Supplementary materials for "Multiway sparse distance weighted discrimination"

Bin Guo[1], Lynn E. Eberly[1,2], Pierre-Gilles Henry[2], Christophe Lenglet[2], Eric F. Lock[1]

[1] Division of Biostatistics, School of Public Health
[2] Center for Magnetic Resonance Research
University of Minnesota

Abstract

In this supplementary document, Sections S1-S8 describe the results of additional simulation studies to assess different aspects of the method. Section S9 gives more detail on the selection of tuning parameters for the application to MRS data in a study of a mouse model for Friedreich's ataxia, and Section S10 discusses computing time for the applications.

S1 Evaluation of cross-validation method for selecting tuning parameters

We conducted a simulation study to assess how the choices of tuning parameters λ_1 and λ_2 affect the performance of the proposed method. A dataset with sample size $N = 100$ was generated, with two classes of equal size ($N_0 = N_1 = 50$). The predictors have the form of a three-way array of dimensions $\mathbb{X} : P_1 \times P_2 \times P_3$ where $P_1 = 4, P_2 = 5$, and $P_3 = 15$. The N_0 samples corresponding to class -1 were generated from a multivariate normal distribution $\mathrm{N}(\mathbf{0},\mathbf{I}_{P_1P_2P_3\times P_1P_2P_3})$. The other N_1 samples corresponding to class 1 were generated from a multivariate normal distribution $\mathrm{N}(\boldsymbol{\mu}_1,\mathbf{I}_{P_1P_2P_3\times P_1P_2P_3})$ where $\boldsymbol{\mu}_1 = \mathbf{u}_1 \circ \mathbf{u}_2 \circ \mathbf{u}_3$. Here $\mathbf{u}_1$ and $\mathbf{u}_2$ were generated from $N(\mathbf{0}, \mathbf{I}_{P_k\times P_k})$, $k = 1, 2$. For $\mathbf{u}_3$ 5 values are set to zero and 10 nonzero values are generated from $N(\mathbf{0}, \mathbf{I}_{10\times 10})$. We consider 4 candidates for λ_1 $(10^{-4}, 0.001, 0.005, 0.01)$ and 3 candidates for λ_2 $(1, 0.5, 0.1)$. We applied the multiway sparse DWD method to the simulated data with fixed parameters and selected parameters by cross-validation. The simulation is repeated 100 times. The

results are shown in Table S1 and Table S2 with average correlations between true values and estimates for $\mathbf{u}_k$ and average percentages of zero or non-zero coefficients in $\mathbf{u}_3$ that are correctly estimated. The accuracy of classification decreases as λ_2 decreases. For fixed λ_2, as λ_1 becomes larger, more zero coefficients for $\mathbf{u}_3$ are correctly shrunk to 0. Table S2 shows simulations where the correlations between true values and estimates are very high, and the classification is accurate using the cross-validation method to select parameters, although the selected parameters might be slightly different for each replicate.

Table S1: Simulation results based on multiway sparse DWD with prespecified λ_1 and λ_2 candidates. "Cor($\mathbf{u}_k$)" is the correlation between the estimated linear hyperplane and the true hyperplane for k^{th} dimension. "TP($\mathbf{u_3}$)" is the true positive rate that is the proportion of non-zero coefficients in $\mathbf{u}_3$ are correctly estimated to be non-zero."TN($\mathbf{u}_3$)" is the true negative rate that is the proportion of zero coefficients in $\mathbf{u}_3$ are correctly estimated to be zero. The results are the mean over 100 replicates.

λ_2	1				0.5				0.1			
λ_1	1e-4	0.001	0.005	0.01	1e-4	0.001	0.005	0.01	1e-4	0.001	0.005	0.01
Cor($\mathbf{u}_1$)	0.96	0.96	0.97	0.97	0.93	0.93	0.93	0.93	0.89	0.91	0.93	0.92
Cor($\mathbf{u}_2$)	1.00	1.00	1.00	0.99	1.00	0.99	0.99	0.98	0.95	0.95	0.95	0.93
Cor($\mathbf{u}_3$)	0.98	0.99	0.98	0.98	0.97	0.97	0.97	0.96	0.94	0.93	0.93	0.92
TP ($\mathbf{u}_3$)	1.00	0.98	0.81	0.85	1.00	0.98	0.90	0.83	0.99	0.94	0.80	0.72
TN ($\mathbf{u}_3$)	0.03	0.37	0.81	0.91	0.03	0.41	0.81	0.92	0.09	0.59	0.90	0.95

Table S2: Simulation results based on multiway sparse DWD with selected penalty parameters by cross-validation. λ_1^G and λ_2^G denote the geometric means of selected parameters over 100 replicates.

	$\lambda_1^G = 0.0008, \lambda_2^G = 1.56$
Cor($\mathbf{u}_1$)	0.98
Cor($\mathbf{u}_2$)	0.99
Cor($\mathbf{u}_3$)	0.99
TP($\mathbf{u}_3$)	0.39
TN($\mathbf{u}_3$)	0.93

S2 Simulation results based on the objective function with separable L_2 penalty

In this section, we consider a simulation study to evaluate the performance of the higher rank model based on the objective function with separable L_2 penalty (i.e. $P^U_{\lambda_1,\lambda_2}(\mathbb{B})$ defined in the main article). We observed that the separable L_2 penalty tends to reduce the rank of the estimated coefficient array, motivating us to use the non-separable L_2 penalty (i.e. $P_{\lambda_1,\lambda_2}(\mathbb{B})$) for the higher rank model. We simulated 200 datasets with high dimensions ($30 \times 15 \times 15$) under the case of no sparsity. The process of generating

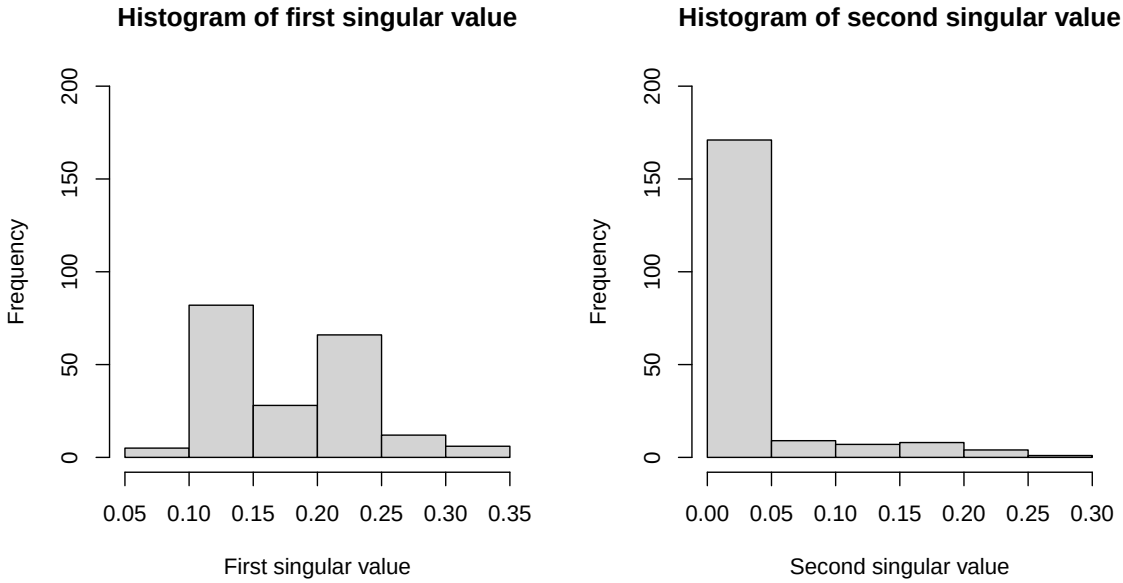

Figure S1: Histograms of the two singular values computed based on the estimated coefficient matrix.

multiway data is identical to that for the higher rank model ($R = 2$) described in Section 5.3 of the main article. Table S3 shows average correlations, misclassification rates and true positive rates over 200 simulation replicates. Multiway sparse DWD (M-SDWD) with $R = 2$ has similar performance to M-SDWD with R=1, which implies M-SDWD (R=2) may not recover the true rank specified in the model. To measure the rank of the estimated model, we conduct SVD to obtain singular values for the estimated coefficient matrix. In most simulations, the second singular values of the estimated coefficient matrix are zero or nearly zero, as shown in Figure S1, which indicates the solutions of higher rank (R=2) model are often shrunk to a lower rank. The proportion of simulations that give estimates with true rank $R = 2$ is only 14.5%. Table S4 shows the results of simulations that can truly detect rank 2 components, and we can see the model with R=2 performs much better than rank-1 multiway models and full SDWD.

S3 Simulation results under different signal to noise ratios

We conducted more simulation studies to compare the proposed method with other methods under different signal to noise ratios (SNR= (0.1,0.2,0.3,0.5)). We considered the high-dimensional, more sparsity case of Section 5.1 in the main article for this simulation. Tables S5 and S6 show the results based on all simulations and those simulations with

Table S3: Simulation results based on all simulations under the high dimensional scenario: "Cor" is the correlation between the estimated linear hyperplane and the true hyperplane. "Mis" is the average misclassification rate. "TP" is the true positive rate, i.e., the proportion of non-zero coefficients that are correctly estimated to be non-zero.The margins of error (2* standard errors across 200 replicates) for each statistic are also listed following the ±.

Methods	Cor	Mis	TP
M-SDWD (R=2)	0.833±0.012	0.000±0.000	0.883±0.029
M-SDWD (R=1)	0.792±0.011	0.000±0.000	0.872±0.029
M-SDWD ($\lambda_1 = 0$, R=1)	0.800±0.010	0.000±0.000	1.000±0.000
M-DWD	0.771±0.014	0.000±0.000	1.000±0.000
Full SDWD	0.781±0.006	0.000±0.000	0.419±0.036

Table S4: Simulation results among simulations that give estimates with true rank (R=2) under the high-dimensional scenario: "Cor" is the correlation between the estimated linear hyperplane and the true hyperplane. "Mis" is the average misclassification rate. "TP" is the true positive rate, i.e., the proportion of non-zero coefficients that are correctly estimated to be non-zero."TN" is the true negative rate, i.e., the proportion of zero variables that are correctly estimated to be zero. The margins of error (2* standard errors across 200 replicates) for each statistic are also listed following the ± symbol.

Methods	Cor	Mis	TP	% rank 2
M-SDWD (R=2)	0.984±0.008	0.000±0.000	0.955±0.040	0.145
M-SDWD (R=1)	0.709±0.033	0.000±0.000	0.844±0.088	-
M-SDWD ($\lambda_1 = 0$, R=1)	0.726±0.023	0.000±0.000	1.000±0.000	-
M-DWD	0.676±0.024	0.000±0.000	1.000±0.000	-
Full SDWD	0.782±0.016	0.000±0.000	0.492±0.088	-

correlations larger than 0.5, respectively. The multiway sparse DWD model performs better than other methods in terms of its correlation with the true hyperplane, and has competive misclassification rates across the different SNR levels. The last column in Table S6 show the proportions of simulations with correlations larger than 0.5 for different methods under different SNRs. As the SNR increases, the proportion increases as well, as the algorithm tends to converge to the true solution with more signal.

S4 Assessment of convergence

We conducted more simulations to explore how the signal to noise ratio (SNR) affects the convergence of the proposed methods and other alternatives. Figure S2 and S3 show the distributions of correlations between the true hyperplane and the estimated hyperplane for SNR = 0.1 and SNR=0.3, under the more sparsity case based on high dimensional

Table S5: Simulation results for different signal to noise ratios: "Cor" is the correlation between the estimated linear hyperplane and the true hyperplane. "Mis" is the average misclassification rate. "TP" is the true positive rate, i.e., the proportion of non-zero coefficients that are correctly estimated to be non-zero."TN" is the true negative rate, i.e., the proportion of zero variables that are correctly estimated to be zero. The margins of error (2* standard errors across 200 replicates) for each statistic are also listed following the ± symbol.

SNR	Methods	Cor	Mis	TP	TN
0.1	M-SDWD	0.717±0.053	0.225±0.029	0.484±0.050	0.837±0.038
	M-SDWD ($\lambda_1 = 0$)	0.594±0.054	0.220±0.028	1.000±0.000	0.000±0.000
	M-DWD	0.560±0.058	0.235±0.030	1.000±0.000	0.000±0.000
	Full SDWD	0.469±0.040	0.251±0.025	0.168±0.028	0.917±0.029
0.2	M-SDWD	0.849±0.038	0.089±0.020	0.668±0.040	0.806±0.041
	M-SDWD ($\lambda_1 = 0$)	0.796±0.040	0.101±0.021	1.000±0.000	0.000±0.000
	M-DWD	0.766±0.049	0.121±0.025	1.000±0.000	0.000±0.000
	Full SDWD	0.636±0.035	0.136±0.022	0.172±0.023	0.957±0.022
0.3	M-SDWD	0.879±0.035	0.066±0.018	0.720±0.036	0.776±0.045
	M-SDWD ($\lambda_1 = 0$)	0.838±0.037	0.077±0.020	1.000±0.000	0.000±0.000
	M-DWD	0.831±0.042	0.084±0.021	1.000±0.000	0.000±0.000
	Full SDWD	0.690±0.031	0.104±0.020	0.165±0.017	0.982±0.011
0.5	M-SDWD	0.927±0.026	0.039±0.014	0.777±0.033	0.762±0.048
	M-SDWD ($\lambda_1 = 0$)	0.890±0.030	0.048±0.016	1.000±0.000	0.000±0.000
	M-DWD	0.884±0.036	0.053±0.018	1.000±0.000	0.000±0.000
	Full SDWD	0.760±0.027	0.066±0.017	0.189±0.017	0.993±0.002

data ($30 \times 15 \times 15$) with sample size $N = 100$. Combining with results for SNR=0.2 shown in the main article we can see as the SNR increases, the number of correlations with small values decreases, and the convergence issue becomes less severe.

S5 Simulations with Rank= 5

We conducted more simulations to evaluate the performances of the rank-R multiway sparse model. These simulations were analogous to those in Section 5.3 of the main article, with different manipulated conditions. Table S7 gives the results under a lower dimensional setting ($\mathbb{X} : 15 \times 4 \times 5$). Table S8 gives results for both the lower-dimensional and higher-dimensional settings for a model with rank $R = 5$. From these results we can conclude that the rank-R model can generally perform well when the data were generated under a true rank-R model, although lower-rank approximations perform comparably well when the dimension is small and the signal is sparse.

Table S6: Simulation results among simulations with correlation greater than 0.5 under different signal to noise ratios: "Cor" is the correlation between the estimated linear hyperplane and the true hyperplane. "Mis" is the average misclassification rate. "TP" is the true positive rate, i.e., the proportion of non-zero coefficients that are correctly estimated to be non-zero."TN" is the true negative rate, i.e., the proportion of zero variables that are correctly estimated to be zero. The margins of error (2* standard errors across 200 replicates) for each statistic are also listed following the ± symbol.

SNR	Methods	Cor	Mis	TP	TN	% Cor>0.5
0.1	M-SDWD	0.916±0.014	0.077±0.015	0.628±0.046	0.869±0.040	0.645
	M-SDWD ($\lambda_1 = 0$)	0.858±0.017	0.088±0.016	1.000±0.000	0.000±0.000	0.66
	M-DWD	0.878±0.018	0.080±0.016	1.000±0.000	0.000±0.000	0.62
	Full SDWD	0.718±0.023	0.096±0.017	0.124±0.014	0.993±0.002	0.52
0.2	M-SDWD	0.935±0.010	0.048±0.010	0.697±0.038	0.831±0.042	0.905
	M-SDWD ($\lambda_1 = 0$)	0.899±0.012	0.051±0.010	1.000±0.000	0.000±0.000	0.88
	M-DWD	0.926±0.010	0.042±0.009	1.000±0.000	0.000±0.000	0.825
	Full SDWD	0.760±0.017	0.060±0.011	0.156±0.016	0.994±0.001	0.765
0.3	M-SDWD	0.948±0.008	0.033±0.008	0.745±0.034	0.795±0.046	0.925
	M-SDWD ($\lambda_1 = 0$)	0.921±0.010	0.035±0.009	1.000±0.000	0.000±0.000	0.905
	M-DWD	0.941±0.008	0.030±0.008	1.000±0.000	0.000±0.000	0.88
	Full SDWD	0.781±0.016	0.048±0.010	0.170±0.016	0.994±0.002	0.825
0.5	M-SDWD	0.963±0.006	0.021±0.007	0.792±0.030	0.767±0.049	0.96
	M-SDWD ($\lambda_1 = 0$)	0.943±0.009	0.022±0.007	1.000±0.000	0.000±0.000	0.94
	M-DWD	0.958±0.007	0.018±0.006	1.000±0.000	0.000±0.000	0.92
	Full SDWD	0.816±0.015	0.030±0.009	0.202±0.018	0.994±0.001	0.89

S6 Rank estimation and misspecification

Table S9 shows the mean correlations between the estimated coefficients and the true coefficients using the proposed method (M-SDWD) by assumed rank ($\hat{R} = 1, 2, 3, 4, 5$) for different true ranks under the low dimensional scenario with less sparsity when N = 40. In general, the correlation is maximized when the assumed rank is equal to the true rank; however, performance tends to be robust to small deviations from the precise value of the true rank, especially when the true rank is higher. Table S9 also shows the mean correlation when the rank is estimated by maximizing the t-statistic of the validation scores between the two classes under 10-fold cross-validation. This approach performs better than any fixed rank over all scenarios, including the true rank of the underlying signal.

S7 Simulations with correlated predictors

In this section, we considered a simulation study in which the predictors are correlated. The covariance matrix for the predictor array is assumed to have a separable structure

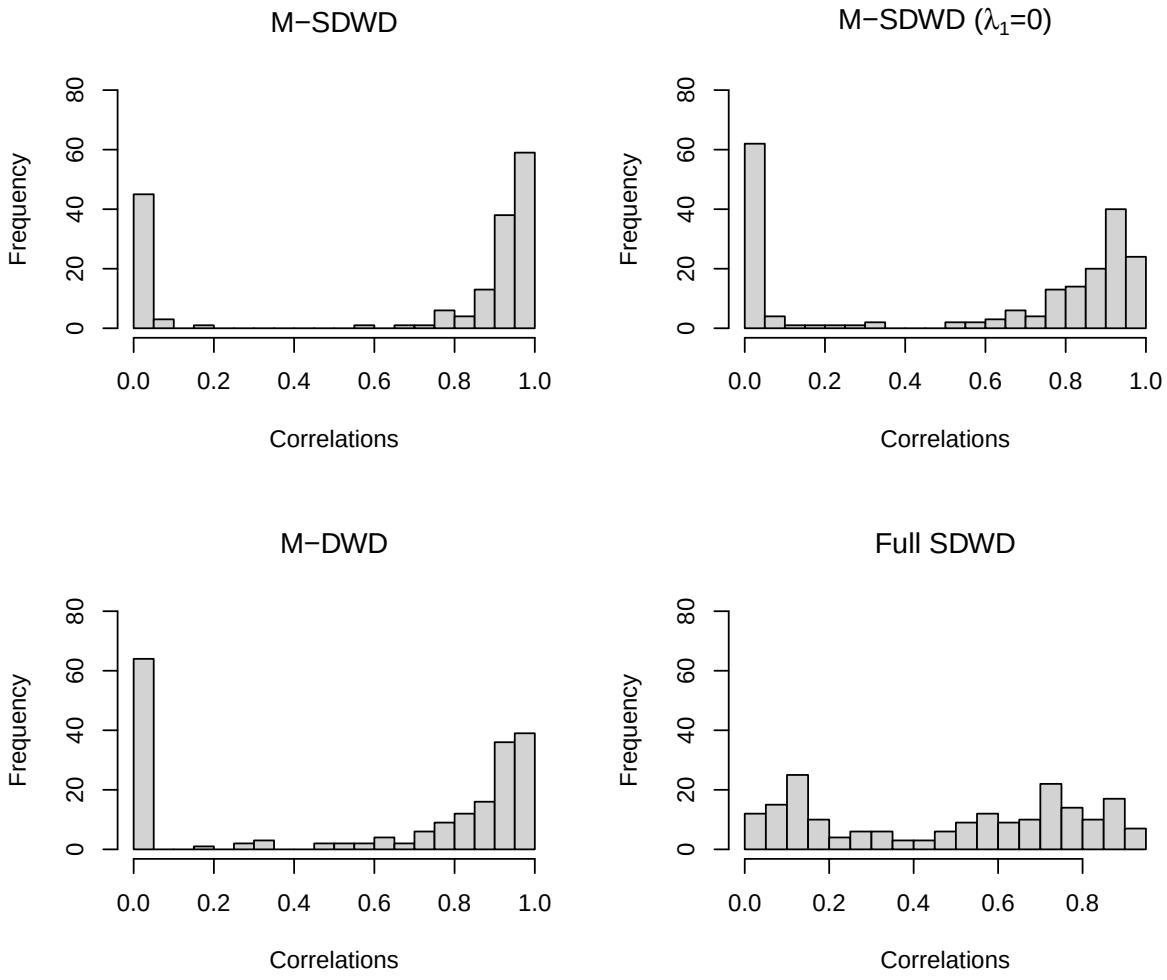

Figure S2: Histogram of correlations between true hyperplane and estimates with four classification methods under high-dimensional scenario with SNR=0.1 and sample size N= 100.

$\Sigma = \Sigma_1 \otimes \Sigma_2 \otimes \Sigma_3$, where $\Sigma_k : P_k \times P_k$ is the covariance along dimension k. Each of Σ_1, Σ_2, and Σ_3 have an AR(1) structure, with correlation determined by a shared parameter ρ:

$$\Sigma_k = \begin{bmatrix} 1 & \rho & \rho^2 & \dots \\ & 1 & \rho & \dots \\ & & \ddots & \\ & & & 1 \end{bmatrix} \quad \text{for } k = 1, 2, 3.$$

Thus, ρ controls the overall level of correlation in the predictors. The samples for class -1 are generated via $\text{vec}(\mathbb{X}_i) = \text{Normal}(\mathbf{0}, \Sigma)$ and the samples for class +1 are generated

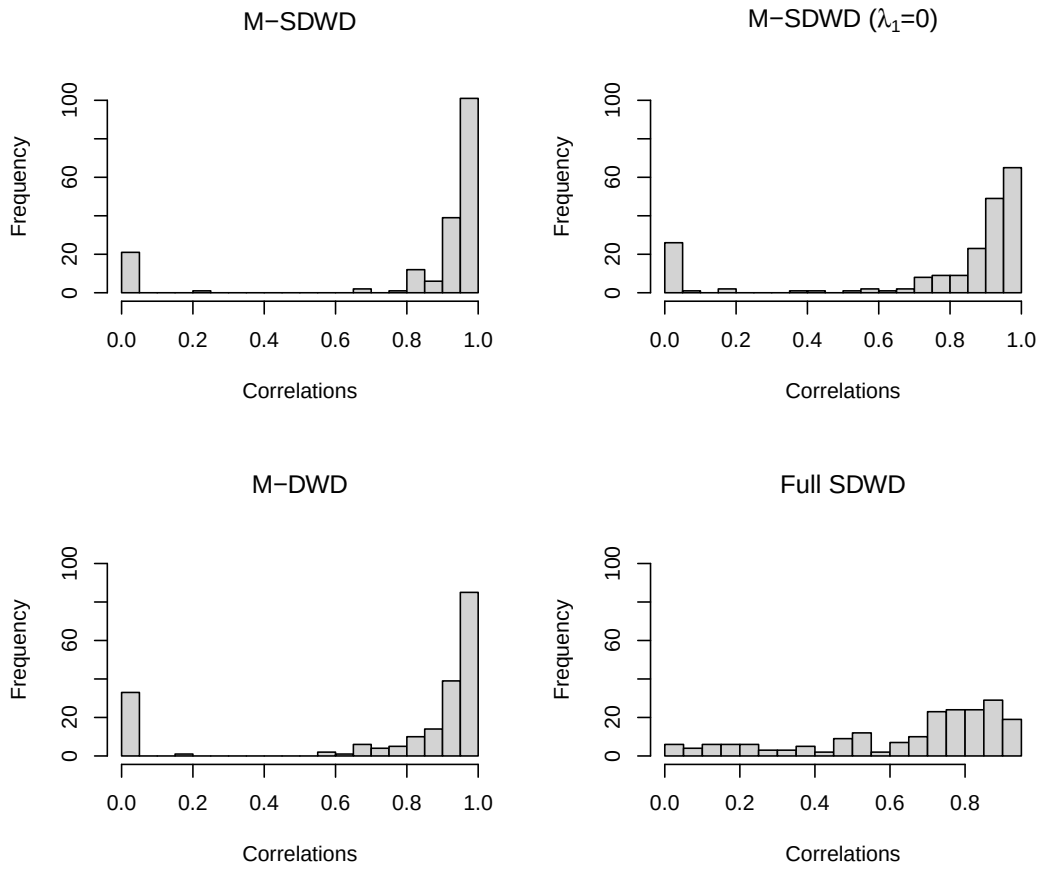

Figure S3: Histogram of correlations between true hyperplane and estimates with four classification methods under high-dimensional scenario with SNR=0.3 and sample size N= 100.

via $\text{vec}(\mathbb{X}_i) = \text{Normal}(\text{vec}(\sqrt{\alpha}\mu_1), \Sigma)$, with $\sqrt{\alpha}\mu_1$ generated as in Section 5.1 of the main article. As a representative scenario we consider the high-dimensional $(30 \times 15 \times 15)$, more sparsity, small sample size (N=40) case of Section 5.1 in the main article, and generate $\sqrt{\alpha}\mu_1$ under those conditions. The CATCH method is considered as a competing approach here because it assumes a separable multiway residual covariance structure analogous to that in our data generating process. However, the CATCH model does not assume a low-rank or any other multiway structure on the mean signal distinguishing the two groups, μ_1, and thus is not well-suited for the other simulation scenarios.

Table S10 shows the results over 200 replications across different methods and different correlation levels ρ. This shows that the M-SDWD method performs relatively well for scenarios with no correlation or mild correlation, but less well for scenarios with high correlation. As expected, CATCH performs relatively well for scenarios with higher correlation but has no advantage when there is no correlation. We find that the multiway

Table S7: Simulation results under the low dimensional scenario ($15 \times 4 \times 5$) when the true model is rank-2. In the Sparsity column, the numbers in parentheses indicate the number of non-zero variables in each dimension. "Cor" is the correlation between the estimated linear hyperplane and the true hyperplane. "Mis" is the average misclassification rate. "TP" is the true positive rate, i.e., the proportion of non-zero coefficients that are correctly estimated to be non-zero."TN" is the true negative rate, i.e., the proportion of zero coefficients that are correctly estimated to be zero. The margins of error (2* standard errors across 200 replicates) for each statistic are also listed following the ± symbol.

N	Sparsity	Methods	Cor	Mis	TP	TN
40	More ($5 \times 2 \times 2$)	M-SDWD (R=2)	0.716±0.034	0.191±0.027	0.659±0.052	0.678±0.048
		M-SDWD ($\lambda_1 = 0$, R=2)	0.643±0.032	0.183±0.023	1.000±0.000	0.000±0.000
		M-SDWD (R=1)	**0.746±0.038**	**0.157±0.024**	0.676±0.052	0.711±0.056
		M-SDWD ($\lambda_1 = 0$, R=1)	0.713±0.036	0.167±0.024	1.000±0.000	0.000±0.000
		M-DWD	0.724±0.041	0.168±0.025	1.000±0.000	0.000±0.000
		Full SDWD	0.618±0.033	0.188±0.023	0.410±0.034	0.861±0.033
	Less ($5 \times 4 \times 5$)	M-SDWD (R=2)	**0.875±0.014**	0.038±0.010	0.786±0.036	0.440±0.052
		M-SDWD ($\lambda_1 = 0$, R=2)	0.870±0.015	**0.036±0.009**	1.000±0.000	0.000±0.000
		M-SDWD (R=1)	0.810±0.017	0.043±0.010	0.698±0.042	0.530±0.054
		M-SDWD ($\lambda_1 = 0$, R=1)	0.818±0.015	0.040±0.009	1.000±0.000	0.000±0.000
		M-DWD	0.824±0.017	0.039±0.010	1.000±0.000	0.000±0.000
		Full SDWD	0.703±0.019	0.064±0.014	0.334±0.032	0.847±0.033
	No ($15 \times 4 \times 5$)	M-SDWD (R=2)	**0.958±0.007**	0.003±0.003	1.000±0.000	-
		M-SDWD ($\lambda_1 = 0$, R=2)	**0.959±0.006**	0.003±0.002	1.000±0.000	-
		M-SDWD (R=1)	0.840±0.014	0.006±0.003	0.839±0.034	-
		M-SDWD ($\lambda_1 = 0$, R=1)	0.849±0.012	0.005±0.003	1.000±0.000	-
		M-DWD	0.843±0.014	0.005±0.003	1.000±0.000	-
		Full SDWD	0.773±0.012	0.011±0.005	0.469±0.038	-
100	More ($5 \times 2 \times 2$)	M-SDWD (R=2)	0.853±0.021	0.144±0.023	0.764±0.046	0.702±0.051
		M-SDWD ($\lambda_1 = 0$, R=2)	0.791±0.026	0.156±0.023	1.000±0.000	0.000±0.000
		M-SDWD (R=1)	**0.862±0.022**	**0.141±0.020**	0.735±0.052	0.763±0.059
		M-SDWD ($\lambda_1 = 0$, R=1)	0.840±0.022	0.148±0.021	1.000±0.000	0.000±0.000
		M-DWD	0.834±0.031	0.148±0.022	1.000±0.000	0.000±0.000
		Full SDWD	0.798±0.027	0.160±0.025	0.395±0.037	0.934±0.025
	Less ($5 \times 4 \times 5$)	M-SDWD (R=2)	**0.941±0.009**	**0.027±0.007**	0.858±0.029	0.410±0.050
		M-SDWD ($\lambda_1 = 0$, R=2)	0.938±0.009	0.027±0.007	1.000±0.000	0.000±0.000
		M-SDWD (R=1)	0.860±0.012	0.030±0.007	0.772±0.033	0.518±0.053
		M-SDWD ($\lambda_1 = 0$, R=1)	0.862±0.012	0.031±0.007	1.000±0.000	0.000±0.000
		M-DWD	0.868±0.012	0.031±0.007	1.000±0.000	0.000±0.000
		Full SDWD	0.838±0.011	0.037±0.009	0.371±0.025	0.893±0.019
	No ($15 \times 4 \times 5$)	M-SDWD (R=2)	0.976±0.004	**0.004±0.003**	0.964±0.012	-
		M-SDWD ($\lambda_1 = 0$, R=2)	**0.979±0.004**	**0.004±0.003**	1.000±0.000	-
		M-SDWD (R=1)	0.849±0.013	0.007±0.003	0.880±0.026	-
		M-SDWD ($\lambda_1 = 0$, R=1)	0.854±0.012	0.007±0.003	1.000±0.000	-
		M-DWD	0.846±0.015	0.007±0.003	1.000±0.000	-
		Full SDWD	0.860±0.009	0.009±0.005	0.543±0.037	-

DWD approaches are more prone to convergence to a local optima with high levels of correlation. Table S11 shows the results for those replications that converge to a reasonable

Table S8: Simulation results when the true model is rank-5 (R=5) under high and low dimensional scenarios and sample size $N = 40$. In the Sparsity column, the numbers in parentheses indicate the number of non-zero variables in each dimension. "Cor" is the correlation between the estimated linear hyperplane and the true hyperplane. "Mis" is the average misclassification rate. "TP" is the true positive rate, i.e., the proportion of non-zero coefficients that are correctly estimated to be non-zero. "TN" is the true negative rate, i.e., the proportion of zero coefficients that are correctly estimated to be zero. The margins of error (2* standard errors across 200 replicates) for each statistic are also listed following the ± symbol.

Dimensions	Sparsity	Methods	Cor	Mis	TP	TN
High ($30 \times 15 \times 15$)	More ($5 \times 5 \times 5$)	M-SDWD (R=5)	**0.845±0.013**	**0.002±0.002**	0.927±0.016	0.623±0.048
		M-SDWD ($\lambda_1 = 0$,R=5)	0.774±0.017	0.007±0.004	1.000±0.000	0.000±0.000
		M-SDWD (R=1)	0.719±0.015	0.005±0.003	0.718±0.037	0.736±0.051
		M-SDWD ($\lambda_1 = 0$,R=1)	0.704±0.018	0.007±0.006	1.000±0.000	0.000±0.000
		M-DWD	0.702±0.020	0.010±0.008	1.000±0.000	0.000±0.000
		Full SDWD	0.732±0.015	0.007±0.004	0.262±0.016	0.993±0.002
	Less ($10 \times 15 \times 15$)	M-SDWD (R=5)	0.965±0.006	0.000±0.000	0.965±0.014	0.374±0.062
		M-SDWD ($\lambda_1 = 0$,R=5)	**0.981±0.002**	0.000±0.000	1.000±0.000	0.000±0.000
		M-SDWD (R=1)	0.598±0.011	0.000±0.000	0.854±0.031	0.450±0.063
		M-SDWD ($\lambda_1 = 0$,R=1)	0.603±0.010	0.000±0.000	1.000±0.000	0.000±0.000
		M-DWD	0.574±0.013	0.000±0.000	1.000±0.000	0.000±0.000
		Full SDWD	0.813±0.005	0.000±0.000	0.420±0.024	0.906±0.020
	No ($30 \times 15 \times 15$)	M-SDWD (R=5)	0.976±0.003	0.000±0.000	1.000±0.000	-
		M-SDWD ($\lambda_1 = 0$,R=5)	**0.981±0.003**	0.000±0.000	1.000±0.000	-
		M-SDWD (R=1)	0.573±0.010	0.000±0.000	0.794±0.037	-
		M-SDWD ($\lambda_1 = 0$,R=1)	0.590±0.008	0.000±0.000	1.000±0.000	-
		M-DWD	0.557±0.012	0.000±0.000	1.000±0.000	-
		Full SDWD	0.828±0.006	0.000±0.000	0.556±0.037	-
Low ($15 \times 4 \times 5$)	More ($5 \times 2 \times 2$)	M-SDWD (R=5	0.733±0.021	0.128±0.022	0.816±0.042	0.472±0.045
		M-SDWD ($\lambda_1 = 0$,R=5)	0.664±0.022	0.120±0.017	1.000±0.000	0.000±0.000
		M-SDWD (R=1)	**0.755±0.022**	**0.093±0.015**	0.671±0.047	0.703±0.053
		M-SDWD ($\lambda_1 = 0$,R=1)	0.731±0.024	0.103±0.016	1.000±0.000	0.000±0.000
		M-DWD	0.739±0.026	0.102±0.017	1.000±0.000	0.000±0.000
		Full SDWD	0.729±0.023	0.096±0.016	0.442±0.025	0.909±0.023
	Less ($5 \times 4 \times 5$)	M-SDWD (R=5)	**0.892±0.011**	0.010±0.007	0.922±0.021	0.269±0.045
		M-SDWD ($\lambda_1 = 0$, R=5)	0.889±0.010	**0.005±0.004**	1.000±0.000	0.000±0.000
		M-SDWD (R=1)	0.717±0.015	0.010±0.004	0.739±0.040	0.515±0.057
		M-SDWD ($\lambda_1 = 0$,R=1)	0.722±0.014	0.011±0.004	1.000±0.000	0.000±0.000
		M-DWD	0.717±0.015	0.010±0.004	1.000±0.000	0.000±0.000
		Full SDWD	0.795±0.010	0.008±0.004	0.468±0.031	0.841±0.029
	No ($15 \times 4 \times 5$)	M-SDWD (R=5)	0.857±0.009	0.000±0.000	0.950±0.018	-
		M-SDWD ($\lambda_1 = 0$, R=5)	**0.864±0.007**	0.000±0.000	1.000±0.000	-
		M-SDWD (R=1)	0.693±0.013	0.000±0.000	0.836±0.033	-
		M-SDWD ($\lambda_1 = 0$,R=1)	0.701±0.012	0.000±0.000	1.000±0.000	-
		M-DWD	0.687±0.014	0.000±0.000	1.000±0.000	-
		Full SDWD	0.841±0.005	0.000±0.000	0.573±0.035	-

solution (with correlation greater than 0.5 with the true vector), and the multiway DWD approaches perform better under all levels of residual correlation ρ for these replications.

Table S9: Mean correlations between the estimated linear hyperplane and the true hyperplane by assumed rank $\hat{R}$ for different true ranks R for N=40 under the low dimensional scenario $15 \times 4 \times 5$ with less sparsity ($5 \times 4 \times 5$ of variables has signals have signal discriminating the classes). The mean correlation when the rank is selected via cross validation is shown under $\hat{R}_{\text{CV}}$.

	$\hat{R}=1$	$\hat{R}=2$	$\hat{R}=3$	$\hat{R}=4$	$\hat{R}=5$	$\hat{R}_{\text{CV}}$
$R=1$	0.753	0.705	0.657	0.664	0.639	0.822
$R=2$	0.809	0.855	0.837	0.829	0.821	0.896
$R=3$	0.777	0.858	0.858	0.858	0.853	0.914
$R=4$	0.737	0.845	0.871	0.865	0.868	0.918
$R=5$	0.719	0.836	0.875	0.880	0.887	0.919

S8 Simulation with component-wise sparsity

We conducted a simulation study to illustrate recovery of a higher rank signal in which each rank-1 component has a different sparsity structure. Such scenarios are plausible for our motivating applications. For example, if different subsets of metabolites discriminate sample groups in different brain regions, that is efficiently captured by multiple sparse rank-1 components.

We simulate data under a moderate dimensional scenario with $P_1 = 20$, $P_2 = 4$, and $P_3 = 4$, and sample sizes $N_0 = N_1 = 50$. For the N_0 samples corresponding to class -1, the entries of $\mathbb{X}_i$ were generated independently from a $N(0,1)$ distribution. For the other N_1 samples corresponding to class 1, the entries of $\mathbb{X}_i$ were generated independently from a normal distribution with variance 1 and the mean for each entry given by the rank-2 array $\sqrt{0.2}\mu_1$ where $\mu_1 = \mathbf{u}_{11} \circ \mathbf{u}_{12} \circ \mathbf{u}_{13} + \mathbf{u}_{21} \circ \mathbf{u}_{22} \circ \mathbf{u}_{23}$. The loadings for the first rank-1 component are

$$\mathbf{u}_{11}[i] = \begin{cases} 1 \text{ for } i = 1, \ldots, 10 \\ 0 \text{ for } i = 11, \ldots, 20 \end{cases}, \ \mathbf{u}_{12}[i] = \begin{cases} 1 \text{ for } i = 1, 2 \\ 0 \text{ for } i = 3, 4 \end{cases}, \ \mathbf{u}_{13}[i] = \begin{cases} 1 \text{ for } i = 1, 2 \\ 0 \text{ for } i = 3, 4 \end{cases}$$

and for the second rank-1 component are

$$\mathbf{u}_{21}[i] = \begin{cases} 0 \text{ for } i = 1, \ldots, 10 \\ 1 \text{ for } i = 11, \ldots, 20 \end{cases}, \ \mathbf{u}_{22}[i] = \begin{cases} 0 \text{ for } i = 1, 2 \\ 1 \text{ for } i = 3, 4 \end{cases}, \ \mathbf{u}_{23}[i] = \begin{cases} 0 \text{ for } i = 1, 2 \\ 1 \text{ for } i = 3, 4 \end{cases}.$$

Thus, every dimension of every mode has some discriminating signal, but the two rank-1 components that make up the signal have separate non-zero support that does not overlap.

Table S12 shows the performance of different methods over 100 replications of this simulation scenario. As expected, the sparse ($\lambda_1 > 0$) rank-2 model that best matches the data generation tends to have the performance in terms of test misclassification rate and correlation with the true mean difference. The overall sparsity structure is also recovered relatively well under this approach, as shown by the true positive and true

Table S10: Simulation results under the high dimensional ($30 \times 15 \times 15$), more sparsity, small sample size (N=40) with different residual correlation levels ρ. "Cor" is the correlation between the estimated linear hyperplane and the true hyperplane. "Mis" is the average misclassification rate. "TP" is the true positive rate, i.e., the proportion of non-zero coefficients that are correctly estimated to be non-zero. "TN" is the true negative rate, i.e., the proportion of zero coefficients that are correctly estimated to be zero. The margins of error (2* standard errors across 200 replicates) for each statistic are also listed following the ± symbol.

ρ	Methods	Cor	Mis	TP	TN
0	M-SDWD	**0.562±0.061**	**0.204±0.033**	0.523±0.049	0.724±0.051
	M-SDWD ($\lambda_1 = 0$)	0.519±0.058	0.205±0.031	1.000±0.000	0.000±0.000
	M-DWD	0.516±0.060	0.226±0.034	1.000±0.000	0.000±0.000
	Full SDWD	0.347±0.037	0.263±0.027	0.197±0.033	0.884±0.034
	CATCH	0.455±0.043	0.208±0.027	0.045±0.006	0.999±0.000
0.3	M-SDWD	**0.454±0.062**	0.255±0.033	0.576±0.052	0.649±0.058
	M-SDWD ($\lambda_1 = 0$)	0.436±0.059	0.260±0.032	1.000±0.000	0.000±0.000
	M-DWD	0.310±0.058	0.337±0.032	1.000±0.000	0.000±0.000
	Full SDWD	0.355±0.036	0.253±0.026	0.193±0.032	0.884±0.034
	CATCH	0.443±0.041	**0.221±0.029**	0.042±0.006	0.999±0.000
0.6	M-SDWD	0.323±0.062	0.325±0.035	0.456±0.057	0.680±0.058
	M-SDWD ($\lambda_1 = 0$)	0.292±0.060	0.356±0.033	1.000±0.000	0.000±0.000
	M-DWD	0.067±0.032	0.474±0.019	1.000±0.000	0.000±0.000
	Full SDWD	0.375±0.039	0.267±0.029	0.177±0.034	0.900±0.035
	CATCH	**0.413±0.041**	**0.220±0.029**	0.041±0.006	0.999±0.000
0.9	M-SDWD	0.434±0.069	0.256±0.040	0.393±0.060	0.791±0.054
	M-SDWD ($\lambda_1 = 0$)	0.427±0.072	0.288±0.040	1.000±0.000	0.000±0.000
	M-DWD	0.044±0.027	0.477±0.017	1.000±0.000	0.000±0.000
	Full SDWD	**0.561±0.043**	0.177±0.032	0.193±0.040	0.897±0.043
	CATCH	0.497±0.036	**0.129±0.029**	0.071±0.007	0.999±0.000

negative rates. We also assess how well the two different rank-1 components are identified under the sparse rank-2 model. If necessary we permute the order of the estimated rank-1 components so that the "first" estimated component has the highest correlation with the true first component. The mean correlation of each estimated component with the true component $\text{cor}(\mathbf{u}_{i1} \circ \mathbf{u}_{i2} \circ \mathbf{u}_{i3}, \hat{\mathbf{u}}_{i1} \circ \hat{\mathbf{u}}_{i2} \circ \hat{\mathbf{u}}_{i3})$ was 0.923, indicating that the two separate components are well-identified. Further, the mean true negative rate for identifying zero values in $\mathbf{u}_{i1} \circ \mathbf{u}_{i2} \circ \mathbf{u}_{i3}$ from $\hat{\mathbf{u}}_{i1} \circ \hat{\mathbf{u}}_{i2} \circ \hat{\mathbf{u}}_{i3}$ was 0.998 and the mean true positive rate was 0.706, demonstrating that the sparsity structure in each component can also be recovered fairly accurately.

Table S11: Simulation results among replications with "Cor">0.5 under the high dimensional ($30 \times 15 \times 15$), more sparsity, small sample size (N=40) with different residual correlation levels ρ. "Cor" is the correlation between the estimated linear hyperplane and the true hyperplane. "Mis" is the average misclassification rate. "TP" is the true positive rate, i.e., the proportion of non-zero coefficients that are correctly estimated to be non-zero."TN" is the true negative rate, i.e., the proportion of zero coefficients that are correctly estimated to be zero. The margins of error (2* standard error across 200 replicates) for each statistic are also listed following the ± symbol.

ρ	Methods	Cor	Mis	TP	TN	%Cor> 0.5
0	M-SDWD	**0.897±0.014**	**0.027±0.007**	0.582±0.050	0.832±0.051	0.62
	M-SDWD ($\lambda_1 = 0$)	0.841±0.018	0.036±0.010	1.000±0.000	0.000±0.000	0.60
	M-DWD	0.864±0.018	0.036±0.010	1.000±0.000	0.000±0.000	0.59
	Full SDWD	0.673±0.029	0.043±0.015	0.131±0.018	0.993±0.002	0.32
	CATCH	0.698±0.025	0.066±0.016	0.069±0.009	0.999±0.000	0.54
0.3	M-SDWD	**0.901±0.016**	**0.022±0.008**	0.601±0.058	0.840±0.056	0.49
	M-SDWD ($\lambda_1 = 0$)	0.850±0.021	0.036±0.011	1.000±0.000	0.000±0.000	0.49
	M-DWD	0.893±0.022	0.028±0.011	1.000±0.000	0.000±0.000	0.33
	Full SDWD	0.674±0.028	0.042±0.016	0.121±0.015	0.994±0.001	0.32
	CATCH	0.698±0.020	0.039±0.010	0.076±0.008	0.999±0.000	0.47
0.6	M-SDWD	0.903±0.022	**0.009±0.008**	0.468±0.079	0.912±0.057	0.35
	M-SDWD ($\lambda_1 = 0$)	0.906±0.023	0.026±0.015	1.000±0.000	0.000±0.000	0.30
	M-DWD	**0.956±0.032**	0.012±0.024	1.000±0.000	0.000±0.000	0.05
	Full SDWD	0.679±0.028	0.050±0.016	0.123±0.018	0.995±0.001	0.36
	CATCH	0.666±0.019	0.054±0.016	0.069±0.009	0.999±0.000	0.50
0.9	M-SDWD	0.877±0.026	**0.002±0.002**	0.475±0.075	0.934±0.042	0.49
	M-SDWD ($\lambda_1 = 0$)	**0.935±0.019**	0.018±0.013	1.000±0.000	0.000±0.000	0.45
	M-DWD	0.885±0.131	0.083±0.110	1.000±0.000	0.000±0.000	0.04
	Full SDWD	0.730±0.018	0.056±0.016	0.117±0.011	0.996±0.001	0.69
	CATCH	0.659±0.021	0.005±0.007	0.100±0.009	0.999±0.000	0.64

S9 Tuning parameters selection for MRS data application

Here we expand on the selection of the tuning parameters for the application of multiway sparse DWD to classify dox treated mice (dox group) and controls (no dox group) in Section 6.1 of the main article. We computed the predicted DWD scores for all subjects by 10-fold cross-validation for a grid of λ_1 $(0, 0.0001, 0.001, 0.005, 0.01, 0.025, 0.05)$ and $\lambda_2(0.25, 0.5, 0.75, 1, 3, 5)$, and select the best combinations of parameters with the maximum t-test statistics that testing the differences between the scores and the two classes. Table S13 shows the t-test statistics between dox and no-dox groups for a grid of λ_1 and λ_2. The best pair of tuning parameters are $\lambda_1 = 0.01$ and $\lambda_2 = 5$ with the maximum test statistic. Table S14 shows the misclassification rates that classify the TG-dox group and other mice. Using t-test statistic as index for selecting parameters gives a unique

Table S12: Simulation results under moderate dimensional (20×4×4) rank-2 scenario with component-wise sparsity more sparsity. "Cor" is the correlation between the estimated linear hyperplane and the true hyperplane. "Mis" is the average misclassification rate. "TP" is the true positive rate, i.e., the proportion of non-zero coefficients that are correctly estimated to be non-zero."TN" is the true negative rate, i.e., the proportion of zero coefficients that are correctly estimated to be zero. The margins of error (2* standard error across 200 replicates) for each statistic are also listed following the ± symbol.

Methods	Cor	Mis	TP	TN
M-SDWD ($R = 2$)	**0.909±0.004**	**0.034±0.003**	0.562±0.051	0.998±0.001
M-SDWD ($\lambda_1 = 0, R = 2$)	0.896±0.004	0.035±0.003	0.000±0.000	1.000±0.000
M-SDWD ($R = 1$)	0.605±0.004	0.097±0.004	0.709±0.049	0.593±0.026
M-SDWD ($\lambda_1 = 0, R = 1$)	0.596±0.004	0.099±0.004	0.000±0.000	1.000±0.000
Full SDWD	0.673±0.005	0.079±0.004	0.442±0.042	0.919±0.011
Full SDWD ($\lambda_1 = 0$)	0.664±0.004	0.081±0.004	0.000±0.000	1.000±0.000

optimal pair of parameters, which also located in the region of parameters with minimal misclassification rates.

Table S13: T-test statistics that test the differences of predicted DWD scores between dox and no-dox groups for a grid of λ_1 and λ_2

λ_2/λ_1	0	1e-04	0.001	0.005	0.01	0.025	0.05	0.1	0.25
0.25	2.098	2.105	2.194	2.537	2.789	2.759	2.823	3.189	3.549
0.50	1.904	1.906	1.932	2.060	2.303	2.712	2.778	2.948	3.090
0.75	1.556	1.564	1.644	1.950	2.350	2.678	2.663	2.658	3.020
1.00	2.676	2.677	2.684	2.694	2.762	3.032	3.024	3.069	3.455
3.00	3.251	3.253	3.266	3.308	3.377	3.483	3.525	3.573	**3.660**
5.00	3.172	3.172	3.180	3.216	3.264	3.327	3.384	3.511	3.427

Table S14: Misclassification rates that classify TG-dox and other groups for a grid of λ_1 and λ_2

λ_2/λ_1	0	1e-04	0.001	0.005	0.01	0.025	0.05	0.1	0.25
0.25	0.333	0.333	0.333	0.333	0.286	0.286	0.286	0.286	0.190
0.50	0.286	0.286	0.286	0.286	0.286	0.143	0.238	0.143	0.190
0.75	0.333	0.333	0.333	0.286	0.286	0.238	0.238	0.190	0.143
1.00	0.143	0.143	0.095	0.143	0.095	0.095	0.095	0.095	0.095
3.00	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	0.095	**0.048**
5.00	0.095	0.095	0.095	**0.048**	**0.048**	**0.048**	**0.048**	0.095	0.095

S10 Computing time for data applications

Table S15 gives the computing time of the M-SDWD approaches for the two data applications presented in main manuscript. These results were obtained on a 2017 Macbook Pro with a 2.9 GHz Intel Core i7 processor and 16 GB of RAM. An application of the algorithm with fixed tuning parameters is relatively quick (1 second), while estimation via cross-validation and multiple random initializations is more computationally intensive, especially for higher ranks.

Table S15: Computing time in seconds for the MRS and gene time course (GENE) applications, under different selected ranks and 10-fold cross-validation with 5 random initializations (CV multi-start) to select λ_1 and λ_2, and for a single run of the algorithm (Single-start) or with multiple initializations (Multi-start) using the selected parameters.

	MRS			**GENE**		
	R=1	R=2	R=3	R=1	R=2	R=3
Single start	0.5	0.3	0.5	0.2	1.6	2.7
Multi-start	0.9	1.0	2.0	0.7	4.3	6.9
CV multi-start	125.8	265.3	585.1	141.6	608.2	1103.0