Tensor classification methods for multi-way biological data

Eric F. Lock

University of Minnesota

Classification Society Annual Meeting
Rochester, NY, 06/16/2023

Binary classification

- Predict y_i from $\mathbf{x}_i$ for $i = 1, \ldots, n$
  - y_i is a binary outcome, $y_i \in \{0, 1\}$
  - $\mathbf{x}_i$ a vector of p features
- In molecular biology, typically $p >> n$
  - n individuals
  - y_i is health condition (yes/no)
  - p molecular features (proteins, metabolites, genes)

Binary classification

- Linear classification model: $\hat{y}_i = f(\mathbf{x}_i \cdot \mathbf{b})$
  - Logistic regression, Fisher's LDA, SVM, PLS-DA, etc.
- Distance weighted discrimination (DWD)
  - Marron et al., JASA, 2007
  - $\hat{y}_i = \text{sign}(\mathbf{x}_i \cdot \mathbf{b} + b_0)$, minimizing

$$\sum_{i=1}^{N} \frac{1}{r_i} + C\xi_i$$

$$r_i = y_i(\mathbf{x}_i^T \mathbf{b} + b_0) + \xi_i \geq 0,\ \xi_i \geq 0 \text{ for } i = 1, \ldots, N,\ ||\mathbf{b}|| \leq 1.$$

  - Performs well in settings where $p >> n$

Application 1: SCA1 classification

- Data for 40 individuals:
  - 16 patients with spinocerebellar ataxia type 1 (SCA1)
  - 24 controls
- Magnetic resonance spectroscopy (MRS)
  - Non-invasive MRI-based method to quantify neurochemicals
  - Quantifications available for 13 metabolites
  - Data from 3 brain regions:
    - Cerebellar vermis, pons, cerebellar hemispheres
- Classify SCA1 individuals using data from all regions

Matrix classification framework

- Predict y_i from $\mathbf{X}_i$ for $i = 1, \ldots, n$
  - y_i is a binary outcome
  - $\mathbf{X}_i : P_1 \times P_2$ is a matrix of predictors
- Linear classification model: $\hat{y}_i = f(\mathbf{X}_i \cdot \mathbf{B})$
- Rank 1: $\mathbf{B} = \mathbf{u}_1 \mathbf{u}_2{}^T$ where $\mathbf{u}_1 : P_1 \times 1$ and $\mathbf{u}_2 : P_2 \times 1$:

$$\mathbf{X}_i \cdot \mathbf{B} = \sum_{p_1=1}^{P_1} \sum_{p_2=1}^{P_2} \mathbf{u}_1[p_1] \cdot \mathbf{u}_2[p_2] \cdot \mathbf{X}_i[p_1, p_2]$$

$$= \sum_{p_1=1}^{P_1} \mathbf{u}_1[p_1] \left(\sum_{p_2=1}^{P_2} \mathbf{u}_2[p_2] \cdot \mathbf{X}_i[p1, p2] \right)$$

$$= \sum_{p_2=1}^{P_2} \mathbf{u}_2[p_2] \left(\sum_{p_1=1}^{P_1} \mathbf{u}_1[p_1] \cdot \mathbf{X}_i[p1, p2] \right)$$

Multi-way classification framework

- Rank R: $\mathbf{B} = \mathbf{U}_1\mathbf{U}_2^T$ where $\mathbf{U}_1 : P_1 \times R$ and $\mathbf{U}_2 : P_2 \times R$:

$$\mathbf{X}_i \cdot \mathbf{B} = \sum_{r=1}^{R} \sum_{p_1=1}^{P_1} \sum_{p_2=1}^{P_2} \mathbf{U}_1[p_1, r] \cdot \mathbf{U}_2[p_2, r] \cdot \mathbf{X}_i[p_1, p_2]$$

- Iteratively solve objective $\mathrm{h}(\{y_i, \mathbf{X}_i \cdot \mathbf{B}\}_{i=1}^n)$
  - Update $\mathbf{U}_1$ with $\mathbf{U}_2$ fixed to minimize $h(\cdot)$
  - Update $\mathbf{U}_2$ with $\mathbf{U}_1$ fixed to minimize $h(\cdot)$

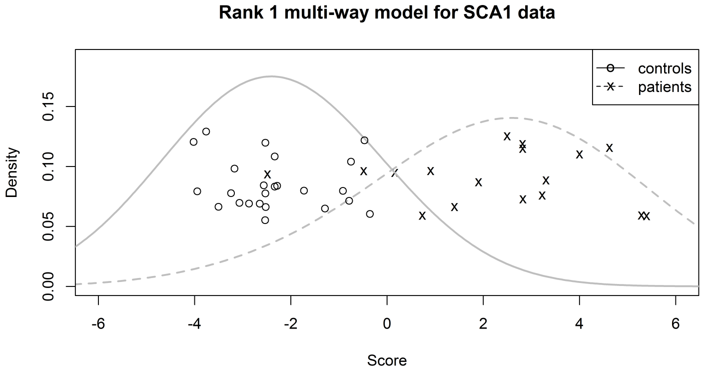
- We will use the DWD objective for $\mathrm{h}(\cdot)$

Multi-way classification framework

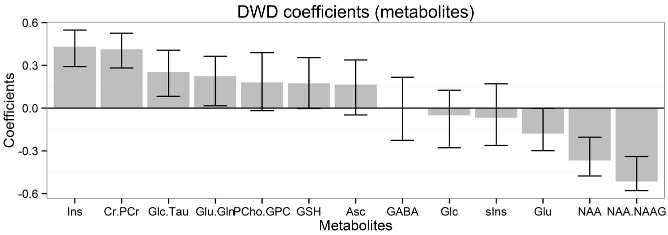
- Parameters for rank R model: $R(P_1 + P_2)$
- Parameters for "full" model: $P_1 P_2$
- Estimate rank R via cross-validation
- Assess uncertainty in coefficient estimates via bootstrapping

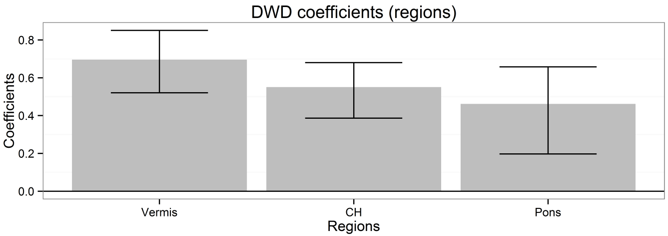
Application 1: Results

- Cross-validation scores for ataxia and control individuals:



Application 1: Results





Application 2: FRDA mouse model

- Transgenic (TG) mouse model:
  - Dose of Doxycycline (Dox) knocks out fratraxin protein
  - Mimics neurodegenerative condition Friedrich's Ataxia (FRDA)
- Data for 21 mice
  - 11 given DOX (6 TG, 5 WT)
  - 10 controls (5 TG, 5 WT)
- MRS data available for
  - 13 metabolites
  - 3 regions (cerebellum, cortex, spine)
  - 3 time points (0 weeks, 12 weeks, 24 weeks post)
- Distinguish mice given Dox from control mice

Multi-way classification framework

- Predict y_i from $\mathbb{X}_i$ for $i = 1, \ldots, n$
  - y_i is a binary outcome
  - $\mathbb{X}_i : P_1 \times P_2 \times \cdots \times P_K$ is a tensor of predictors
- Linear classification model: $y_i = f(\mathbb{X}_i \cdot \mathbb{B})$
- Rank-1 restriction: $\mathbb{B} = \mathbf{u}_1 \circ \mathbf{u}_2 \circ \cdots \circ \mathbf{u}_K$

$$\mathbb{X}_i \cdot \mathbb{B} = \sum_{i_1=1}^{P_1} \cdots \sum_{i_K=1}^{P_K} \mathbb{X}[i, i_1, \cdots, i_K] \mathbf{u}_1[i_1] \mathbf{u}_2[i_2] \cdots \mathbf{u}_K[i_K].$$

- Rank R Candecomp/Parafac (CP) model:

$$\mathbb{B} = [[\mathbf{U}_1, \ldots, \mathbf{U}_K]] = \sum_{r=1}^{R} \mathbf{u}_{1r} \circ \cdots \circ \mathbf{u}_{Kr}$$

where $\mathbf{U}_k : P_k \times R$ for $k = 1, \ldots, K$

Sparse (vector) DWD

- DWD objective can be expressed as [Liu et al., JASA, 2011]

$$h(\mathbf{y}, \mathbf{x}; \mathbf{b}, b_0) = \frac{1}{N}\sum_{i=1}^{N} V(\mathbf{y}_i(b_0 + \mathbf{x}\cdot\mathbf{b})) + \frac{\lambda_2}{2}||\mathbf{b}||_2^2,$$

where

$$V(u) = \left\{ \begin{array}{ll} 1-u, & \text{if } u \leq 1/2 \\ 1/(4u), & \text{if } u > 1/2. \end{array} \right\}$$

- Modify to induce sparsity [Wang & Zou, JCGS, 2016]:

$$\frac{1}{N}\sum_{i=1}^{N} V(\mathbf{y}_i(b_0 + \mathbf{x}\cdot\mathbf{b})) + \lambda_1||\mathbf{b}||_1 + \frac{\lambda_2}{2}||\mathbf{b}||_2^2$$

Sparse multiway DWD

- Sparse multiway DWD objective function:

$$h(\mathbf{y}, \mathbb{X}; \mathbb{B}, b_0) = \frac{1}{N}\sum_{i=1}^{N} V\left(\mathbf{y}_i(b_0 + \mathbb{X}_i \cdot \mathbb{B})\right) + P_{\lambda_1,\lambda_2}(\mathbb{B}),$$

where where $\mathbb{B} = \sum_{r=1}^{R} \mathbf{u}_{1r} \circ \cdots \circ \mathbf{u}_{Kr}$ and

$$P_{\lambda_1,\lambda_2}(\mathbb{B}) = \lambda_1 \sum_{r=1}^{R}\prod_{k=1}^{K} \|\mathbf{u}_{kr}\|_1 + \frac{\lambda_2}{2}\|\mathbb{B}\|_2^2.$$

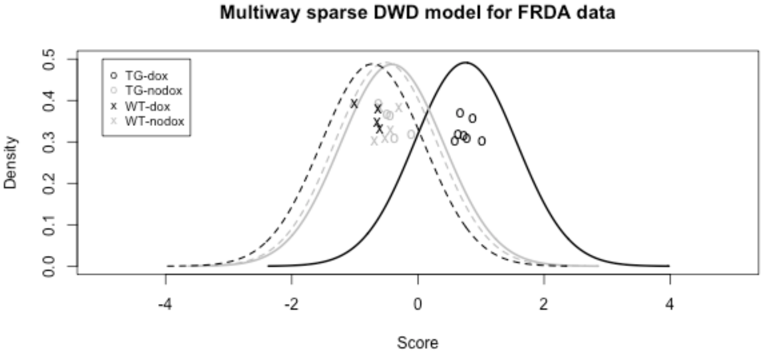
- Solved by a majorization-minimization (MM) algorithm.

Sparse multiway DWD

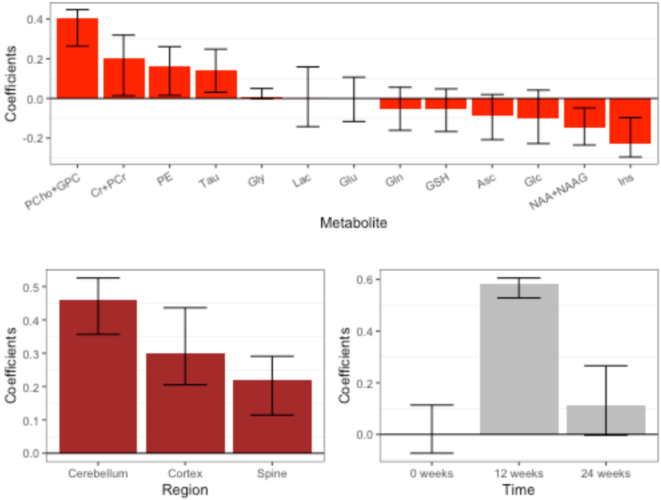
- Potential local optima
  - Multiple initialization and pruning
  - Tempering of regularization parameters λ_1, λ_2
- Estimate R, λ_1, λ_2 via double cross-validation
- Bootstrapping to assess uncertainty

Application 2: Results

- Cross-validation scores for ataxia and control individuals:



Application 2: Results



Application 3: ID monkey model

- Prospective cohort of 12 infant rhesus monkeys
  - 6 develop iron deficiency (ID) anemia after 6 months
  - 6 remain iron sufficient (IS)
- Serum 'omics measurements taken at 4 months and 6 months
- Proteomics (205 proteins)
- metabolomics (238 metabolites)
- Distinguish ID monkeys from IS monkeys

Multi-source multi-way classification framework

- Predict y_i from $\mathbf{X}_i$ for $i = 1, \ldots, n$
  - y_i is a binary outcome
  - $\mathbf{X}_i : [P^{[1]}, \ldots, P^{[M]}] \times D$ is a matrix with M data sources and shared dimension D
- Linear classification model: $y_i = f(\mathbf{X}_i \cdot \mathbf{B})$
- Rank R model:

$$\mathbf{B} = \mathbf{U}\mathbf{V}^T = [\mathbf{U}^{[1]}, ..., \mathbf{U}^{[M]}]\mathbf{V}^T$$

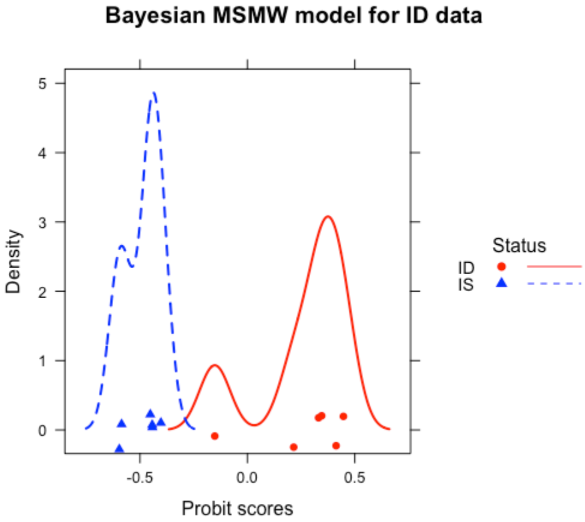
where each $\mathbf{U}^{[m]} : P^{[m]} \times R$ and $\mathbf{V} : D \times R$

Bayesian probit regression model

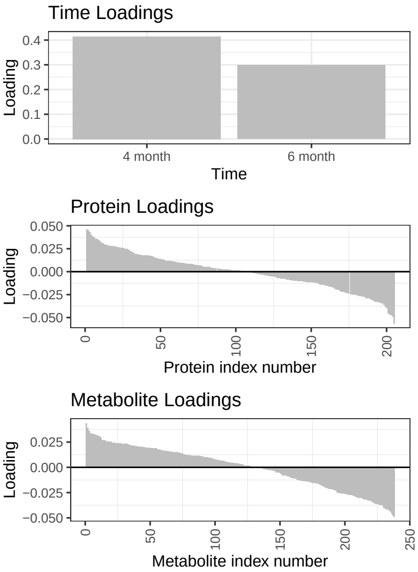
- $Pr(y_i = 1|\mathbf{X}_i) = \Phi(\mathbf{X}_i \cdot \mathbf{B})$, where Φ is cdf for $N(0,1)$
- $\mathbf{U}^{[m]} \overset{iid}{\sim} N(0, \tau_m)$ for $m = 1, \ldots, M$
- $\mathbf{V} \overset{iid}{\sim} N(0,1)$
- $\tau_m \sim IG(\alpha_0, \beta_0)$
- Infer full posterior via Gibbs sampling

Application 3: Results

- Scores (leave-one-out CV) for ID and IS monkeys



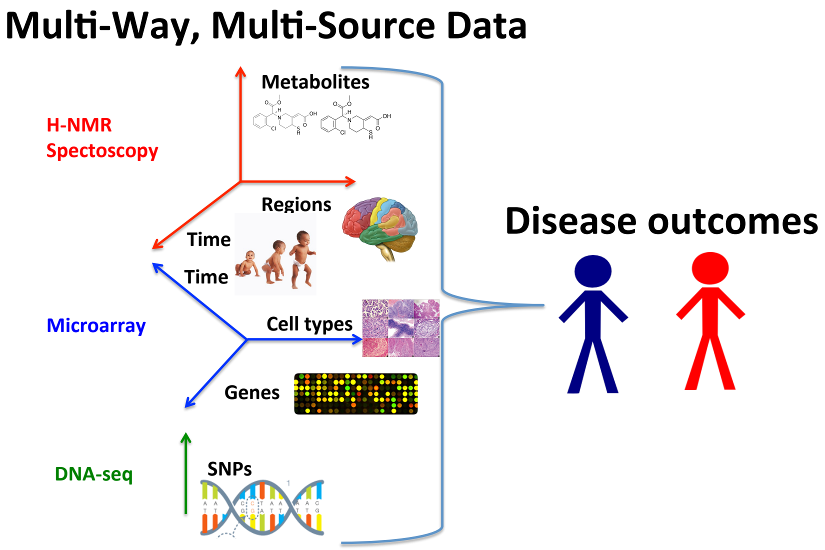
Application 3: Results



Application 3: further multi-source multi-way data

- Available for a larger cohort of IS/ID monkeys:
  - Proteomic data at multiple timepoints, in CSF & serum:
    Proteins $\times$ *Time* $\times$ *Fluid*
  - Metabolite data at the same timepoints, in CSF & serum
    Metabolites $\times$ *Time* $\times$ *Fluid*
  - Hematology panel at the same timepoints :
    Hematology parameters $\times$ *Time*
  - Cross-sectional brain MRI:
    $X \times Y \times Z$

- Distinguish IS from ID monkeys using available data.
- Work in progress...


Multi-Way, Multi-Source Data
Metabolites
H-NMR
Spectoscopy
Regions
Time
Time
Microarray
Cell types
Genes
DNA-seq
SNPs
Disease outcomes

Software

- R package `MultiwayClassification`
  - https://github.com/lockEF/MultiwayClassification
  - Performs multiway DWD, multiway SVM, and multiway sparse DWD
- R implementation `BayesMSMW`
  - https://github.com/BiostatsKim/BayesMSMW
  - Performs Bayesian multi-source multi-way classification

Thank you!

- Support: NIH NIGMS R01-GM130622
- Application 1 (Multiway DWD):
  - T Lyu, EF Lock, and LE Eberly. Discriminating sample groups with multi-way data. *Biostatistics*, 18 (3): 434-450, 2017.
  - Methods collaborators: Lynn Eberly and Tianmeng Lyu
  - Data: Dinesh Deelchand & Gulin Oz, UMN CMRR
- Application 2 (Multiway Sparse DWD)
  - B Guo, LE Eberly, PG Henry, C Lenglet, and EF Lock. Multiway sparse distance weighted discrimination. *JCGS*, 33 (2): 730-743, 2023.
  - Methods collaborators: Lynn Eberly and Bin Guo
  - Data: Pierre-Gilles Henry & Christophe Lenglet, UMN CMRR
- Application 3 (Bayesian multi-source multi-way classification)
  - J Kim, BJ Sandri, RB Rao, EF Lock. Bayesian predictive modeling of multi-source multi-way data. *CSDA*, 186: 107783, 2023.
  - Methods collaborator: Jonathan Kim
  - Data: Raghu Rao and Brian Sandri, UMN Pediatrics