

Unsupervised learning for data integration studies

Benjamin B. Risk

Department of Biostatistics & Bioinformatics
Rollins School of Public Health
Emory University

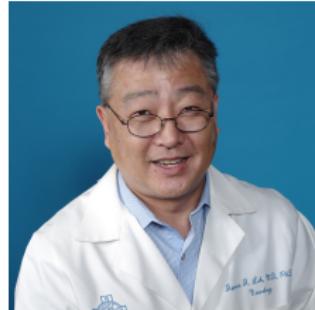
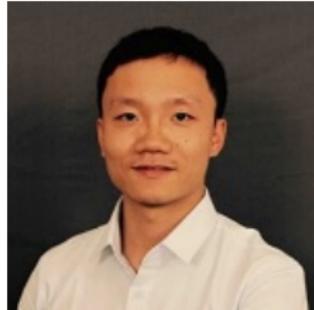
benjamin.risk@emory.edu

github.com/thebrisklab



Joint work

Ganzhong Gavin Tian (PhD student in Biostatistics, Emory, now at Amazon), Raphiel Murden (Assistant Professor of Biostatistics, Emory), James Lah (MD, Department of Neurology, Emory University School of Medicine), John Hanfelt (Professor in Biostatistics, Emory) and Benjamin Risk.



Introduction: AD Prevalence

Alzheimer's Disease (AD):

The most common type of dementia, accounting for 60% – 80% of cases.

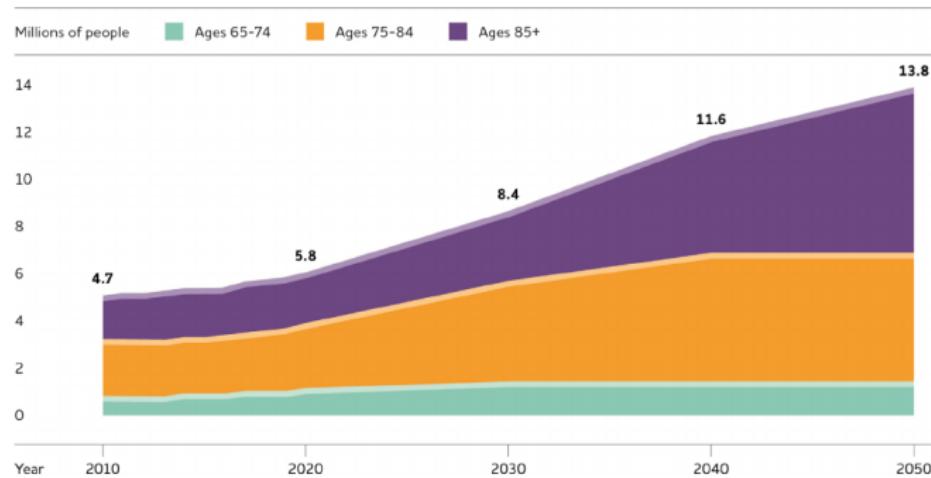


Figure: Projected number of people in the U.S. population with Alzheimer's (Report, 2020)

Introduction: AD Continuum

Alzheimer's Disease (AD):

Slowly destroys memory, thinking skills, and ability to carry out basic functions.

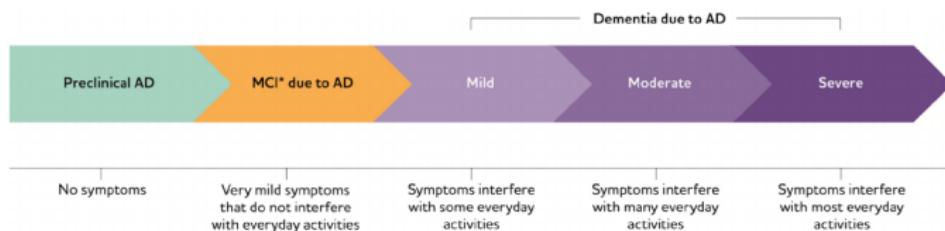


Figure: Alzheimer's disease (AD) continuum. *MCI is the acronym for mild cognitive impairment Report (2020)

Preclinical AD: No cognitive symptoms. Long-term goal: identify individuals likely to develop AD to implement preventative treatment programs. E.g., use cerebrospinal fluid (CSF) biomarkers (amyloid beta and tau) and MRI.

Introduction: No gold standard for AD

No definitive diagnosis: A definitive diagnosis of Alzheimer's disease (AD) is only possible from a brain tissue autopsy (Dubois et al., 2007).

Clinical diagnosis: In Alzheimer's Disease Neuroimaging Initiative (ADNI), primarily neurocognitive assessment; more generally, complicated combination of cognitive exams, brain imaging, CSF biomarkers.

Low amyloid beta and high tau protein CSF biomarker associated with AD. (Report, 2020; Meyer et al., 2010)).

MRI brain morphometry: low cortical thickness and large ventricles associated with brain deterioration.

Unsupervised methods: Clustering can reveal new insights when gold standard not available (Collins and Huynh, 2014; Meyer et al., 2010). We focus on integrating CSF biomarkers and brain morphometry from MRIs.

Background: Model-based clustering

Finite Mixture Models (FMM): probabilistic clustering to find latent classes (McLachlan and Peel, 2000).

Gaussian Mixture Models (GMM) are a popular approach to modeling the dependence structure among features (manifest variables) (Scrucca et al., 2016a).

Integrating multiple datasets of possibly high dimension is challenging. Two datasets, $\mathbf{y}_{i1} \in \mathbb{R}^{p_1}$ and $\mathbf{y}_{i2} \in \mathbb{R}^{p_2}$, then a $(p_1 + p_2) \times (p_1 + p_2)$ covariance matrix has to be estimated for each cluster. Can not estimate without additional constraints.

Background: M-Clust and covariance structures

Estimating covariance matrices in GMMs is challenging.

Model	Σ_k	Distribution	Volume	Shape	Orientation
EII	λI	Spherical	Equal	Equal	—
VII	$\lambda_k I$	Spherical	Variable	Equal	—
EEI	λA	Diagonal	Equal	Equal	Coordinate axes
VEI	$\lambda_k A$	Diagonal	Variable	Equal	Coordinate axes
EVI	λA_k	Diagonal	Equal	Variable	Coordinate axes
VVI	$\lambda_k A_k$	Diagonal	Variable	Variable	Coordinate axes
EEE	λDAD^\top	Ellipsoidal	Equal	Equal	Equal
EVE	$\lambda DA_k D^\top$	Ellipsoidal	Equal	Variable	Equal
VEE	$\lambda_k DAD^\top$	Ellipsoidal	Variable	Equal	Equal
VVE	$\lambda_k DA_k D^\top$	Ellipsoidal	Variable	Variable	Equal
EEV	$\lambda D_k AD_k^\top$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_k D_k AD_k^\top$	Ellipsoidal	Variable	Equal	Variable
EVV	$\lambda D_k A_k D_k^\top$	Ellipsoidal	Equal	Variable	Variable
VVV	$\lambda_k D_k A_k D_k^\top$	Ellipsoidal	Variable	Variable	Variable

Table 3: Parameterisations of the within-group covariance matrix Σ_k for multidimensional data available in the **mclust** package, and the corresponding geometric characteristics.

Figure: Covariance structure specification in the R package *mclust* (Scrucca et al., 2016b)

Background: Data integration methods

Setting: same subjects in multiple datasets.

Canonical-correlation analysis (CCA) is a popular method to find maximally correlated latent variables from two datasets (Hotelling, 1992).

Recently, methods such as Joint and Individual Variance Explained (JIVE) have been proposed that also extract latent factors unique to each of the datasets (Lock et al., 2013; Feng et al., 2018).

Joint and Individual Clustering (JIC) applies k-means to the joint and individual latent scores to find subgroups (Hellton and Thoresen, 2016).

We develop a statistical model for simultaneous clustering and dimension reduction from multiple datasets (ADNI dataset with CSF ($p_1 = 3$) and MRI biomarkers ($p_2 = 245$)).

Scalable covariance matrix estimation.

Background: Motivation

Probabilistic Principal Component Analysis (PPCA) formulates a probabilistic model for PCA (Tipping and Bishop, 1999b).

Extended PPCA to Mixture of PPCA (MixPPCA) to perform clustering and dimension reduction (Tipping and Bishop, 1999a).

Murden et al developed probabilistic JIVE (ProJIVE) as a statistical model-based extension of JIVE for data integration.

We propose a **Mixture of Probabilistic JIVE (ProJIVE-Mix)** for simultaneous clustering and dimension reduction with a computationally feasible EM algorithm.

First, background on ProJIVE

Probabilistic JIVE (Murden et al. in prep):

$$\mathbf{y}_{ik} = \boldsymbol{\mu}_k + \mathbf{W}_{Jk}\mathbf{a}_i + \mathbf{W}_{Ik}\mathbf{b}_{ik} + \epsilon_{ik}, \quad i \in \{1, \dots, n\}, \quad k \in \{1, \dots, K\}, \quad (1)$$

$i = 1, \dots, n$ participant,

$k = 1, \dots, K$ feature block (dataset),

$\mathbf{W}_{Jk} \in \mathbb{R}^{p_k \times q_J}$ and $\mathbf{W}_{Ik} \in \mathbb{R}^{p_k \times q_k}$ (with $q_J + q_k < p_k$) are full-rank loading matrices,

$\mathbf{a}_i \in \mathbb{R}^{q_J}$, $\mathbf{b}_{ik} \in \mathbb{R}^{q_k}$ are **joint** and **individual latent scores** for the i th observation, $(\mathbf{a}_i^\top, \mathbf{b}_{i1}^\top, \mathbf{b}_{i2}^\top, \dots, \mathbf{b}_{iK}^\top)^\top \stackrel{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_{q \times q})$

$\epsilon_{ik} \stackrel{iid}{\sim} \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{I}_{p_k \times p_k})$ are noise,

$\epsilon_{ik} \perp (\mathbf{a}_i^\top, \mathbf{b}_{ik}^\top)^\top, \forall k, i.$

ProJIVE: A Generalization of PPCA

(1) can be written as a generalized form of PPCA:

$$\mathbf{y}_i = \begin{pmatrix} \mathbf{y}_{i1} \\ \vdots \\ \mathbf{y}_{iK} \end{pmatrix} = \mathbf{W}\boldsymbol{\theta}_i + \begin{pmatrix} \boldsymbol{\epsilon}_{i1} \\ \vdots \\ \boldsymbol{\epsilon}_{iK} \end{pmatrix} \quad (2)$$

where

$\boldsymbol{\theta}_i = (\mathbf{a}_i^\top, \mathbf{b}_{i1}^\top, \mathbf{b}_{i2}^\top, \dots, \mathbf{b}_{iK}^\top)^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{q \times q})$, and $q = q_J + \sum_{k=1}^K q_k$,

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{J1} & \mathbf{W}_{I1} & 0 & \dots & 0 \\ \mathbf{W}_{J2} & 0 & \mathbf{W}_{I2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{W}_{JK} & 0 & 0 & \dots & \mathbf{W}_{IK} \end{pmatrix} \begin{pmatrix} \mathbf{a}_i \\ \mathbf{b}_{i1} \\ \mathbf{b}_{i2} \\ \vdots \\ \mathbf{b}_{iK} \end{pmatrix}$$

Therefore, $\mathbf{y}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, where $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \mathbf{D}$,

$\mathbf{D} = \text{diag}(\sigma_1^2 \mathbf{1}_{p_1}^\top, \sigma_2^2 \mathbf{1}_{p_2}^\top, \dots, \sigma_K^2 \mathbf{1}_{p_K}^\top)$.

Our Model: Mixture of ProJIVE

Extend the ProJIVE model to a **finite mixture model** as:

$$\mathbf{y}_i|g = \begin{pmatrix} \mathbf{y}_{i1}|g \\ \vdots \\ \mathbf{y}_{iK}|g \end{pmatrix} = \boldsymbol{\mu}_g + \mathbf{W}_g \boldsymbol{\theta}_{ig} + \begin{pmatrix} \epsilon_{i1}|g \\ \vdots \\ \epsilon_{iK}|g \end{pmatrix}, \quad g \in \{1, \dots, G\} \quad (3)$$

where $\boldsymbol{\theta}_{ig} = (\mathbf{a}_{ig}^\top, \mathbf{b}_{i1g}^\top, \mathbf{b}_{i2g}^\top, \dots, \mathbf{b}_{iKg}^\top) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$,

$$\mathbf{W}_g = \begin{pmatrix} \mathbf{W}_{J1g} & \mathbf{W}_{I1g} & 0 & \dots & 0 \\ \mathbf{W}_{J2g} & 0 & \mathbf{W}_{I2g} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{W}_{JKg} & 0 & 0 & \dots & \mathbf{W}_{IKg} \end{pmatrix} \begin{pmatrix} \mathbf{a}_{ig} \\ \mathbf{b}_{i1g} \\ \mathbf{b}_{i2g} \\ \dots \\ \mathbf{b}_{iKg} \end{pmatrix}$$

Therefore, $\mathbf{y}_i|g \sim \mathcal{N}(\boldsymbol{\mu}_g, \mathbf{C}_g)$, where: $\mathbf{C}_g = \mathbf{W}_g \mathbf{W}_g^\top + \mathbf{D}_g$,

$\mathbf{D}_g = \text{diag}(\sigma_{g1}^2 \mathbf{1}_{p_1}^\top, \sigma_{g2}^2 \mathbf{1}_{p_2}^\top, \dots, \sigma_{gK}^2 \mathbf{1}_{p_K}^\top)$.

Our Model: Diagram

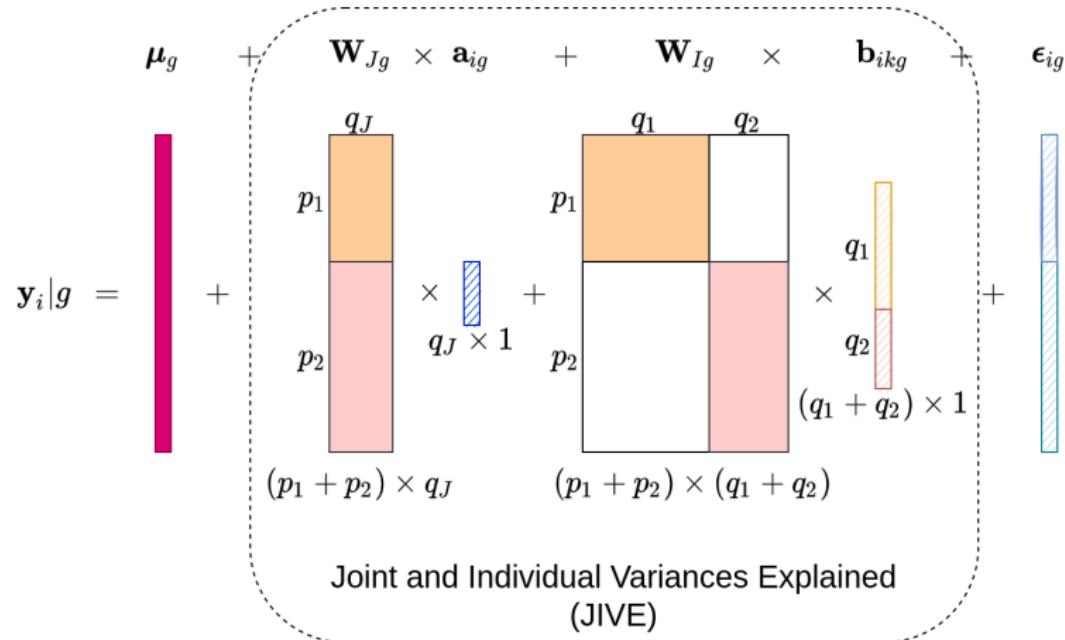


Figure: Diagram of the signal decomposition of $\mathbf{y}_i|g \sim \mathcal{N}(\mu_g, \mathbf{W}_g \mathbf{W}_g^\top + \mathbf{D}_g)$, in an example where we have 2 feature blocks of dimensions p_1 and p_2 .

Our model: Likelihood

Assuming latent class indicators \mathbf{Z} and scores $\boldsymbol{\Theta}$ are observed, the complete data log-likelihood function is:

$$\begin{aligned}\ell_c(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\Theta}; \boldsymbol{\Psi}) &= \sum_{i=1}^n \sum_{g=1}^G z_{ig} [\log \omega_g + \log f_g(\mathbf{y}_i, \boldsymbol{\theta}_{ig}; \boldsymbol{\psi}_g)] \\ &= \sum_{i=1}^n \sum_{g=1}^G z_{ig} \left\{ \log \omega_g - \frac{p+q}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{D}_g| \right. \\ &\quad \left. - \frac{1}{2} [(\mathbf{y}_i - \mathbf{W}_g \boldsymbol{\theta}_{ig} - \boldsymbol{\mu}_g)^\top \mathbf{D}_g^{-1} (\mathbf{y}_i - \mathbf{W}_g \boldsymbol{\theta}_{ig} - \boldsymbol{\mu}_g) \right. \\ &\quad \left. + \boldsymbol{\theta}_{ig}^\top \boldsymbol{\theta}_{ig}] \right\} \quad (4)\end{aligned}$$

where $f_g(\mathbf{y}_i, \boldsymbol{\theta}_{ig}; \boldsymbol{\psi}_g)$ is the joint multivariate Gaussian density.

$\langle z_{ig} \rangle \equiv \mathbb{E}_{\boldsymbol{\Psi}^{(\nu)}}(z_{ig} | \mathbf{y}_i)$ evaluated at the ν th iteration.

$\langle \boldsymbol{\theta}_{ig} \rangle \equiv \mathbb{E}_{\boldsymbol{\Psi}^{(\nu)}}(\boldsymbol{\theta}_{ig} | \mathbf{y}_i)$ and $\langle \boldsymbol{\theta}_{ig} \boldsymbol{\theta}_{ig}^\top \rangle \equiv \mathbb{E}_{\boldsymbol{\Psi}^{(\nu)}}(\boldsymbol{\theta}_{ig} \boldsymbol{\theta}_{ig}^\top | \mathbf{y}_i)$.

Our Model: Likelihood

Taking conditional expectation of the complete data log-likelihood w.r.t z_{ig} given \mathbf{y}_i , and $\boldsymbol{\theta}_{ig}$ given \mathbf{y}_i :

$$\begin{aligned} \mathcal{Q}_c(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(\nu)}) \propto & \sum_{i=1}^n \sum_{g=1}^G \langle z_{ig} \rangle \left\{ \log \omega_g - \frac{1}{2} \sum_{k=1}^K p_k \log(\sigma_{kg}^2) \right. \\ & - \frac{1}{2} \left[\sum_{k=1}^K \sigma_{kg}^{-2} (\mathbf{y}_{ik} - \boldsymbol{\mu}_{kg})^\top (\mathbf{y}_{ik} - \boldsymbol{\mu}_{kg}) \right. \\ & - 2 \sum_{k=1}^K \sigma_{kg}^{-2} \langle \boldsymbol{\theta}_{ikg} \rangle^\top \mathbf{W}_{kg}^\top (\mathbf{y}_{ik} - \boldsymbol{\mu}_{kg}) + \text{tr}(\langle \boldsymbol{\theta}_{ig} \boldsymbol{\theta}_{ig}^\top \rangle) \\ & \left. \left. + \sum_{k=1}^K \sigma_{kg}^{-2} \text{tr}(\mathbf{W}_{kg}^\top \mathbf{W}_{kg} \langle \boldsymbol{\theta}_{ikg} \boldsymbol{\theta}_{ikg}^\top \rangle) \right] \right\} \end{aligned} \quad (5)$$

Where $\boldsymbol{\Psi} = (\pi_1, \dots, \pi_{G-1}, \psi_1, \dots, \psi_G)$ denote the vector of parameters to be estimated; $\psi_g = (\boldsymbol{\mu}_g, \text{vec}(\mathbf{W}_g), \sigma_{g1}^2, \dots, \sigma_{gK}^2)$.

Our Model: EM algorithm E-step

E-step, update the conditional expectations:

$$\langle z_{ig} \rangle \equiv \mathbb{E}_{\Psi^{(\nu)}}(Z_{ig} | \mathbf{y}_i) = \frac{\omega_g^{(\nu)} \phi(\mathbf{y}_i; \boldsymbol{\psi}_g^{(\nu)})}{\sum_{h=1}^G \omega_h^{(\nu)} \phi(\mathbf{y}_i; \boldsymbol{\psi}_h^{(\nu)})} \quad (6)$$

$$\langle \boldsymbol{\theta}_{ikg} \rangle \equiv \mathbb{E}_{\Psi^{(\nu)}}(\boldsymbol{\theta}_{ikg} | \mathbf{y}_i) = \mathbf{B}_k \mathbf{M}_g^{(\nu)} \mathbf{W}_g^{(\nu)\top} \mathbf{D}_g^{(\nu)-1} (\mathbf{y}_i - \boldsymbol{\mu}_g^{(\nu)}) \quad (7)$$

$$\langle \boldsymbol{\theta}_{ikg} \boldsymbol{\theta}_{ikg}^\top \rangle \equiv \mathbb{E}_{\Psi^{(\nu)}}(\boldsymbol{\theta}_{ikg} \boldsymbol{\theta}_{ikg}^\top | \mathbf{y}_i) = \mathbf{B}_k \mathbf{M}_g^{(\nu)} \mathbf{B}_k^\top + \langle \boldsymbol{\theta}_{ikg} \rangle \langle \boldsymbol{\theta}_{ikg} \rangle^\top \quad (8)$$

where $\mathbf{M}_g = (\mathbf{I} - \mathbf{W}_g^\top \mathbf{D}_g \mathbf{W}_g)^{-1}$ are the conditional variances of $\boldsymbol{\theta}_{ig}$ conditioning on \mathbf{y}_i , given the g th cluster component; \mathbf{B}_k is a $(q_J + q_k) \times q$ selection matrix such that: $\boldsymbol{\theta}_{ikg} = \begin{pmatrix} \mathbf{a}_{ig} \\ \mathbf{b}_{ikg} \end{pmatrix} = \mathbf{B}_k \boldsymbol{\theta}_{ig}$. Similar, we can define \mathbf{A}_k as a $p_k \times p$ selection matrix such that $\mathbf{y}_{ik} = \mathbf{A}_k \mathbf{y}_i$

Our Model: EM algorithm M-step

M-step, by maximization of $\mathcal{Q}_c(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(\nu)})$:

$$\omega_g^{(\nu+1)} = \frac{\sum_{i=1}^n \langle z_{ig} \rangle}{n} \quad (9)$$

$$\boldsymbol{\mu}_{kg}^{(\nu+1)} = \mathbf{A}_k \frac{\sum_{i=1}^n \langle z_{ig} \rangle \mathbf{y}_i}{\sum_{i=1}^n \langle z_{ig} \rangle} \quad (10)$$

$$\mathbf{W}_{kg}^{(\nu+1)} = \left\{ \sum_{i=1}^n \langle z_{ig} \rangle (\mathbf{y}_{ik} - \boldsymbol{\mu}_{kg}^{(\nu+1)}) \langle \boldsymbol{\theta}_{ikg} \rangle^\top \right\} \left\{ \sum_{i=1}^n \langle z_{ig} \rangle \langle \boldsymbol{\theta}_{ikg} \boldsymbol{\theta}_{ikg}^\top \rangle \right\}^{-1} \quad (11)$$

$$\begin{aligned} \sigma_{kg}^{(2)} &= \frac{1}{p_k \sum_{i=1}^n \langle z_{ig} \rangle} \sum_{i=1}^n \langle z_{ig} \rangle [(\mathbf{y}_{ik} - \boldsymbol{\mu}_{kg}^{(\nu+1)})^\top (\mathbf{y}_{ik} - \boldsymbol{\mu}_{kg}^{(\nu+1)}) \\ &\quad - 2 \langle \boldsymbol{\theta}_{ikg} \rangle^\top \mathbf{W}_{kg}^\top (\mathbf{y}_{ik} - \boldsymbol{\mu}_{kg}^{(\nu+1)}) + \text{tr}(\mathbf{W}_{kg}^\top \mathbf{W}_{kg} \langle \boldsymbol{\theta}_{ikg} \boldsymbol{\theta}_{ikg}^\top \rangle)] \end{aligned} \quad (12)$$

Simulation Cases

Simulations generated from ProJIVE-Mix with $q_J = 1$, $q_1 = 1$, $q_2 = 1$, two datasets ($K = 2$) and two clusters ($G = 2$).

		Setting-1		Setting-2		Setting-3	
		$p_1 = 2, p_2 = 50$		$p_1 = 10, p_2 = 250$		$p_1 = 50, p_2 = 1,250$	
		$n = 1,000$		$n = 1,000$		$n = 1,000$	
Data Block		$k = 1$	$k = 2$	$k = 1$	$k = 2$	$k = 1$	$k = 2$
Cluster	$g = 1, \omega_1 = 0.3$	0.67(0.67,0.67)	0.31(0.29,0.33)	0.67(0.67,0.67)	0.31(0.29,0.33)	0.67(0.67,0.67)	0.31(0.29,0.333)
	$g = 2, \omega_2 = 0.7$	0.86(0.86,0.86)	0.46(0.43,0.48)	0.86(0.86,0.86)	0.46(0.43,0.48)	0.69(0.69,0.69)	0.32(0.30,0.34)

Table: Signal variance (proportions of the features' explainable variances) with differing number of features.

Bhattacharyya distance (D_B): measures the overlap of two distributions. This represents are moderately challenging setting.

Simulation Results

		Setting-1 $p_1 = 2, p_2 = 50$ $n = 1,000$ $D_B = 1.998$		Setting-2 $p_1 = 10, p_2 = 250$ $n = 1,000$ $D_B = 1.971$		Setting-3 $p_1 = 50, p_2 = 1,250$ $n = 1,000$ $D_B = 1.975$		
Methods	Train/Test	model	df	ARI	df	ARI	df	ARI
ProJIVE-Mix	Training	317		0.884(0.023)	1,565	0.835(0.022)	7,805	0.848(0.024)
	Testing	317		0.888(0.023)	1,565	0.809(0.027)	7,805	0.717(0.032)
MixPPCA	Training	419		0.759(0.048)	2,083	0.650(0.040)	10,403	0.710(0.030)
	Testing	419		0.785(0.034)	2,083	0.602(0.041)	10,403	0.388(0.034)
PCA+GMM	Training	461		0.850(0.027)	1,721	0.692(0.040)	10,301	0.991(0.009)
	Testing	461		0.851(0.024)	1,721	0.655(0.036)	10,301	0.630(0.034)
Full GMM	Training	2,861		0.896(0.026)	68,381	-	1,693,901	-
	Testing	2,861		0.614(0.046)	68,381	-	1,693,901	-
Diag GMM	Training	209		0.568(0.048)	1,041	0.216(0.032)	5,201	0.168(0.029)
	Testing	209		0.577(0.040)	1,041	0.219(0.033)	5,201	0.166(0.023)
JIC.joint	Training	-		0.589(0.038)	-	0.231(0.035)	-	0.284(0.039)
	Testing	-		0.592(0.036)	-	0.231(0.033)	-	0.281(0.035)
K-means	Training	105		0.514(0.045)	521	0.187(0.028)	2601	0.149(0.028)
	Testing	105		0.524(0.037)	521	0.191(0.030)	2601	0.148(0.022)

Table: ARI measured on a 1,000-point training set and 1,000-point independent test dataset, with mean \pm sd reported in 101 simulations.

Simulation Results

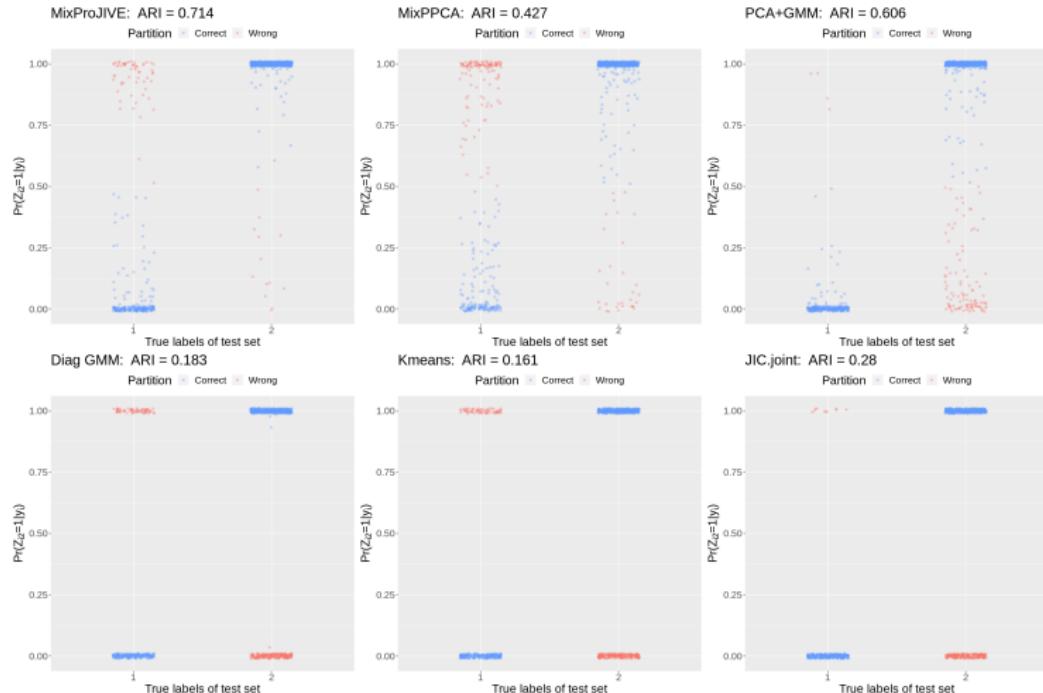


Figure: Scatter plot of posterior probabilities in a ‘median’ ARI case testing set, under the third simulated setting with high dimensions ($p_1 = 50, p_2 = 1250, n = 1,000$).

Real Data Analysis: ADNI CSF+MRI dataset

	Baseline Diagnosis:				P-value
	AD (N = 220)	MCI (N = 608)	CN (N = 360)	Total (N = 1,188)	
Abeta					< 0.001
- Mean (SD)	692.776 (420.054)	1013.824 (549.302)	1354.716 (649.508)	1057.671 (605.693)	
- Range	212.300 - 3139.000	210.900 - 3331.000	203.000 - 3592.000	203.000 - 3592.000	
ttau					< 0.001
- Mean (SD)	368.278 (145.529)	285.022 (128.964)	238.455 (89.113)	286.328 (129.440)	
- Range	133.300 - 851.600	97.890 - 851.800	88.690 - 590.100	88.690 - 851.800	
ptau					< 0.001
- Mean (SD)	36.764 (15.797)	27.665 (14.626)	21.837 (9.106)	27.584 (14.358)	
- Range	10.770 - 94.860	8.210 - 103.000	8.260 - 59.990	8.210 - 103.000	
APOE4					< 0.001
- 0	70 (31.8%)	309 (50.8%)	261 (72.5%)	640 (53.9%)	
- 1	104 (47.3%)	235 (38.7%)	90 (25.0%)	429 (36.1%)	
- 2	46 (20.9%)	64 (10.5%)	9 (2.5%)	119 (10.0%)	
Age					< 0.001
- Mean (SD)	74.519 (8.149)	72.408 (7.514)	73.683 (5.954)	73.185 (7.250)	
- Range	55.600 - 90.300	54.400 - 91.400	56.200 - 89.600	54.400 - 91.400	
Gender					< 0.001
- Female	91 (41.4%)	248 (40.8%)	190 (52.8%)	529 (44.5%)	
- Male	129 (58.6%)	360 (59.2%)	170 (47.2%)	659 (55.5%)	
Edu					< 0.001
- Mean (SD)	15.432 (2.913)	16.079 (2.773)	16.353 (2.628)	16.042 (2.772)	
- Range	4.000 - 20.000	6.000 - 20.000	6.000 - 20.000	4.000 - 20.000	
Race					< 0.001
- Black	4 (1.8%)	15 (2.5%)	22 (6.1%)	41 (3.5%)	
- White	211 (95.9%)	573 (94.2%)	327 (90.8%)	1111 (93.5%)	
- Other	5 (2.3%)	20 (3.3%)	11 (3.1%)	1111 (93.5%)	
Last Diagnosis					< 0.001
- AD	219 (99.5%)	199 (32.7%)	10 (2.8%)	428 (36.0%)	
- MCI	1 (0.5%)	373 (61.3%)	46 (12.8%)	420 (35.4%)	
- CN	0 (0.0%)	36 (5.9%)	304 (84.4%)	340 (28.6%)	

Table: TADPOLE ADNI CSF+MRI data ($p_1 = 3$, $p_2 = 245$, $n = 1,188$) by baseline diagnosis. **Dataset 2: 245 MRI features from Freesurfer (Desikan), cortical thickness, surface area, volume.**

Preliminary Results: Model selection

	G = 1	G = 2	G = 3	G = 4
$q_J = 1, q_1 = 1, q_2 = 1$	698.33	690.13	686.93	687.30
$q_J = 1, q_1 = 1, q_2 = 2$	684.54	678.79	679.39	681.92
$q_J = 1, q_1 = 1, q_2 = 3$	677.06	672.75	674.35	678.36
$q_J = 1, q_1 = 1, q_2 = 4$	670.47	667.54	671.25	677.50
$q_J = 1, q_1 = 1, q_2 = 5$	665.21	663.87	669.39	676.95
$q_J = 1, q_1 = 1, q_2 = 6$	660.22	661.54	668.29	677.15
$q_J = 1, q_1 = 1, q_2 = 7$	655.97	659.28	668.31	678.57
$q_J = 1, q_1 = 1, q_2 = 8$	653.19	658.08	668.32	679.93

Table: Selection of optimal ranks of latent scores and number of mixture components using BIC-ICL on ADNI CSF+MRI data ($p_1 = 3, p_2 = 245, n = 1,188$).

The numbers are representing $\times 10^3$.

Preliminary Results: Cluster-1 (AD-like 38.9%)

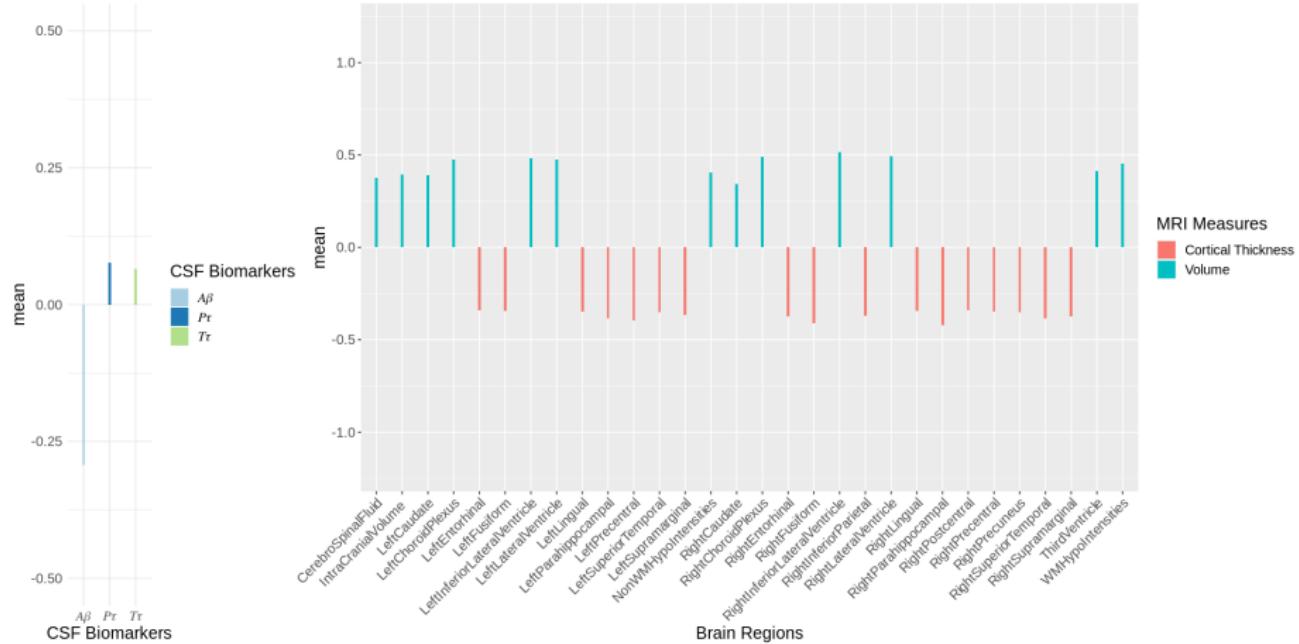


Figure: Mean in cluster 1 of CSF biomarkers and top 30 MRI features ($p_1 = 3$, $p_2 = 245$, $n = 1,188$).

Preliminary Results: Cluster-2 (non-AD 61.1%)

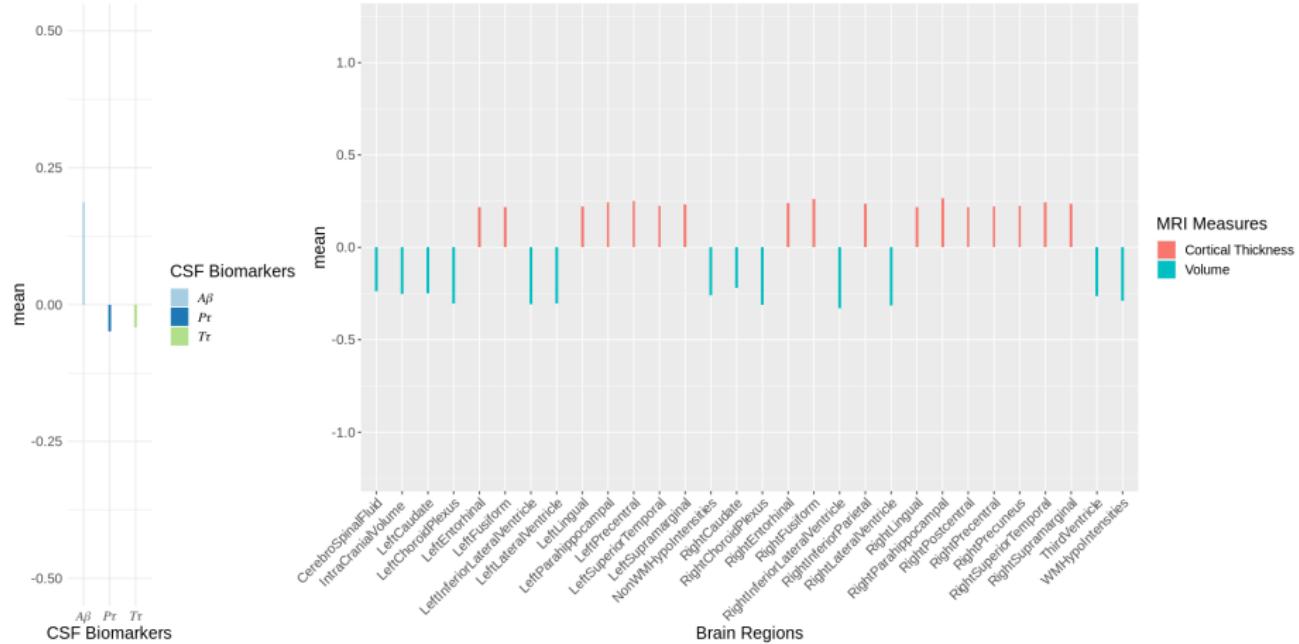


Figure: Mean in cluster 2 of CSF biomarkers and top 30 MRI features ($p_1 = 3$, $p_2 = 245$, $n = 1,188$).

Preliminary Results: Modeled cluster with $G = 2$, $q_1 = 1$, $q_2 = 5$

Cluster-1: AD-like		Last Available Diagnosis:		
	AD	MCI	CN	Total
Baseline Diagnosis:				
AD (%)	134 (100%)	0 (0%)	0 (0%)	134
MCI (%)	100 (43.3%)	126 (54.5%)	5 (2.2%)	231
CN (%)	4 (4.1%)	21 (21.6%)	72 (74.2%)	97
Total (%)	238 (51.5%)	147 (31.8%)	77 (16.7%)	n = 462
Cluster-2: Control-like		Last Available Diagnosis:		
	AD	MCI	CN	Total
Baseline Diagnosis:				
AD (%)	85 (98.8%)	1 (1.1%)	0 (0%)	86
MCI (%)	99 (26.3%)	247 (65.5%)	31 (8.2%)	377
CN (%)	6 (2.3%)	25 (9.5%)	232 (88.2%)	263
Total (%)	190 (26.2%)	273 (37.6%)	263 (36.2%)	n = 726

Table: **Clusters based on baseline data improve prediction of future conversion.** OR: MCI→AD, AD-like versus non-AD like: 2.1. CN→MCI, AD-like versus non-AD like: 2.6.

Conclusions

In this study, we:

proposed a probabilistic JIVE model with Gaussian mixtures for joint clustering from multiple datasets using an EM algorithm;

parsimonious covariance matrix for improved clustering;

identified two clusters that align with an AD-like group and control-like group;

Limitations and future research:

model selection is challenging due to trade-off between number of clusters and the rank of the covariance matrices.

penalized approaches with regularization of the covariance structure may be helpful in data with even higher dimensions.

Acknowledgments

Thank you!

Post doc opportunity available – email me for information!

Data provided by the ADNI Consortium <https://adni.loni.usc.edu/tadpole-challenge-dataset-available/>

BR was supported by R21 AG066970. JH was supported by R01 AG055634 and P50 AG025688. JL was supported by R01 AG070937 and P50 AG025688. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Additional funding for this work was provided by a generous gift from the Goizueta Foundation.



References I

- Collins, J. and Huynh, M. (2014). Estimation of diagnostic test accuracy without full verification: a review of latent class methods. *Statistics in medicine*, 33(24):4141.
- Dubois, B., Feldman, H. H., Jacova, C., DeKosky, S. T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G., Meguro, K., O'Brien, J., Pasquier, F., Robert, P., Rossor, M., Salloway, S., Stern, Y., Visser, P. J., and Scheltens, P. (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *The Lancet Neurology*, 6(8):734–746.
- Feng, Q., Jiang, M., Hannig, J., and Marron, J. S. (2018). Angle-based joint and individual variation explained. *Journal of Multivariate Analysis*, 166:241–265.
- Hellton, K. H. and Thoresen, M. (2016). Integrative clustering of high-dimensional data with joint and individual clusters. *Biostatistics*, 17(3):537–548.
- Hotelling, H. (1992). Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer.
- Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, 7(1):523.
- McLachlan, G. J. and Peel, D. (2000). *Finite mixture models*. Wiley.
- Meyer, G. D., Shapiro, F., Vanderstichele, H., Vanmechelen, E., Engelborghs, S., Deyn, P. P. D., Coart, E., Hansson, O., Minthon, L., Zetterberg, H., Blennow, K., Shaw, L., Trojanowski, J. Q., and Initiative, A. D. N. (2010). Diagnosis-Independent Alzheimer Disease Biomarker Signature in Cognitively Normal Elderly People. *Archives of Neurology*, 67(8):949–956.
- Report, A. A. (2020). 2020 alzheimer's disease facts and figures. *Alzheimer's and Dementia*, 16.
- Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016a). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317.
- Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016b). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R journal*, 8(1):289.
- Tipping, M. E. and Bishop, C. M. (1999a). Mixtures of Probabilistic Principal Component Analyzers. *Neural Computation*, 11(2):443–482.
- Tipping, M. E. and Bishop, C. M. (1999b). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.