Biometrics 0, 1–17 DOI: 000
Feb 2020

Supporting Information for Sparse Linear Discriminant Analysis for Multi-view Structured Data

Sandra E. Safo[1]*, Eun Jeong Min[2], and Lillian Haine[1]

[1]Division of Biostatistics, University of Minnesota, Minneapolis, MN

[2]Department of Medical Life Sciences, College of Medicine

The Catholic University of Korea, Seoul, Republic of Korea

**email:* ssafo@umn.edu (corresponding author), ej.min@catholic.ac.kr, haine108@umn.edu

This paper has been submitted for consideration for publication in *Biometrics*

1. Unique features of proposed method

Table 1 highlights the unique features of our proposed methods compared to existing works.

[Table 11.2 here]

2. Proof of Theorem 1

The Lagrangian

$$\begin{aligned} L(\mathbf{A},\mathbf{B},\lambda_1,\lambda_2) &= \rho\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{S}_b^1\mathbf{A}+\mathbf{B}^{\mathrm{T}}\mathbf{S}_b^2\mathbf{B})+(1-\rho)\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{S}_{12}\mathbf{B}\mathbf{B}^{\mathrm{T}}\mathbf{S}_{12}^{\mathrm{T}}\ \mathbf{A}) \\ &- \lambda_1(\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{S}_w^1\mathbf{A})-(K-1))-\lambda_2(\mathrm{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{S}_w\mathbf{B})-(K-1)) \end{aligned}$$

Let $\mathbf{\Omega}^1=\mathbf{S}_{12}\mathbf{B}\mathbf{B}^{\mathrm{T}}\mathbf{S}_{12}^{\mathrm{T}}$ and $\mathbf{\Omega}^2=\mathbf{S}_{12}^{\mathrm{T}}\mathbf{A}\mathbf{A}^{\mathrm{T}}\mathbf{S}_{12}$.

The first order stationary solutions for $\mathbf{A}$ and $\mathbf{B}$ are

$$\begin{aligned} \frac{\partial L(\mathbf{A},\mathbf{B},\lambda_1,\lambda_2)}{\partial\mathbf{A}} &= \rho(\mathbf{S}_b^1+\mathbf{S}_b^{1^{\mathrm{T}}})\mathbf{A}+(1-\rho)(\mathbf{\Omega}^1+\mathbf{\Omega}^{1^{\mathrm{T}}})\mathbf{A}-\lambda_1(\mathbf{S}_w^1+\mathbf{S}_w^{1^{\mathrm{T}}})\mathbf{A}=\mathbf{0} \\ \frac{\partial L(\mathbf{A},\mathbf{B},\lambda_1,\lambda_2)}{\partial\mathbf{B}} &= \rho(\mathbf{S}_b^2+\mathbf{S}_b^{2^{\mathrm{T}}})\mathbf{B}+(1-\rho)(\mathbf{\Omega}^2+\mathbf{\Omega}^{2^{\mathrm{T}}})\mathbf{B}-\lambda_1(\mathbf{S}_w^2+\mathbf{S}_w^{2^{\mathrm{T}}})\mathbf{B}=\mathbf{0} \end{aligned}$$

Rearranging, we obtain the eigensystems for $\mathbf{A}$ and $\mathbf{B}$ respectively as

$$\begin{aligned} \Big(\rho(\mathbf{S}_b^1+\mathbf{S}_b^{1^{\mathrm{T}}})+(1-\rho)(\mathbf{\Omega}^1+\mathbf{\Omega}^{1^{\mathrm{T}}})\Big)\mathbf{A} &= \lambda_1(\mathbf{S}_w^1+\mathbf{S}_w^{1^{\mathrm{T}}})\mathbf{A} \qquad (1) \\ \Big(\rho(\mathbf{S}_b^2+\mathbf{S}_b^{2^{\mathrm{T}}})+(1-\rho)(\mathbf{\Omega}^2+\mathbf{\Omega}^{2^{\mathrm{T}}})\Big)\mathbf{B} &= \lambda_2(\mathbf{S}_w^2+\mathbf{S}_w^{2^{\mathrm{T}}})\mathbf{B} \qquad (2) \end{aligned}$$

For $\mathbf{B}$ fixed in $\mathbf{\Omega}^1$, equation (1) can be solved for the nonzero eigenvalues of $(\mathbf{S}_w^1+\mathbf{S}_w^{1^{\mathrm{T}}})^{-1}(\rho(\mathbf{S}_b^1+\mathbf{S}_b^{1^{\mathrm{T}}})+(1-\rho)(\mathbf{\Omega}^1+\mathbf{\Omega}^{1^{\mathrm{T}}}))$. Denote the corresponding eigenvectors as $\widetilde{\mathbf{A}}=[\tilde{\boldsymbol{\alpha}}_1,\ldots,\tilde{\boldsymbol{\alpha}}_r]$. Similarly, with $\mathbf{A}$ fixed in $\mathbf{\Omega}^2$, we can solve for the nonzero eigenvalues in equation (2) from $(\mathbf{S}_w^2+\mathbf{S}_w^{2^{\mathrm{T}}})^{-1}(\rho(\mathbf{S}_b^2+\mathbf{S}_b^{2^{\mathrm{T}}})+(1-\rho)(\mathbf{\Omega}^2+\mathbf{\Omega}^{2^{\mathrm{T}}}))$. Let $\widetilde{\mathbf{B}}=[\tilde{\boldsymbol{\beta}}_1,\ldots,\tilde{\boldsymbol{\beta}}_r]$. We iterate over $\mathbf{A}$ and $\mathbf{B}$

in equations (1) and (2) until convergence (both $\|\widetilde{\mathbf{A}}_{new} - \widetilde{\mathbf{A}}_{old}\|_F < \epsilon$ and $\|\widetilde{\mathbf{B}}_{new} - \widetilde{\mathbf{B}}_{old}\|_F < \epsilon$). At which point we set $\widehat{\mathbf{A}} = \widetilde{\mathbf{A}}$ and $\widehat{\mathbf{B}} = \widetilde{\mathbf{B}}$.

Remark 1: In high-dimensional examples where $p > n$, we make $\mathbf{S}_w^1$ and $\mathbf{S}_w^2$ positive definite by adding a small multiple of the identity. We could estimate $\mathbf{S}_w^1$ and $\mathbf{S}_w^2$ using techniques proposed in Cai et al. (2011) and Bickel and Levina (2008) but that would add a layer of complexity. To reduce computations, we use techniques described in Hastie and Tibshirani (2004) to avoid inverting the $p \times p$ (or $q \times q$) matrices $\mathbf{S}_w^{1^{1/2}}$ and $\mathbf{S}_w^{2^{1/2}}$; instead, we invert a $n \times n$ matrix, and $n \ll p$ (or q).

3. More comments on association component of proposed method

Note that the cross-covariance matrix between $\mathbf{X}^1$ and $\mathbf{X}^2$, (i.e., $\mathbf{S}_{12}$) can be decomposed as $\mathbf{S}_{12} = \mathbf{S}_w^{12} + \mathbf{S}_b^{12}$, where

$$\mathbf{S}_w^{12} = \sum_{k=1}^{K}\sum_{i=1}^{n}(\mathbf{x}_{ik}^1 - \hat{\boldsymbol{\mu}}_k^1)(\mathbf{x}_{ik}^2 - \hat{\boldsymbol{\mu}}_k^2)^{\mathrm{T}}; \qquad \mathbf{S}_b^{12} = \sum_{k=1}^{K} n_k(\hat{\boldsymbol{\mu}}_k^1 - \hat{\boldsymbol{\mu}}^1)(\hat{\boldsymbol{\mu}}_k^2 - \hat{\boldsymbol{\mu}}^2)^{\mathrm{T}}.$$

Here, $\hat{\boldsymbol{\mu}}_k^j = (1/n_k)\sum_{i=1}^{n_k} \mathbf{x}_{ik}^j, j = 1, 2$ and $\hat{\boldsymbol{\mu}}^j$ is the combined class mean vector for View j, $j = 1, 2$, and is defined as $\hat{\boldsymbol{\mu}}^j = (1/n) \sum\limits_{k=1}^{K} n_k \hat{\boldsymbol{\mu}}_k^j$. In keeping with terminology in LDA, we term $\mathbf{S}_w^{12}$ as *within-class cross-covariance* and $\mathbf{S}_b^{12}$ as *between-class cross-covariance*. $\mathbf{S}_b^{12}$ measures the cross-covariance within the classes, and $\mathbf{S}_b^{12}$ measures the cross-covariance between the classes. So in using $\mathbf{S}_{12}$ in our method, we are capturing the cross-covariances between and within the classes.

Similar to LDA where we maximize separation between classes while minimizing variation within the classes for one data type, we also explored a different formulation of the association part of our optimization problem:

$$\begin{aligned} \max_{\mathbf{A},\mathbf{B}} \rho \overbrace{\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{S}_b^1\mathbf{A} + \mathbf{B}^{\mathrm{T}}\mathbf{S}_b^2\mathbf{B})}^{\text{separation}} \quad &+ \quad (1-\rho)\,\overbrace{\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{S}_b^{12}\mathbf{B}\mathbf{B}^{\mathrm{T}}\mathbf{S}_b^{12^{\mathrm{T}}}\mathbf{A})}^{\text{association}} \\ \text{subject to } \ \mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{S}_w^1\mathbf{A})/(K-1) \quad &= \quad 1, \ \ \mathrm{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{S}_w\mathbf{B})/(K-1) = 1. \end{aligned} \tag{3}$$

That is, we maximize simultaneously separation for each data type and the square of the between-class cross-covariance (instead of the cross-covariance $\mathbf{S}_{12}$). Table 17 gives simulation results for Scenario four in Section 10.4. We did observe an improvement in classification accuracy when we used $\mathbf{S}_b^{12}$ but the estimated RV coefficients where similar or lower when compared to estimates from $\mathbf{S}_{12}$.

4. Rank Determination

In the classical LDA problem, the rank (maximum number of eigenvalues) is $K-1$, where K is the number of classes. This coincides with $\text{rank}(\mathbf{S}_b^1)$ (or $\text{rank}(\mathbf{S}_b^2)$). For a fixed $\mathbf{B}^*$,

$$
\begin{aligned}
&\mathit{rank}\left((\mathbf{S}_w^1+\mathbf{S}_w^{1^{\mathrm{T}}})^{-1}(\mathbf{S}_b^1+\mathbf{S}_b^{1^{\mathrm{T}}}+\boldsymbol{\Omega}^1+\boldsymbol{\Omega}^{1^{\mathrm{T}}})\right)\\
&\qquad\qquad\leqslant\ K-1+\min\left(\text{rank}(\mathbf{S}_w^{1^{-1}}),\text{rank}(\mathbf{S}_{12}),\text{rank}(\mathbf{B})\right).
\end{aligned}
$$

This suggests that for the integrative LDA problem, there could be more than $K-1$ eigenvalue-eigenvector pairs. In practice, one could use a scree-plot to choose the rank. However, in our simulations and real data analyses, we find that the first $K-1$ eigenvalues dominate the rest of the eigenvalues. Thus, we set the maximum number of eigenvalues to be $K-1$, similar to the classical LDA.

5. More Comments on Laplacian

One could use the Laplacian (not normalized) defined as:

$$
\mathcal{L}(u,v) \;=\; \begin{cases} r_v - w(u,v) & \text{if } \; u=v \\ -w(u,v) & \text{if } u\neq v \text{ and } \; \text{variables } u \; \text{ and } \; v \; \text{ are adjacent} \\ 0 & \;\text{otherwise} \end{cases} \tag{4}
$$

instead of the normalized Laplacian defined in equation (10) in the main text. However, the Laplacian in equation (4) encourages variables in the network to have the same coefficients. This is true since $w(u,v)$ is the same for variables that are connected. We believe variables that are connected will often have different coefficients that capture their contributions to overall dependency structure and class separation. As such, we use the normalized Laplacian, which normalizes the connected variables by their degrees, thus encouraging different coefficients.

6. Extension to multiple views of data

We extend the proposed methods to more than two views of data. Let $\mathbf{X}^d = [\mathbf{X}_1^d, \mathbf{X}_2^d, \ldots, \mathbf{X}_K^d]$, $\mathbf{X}^d \in \Re^{n \times p_d}, \mathbf{X}_k^d \in \Re^{n_k \times p_d}, k = 1, \ldots, K,\ d = 1, 2, \ldots, D$ be a concatenation of the K classes in the d-th view. Let $\mathbf{S}_b^d$ and $\mathbf{S}_w^d$ be the between-class and within-class covariances for the d-th view. Let $\mathbf{S}_{dj}, j < d$ be the cross-covariance between the d-th and j-th views. Define $\mathcal{M}^d = \mathbf{S}_w^{d^{-1/2}} \mathbf{S}_b^d \mathbf{S}_w^{d^{-1/2}}$ and $\mathcal{N}_{dj} = \mathbf{S}_w^{d^{-1/2}} \mathbf{S}_{dj} \mathbf{S}_w^{j^{-1/2}}$. We solve the optimization problem for multiple views of data:

$$\max_{\mathbf{\Gamma}^1, \cdots, \mathbf{\Gamma}^D} \rho \sum_{d=1}^{D} \mathrm{tr}(\mathbf{\Gamma}^{d\mathrm{T}} \mathcal{M}^d \mathbf{\Gamma}^d) + \frac{2(1-\rho)}{D(D-1)} \sum_{d=1, d \neq j}^{D} \mathrm{tr}(\mathbf{\Gamma}^{d\mathrm{T}} \mathcal{N}_{dj} \mathbf{\Gamma}^j \mathbf{\Gamma}^{j\mathrm{T}} \mathcal{N}_{jd} \mathbf{\Gamma}^d) \text{ s.t } \mathrm{tr}(\mathbf{\Gamma}^{d\mathrm{T}} \mathbf{\Gamma}^d) = K - 1.$$

As before, ρ controls the influence of separation or association in the optimization problem. The second term essentially sums all of these unique pairwise squared correlations and weight them by $\frac{D(D-1)}{2}$ so that the sum of the squared correlations is one. As in Proposition 1 in the main text, the nonsparse basis discriminant directions for the d-th view, $\widetilde{\mathbf{\Gamma}}^d$, are given by the eigenvectors corresponding to the eigenvalues that iteratively solve the following eigensystems:

$$\begin{aligned} \left(c_1 \mathcal{M}^1 + c_1 \mathcal{M}^{1^\mathrm{T}} + c_2 \overline{\mathcal{N}}_{1j} + c_2 \overline{\mathcal{N}}_{1j}^\mathrm{T}\right) \mathbf{\Gamma}^1 &= \mathbf{\Lambda}_1 \mathbf{\Gamma}^1, \\ &\vdots \\ \left(c_1 \mathcal{M}^D + c_1 \mathcal{M}^{1^\mathrm{T}} + c_2 \overline{\mathcal{N}}_{Dj} + c_2 \overline{\mathcal{N}}_{Dj}^\mathrm{T}\right) \mathbf{\Gamma}^D &= \mathbf{\Lambda}_D \mathbf{\Gamma}^D, \end{aligned} \quad (5)$$

where we set $c_1 = \rho$ and $c_2 = \frac{2(1-\rho)}{D(D-1)}$, and $\overline{\mathcal{N}}_{dj} = \sum_{d,j}^{D} \mathcal{N}_{dj} \mathbf{\Gamma}^j \mathbf{\Gamma}^{j^\mathrm{T}} \mathcal{N}_{jd},\ d, j = 1, \ldots D, j \neq d$ sums all unique pairwise correlations of the d-th and the j-th views. For sparsity or smoothness we solve the following optimization problems:

$$
\begin{aligned}
&\min_{\mathbf{\Gamma}^1} \mathcal{P}(\mathbf{\Gamma}^1) \qquad \text{s.t} \qquad \|(c_1\mathcal{M}^1 + c_1\mathcal{M}^{1^{\mathrm{T}}} + c_2\overline{\mathcal{N}}_{1j} + c_2\overline{\mathcal{N}}_{1j}^{\mathrm{T}})\widetilde{\mathbf{\Gamma}}^1 - \widetilde{\mathbf{\Lambda}}_1\mathbf{\Gamma}^1\|_\infty \leqslant \tau_1 \\
&\qquad\qquad\vdots \\
&\min_{\mathbf{\Gamma}^D} \mathcal{P}(\mathbf{\Gamma}^D) \qquad \text{s.t} \qquad \|(c_1\mathcal{M}^D + c_1\mathcal{M}^{D^{\mathrm{T}}} + c_2\overline{\mathcal{N}}_{Dj} + c_2\overline{\mathcal{N}}_{Dj}^{\mathrm{T}})\widetilde{\mathbf{\Gamma}}^D - \widetilde{\mathbf{\Lambda}}_D\mathbf{\Gamma}^D\|_\infty \leqslant \tau_D. \qquad (6)
\end{aligned}
$$

The penalty term $\mathcal{P}(\mathbf{\Gamma}^d)$ is either set respectively to equations (6) or (9) in the main text, depending on whether sparsity or smoothness (with sparsity) is desired.

7. More Comments on initialization, tuning parameters, and algorithm

The optimization problems in equations (8) and (12) in the main text are biconvex. With $\mathbf{\Gamma}^d$ fixed at $\mathbf{\Gamma}^{d^*}$, the problem of solving for $\widehat{\mathbf{\Gamma}}^j$, $j \neq d$ is convex, and may be solved easily with any-off-the shelf convex optimization software. The technique of solving biconvex problems by fixing parameters and then solving the resulting convex problems is popularly used in the statistical literature. Since the problem is biconvex, alternating minimization does not guarantee a global solution, but instead a solution where the cost function (objective) at previous and current iterations is within a specified tolerance value. At the first iteration, we fix $\mathbf{\Gamma}^{d^*}$ as the classical LDA solution from applying LDA on $\mathbf{X}^d$. We can initiate $\mathbf{\Gamma}^{d^*}$ with random orthonormal matrices, but we choose to initialize with regular LDA solutions because the algorithm converges faster. At subsequent solutions, we fix $\mathbf{\Gamma}^{d^*}$ as the solution from previous iteration, and iterate until convergence. We use the following criteria for convergence, which ever occurs first: (i)max $\left(\frac{\|\mathbf{\Gamma}^1_{new}-\mathbf{\Gamma}^1_{old}\|_F^2}{\|\mathbf{\Gamma}^1_{old}\|_F^2}, \ldots, \frac{\|\mathbf{\Gamma}^D_{new}-\mathbf{\Gamma}^D_{old}\|_F^2}{\|\mathbf{\Gamma}^D_{old}\|_F^2}\right)$, (ii) $\frac{\sum_{d=1}^D \|\mathbf{\Gamma}^d_{new}-\mathbf{\Gamma}^d_{old}\|_F^2\|}{\sum_{d=1}^D \|\mathbf{\Gamma}^d_{old}\|_F^2}$. Here, $\mathbf{\Gamma}^d_{new}$ and $\mathbf{\Gamma}^d_{old}$, $d = 1, \ldots, D$, are the current and old previous iteration solutions. In our simulations and real data applications, we observed that the algorithm usually converged within $4 \sim 7$ iterations. Algorithm 1 gives an outline of our proposed methods.

The optimization problems depend on tuning parameters τ_d, which need to be chosen. We

```
Input: training data $(\mathbf{X}^d, \mathbf{y})$; tuning parameters $\tau_d, d = 1, \dots, D$; edge matrix, $E^d$
  and edge weight, $W^d$ (for SIDANet)
                                        ▷ $\tau_D = 0$ if covariates ($D$-th view) available
Output: estimated sparse discriminant vectors $\widehat{\mathbf{\Gamma}}^d$.
Initialize: $\mathbf{\Gamma}^d$, $d = 1, \dots, D$ .
                     ▷ Use random orthonormal matrices or solution from classical LDA
repeat
    for $d = 1, \dots, D$ do
        Fix $\widetilde{\mathbf{\Gamma}}^d$ and $\widetilde{\mathbf{\Lambda}}_d$.
                     ▷ Use solutions from generalized eigenvalue systems (equation 5)

        Solve
             $\min_{\mathbf{\Gamma}^d} \mathcal{P}(\mathbf{\Gamma}^d)$   s.t   $\|(c_1\mathcal{M}^d + c_1{\mathcal{M}^d}^{\mathrm{T}} + c_2\bar{\mathcal{N}}_{dj} + c_2\bar{\mathcal{N}}_{dj}^{\mathrm{T}})\widetilde{\mathbf{\Gamma}}^d - \widetilde{\mathbf{\Lambda}}_d\mathbf{\Gamma}^d\|_\infty \leqslant \tau_d$
             ▷ $\mathcal{P}(\mathbf{\Gamma}^d)$ is defined in equation (6) for SIDA and (9) for SIDANet in main text.
    end
until convergence
```

Algorithm 1: Algorithm for obtaining sparse (and network-constrained) integrative discriminant vectors for multi-view data.

fix $w = 0.5$ to provide equal weight on separation and association. Without loss of generality, assume the D-th (last) view is the covariates, if available. We fix $\tau_D = 0$ and select the optimal tuning parameters for the other views from a range of tuning parameters. Note that searching the tuning parameters hyperspace can be computationally intensive. For instance, if there are two views (excluding covariates) each having 10 grid points, then one needs to search a 10×10 grid space, representing 100 grid values to choose the optimal combination. For $d = 1, .., D-1$, we need to search a large hyperparameter space $[(G_1 \times G_2 \times \cdots \times G_{D-1})$ grid values assuming G_d is the number of grid points for the d-th view]. This obviously is computationally taxing. To overcome this computational bottleneck, we follow ideas in Bergstra and Bengio (2012) and randomly select some grid points (from the entire grid space) to search for the optimal tuning parameters; we term this approach *random search*. This technique has been shown to yield good results (Bergstra and Bengio, 2012) when compared to searching the entire space (*grid search*). In fact, our own simulations with *random search* produced satisfactory results when compared to *grid search*. A detailed comparison of *random search* and *grid search* in terms of error rates, estimated correlations, variables selected, and computational time is found here and in the main text.

We provide upper and lower bounds for τ_d. Let $d = 1$. Note that $\tau_1 > \|(c_1\mathcal{M}^1 + c_1\mathcal{M}^{1^{\mathrm{T}}} + c_2\bar{\mathcal{N}}_{1j} + c_2\bar{\mathcal{N}}_{1j}^{\mathrm{T}})\|_\infty$ results in trivial solution vectors, i.e., $\widehat{\mathbf{\Gamma}}^1 = \mathbf{0}$. Hence, we set the upper bound for τ_1 as $\tau_{1\,\max} = \|(c_1\mathcal{M}^1 + c_1\mathcal{M}^{1^{\mathrm{T}}} + c_2\bar{\mathcal{N}}_{1j} + c_2\bar{\mathcal{N}}_{1j}^{\mathrm{T}})\|_\infty$. Similar results hold for the other views. Instead of using a lower bound of 0, we use a lower bound dependent on the dimensions of each view (specifically $\tau_{d\,\min} = (\sqrt{\log p^d/n}) \cdot \tau_{d\,\max}$) to encourage sparsity. We choose the optimal tuning parameters from the range of tuning parameters using K-fold cross validation ($K = 5$ in our simulations and real data applications) to minimize average classification error.

An ideal situation is to select the τ and η that result in sparse estimates leading to optimal separation and association given a fixed data. But overlaying the η selection on top of that of τ will result in another layer of complexity. Our simulations comparing RV coefficients using true and estimated discriminant vectors show that the optimal tuning parameters τ (with $\eta = 0.5$) are been selected to produce reasonable association or correlation between views. Please refer to Table 2 for these results.

[Table 2 here]

Our classification approach is found in Section 8.

8. Using SIDA and SIDANet for classification

Once the SIDA or SIDANet discriminant functions have been obtained, one can make future class assignments by either 1) pooling the discriminant scores for each view $\mathbf{X}^d, d = 1 \ldots D$, or 2) using individual discriminant scores from each view. The latter option, which we term separate class assignment, is appealing if for some reasons some of the views are not available for future observations. In such instances, future class assignments can be carried out using the discriminant functions for available views. In either the pooled or separate class assignments, we use nearest centroid for classification.

The discriminant scores are defined to be $\mathbf{U}^d = \mathbf{X}^d\widehat{\mathbf{\Gamma}}^d, d = 1, ..., D$, where $\widetilde{\mathbf{\Gamma}}^d$ is a $p^d \times (K-1)$ matrix of basis vectors obtained from SIDA or SIDANet. Let $\mathbf{z}^d = (z_1^d, ..., z_p^d)^{\mathrm{T}}$ be the available measurement for a new (future) observation for the d-th view. Consider projecting these future observations onto the estimated discriminant vectors $\widehat{\mathbf{\Gamma}}^d$ for the d-th view (i.e., $\mathbf{v}^d = \mathbf{z}^{d\mathrm{T}}\widehat{\mathbf{\Gamma}}^d$) and concatenating the scores for all d views; i.e $\mathbf{v} = [\mathbf{z}^{1\mathrm{T}}\widehat{\mathbf{\Gamma}}^1, \mathbf{z}^{2\mathrm{T}}\widehat{\mathbf{\Gamma}}^2, \cdots, \mathbf{z}^{D^{\mathrm{T}}}\widehat{\mathbf{\Gamma}}^D]^{\mathrm{T}} \in \Re^{D(K-1)}$. For pooled class assignment, we assign $\mathbf{z} = [\mathbf{z}^1, \cdots, \mathbf{z}^D]$ to class k if the distance from $\mathbf{v}$ to $\hat{\mathbf{u}}_k$ is minimum, that is,

$$\min_k \|\mathbf{v} - \hat{\mathbf{u}}_k\|_2, \quad k = 1, ..., K$$

where $\hat{\mathbf{u}}_k^{\mathrm{T}} \in \Re^{D(K-1)}$ is the pooled mean for class k obtained from the pooled scores $\mathbf{U} = [\mathbf{U}^1, \cdots, \mathbf{U}^D] \in \Re^{D(K-1)}$. For separate class assignments, we assign $\mathbf{z}^d$ to the population whose class mean is closest to $\mathbf{v}^d$, i.e.,

$$\min_k \|\mathbf{v}^d - \hat{\mathbf{u}}_k^d\|_2, \quad k = 1, ..., K, d = 1, \cdots, D$$

9. Time Comparisons

We compare the run times of *random* and *grid search*. We consider a $K = 3$ class and $D = 2$ views problem and simulate data according to Scenario One in the main text when no prior information exists. In *grid search*, we choose tuning parameters over a 8×8 grid (or 64 grid points). *Random search* randomly selects 15% of the grid points to optimize. We compare run times for $N < p$ and $N > p$, and when the cross validation task for choosing optimal tuning parameters is executed in parallel (using 4 workers) or not. All comparisons are carried out with the Matlab codes for SIDA on an Intel (R) Core (TM) i7-7700 3.60 GHz processor. Table 3 gives timings in minutes averaged over three runs. We observe that *random search* is considerably faster than *grid search*. SIDA with *random search*, with or without parallelization is faster than JACA especially when $N < p$.

[Table 3 here]

10. Simulation Results

In our simulations and real data applications, for two views (excluding covariates), we set 8 grid points each, and randomly select 20% of the grid values in the hyperparameter space to optimize. For $d > 2$, we set the number of grid points to 5, and randomly select 15% of the grid values in the hyperparameter space to optimize. Figure 1 is a visual representation of random data projected onto the true integrative discriminant vectors for different combinations of c, p_1 and p_2. In the top panel, $(\rho_1 = 0.9, \rho_2 = 0.7, c = 0.5)$. In the middle panel, $(\rho_1 = 0.4, \rho_2 = 0.2, c = 0.2)$. In the bottom panel, $(\rho_1 = 0.15, \rho_2 = 0.05, c = 0.12)$.

[Figure 1 about here.]

10.1 *Scenario One, degree of separation varies between the two views*

We consider a setting similar to Setting 1 in Scenario One, but we allow the degree of separation to vary between the two views. Here, the first column of $\mathbf{A}^1 \in \Re^{p\times 2}$ is set to $(c_1\mathbf{1}_{10}, \mathbf{0}_{p-10})$; the second column is set to $(\mathbf{0}_{10}, -c_1\mathbf{1}_{10}, \mathbf{0}_{p-20})$. The first column of $\mathbf{A}^2 \in \Re^{p\times 2}$ is set to $(c_2\mathbf{1}_{10}, \mathbf{0}_{p-10})$; the second column is set to $(\mathbf{0}_{10}, -c_2\mathbf{1}_{10}, \mathbf{0}_{p-20})$. We fix $c_1 = 0.7$ and

$c_2 = 0.2, 0.8$.

[Table 4 here]

10.2 *Scenario Two (Binary class, equal covariance within class):*

We consider a $D = 2$ high-dimensional and $K = 2$ class problem. The covariance matrices for each class follow Scenario One. The mean matrices follow Scenario One in the main text but with this exception: $\mathbf{A}^1 \in \Re^p$ is set to $(c\mathbf{1}_{20}, \mathbf{0}_{p-20})$. $\mathbf{A}^2$ is defined similarly. As before, we vary c to assess separation between the two classes.

[Table 5 here]

10.3 *Scenario Three (Multi-class, unequal covariance within class)*

In Scenario One (main text) we considered an example where the LDA assumption holds, i.e., the within-class covariance is the same for each class. In this setting, we relax this assumption. The covariance matrices for the three classes within $\mathbf{X}^1$ and $\mathbf{X}^2$ are each given as follows: for class 1, the covariance matrix has the same form as in Model 1; for class 2, the covariance matrix has entries $\sigma_{ij} = 0.6^{|i-j|}$; for class 3, the covariance matrix is the identity matrix, $\mathbf{I}_{(p \text{ or } q)}$. Table 6 show results of our method compared to the other methods considered.

[Table 6 here]

10.4 *Scenario Four (Binary class with covariates, degree of separation varies)*

We consider a simulation scenario with covariates and two other views where the classes are well separated in the covariates data. We assess the performance of the methods when we integrate the two views, and when we integrate all three views (including covariates). For the proposed method, SIDA, we also consider the situation where we regularize the coefficients of the covariates (i.e., we select covariates) to evaluate whether our proposal to not regularize the coefficients of covariates (Remark 3) is reasonable.

The first view $\mathbf{X}^1$ is genetic (single nucleotide polymorphisms [SNPs]) data from the Emory Predictive Health Institute. We began with about 8.2 million imputed SNPs, and we extracted SNPs in genes belonging to the immune regulation pathway; this pathway is suggested to play a role in cardiovascular disease risk. After preprocessing, there remained 225,182 genetic variants (see Figure 11.2) and we randomly selected 15% of these SNPs for view 1; there were 2,925 genetic variants and 567 samples. The genetic data are coded 0, 1, 2, where 0 is homozygous normal, 1 is heterozygous with 1 risk allele and 1 normal allele, and 2 is homozygous risk (minor allele in our data). We obtained class membership comprising of two groups as follows. We randomly selected twenty SNPs and summed the number of alleles. Samples with total alleles < 10 were in Class 1, and those with alleles $\geqslant 10$ were in Class 2. The proportions of 1's and 2's were 55.73% and 44.27% respectively. The second view $\mathbf{X}^2$ was set to $\mathbf{X}^2 = 5\mathbf{X}^1 + \mathbf{E}$, where each element in the error term $\mathbf{E}$ was drawn from a normal distribution with mean 0 and variance 1. By this, $\mathbf{X}^1$ and $\mathbf{X}^2$ are correlated. The third view $\mathbf{X}^3$ comprised of 5 variables representing covariates. Data for covariates were generated to fall within the range of 5 continuous variables in our real data: age (40-78 years), systolic blood pressure (83 - 171 mmHg), low-density lipoprotein (14 - 219 mg/dL), body mass index (17.9- 45 kg/m^2), and triglycerides (31- 330 mg/dL). We allowed for different class means so that the classes were well separated in this view. Samples with values below the normal ranges were assigned to class 1 and samples with values on or above the normal ranges were assigned to class 2. For the age variable, samples with age below 60 were assigned to Class 1. Denote this data by $\mathbf{X}^{3*}$. Finally, we let each covariate to have moderate association with one SNP from the 20 SNPs used to construct the two classes. Specifically, we set as our final covariates data: $\mathbf{X}^3 = \mathbf{X}^{3*} + 10\mathbf{X}^{1*}$ where $\mathbf{X}^{1*}$ is a data matrix of 5 SNPs. Twenty bootstrap training datasets where generated and each method was trained using the training sets. The out-of-bag samples (testing sets) were used to evaluate the performance of each method.

Table 7 shows results of our method compared to the other methods considered. Of note we did not compare with sCCA since the method is applicable to two views, but we have three views in this example. Further, MGSDA (stack) did not select any variable out of all 20 bootstrap datasets. We first compare methods when we integrate $\mathbf{X}^1$ and $\mathbf{X}^2$, and when we integrate all three views (i.e., add covariates). We note that the misclassification rates of all methods are considerably lower and even more so for SIDA when we add covariates. This is not surprising since in this example, the data are simulated such that the classes are well-separated in the covariates data, thus improving the performance when all views are integrated. This goes to suggest that if clinical covariates are available, they may help improve classification accuracy. Compared to JACA, SIDA had lower error rates and a higher correlations. When we did regularize the coefficients of the covariates, the error rate was somewhat higher than when we did not regularize; but this was lower than that of JACA. This suggests that in some applications, it is enough to add covariates to guide the selection of variables and estimation of discriminant vectors, and not regularize their coefficients. Our proposed methods give users this option. In terms of variable selectivity, we selected fewer variables compared to JACA, and still achieved lower error rates and higher or comparable correlation estimates.

[Figure 2 about here.]

[Table 7 here]

[Table 8 here]

11. Real Data Analysis

Data preprocessing: The gene expressions data consist of $38,694$ probes, and the metabolomics data consist of $\sim 6,000$ mass to ion (m/z) features. We preprocess and preselect genes as follows. We remove probes with gene names not found in KEGG database. We also remove probes with variance and entropy expression values that are respectively less than the 90th and 20th percentile, resulting in $1,658$ genes. For the metabolomics data, we removed m/z features with at least 50% zeros, and features with coefficient of variation $\geqslant 50\%$; this resulted in 2,416 features for the analyses. Because of the skewed distributions of most metabolomic levels, we log2 transformed each feature.

Genes and m/z features selected by methods: Tables 9, 10 and 11 give the genes and m/z features selected by the proposed and competing methods in at least 60% (12 times) of the 20 resampled datasets. SIDANet and JACA selected 28 and 45 genes respectively, of which 17 overlap; 6 m/z features overlapped between SIDANet and JACA. Additionally, all genes identified by SIDA were also selected by SIDANet; there were 6 overlapping m/z features selected by SIDA and SIDANet. sLDA (Ens) and sLDA (Stack) did not identify any gene and m/z feature.

We also used ToppGene Suite (Chen et al., 2009) to investigate the biological relationships of these "stable" genes. These genes were taken as input in ToppGene online tools for pathway enrichment analysis. The pathways that are significantly enriched (Bonferonni p-value <= 0.05) in the 28 genes selected by SIDANet are listed in Table 12.

[Table 9 here]

[Table 10 here]

[Table 11 here]

[Table 12 here]

11.1 *Genes or m/z features from SIDA and SIDANet plus established risk factors predict ASCVD better*

: Our aim here is to assess whether including the genes and/or m/z features identified by our methods is any better than a model with only age and gender. Given the sample size of 71 in each of the 20 testing resampled datasets, we can only include a few variables to increase power of detecting differences in low vs high-risk ASCVD. We include the demographic variables age and gender in model one (M1). In model two we further include a risk score calculated with the genes or m/z features identified by the methods using the testing datasets. Specifically, we run a logistic regression model on the training data to obtain effect sizes (logarithm of the odds ratio of the probability that ASCVD risk group is high) for each gene or m/z feature. The genetic risk score (GRS) or metabolomic risk score (MRS) are each obtained as a sum of the genes or m/z features in the testing data set, weighted by the effect sizes. In Model 3 (M3), we include both GRS and MRS. We summarize the area under the curves (AUCs) from the receiver operating characteristic in Table 16. We observe that including genes and/or m/z features identified by our methods to a model with age and gender results in better discrimination of the ASCVD risk groups compared to association or classification-based methods, and when compared to a model with only age and gender. By integrating gene expression and m/z features and simultaneously discriminating ASCVD risk group, we have identified biomarkers that potentially may be used to predict ASCVD risk, in addition to a few established ASCVD risk factors.

[Table 16 here]

11.2 *Comparison of Genes and m/z features selected by SIDA and SIDANet for both random and grid search*

We compare genes and m/z features identified by SIDA and SIDANet using both random search and grid search for tuning parameter optimizations. Table **??** gives the average error rate on the testing data, average estimated correlation on the training data, and average number of genes and m/z features. Averages are over 20 resampled datasets. SIDA with *random search* and *grid search* yield similar error rates, and estimated correlations. This is also true for SIDANet. Table 13 gives variable stability results using the criteria discussed in the main text. Seven genes and five m/z features overlap between SIDA with *random* and *grid* search (Table 14). Comparing SIDANet (RS) with SIDANet (GS), 22 genes identified by SIDANet (GS) are also identified by SIDANet (RS) [Table 14]. The m/z features identified by SIDANet (GS) are contained in those selected by SIDANet (RS) [Table 15]. Table 16 compares the AUC's for the three models under consideration. The results are similar for both RS and GS. These findings suggest that we can choose optimal tuning parameters at a lower computational cost (see Table 3) by randomly selecting grid points from the entire tuning parameter hyperspace and searching over those grid values (instead of searching over the entire grid space) and still achieve competitive performance. In our algorithm, the default method to obtain optimal tuning parameter is *random search*. However, we make it as an option for the interested user to choose tuning parameters using *grid search*.

[Table 13 here]

[Table 14 here]

[Table 15 here]

References

Bergstra, J. and Y. Bengio (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research 13(Feb), 281–305.

Bickel, P. and E. Levina (2008, 04). Regularized estimation of large covariance matrices. Annals of Statistics 36.

Cai, T., W. Liu, and X. Luo (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. JASA Theory and Methods 106(494), 594–607.

Chen, J., E. E. Bardes, B. J. Aronow, and A. G. Jegga (2009). Toppgene suite for gene list enrichment analysis and candidate gene prioritization. Nucleic acids research 37(suppl_2), W305–W311.

Hastie, T. and R. Tibshirani (2004). Efficient quadratic regularization for expression arrays. Biostatistics 5(3), 329–340.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

[Table 7 about here.]

[Table 8 about here.]

[Table 9 about here.]

[Table 10 about here.]

[Table 11 about here.]

[Table 12 about here.]

[Table 13 about here.]

[Table 14 about here.]

Received Feb 2020. *Revised Month* 2020.

Accepted Month 2020.

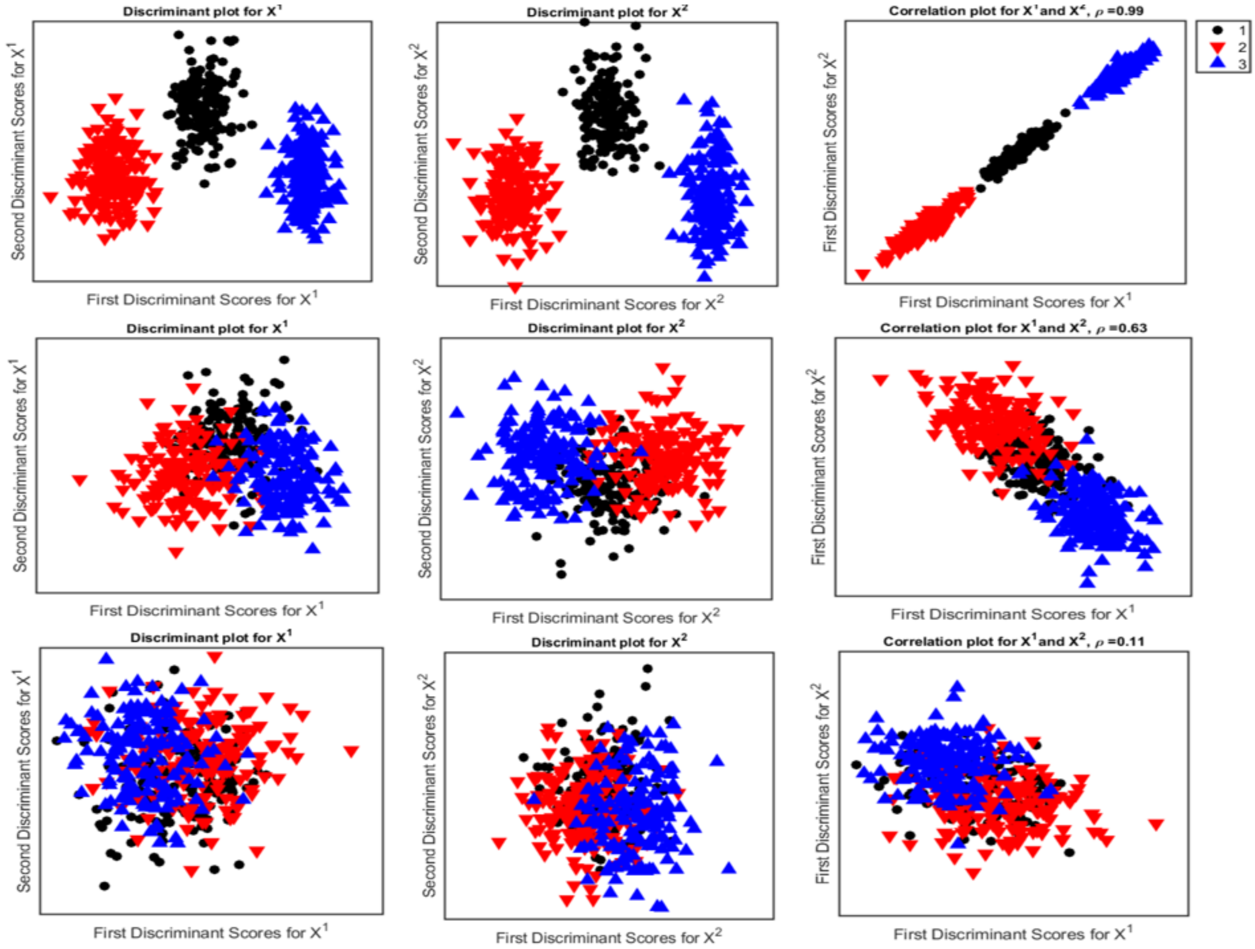

Figure 1. Projection of random data simulated from Scenario One onto true integrative discriminant direction vectors. Top panel: good separation of classes, and strong association between views. Middle pane: moderate separation and moderate association. Bottom panel: weak separation and weak association.

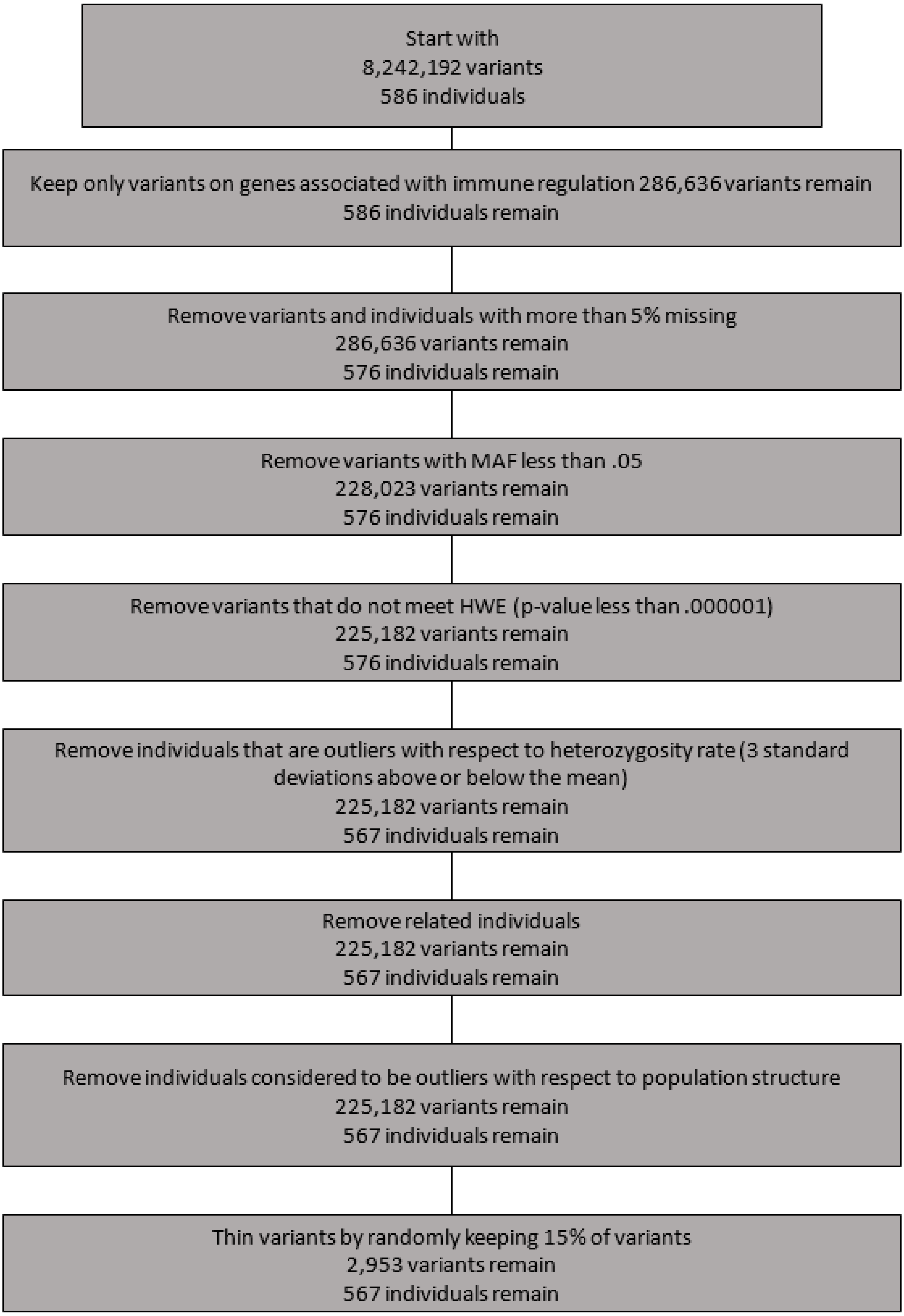
Start with
8,242,192 variants
586 individuals
Keep only variants on genes associated with immune regulation 286,636 variants remain
586 individuals remain
Remove variants and individuals with more than 5% missing
286,636 variants remain
576 individuals remain
Remove variants with MAF less than .05
228,023 variants remain
576 individuals remain
Remove variants that do not meet HWE (p-value less than .000001)
225,182 variants remain
576 individuals remain
Remove individuals that are outliers with respect to heterozygosity rate (3 standard deviations above or below the mean)
225,182 variants remain
567 individuals remain
Remove related individuals
225,182 variants remain
567 individuals remain
Remove individuals considered to be outliers with respect to population structure
225,182 variants remain
567 individuals remain
Thin variants by randomly keeping 15% of variants
2,953 variants remain
567 individuals remain

Figure 2. Genetic data preprocessing criteria.

Property/ Method	Classification- Based	Association- Based	JACA	CCA- Regression	**SIDA**	**SIDANet**
Association		✓	✓	✓	✓	✓
Classification	✓		✓	✓*	✓	✓
Variable Selection	✓	✓	✓	✓	✓	✓
Smoothness	✓	✓				✓
Covariates					✓	✓

Table 1
*Unique features of SIDA and SIDANet compared to other methods. *CCA-regression is not applicable when there are more than two classes.*

	Equal	Class	Variance		Binary	Class	Unequal	Class	Variance
Method	S1	S2	S3	S1	S2	S3	S1	S2	S3
True	0.99	0.63	0.16	0.97	0.46	0.11	0.97	0.42	0.10
SIDA (RS)	0.99	0.58	0.14	0.91	0.37	0.09	0.97	0.40	0.03

Table 2

Comparison of test RV coefficients from using true and estimated discriminant vectors. In most situations, the estimated RV coefficients are comparable to the truth.

	SIDA (RS, P)	SIDA (GS, P)	SIDA (RS, NP)	SIDA (GS, NP)	JACA
(N, p/q)					
(240, 200/200)	1.49	6.80	8.43	39.79	1.31
(240, 2000/2000)	3.39	13.32	12.90	61.51	22.31
(1000, 200/200)	1.36	6.52	10.24	35.00	3.22
(1000, 2000/2000)	5.61	26.35	12.81	66.31	69.53

Table 3

Timings (in minutes). Average time for five fold cross-validation. RS and GS denote random and grid search respectively. P is parallel computing (4 workers), and NP is no parallel computing. N is the sample size, and p/q are the dimensions for the two views of data.

Method	Error (%)	$\hat{\rho}$	TPR-1	TPR-2	FPR-1	FPR-2	F-1	F-2
$(\rho_1 = 0.9, \rho_2 = 0.7, c_1 = 0.5, c_2 = 0.2)$								
SIDA (RS)	0.77	0.97	100.00	100.00	0.02	0.06	99.08	98.08
SIDA (GS)	0.76	0.97	100.00	100.00	0.02	0.06	99.08	98.08
sCCA	1.75	0.98	100.00	100.00	0.11	0.02	96.09	99.21
JACA	1.25	0.96	100.00	100.00	1.89	7.50	57.15	22.37
MGSDA (Stack)	0.82	0.00	30.00	0.00	0.04	0.03	41.99	-*
MGSDA (Ens)	2.10	0.94	28.75	33.50	0.12	0.02	39.41	48.68
$(\rho_1 = 0.9, \rho_2 = 0.7, c_1 = 0.5, c_2 = 0.8)$								
SIDA (RS)	0.00	0.99	100.00	100.00	0.00	0.00	100.00	100.00
SIDA (GS)	0.00	0.99	100.00	100.00	0.00	0.00	100.00	100.00
sCCA	0.00	0.99	100.00	100.00	0.03	0.02	98.72	99.07
JACA	0.00	0.93	100.00	100.00	7.52	3.19	22.25	42.38
MGSDA (Stack)	0.00	0.86	0.50	11.25	0.00	0.00	9.52	20.16
MGSDA (Ens)	0.04	0.96	13.25	11.25	0.00	0.00	23.25	20.16

Table 4

Scenario One: Multi-class, equal covariance within class. RS; randomly select tuning parameters space to search. GS; search entire tuning parameters space. MGSDA (Ens) applies sparse LDA method on separate views and peform classification on the pooled discriminant vectors. MGSDA (Stack) applies sparse LDA on stacked views. TPR-1; true positive rate for $\mathbf{X}^1$*. Similar for TPR-2. FPR; false positive rate for* $\mathbf{X}^2$*. Similar for FPR-2; F-1 is F-measure for* $\mathbf{X}^1$*. Similar for F-2.* ρ_1 *and* ρ_2 *controls the strength of association between* $\mathbf{X}^1$ *and* $\mathbf{X}^2$*.* c_1 *and* c_2 *control the between-class variability in views 1 and 2 respectively. *MGSDA (Stack) had 0 TPR for the second dataset, which results in a division by zero in the definition of F-2.*

Method	Error (%)	$\hat{\rho}$	TPR-1	TPR-2	FPR-1	FPR-2	F-1	F-2
$(\rho_1 = 0.9, \rho_2 = 0.7, c = 0.25)$								
SIDA (RS)	0.77	0.91	100.00	81.50	0.13	0.00	96.14	89.01
SIDA (GS)	0.83	0.90	99.50	71.50	0.13	0.00	95.84	82.21
sCCA	1.08	0.96	97.75	100.00	0.06	0.01	96.41	100.00
JACA	0.95	0.96	100.00	100.00	0.34	0.35	89.14	89.46
MGSDA (Stack)	1.78	0.83	17.25	17.25	0.02	0.01	39.57	27.87
MGSDA (Ens)	1.36	0.87	34.00	25.00	0.01	0.02	49.23	37.18
$(\rho_1 = 0.4, \rho_2 = 0.2, c = 0.2)$								
SIDA (RS)	9.19	0.37	58.00	57.00	1.39	0.68	56.46	59.50
SIDA (GS)	9.28	0.37	60.75	58.75	1.55	1.38	51.75	56.80
sCCA	9.81	0.37	56.75	60.75	0.00	0.01	71.35	73.53
JACA	9.97	0.40	74.50	79.00	2.95	2.56	40.85	47.27
MGSDA (Stack)	10.75	0.32	18.00	17.25	0.13	0.12	27.00	25.88
MGSDA (Ens)	12.95	0.34	21.00	23.50	0.10	0.23	31.34	31.66
$(\rho_1 = 0.15, \rho_2 = 0.05, c = 0.12)$								
SIDA (RS)	23.83	0.09	50.00	49.00	1.87	3.75	47.25	33.09
SIDA (GS)	23.38	0.09	51.00	50.25	2.63	3.14	41.21	38.00
sCCA	27.69	0.07	37.50	41.50	5.30	0.07	49.75	58.54
JACA	22.63	0.10	43.00	42.50	0.38	0.16	52.12	54.36
MGSDA (Stack)	24.77	0.08	13.00	10.75	0.12	0.12	21.15	18.04
MGSDA (Ens)	26.95	0.08	13.00	10.75	0.35	0.14	18.28	17.43

Table 5

Scenario Two: Binary class, equal covariance within class. RS; randomly select tuning parameters space to search. GS; search entire tuning parameters space. MGSDA (Ens) applies sparse LDA method on separate views and peform classification on the pooled discriminant vectors. MGSDA (Stack) applies sparse LDA on stacked views. TPR-1; true positive rate for $\mathbf{X}^1$*. Similar for TPR-2. FPR; false positive rate for* $\mathbf{X}^2$*. Similar for FPR-2; F-1 is F-measure for* $\mathbf{X}^1$*. Similar for F-2.* ρ_1 *and* ρ_2 *controls the strength of association between* $\mathbf{X}^1$ *and* $\mathbf{X}^2$*. c controls the between-class variability within each view.*

Method	Error (%)	$\hat{\rho}$	TPR-1	TPR-2	FPR-1	FPR-2	F-1	F-2
$(\rho_1 = 0.9, \rho_2 = 0.7, c = 0.5)$								
SIDA (RS)	2.16	0.97	83.75	87.17	0.19	0.03	85.40	92.06
SIDA (GS)	2.25	0.97	84.38	87.17	0.20	0.03	85.64	92.06
sCCA	3.61	0.96	83.54	88.04	1.28	6.23	60.72	52.73
JACA	2.08	0.98	83.96	87.61	1.73	1.91	56.06	54.12
MGSDA (Stack)	2.59	0.83	33.75	31.09	0.01	0.02	49.04	45.97
MGSDA (Ens)	3.31	0.93	46.25	45.87	0.09	0.08	59.25	59.33
$(\rho_1 = 0.4, \rho_2 = 0.2, c = 0.2)$								
SIDA (RS)	22.80	0.40	85.45	82.95	0.17	0.18	85.64	85.60
SIDA (GS)	22.32	0.40	88.18	87.27	1.09	1.03	74.25	74.37
sCCA	28.49	0.49	84.77	85.68	1.49	1.28	59.31	60.43
JACA	20.77	0.49	91.14	91.14	1.01	0.95	72.83	74.24
MGSDA (Stack)	25.55	0.34	47.95	45.91	0.07	0.11	61.50	58.65
MGSDA (Ens)	27.97	0.39	57.73	57.73	0.18	0.39	66.36	62.03
$(\rho_1 = 0.15, \rho_2 = 0.05, c = 0.12)$								
SIDA (RS)	48.84	0.03	31.82	44.29	0.59	2.29	33.79	33.28
SIDA (GS)	47.69	0.03	30.45	44.76	0.49	1.72	34.02	34.40
sCCA	50.02	0.03	29.55	42.14	0.47	1.31	33.19	36.54
JACA	40.42	0.07	63.64	66.67	1.03	0.95	56.00	55.51
MGSDA (Stack)	47.72	0.03	22.50	25.48	0.37	0.41	30.77	33.18
MGSDA (Ens)	49.77	0.04	26.36	34.05	0.74	1.13	32.70	36.17

Table 6

Scenario Three: We assume unequal covariances in each class. This violates the LDA assumption. RS; randomly select tuning parameters space to search. GS; search entire tuning parameters space. MGSDA (Ens) applies sparse LDA method on separate views and perform classification on the pooled discriminant vectors. MGSDA (Stack) applies sparse LDA on stacked views. TPR-1; true positive rate for $\mathbf{X}^1$. Similar for TPR-2. FPR; false positive rate for $\mathbf{X}^2$. Similar for FPR-2; F-1 is F-measure for $\mathbf{X}^1$. Similar for F-2. ρ_1 and ρ_2 control the strength of association between $\mathbf{X}^1$ and $\mathbf{X}^2$. c controls the between-class variability within each view.

Method	Error (%)	$\hat{\rho}$	average # of	variables	selected	
			View 1	View 2	View 1 (20 SNPs)	View 2 (20 SNPs)
SIDA(RS)	0.43	0.40	13.40	18.10	9.25	8.10
SIDA(RS)*	0.49	0.36	20.00	15.85	7.30	6.75
SIDA(RS)$^+$	29.19	0.43	323.25	95.35	16.95	11.30
SIDA(GS)	0.31	0.36	7.50	8.35	6.350	5.25
SIDA(GS)*	0.52	0.36	21.45	14.00	8.70	6.20
SIDA(GS)$^+$	29.82	0.41	349.05	130.35	16.80	11.40
JACA	7.15	0.24	1544.85	1561.20	19.35	19.30
JACA$^+$	33.45	0.64	1853	1863.40	19.65	19.45

Table 7

Scenario Four: Simulation setting with covariates and different separation of classes within views. sCCA is applicable to two views but we have three views in this example. View 1 (20 SNPs) denote the average number of SNPs selected in View 1 from the twenty genetic variants that was used to construct the two classes. Similarly for View 2 (20 SNPs). * *denotes results when we regularize the coefficients of View 3.* $^+$ *is the result when we integrate two views without the covariates.*

Method	Error (%)	$\hat{\rho}$	average # of	variables	selected	
			View 1	View 2	View 1 (20 SNPs)	View 2 (20 SNPs)
SIDA(RS)	0.31	0.40	14.40	15.85	9.10	7.75
SIDA(RS)*	0.45	0.36	21.65	17.20	7.70	7.15
SIDA(RS)$^+$	29.27	0.31	332.85	93.30	16.55	10.60
SIDA(GS)	0.28	0.36	8.10	6.90	6.40	5.10
SIDA(GS)*	0.38	0.36	19.65	15.55	8.30	6.80
SIDA(GS)$^+$	28.30	0.37	262.15	124.40	16.60	11.85
JACA	7.15	0.24	1544.85	1561.20	19.35	19.30
JACA$^+$	33.45	0.64	1853.00	1863.40	19.65	19.45

Table 8

Scenario Four: Simulation setting with covariates and different separation of classes within views. We use $\mathbf{S}_{b_{dj}}$ in the association part of equation (1) instead of $\mathbf{S_{dj}}, d, j = 1, 2, 3, d \neq j$. sCCA is applicable to two views but we have three views in this example. View 1 (20 SNPs) denote the average number of SNPs selected in View 1 from the twenty genetic variants that was used to construct the two classes. Similarly for View 2 (20 SNPs). * denotes results when we regularize the coefficients of View 3. $^+$ is the result when we integrate two views without the covariates.

	# Genes	#m/z features
SIDA	7	7
SIDANet	28	7
sCCA	1	185
JACA	45	20
sLDA (Ens)	0	0
sLDA (Stack)	0	0

Table 9

Genes and m/z feature selected at least 60% (12 times out of 20 resampled datasets).

Method	Genes selected
SIDA	CIRBP CLEC1BH4C8 MAGEB4 RASEF SCGB1C1 TNS2
SIDANet	ABHD3 TSPOAP1 CBS CIRBP CLEC1B EMP2 FCER1A FCER1G GJA9 GLYAT GNAI1 GNAQ H3F3A H2BC5 H4C8 HMBOX1 IRX3 MAGEB4 NEURL2 NPVF RASEF RGS18 SCGB1C1 SCUBE1 TNS2 ALKAL1 YPEL5 ZNF667
JACA	ABHD3 ADAD2 ALKBH8 ANKLE1 ARG1 TSPOAP1 CIRBP CLCN3 CYP17A1 ACKR1 DEFB127 DNAJB14 ERV3-1 FCER1G FPR2 GLYAT GNAI1 H3F3A HBE1 H2BC5 H4C8 HLA-DRB4 HMBOX1 KCTD20 LOC402634 LRRC6 MAGEB4 NEURL2 OR2T29 PAIP2 PDGFRB POLR2J PTGS2 RASEF RPA1 SCGB1C1 SCUBE1 SEMA3E SIGLEC16 TNS2 TMEM190 TMEM40 TRAF4 VGF YPEL5
sLDA(Ens)	-
sLDA(Stack)	-

Table 10

Genes feature selected at least 60% (12 times out of 20 resampled datasets). There are six overlapping genes between SIDA and SIDANet.

Method	m/z features (retention times) selected
SIDA	168.9045(73.1430) 212.9862 (373.9647) 216.9397(134.2085) 228.8127 (98.0079) 250.1187 (30.9802) 542.3191 (572.5522) 756.7378 (64.1087)
SIDANet	168.9045 (73.1430) 216.9397 (134.2085) 250.1187 (30.9802) 542.3191 (572.5522) 754.4435 (42.6461) 756.7378 (64.1087)
JACA	87.1004 (104.4254) 102.0666 (140.1219) 131.5336 (209.9927) 140.9912 (30.7790) 146.0601 (117.5612) 168.9045 (73.1430) 201.9908 (131.3211) 212.9862 (373.9647) 216.9397 (134.2085) 250.1187 (30.9802) 282.1301 (574.5190) 283.2631 (31.3793) 317.1609 (20.8866) 342.3191 (37.0602) 363.0753 (516.8183) 604.7894 (61.5482) 652.6285 (67.6081) 682.3266 (52.4711) 754.4435 (42.6461) 1071.7809 (54.6874)
sLDA(Ens)	-
sLDA(Stack)	-

Table 11

m/z features (retention times) selected at least 60% (12 times out of 20 resampled datasets). There are six overlapping features between SIDA and SIDANet.

ID	Pathway	Database	p-value	Bonferonni q-value	Genes in list
SIDANet					
1144995	Sphingolipid signaling pathway	KEGG	2.08E-05	6.71E-03	GNAI1,GNAQ,FCER1A,FCER1G
1269659	RNA Polymerase I Promoter Opening	REACTOME	9.86E-05	3.19E-02	H2BC5,H3-3A,H4C8
1269740	DNA methylation	REACTOME	1.08E-04	3.49E-02	H2BC5,H3-3A,H4C8
585563	Alcoholism	KEGG	1.08E-04	3.50E-02	H2BC5,H3-3A,GNAI1,H4C8
1269513	Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3	REACTOME	1.18E-04	3.82E-02	H2BC5,H3-3A,H4C8
1269738	SIRT1 negatively regulates rRNA Expression	REACTOME	1.29E-04	4.16E-02	H2BC5,H3-3A,H4C8
JACA					
137937	S1P1 pathway	Pathway Interaction	1.116E-5	6.862E-3	GNAI1, PDGFRB, PTGS2
1269866	Meiotic recombination	REACTOME	4.300E-5	2.644E-2	H2BC5, H3-3A, RPA1, H4C8

Table 12

Pathway enrichment analysis of "stable" genes using ToppGene Suite. SIDANet used Random Search for optimal tuning parameters.

	# Genes	# m/z features
SIDA (RS)	7	7
SIDA (GS)	17	6
SIDANet (RS)	28	7
SIDANet (GS)	32	4

Table 13
Genes and m/z feature selected at least 60% (12 times out of 20 resampled datasets).

Method	Genes selected
SIDA (RS)	CIRBP CLEC1B H4C8 MAGEB4 RASEF SCGB1C1 TNS2
SIDA (GS)	ABHD3 TSPOAP1 CBS CIRBP CLEC1B EMP2 H4C8 HMBOX1 IRX3 MAGEB4 NEURL2 RASEF RGS18 SCGB1C1 SCUBE1 TNS2 YPEL5
SIDANet (RS)	ABHD3 TSPOAP1 CBS CIRBP CLEC1B EMP2 FCER1A FCER1G GJA9 GLYAT GNAI1 GNAQ H3F3A H2BC5 H4C8 HMBOX1 IRX3 MAGEB4 NEURL2 NPVF RASEF RGS18 SCGB1C1 SCUBE1 TNS2 ALKAL1 YPEL5 ZNF667
SIDANet (GS)	ABHD3 AICDA TSPOAP1 CBS CIRBP CLCN3 CLEC1B CYP17A1 DEFB127 EMP2 ERV3-1 GNAI1 GNAQ H3F3A HELB H2BC5 H4C8 HMBOX1 IRX3 JAG2 MAGEB4 NEURL2 RASEF RGS18 RPA1 SCGB1C1 SCUBE1 TNS2 TIMELESS ALKAL1 VGF YPEL5

Table 14

Genes feature selected at least 60% (12 times out of 20 resampled datasets). There are overlapping genes.

Method	m/z features (retention times) selected
SIDA (RS)	168.9045(73.1430) 212.9862 (373.9647) 216.9397(134.2085) 228.8127 (98.0079) 250.1187 (30.9802) 542.3191 (572.5522) 754.4435 (42.6461) 756.7378
SIDA (GS)	168.9045(73.1430) 216.9397(134.2085) 228.8127 (98.0079) 250.1187 (30.9802) 754.4435 (42.6461) 756.7378 (64.1087)
SIDANet (RS)	168.9045 (73.1430) 212.9862 (373.9647) 216.9397 (134.2085) 250.1187 (30.9802) 342.3191 (37.0602) 542.3191 (572.5522) 754.4435 (42.6461)
SIDANet (GS)	168.9045 (73.1430) 216.9397 (134.2085) 250.1187 (30.9802) 754.4435 (42.6461)

Table 15

m/z features (retention times) selected at least 60% (12 times out of 20 resampled datasets). There are overlapping features between SIDA (RS) and SIDA (GS).

	minimum	mean	median	maximum
M1	0.71	0.80	0.81	0.89
M2: M1 + GRS				
SIDA (RS)	0.82	0.89	0.90	0.94
SIDA (GS)	0.86	0.93	0.93	0.99
SIDANet (RS)	0.86	0.94	0.94	0.98
SIDANet (GS)	0.86	0.94	0.94	0.98
M3: M1 + MRS				
SIDA (RS)	0.79	0.85	0.84	0.91
SIDA (GS)	0.79	0.85	0.85	0.92
SIDANet (RS)	0.81	0.87	0.87	0.94
SIDANet (GS)	0.81	0.87	0.86	0.95
M4: M1 + GRS + MRS				
SIDA (RS)	0.84	0.91	0.91	0.95
SIDA (GS)	0.89	0.94	0.93	0.99
SIDANet (RS)	0.89	0.95	0.95	1.00
SIDANet (GS)	0.87	0.94	0.94	0.99

Table 16

Comparison of AUCs using genes and m/z features identified: Model 1 (M1): Age + gender; Model 2 (M2): Age + gender + gene risk score (GRS); Model 3 (M3): Age + gender+ metabolomic risk score (MRS). Model 4 (M4): age + gender + metabolomic risk score + gene risk score. The genes and m/z features identified by the methods on the training datasets are used to calculate GRS and MRS.