cortex, 1,314 hippocampus cells (Zeisel, 2015)
- Expression of 10,305 genes measured within individual cells via single-cell RNA-seq (on comparable log-FPKM scale)

Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*.

Cellular gene expression in cortex vs. hippocampus

- From juvenile mice: 1,691 cells sampled from somatosensory cortex, 1,314 hippocampus cells (Zeisel, 2015)
- Expression of 10,305 genes measured within individual cells via single-cell RNA-seq (on comparable log-FPKM scale)
- Standard method to identify differentially expressed genes: assume expression distribution follows parametric family, assess statistical significance of marginal-mean-shifts (eg. Limma)

Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*.

Ritchie M, Phipson B, Wu D, Hu Y, Law CW, et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*.

Cellular gene expression in cortex vs. hippocampus

- From juvenile mice: 1,691 cells sampled from somatosensory cortex, 1,314 hippocampus cells (Zeisel, 2015)
- Expression of 10,305 genes measured within individual cells via single-cell RNA-seq (on comparable log-FPKM scale)
- Standard method to identify differentially expressed genes: assume expression distribution follows parametric family, assess statistical significance of marginal-mean-shifts (eg. Limma)
- Brain regions contain vast diversity of cell subtypes
(mean differences unsatisfactory, there is no "average" cell)
- Single-cell RNA-seq data is highly noisy and does not follow nice parametric distribution

Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*.

Ritchie M, Phipson B, Wu D, Hu Y, Law CW, et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*.

Single-cell RNA-seq differential expression analysis

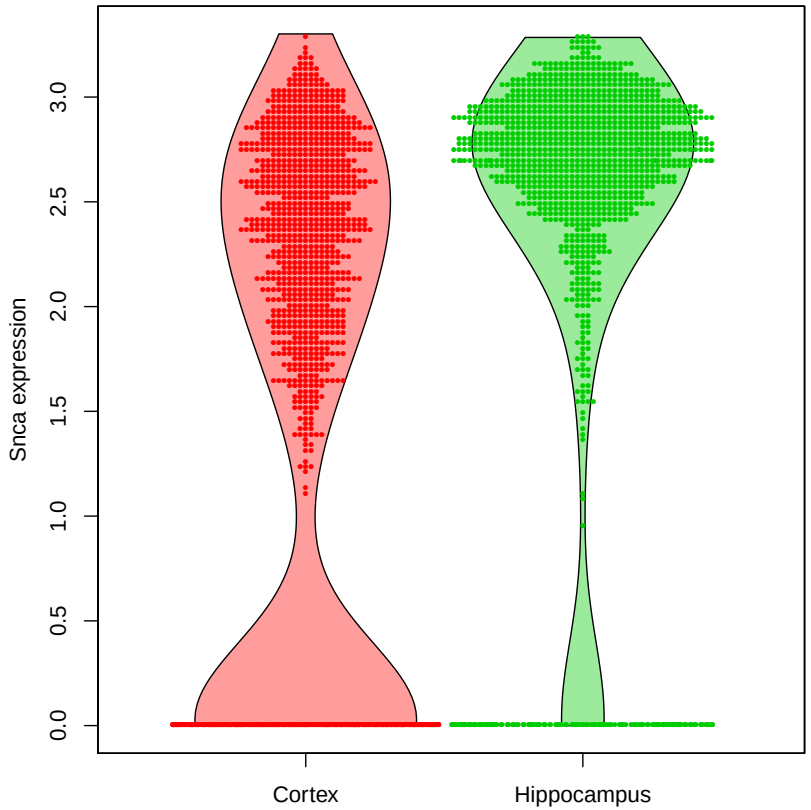
GENE	WEIGHT	DESCRIPTION
Cck	0.0593	Primary distinguishing gene between distinct interneuron classes identified in the cortex and hippocampus
Neurod6	0.0583	General regulator of nervous system development whose induced mutation displays different effects in neocortex vs. the hippocampal region
Stmn3	0.0573	Up-expressed in hippocampus of patients with depressive disorders
Plp1	0.0570	An oligodendrocyte- and myelin-related gene which exhibits cortical differential expression in schizophrenia
Crym	0.0550	Plays a role in neuronal specification
Spink8	0.0536	Serine protease inhibitor specific to hippocampal pyramidal cells
Gap43	0.0511	Encodes plasticity protein important for axonal regeneration and neural growth
Cryab	0.0500	Stress induction leads to reduced expression in the mouse hippocampus
Mal	0.0494	Regulates dendritic morphology and is expressed at lower levels in cortex than in hippocampus
Tspan13	0.0488	Membrane protein which mediates signal transduction events in cell development, activation, growth and motility

Table: Genes with the greatest weight in the projection $\widehat{\beta}$ produced by SPARDA

- *Crym*, *Spink8*, *Neurod6* are also among the top 10 genes identified by LIMMA

Snca

- Presynaptic signaling and membrane trafficking gene whose defects are implicated in both Parkinson and Alzheimer's disease
- Ranks 11th highest in SPARDA analysis, but only 349 by LIMMA differential expression analysis



Marginally-normalized differential expression analysis

- Data from both populations is marginally centered at zero with unit variance (per-gene basis)
- Only major remaining differences are changes in gene-gene relationships between cortex and hippocampus

GENE	WEIGHT	DESCRIPTION
Thy1	0.1245	Plays a role in cell-cell & cell-ligand interactions during synaptogenesis and other processes in the brain
Vsnl1	0.1245	Modulates intracellular signaling pathways of the central nervous system
Stmn3	0.1222	Stathmins form important protein complex with tubulins
Stmn2	0.1188	Note: Tubulins Tubb3 and Tubb2 are ranked 20^{th} and 25^{th} by weight in $\widehat{\beta}$
Tmem59	0.1176	Fundamental regulator of neural cell differentiation. Knock out in the hippocampus results in drastic expression changes of many other genes
Basp1	0.1171	Transcriptional cofactor which can divert the differentiation of cells to a neuronal-like morphology
Snhg1	0.1166	Unclassified non-coding RNA gene
Mllt11	0.1145	Promoter of neurodifferentiation and axonal/dendritic maintenance
Uchl1	0.1137	Loss of function leads to profound degeneration of motor neurons
Cck	0.1131	Targets pyramidal neurons and enables neocortical plasticity allowing for example the auditory cortex to detect light stimuli

Table: Genes with the greatest weight in $\widehat{\beta}$ produced by SPARDA analysis of marginally normalized data

PCA $\rightarrow$ PDA

- Consider setting with *paired* samples $(x^{(i)}, y^{(i)})$

PCA $\rightarrow$ PDA

- Consider setting with *paired* samples $(x^{(i)}, y^{(i)})$
- $\widehat{\beta}_{\mathsf{PCA}} :=$ largest principal component of (uncentered) differences $x^{(i)} - y^{(i)}$

PCA $\rightarrow$ PDA

- Consider setting with *paired* samples $(x^{(i)}, y^{(i)})$
- $\widehat{\beta}_{\text{PCA}} :=$ largest principal component of (uncentered) differences $x^{(i)} - y^{(i)}$
- $\widehat{\beta}_{\text{PDA}} :=$ direction which maximizes projected Wasserstein difference between empirical distribution of $X - Y$ and delta distribution at 0.

PCA $\rightarrow$ PDA

- Consider setting with *paired* samples $(x^{(i)}, y^{(i)})$
- $\widehat{\beta}_{\text{PCA}} :=$ largest principal component of (uncentered) differences $x^{(i)} - y^{(i)}$
- $\widehat{\beta}_{\text{PDA}} :=$ direction which maximizes projected Wasserstein difference between empirical distribution of $X - Y$ and delta distribution at 0.

Fact

$$\widehat{\beta}_{PCA} \equiv \widehat{\beta}_{PDA}$$

Future Work

- Develop structural assumptions to leverage underlying sparsity in differences and improve exponential bounds on convergence in high-dimensional settings (eg. spiked covariance, restricted isometry).

Future Work

- Develop structural assumptions to leverage underlying sparsity in differences and improve exponential bounds on convergence in high-dimensional settings (eg. spiked covariance, restricted isometry).
- Confidence intervals for projection weights $\widehat{\beta}_\ell$ (beyond bootstrap)

Future Work

- Develop structural assumptions to leverage underlying sparsity in differences and improve exponential bounds on convergence in high-dimensional settings (eg. spiked covariance, restricted isometry).
- Confidence intervals for projection weights $\widehat{\beta}_\ell$ (beyond bootstrap)
- Employ multiple successive projections (eg. maximum-entropy)

Future Work

- Develop structural assumptions to leverage underlying sparsity in differences and improve exponential bounds on convergence in high-dimensional settings (eg. spiked covariance, restricted isometry).
- Confidence intervals for projection weights $\widehat{\beta}_\ell$ (beyond bootstrap)
- Employ multiple successive projections (eg. maximum-entropy)
- Adapt approach to non-pairwise comparison of multiple populations

Thanks!

Questions?

Paper: Mueller J, Jaakkola T. Principal Differences Analysis: Interpretable Characterization of Differences between Distributions. *NIPS* 2015.

Code: `http://www.mit.edu/~jonasm/`