Data Fusion and Transfer Learning in Patient Care Cycle: From Diagnosis to Care to System-level Decision Making

Jing Li, PhD

Associate Professor in Industrial Engineering
Co-director of ASU-Mayo Clinic Imaging Informatics Laboratory (AMIIL)
Arizona State University

This overview covers partial research AMIIL has been conducting as of November 2017.
Overview

• Data science research in patient care cycle
• Sample projects in ASU-Mayo Clinic Imaging Informatics Laboratory (AMIIL)
• Topic I
  ▪ A privacy-preserving positive transfer learning (P3TL) approach for telemonitoring of Parkinson’s disease
• Topic II
  ▪ A multi-mode factor mixture model (MFMM) with hierarchically-structured sparsity for imaging-based migraine subtype discovery
• Summary and acknowledgment
Precision Medicine: offers the right medical decision to the right person at the right time
Precision Medicine: offers the right medical decision to the right person at the right time
Health Care System

System-level Decision Making: considers multi-perspectives
Sample Projects in AMIIL Lab in Patient Care Cycle

Funding support: NIH, NSF, Mayo Clinic
Precision Treatment for Brain Cancer (Glioblastoma)

**Medical innovation**: map out tumor cell density by fusing multi-sequence MRI ($T1w$, $T2w$, rCBV, FA, MD, EPI) without invasive biopsy
- Precision neurosurgery
- Precision radiation therapy

**Data science innovation**: integration of bio-mechanistic model and data-driven machine learning model

- **bio-mechanistic model alone**
- **machine learning alone**

**Tumor cell density map**

**Collaborators**: Departments of Radiology and Neurosurgery in Mayo Clinic
Patient care is a team work. Nurses are team coordinators.

Large quantitative datasets collected by newly available Nurse Care Coordination Instrument

**Medical innovation:** identify impact factors for nurse care coordination
- Optimize nurse training and workload assignment
- Design practice environment to improve coordination

**Data science innovation:** a novel multi-response multi-level model (M3)
- Sparse selection of both fixed and random effects
- Joint estimation of multiple responses for transfer learning

**Collaborators:** ASU School of Nursing and Health Innovations
Improving Breast Cancer Detection by a Deep Learning Empowered Diagnostic System

Challenge

• Newly emerging contrast-enhanced digital mammography (CEDM) shows great promise for breast cancer detection.
• Fusion of multi-mode multi-view images is a significant challenge.

Innovation: a dual-mode deep transfer learning (D2TL) System

• Dual-mode deep architecture design
• Transfer learning for model estimation
• No need for precise tumor segmentation
• Quantification for intra- and inter-tumor malignancy heterogeneity

Visualization of malignancy distribution

Collaborators: Department of Radiology in Mayo Clinic
Predicting Mild Cognitive Impairment (MCI) Conversion to Alzheimer’ Disease (AD) using Incomplete Multi-modality Images

Incomplete multi-modality images for most patient cohorts (cost, accessibility, safety)

Medical innovation: accurate prediction of MCI conversation to AD for every patient cohort
- Facilitate AD early detection
- Increase accessibility

Data science innovation: missing-modality transfer learning (MMTL) model
- Joint estimation of cohort-wise models

Collaborators: Banner Alzheimer’s Institute, Banner Health
Biomarker Testing Process Optimization for Alzheimer’s Disease (AD)

**AD biomarkers**
- P-tau in cerebrospinal fluid
- Hippocampal volume on MRI
- Mean cerebral metabolism on FDG-PET

**Challenges**
- Multi-perspectives in optimizing test sequence and biomarker cutoffs
  - Patient-level: personalization, high accuracy
  - System-level: efficiency, resource availability

**Innovation:** a sequential tree-based classifier (STC)
- Green: low risk AD
- Red: high risk AD
- Grey: inconclusive diagnosis

**Collaborators:** Department of Radiology, Technical University of Munich, Germany
A Novel Privacy-Preserving Positive Transfer Learning (P3TL) Approach for Telemonitoring of Parkinson’s Disease
Telemonitoring

Use of electronic devices or simply smart phones to remotely monitor patients

- Logistically convenient
- Cost-effective
- Allows for close monitoring of disease progression
Parkinson’ Disease (PD)

- Parkinson’ disease (1.5 million)

U.S. Prevalence of Selected Neurodegenerative Disorders

- Multiple sclerosis (400,000)
- Huntington’s disease (30,000)
- Alzheimer’s disease (5.4 million)

Graph showing rate per 100,000 US standard population, age-adjusted:

- All ages – age-adjusted all ethnic groups male and female
- 328.7% increase in the death rate between 1976 and 2011
Telemonitoring for Parkinson’ Disease (PD)

• PD progression is **in-clinic** tracked by a Unified PD Rating Scale (UPDRS).
  - Six subscales, 42 questions
  - Requires patient’s presence in a specialized clinic
  - *Costly, logistically inconvenient*
  - *Infrequent examinations lead to delayed knowing of progression and sub-optimal treatment.*

• **At-Home** Testing Device (AHTD)
  - Records patient voice signal
  - Voice impairment is a key symptom for PD.

Build a predictive model between AHTD voice signals and UPDRS.
Challenges

• Need to build **one model for each patient**
  - Each patient has a different voice-UPDRS relationship due to heterogeneity in demographics, genetic risk factors, comorbidity, staging, treatment regimen.

• Conventional machine learning
  - Single learning: *Sample size limitation*
  - Data pooling: *One model fits all*

• Transfer learning (TL)
  - Leverages other patients’ information to make up the sample shortage when modeling a target patient

✓ How NOT to disclose the data of other patients (privacy preservation)?
✓ How to avoid negative transfer (TL from *some* other patients hurts modeling the target patient)?
State of the Art

Machine learning

Transfer learning (TL)

Supervised TL
(X, Y)

Homogenous TL
\( \chi_T = \chi_S \)

Instance transfer
(use source domain data)
Chattopadhyay et al. 2012
Lin et al. 2013
...

Feature transfer
(joined feature repres.)
Glorot et al. 2011
Gong et al. 2012
...

Parameter transfer
(similar model paras.)
Duan and Chang 2012
Long et al. 2014
...

Unsupervised TL
(X)

Heterogeneous TL
\( \chi_T \neq \chi_S \)

Feature transfer
(joined feature repres.)
Dai et al. 2008
Wang et al. 2008
...

Negative transfer conscious
Eaton and Lane 2008
Seah et al. 2016
Gong et al. 2012
Duan et al. 2012
Ge et al. 2014

- Instance and feature transfer do not preserve privacy.
- Black-box methods lack interpretation.
Objective of This Research

• Develop a white-box, parameter transfer model, P3TL, that enables
  • privacy-preserving transfer learning
  • selection of source domains (other patients) to transfer from to guard against negative transfer
• Apply P3TL to telemonitoring of Parkinson's disease with high accuracy
A Bayesian Parameter Transfer (BPT) model \[1\]

Risk of negative transfer not studied

Theoretical study on conditions of negative transfer

Privacy-preserving positive transfer learning (P3TL) model

Simulation studies and real-world application on telemonitoring of Parkinson’s disease

Bayesian Parameter Transfer (BPT) Model

Model for patient \( i \):

\[ Y_i = X_i \times w_i + \varepsilon_i \]

\( i = 1, \ldots, K - 1, K \)

Source domains (other patients)

Target domain (target patient)

Prior for target patient \( K \):

\[ p(w_K | \tilde{W}) \propto \text{Laplace}(w_K; b) \times N(w_K; \mu_K, \Sigma_K) \]

Source domains coeffs: \( \tilde{W} = (w_1, \ldots, w_{K-1}) \)

Posterior for target \( K \):

\[ p(w_K | \{y_K, X_K\}, \tilde{W}) \propto p(w_K | \tilde{W})p(y_K | X_K, w_K) \]

Bayesian Maximum A Posteriori (MAP) estimation

\[ (\hat{w}_K, \hat{\mu}_K, \hat{\Sigma}_K) = \arg\min_{w_K, \mu_K, \Sigma_K} \|y_K - X_Kw_K\|^2 + \lambda_1 \|w_K\|_1 + \lambda_2 \left( \log |\Sigma_K| + (w_K - \mu_K)^T \Sigma_K^{-1} (w_K - \mu_K) \right) \]

Impose sparsity

Enable transfer learning

Special case: uncorrelated sources

\[ \mu_K = \sum_{i=1}^{K-1} w_i \frac{\alpha_{iK}}{\kappa_i} \]

\[ \Sigma_K = \left( \varsigma_K - \sum_{i=1}^{K-1} \frac{\alpha_{iK}^2}{\kappa_i} \right) I \]

\( \alpha_{iK} \): covariance b/w source \( i \) and target \( K \)

Other patients’ data not shared but only model coeffs (privacy preserving)

Impose sparsity

Enable transfer learning

Impose sparsity

Enable transfer learning

Impose sparsity

Enable transfer learning
Tuning Parameter Selection

**Training Set**

\[
\begin{array}{cc}
(\lambda_1, \lambda_2) & \mathbf{w}_K \\
(\cdots) & (\cdots) \\
(\cdots) & (\cdots) \\
(\cdots) & (\cdots)
\end{array}
\]

**Validation Set**

\[
\begin{array}{c}
MSPE_{val}
\end{array}
\]

**Test Set**

\[
\begin{array}{c}
MSPE_{te}(\lambda_1, \lambda_2^*)
\end{array}
\]

**Generalization error**

**Smallest**

**Definition (Negative Transfer):** Let \( \lambda_1 = 0 \) and \( \lambda_2 = \lambda \) (suppress sparsity and focus on TL).

Let \( MSPE_{te}(0) \) be the MSPE of single learning on test set. Let \( MSPE_{te}(\lambda^*) \) be the MSPE of TL on test set. Negative transfer happens when \( MSPE_{te}(\lambda^*) > MSPE_{te}(0) \).

**MSPE:** Mean Square Prediction Error: \( \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n \)
**Theorem 1:** Assume training samples have unity norm, i.e., \( \|x_{tr_i}\|_2^2 = 1, i = 1, \ldots, Q \) (predictors). Let \( \alpha_i \) be the \( i \)-th element of \( \alpha = (\mu_K - w_K) \). A sufficient and necessary condition for negative transfer, i.e., \( \text{MSPE}_{te}(\lambda^*) > \text{MSPE}_{te}(0) \), is

\[
\left( \sum_{i=1}^{Q} \|x_{val_i}\|_2^2 \sum_{i=1}^{Q} \alpha_i^2 \|x_{te_i}\|_2^2 - \sigma^2 \sum_{i=1}^{Q} \|x_{val_i}\|_2^2 \sum_{i=1}^{Q} \|x_{te_i}\|_2^2 - 2 \sum_{i=1}^{Q} \|x_{val_i}\|_2^2 \alpha_i^2 \sum_{i=1}^{Q} \|x_{te_i}\|_2^2 \right) > 0
\]
Corollary 1: A sufficient condition that negative transfer will not happen is \( \|x_{val_i}\|_2^2 = \|x_{te_i}\|_2^2 \) for \( i = 1, \ldots, Q \) (predictors).

\[
\|
\begin{array}{c}
x_{val_i}^2 \\
val. sample 1
\end{array}
\|
\|
\begin{array}{c}
x_{te_i}^2 \\
test sample 1
\end{array}
\]

\( \|
\begin{array}{c}
x_{val_i}^2 \\
val. sample 1
\end{array}
\| \neq \|
\begin{array}{c}
x_{te_i}^2 \\
test sample 1
\end{array}\)  \( \Rightarrow \) Negative transfer will happen!
**Condition for Negative Transfer**

**Theorem 2**: Negative transfer will happen, i.e., $\text{MSPE}_{te}(\lambda) > \text{MSPE}_{te}(0)$, if and only if $\lambda > \eta$, where

$$\eta = \frac{2 \sum_{i=1}^{Q} \|x_{tei}\|^{2} \sigma^{2}}{\sum_{i=1}^{Q} \alpha_{i}^{2} \|x_{tei}\|^{2} - \sum_{i=1}^{Q} \|x_{tei}\|^{2} \sigma^{2}}.$$

---

**Graphs**

- **MSPE on validation set**
- **MSPE on test set**

- **TL with negative transfer**
- **TL without negative transfer**

---

- Solid line: Transfer learning (TL)
- Dashed line: Single learning (SL)
Hint from Theorem 2 on Preventing Negative Transfer

Search the optimal tuning parameter $\lambda^*$ within a small range $(0, \eta')$ prevents negative transfer.
Privacy-Preserving Positive Transfer Learning (P3TL) Algorithm

**Input:** data for the target domain, $X_K$ and $y_K$, split into a training and a validation set; a collection of $m$ subsets of source domains, $S = \{s_1,...,s_m\}$; parameters for P3TL, $\nu$ and $\tau$

**For each subset of source domains, $s_j$, $j = 1,...,m$, do the following:**

<table>
<thead>
<tr>
<th>Estimation of $\mu_K$ and $\Sigma_K$ by a coarse-to-fine strategy:</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) <strong>Coarse step:</strong> Find the smallest $\lambda \in {10^{-14}, 10^{-13}, 10^{-12}, 10^{-11}, ..., 10^{-1}}$ such that $\text{MSPE}(\lambda)<em>{val} &lt; \text{MSPE}(0)</em>{val}$. Denote this $\lambda$ by $\lambda_s$.</td>
</tr>
<tr>
<td>(b) <strong>Fine step:</strong> Find the $\lambda \in [0, \lambda_s \times 10^2]$ such that $\text{MSPE}(\lambda)_{val}$ is the lowest. Denote this $\lambda$ by $\lambda'_s$. Under $\lambda'_s$, obtain estimators for $\mu_K$ and $\Sigma_K$ by applying the BPL algorithm.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Screening rule:</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Find the lowest $\text{MSPE}(\lambda)<em>{val}$ with $\lambda$ searched within $[0, \lambda_s \times 10^2 \times \nu]$. Denote it by $\text{MSPE}(\lambda^*)</em>{val}$.</td>
</tr>
<tr>
<td>(b) Keep $s_j$ if $\text{MSPE}(\lambda^*)_{val}$ is lower than SL by at least $\tau \times 100%$; exclude $s_j$ otherwise.</td>
</tr>
</tbody>
</table>

**Identification of the best subset of source domains for transfer learning:** Among the kept subsets in $S$, the best subset is the one having the lowest $\text{MSPE}(\lambda)_{val}$ with $\lambda$ searched within $[0, \lambda_s \times 10^2]$.

**Output:** Model coefficients of the target domain, $\hat{w}_K$, by transfer learning from the best subset using BPL.
Application on Telemonitoring of PD

• 43 PD patients
• Sustained vowel phonation was measured weekly using AHTD.
• Speech signal processing algorithms extract 16 features, called dysphonia measures.
• UPDRS was administered at one of six medical centers the patients were enrolled at.
• Each patient is considered a “target” domain and all other patients considered “source” domains.
• Each patient has 101-168 samples, divided into training, validation, and test sets.
## Effectiveness of Negative Transfer Prevention

<table>
<thead>
<tr>
<th></th>
<th>Percentage of negative transfer</th>
<th>Average reduction of MSPE over SL on the test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>P3TL</td>
<td>7.1%</td>
<td>26.8%</td>
</tr>
<tr>
<td>2nd competing method (TL with full-range search)</td>
<td>7.1%</td>
<td>21.6%</td>
</tr>
<tr>
<td>1st competing method (transfer from any)</td>
<td>14.3%</td>
<td>15.6%</td>
</tr>
</tbody>
</table>

70% (10%) samples for training (validation)

<table>
<thead>
<tr>
<th></th>
<th>Percentage of negative transfer</th>
<th>Average reduction of MSPE over SL on the test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>P3TL</td>
<td>16.7%</td>
<td>19.9%</td>
</tr>
<tr>
<td>2nd competing method (TL with full-range search)</td>
<td>19.1%</td>
<td>13.9%</td>
</tr>
<tr>
<td>1st competing method (transfer from any)</td>
<td>23.8%</td>
<td>11.2%</td>
</tr>
</tbody>
</table>

50% (30%) samples for training (validation)
Prediction Accuracy on UPDRS

Conventional machine learning
- SL: single learning
- OMFA: one model fits all
A Multi-mode Factor Mixture Model (MFMM) with Hierarchically-Structured Sparsity for Imaging-based Migraine Subtype Discovery
Migraine Facts

• 1 in 4 U.S. households includes someone with migraine.

• Top 20 of the world’s most disabling illnesses

• Current treatment can only temporally alleviate symptoms.
Subtype Discovery and Precision Medicine

• Treatment would be more effective if customized to each migraine subtype, and eventually to each patient.

• Current subtype definition
  ▪ International Classification of Headache Disorders (ICHD-3 beta) criterion
  ▪ Based on patient self-reported symptoms: headache frequency, with and without aura
  ▪ Subjective, error-prone

• Multi-mode diagnostic imaging

  ![Brain MRI](image)

  Segmentation into anatomically defined regions of interest (ROIs)

  **Subtype discovery**
  Accurate, complementary

  **Multi-mode data**

  - Cortical surface area
  - Cortical thickness
  - Cortical volume
Challenges and State of the Art

• Challenges in clustering multi-mode imaging data
  ▪ High-dimensional imaging features
  ▪ Mode and feature selection
  ▪ Latent factors

• Existing latent factor models
  ▪ Factor analysis
    • Factors from a single distribution
  ▪ Factor mixture models (FMMs)
    • Factors from a mixture distribution with different components (clusters)
    • Muthen et al., 2006; Montanari et al., 2010; Baek et al., 2010

• Limitations
  ▪ Single-mode latent factor models
  ▪ No mode and feature selection
Objective of This Research

• Develop a Multi-mode Factor Mixture Model (MFMM) capable of hierarchical selection for modes and features
  • If a mode is uninformative to clustering, all features it includes will be excluded from clustering.
  • Feature selection happens in the modes that remain.
• Apply MFMM for migraine subtype discovery using multi-mode diagnostic MRI imaging data.
MFMM – Mathematical Formulation

Link mode-wise image features with latent factors and covariates:
\[ x_m = H_m f_m + B_m z + \epsilon_m, \quad \epsilon_m \sim N(0, \Psi_m) \]
Mode \( m = 1, ..., M \)

Link latent factors with latent subtypes:
\[ f_m = A_m s + \xi_m, \quad \xi_m \sim N(0, \Sigma_m) \]

Prior distribution of subtypes:
\[ (s_1, ..., s_K)^T \sim \text{multinomial} \left( w_1, ..., w_K \right) \]
Subtype \( k = 1, ..., K \)
MFMM – Mathematical Formulation

• Complete-data (observed & latent) log-likelihood function:

\[
l(\Theta) = \sum_{i=1}^{N} \left\{ \sum_{m=1}^{M} \log \left( f(x_{m,i}|f_{m,i}, z_i; \Theta) \right) + \sum_{m=1}^{M} \log \left( f(f_{m,i}|s_i; \Theta) \right) + \log \left( f(s_i; \Theta) \right) \right\}
\]

• Optimization with double \(l_{21}\) penalization

\[
\min_{\Theta} -l(\Theta) + \lambda_1 \sum_{m=1}^{M} ||A_m||_2 + \lambda_2 \sum_{m=1}^{M} \sum_{j=1}^{P_m} ||h_m^j||_2
\]

subject to \(E(f_m) = 0, Var(f_m) = I\) (identifiability constraints)

**Novel Property of the optimization: hierarchical selection of mode and features**

\[
x_{m}|s_k = 1 \sim N \left( H_m a_{m,k} + B_m z, \ H_m \Sigma_m H_m^T + \Psi_m \right)
\]

• If \(A_m = 0\), then \(a_{m,k} = 0\). Then, \(x_m\) have the same distribution regardless of subtypes, i.e., the \(m^{th}\) mode is uninformative to subtype clustering.

• If \(A_m \neq 0\) and \(h_m^j = 0\), then the \(j^{th}\) feature in \(x_m\) have the same distribution regardless of subtypes, i.e., this feature is uninformative to subtype clustering.
Model Estimation by Integrating EM and an efficient GMD algorithm

• Expectation-Maximization (EM) framework
  ▪ E-step: analytically derive $E_{\{f_m\}_{m=1}^M, s_i|x_m}_{m=1}^M; \Theta (g(\Theta))$
  ▪ M-step: solve two $l_2^1$-penalized sub-optimizations

**OPT 1:**

$$\{A_m^*\}_{m=1}^M = \arg\min_{\{A_m\}_{m=1}^M} \sum_{i=1}^N \sum_{m=1}^M E_{f_m, i | x_{1,i}, \ldots, x_{M,i}; \Theta} (\log f(f_m | s_i; \Theta)) + \lambda_1 \sum_{m=1}^M \|A_m\|_2$$

**OPT 2:**

$$H_m^*, B_m^* = \arg\min_{H_m, B_m} \sum_{i=1}^N E_{f_m, i | x_{1,i}, \ldots, x_{M,i}; \Theta} (\log f(x_m, i | f_m, i, z_i; \Theta)) + \lambda_2 \sum_{j=1}^{P_m} \|h_m^j\|_2$$

- Conventional solvers such as block-wise descent, block coordinate gradient descent, and Nesterov’s method are too slow.
- Group-wise Majorization Descent (GMD) algorithm is 5~10 times faster.
- Use of GMD requires that the optimization problem satisfies a Quadratic Majorization (QM) condition.
Theorem 1 (OPT1 and OPT2 satisfy the QM condition): Let $\mathbf{D}$ denote a dataset and $\boldsymbol{\beta}$ denote the parameters to be estimated in OPT1 or OPT2. $\boldsymbol{\beta}$ is partitioned into $J$ groups, $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(J)}$. OPT1 and OPT2 can be written as:

$$
\arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}|\mathbf{D}) + \lambda \sum_{j=1}^{J} \|\boldsymbol{\beta}^{(j)}\|_2
$$

1. $L(\boldsymbol{\beta}|\mathbf{D})$ is differentiable as a function of $\boldsymbol{\beta}$, i.e., $\nabla L(\boldsymbol{\beta}|\mathbf{D})$ exists everywhere.

2. There exists a $p \times p$ matrix $\boldsymbol{\Lambda}$, which may only depend on the data $\mathbf{D}$, such that for all $\boldsymbol{\beta}, \boldsymbol{\beta}^*$, $L(\boldsymbol{\beta}|\mathbf{D}) \leq L(\boldsymbol{\beta}^*|\mathbf{D}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \nabla L(\boldsymbol{\beta}^*|\mathbf{D}) + \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \boldsymbol{\Lambda} (\boldsymbol{\beta} - \boldsymbol{\beta})$. 
GMD Algorithm

To update the $j$-th group in $\beta^{(\omega)}$ at the $(\omega + 1)^{th}$ iteration:

$$\beta^{(\omega+1)} - \beta^{(\omega)} = \begin{pmatrix} 0, \ldots, 0, & \left(\beta^{(\omega+1)}(j) - \beta^{(\omega)}(j)\right)^T \\ j\text{-th group} \end{pmatrix}^T, \ 0, \ldots, 0.$$  

QM condition:

$$L(\beta^{(\omega+1)}|D) \leq L(\beta^{(\omega)}|D) + \left(\beta^{(\omega+1)}(j) - \beta^{(\omega)}(j)\right)^T \nabla L(j) + \frac{1}{2} \rho_j \left(\beta^{(\omega+1)}(j) - \beta^{(\omega)}(j)\right)^T \left(\beta^{(\omega+1)}(j) - \beta^{(\omega)}(j)\right)$$

Letting $\rho_j = (1 + 10^{-6})\tau_j$ where $\tau_j$ is the largest eigenvalue of $\Lambda^{(j)}$

OPT1 or OPT 2 becomes:

$$\arg\min_{\beta^{(\omega+1)}} L(\beta^{(\omega)}|D) + \left(\beta^{(\omega+1)}(j) - \beta^{(\omega)}(j)\right)^T \nabla L(j) + \frac{1}{2} \rho_j \left(\beta^{(\omega+1)}(j) - \beta^{(\omega)}(j)\right)^T \left(\beta^{(\omega+1)}(j) - \beta^{(\omega)}(j)\right) + \lambda \left\| \beta^{(\omega)}(j) \right\|_2$$

Analytical solution:

$$\beta^{(\omega+1)}(j) = \frac{1}{\rho_j} \left(-\nabla L(j) + \rho_j \beta^{(\omega)}(j)\right) \left(1 - \frac{\lambda}{\left\| \nabla L(j) + \rho_j \beta^{(\omega)}(j) \right\|_2^2}\right)$$
Application in Migraine Subtype Discovery

- 120 patients: Mayo Clinic at Arizona, Washington University School of Medicine in St. Louis
- 2 covariates: age and gender
- Modes and features
  - 3 MRI imaging modes: cortical surface area, thickness, and volume
  - 34 features in each mode corresponding to 34 brain regions of interest (ROIs)
- Add nuisance modes and features
  - To assess mode and feature selection accuracy
  - We previously published a non-penalized version of MFMM that found all 3 modes and 34 features in each mode are informative to identification of two subtypes \[^{[1]}\].
  - 3 nuisance modes
  - 34 nuisance features in each mode.

Results

• Mode and feature selection accuracy
  ▪ All 3 real modes are selected.
  ▪ 100% sensitivity and 96.6% specificity in feature selection

• 2 subtypes are found:
  ▪ A: 67 subjects
  ▪ B: 53 subjects

• Factors contributing to subtype separation
  ▪ 2 from cortical surface area
  ▪ 3 from cortical thickness
  ▪ 2 from cortical volume
Correlate Imaging-based Subtypes with Clinical Symptoms

- Clinical symptom measurements
  - Beck Depression Inventory (BDI)
  - State-Trait Anxiety Inventory (STAI)
  - Alldynia Symptom Checklist 12 (ASC-12)
  - Migraine Disability Assessment (MIDAS)
  - Hyperacusis Questionnaire
  - Photosensitivity Assessment Questionnaire
  - Number of headache days per month, number of years with migraine, aura status

<table>
<thead>
<tr>
<th>Clinical symptoms</th>
<th>Subtype difference</th>
<th>P value</th>
<th>Correlation with imaging also found in literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of years with migraine</td>
<td>A&gt;B</td>
<td>0.01</td>
<td>Chong et al., 2015; Jin et al., 2013</td>
</tr>
<tr>
<td>Migraine-related disability measured by MIDAS score</td>
<td>A&gt;B</td>
<td>0.04</td>
<td>Hubbard et al., 2014; Rogachove et al., 2016; Emmert et al., 2016</td>
</tr>
<tr>
<td>Symptoms of alldynia during migraine attacks</td>
<td>A&gt;B</td>
<td>0.03</td>
<td>Moulton et al., 2008; Schwedt et al., 2014; Russo et al., 2016</td>
</tr>
</tbody>
</table>
• These ROIs have been reported to be related to pain processing (Russo et al., 2016), differentiating migraineurs from healthy controls (Schwedt et al., 2015), and differentiating migraineurs with and without allodynia (Schwedt et al., 2014).
Conclusion

• Big Data in health care creates Big Opportunities for Precision Medicine.
• In AMIIL lab, we collaborate with medical professionals and conduct research at the interface between data science and emerging, challenging health care problems.
• We tackle problems across the patient care life cycle including diagnosis/subtyping, prognosis, treatment, care, and (tele)monitoring, focusing on not only the patient perspective but also the system perspective.
• The disease domains we have experiences with include neurological disorders (Alzheimer’s, Parkinson’s, migraine, traumatic brain injury) and cancer (brain, breast).
• Every real-world problem has its unique properties that can driven new methodological development.
• We envision our research can help improve patient care as well as contribute to methodological advances in statistics, machine learning, and systems engineering.
Acknowledgement

We would like to sincerely acknowledge the patients who consented to share their data for medical research in support of the united multi-disciplinary efforts for improving the health and humanity of the current and next generations to come.
Thank you!