IFCS Data Challenge: Lower Back Pain

Hongzhe Liu, Le Phan

Background

- IFCS cluster analysis challenge
- Lower back pain (LBP) data
  - 928 patients
  - Many missing values
  - 112 variables of mixed data types
    - 64 dichotomous
    - 30 ordinal
    - 9 multistate nominal
    - 8 continuous
    - 1 trichotomous
- Data source: http://ifcs.boku.ac.at/repository/challenge2/

Methodology

- Data cleaning
  - Inferring missing values
  - Imputing missing values & techniques
- Clustering methods
  - Data transformation
  - Choosing optimal number of clusters
  - Select appropriate clustering algorithm
- Validation & key findings
  - Cluster characteristics
  - Interpreting validation variables

Data Cleaning

1. Inferring missing values
   a. What is it?
      i. Data missing not at random
      ii. Fill in missing values using known relationship with other variables
   b. Goal: fill in as many missing values as possible by inference before imputing

Variables with missing values	Reason for missing not at random	Inference approach
fabq60, fabq70, fabq80, fabq90, fab100, fabq110, fab120, fab130, fabq140	Questions involving pain level with respect to work condition; only to be answered if patient is working	Replace NAs with new category, -1, if patient's employment situation, barb0, indicates not working
facetextrot, facetsit, facetwalk, paraspin_debut	Questions only to be asked if patients answer yes to having dominating back pain	Replace NAs with new category, - 1, if patient does not have dominating back pain (domin_bp = 1 for no)

Data Cleaning (cont.)

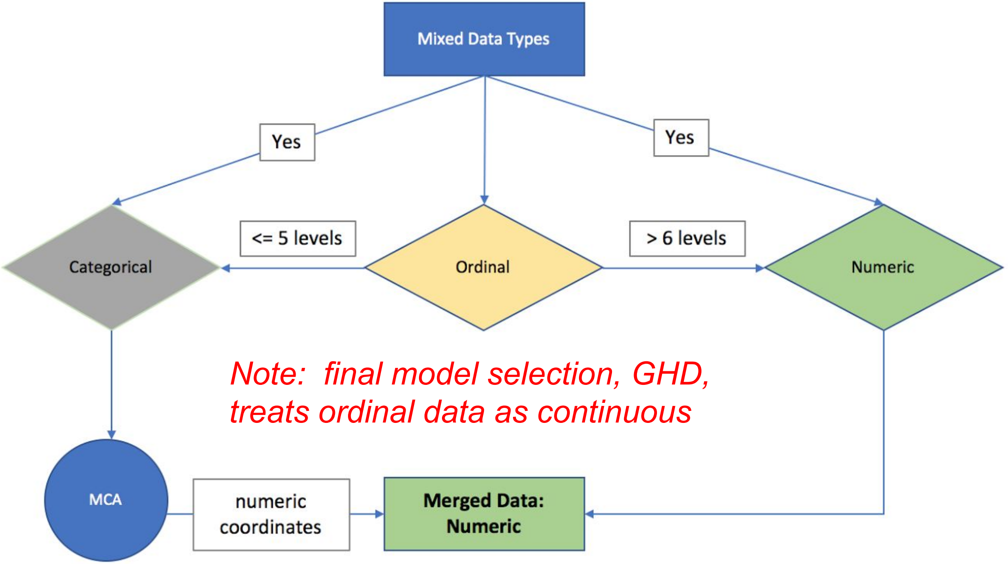
2. Imputing missing values
   a. Fill in values missing at random (MAR)
   b. Method:
      i. Multiple Imputation by Chained Equations (MICE)
      ii. R-package and function: mice::mice
   c. More information on MICE: Stef van Buuren and Karin Groothuis-Oudshoorn, "*mice: Multivariate Imputation by Chained Equations in R*", 2011

MICE algorithm:

1. Initial imputation — random draw from the data as "placeholders".
2. The "placeholders" for the first variable with missing values are set back to missing, this variable is the response in the regression model while the remaining variables are predictors in step 3.
3. The appropriate model (i.e., logistic regression, linear regression, multinomial, etc.) is used to predict the missing values in the response.
4. The missing values in the response is replaced with the predicted values.
5. Step 2-4 is repeated for the next variable with missing values.

Clustering Method – Data Transformation

- **The clustering problem**: mixed data types
  - Numeric data: better with k-means, fuzzy k-means, mixture models
  - Categorical: better with Multiple Correspondence Analysis (MCA)
- **Solution to mixed data**: split-then-join
  - Split into two subsets
    - Pure categorical
    - Pure numeric
  - Transform categorical data with MCA
  - Reduce dimension
  - Append MCA's numeric coordinate to numeric data subset



Note: final model selection, GHD, treats ordinal data as continuous

Clustering Methods – Determine Optimal G Clusters

- What is the optimal number of clusters?
  - K-means
  - Partition Around Medoids (PAM)
  - Fuzzy K-means (FKM)
  - Probabilistic Distance Clustering (PDClust)
  - Gaussian Parsimonious Clustering Models
  - Mixture Generalized Hyperbolic Distributions (MGHD)
- These clustering algorithms suggest 3 or 4

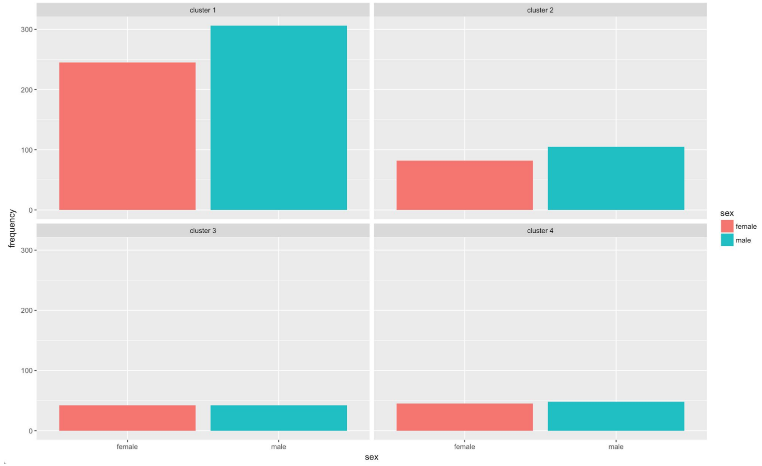
Clustering Methods – Algorithm Comparisons

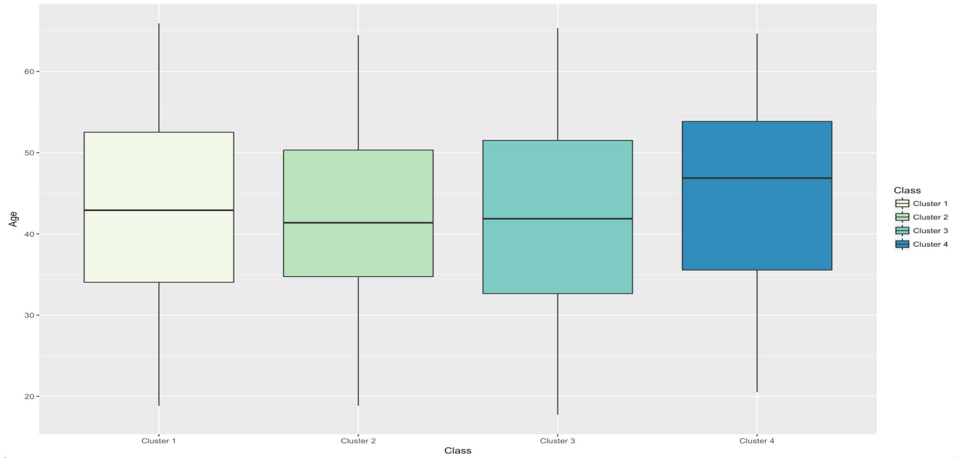
Algorithm Comparisons	Adjusted Rand Index (ARI)
ARI(kmeans.out$cluster, pam.out$clustering)	0.80602
ARI(kmeans.out$cluster, fkm.out$clus[,1])	0.87352
ARI(kmeans.out$cluster, pdc.out$label)	0.52930
ARI(kmeans.out$cluster, mvn.out$clust)	0.03592
ARI(pam.out$clustering, fkm.out$clus[,1])	0.70774
ARI(pam.out$clustering, pdc.out$label)	0.58555
ARI(fkm.out$clus[,1], pdc.out$label)	0.54035
ARI(mvn.out$clust, mst.out$clust)	0.71292
ARI(mvn.out$clust, GHD.out@map)	0.51473
ARI(mst.out$clust, GHD.out@map)	0.68374

Algorithms	Average Silhouette
K-means	0.29
PAM	0.29
FKM	0.28

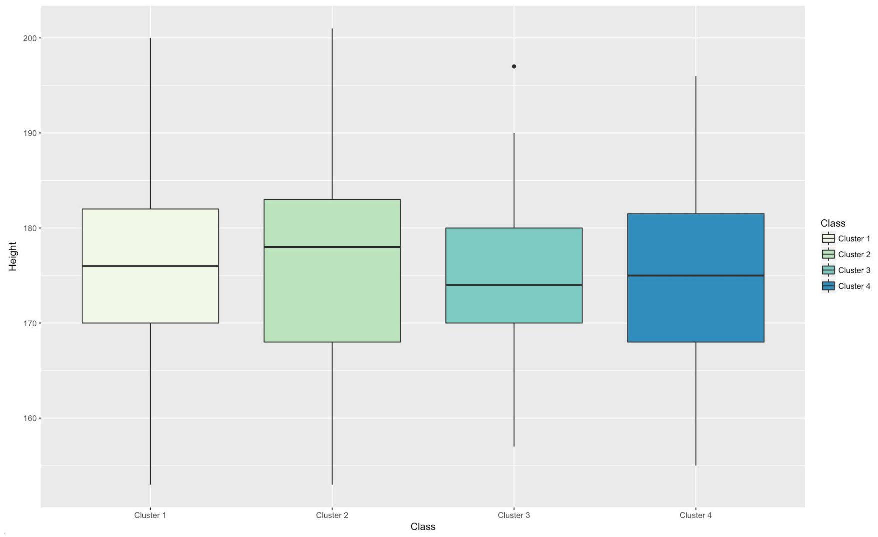
- Narrows down to:
  - K-means
  - GHD
  - Skew-t distribution - will ignore since it is a member of GHD
- Next step: compare with validation variables
- Final decision: GHD model with G=4 due to better separation of clusters in the validation variables

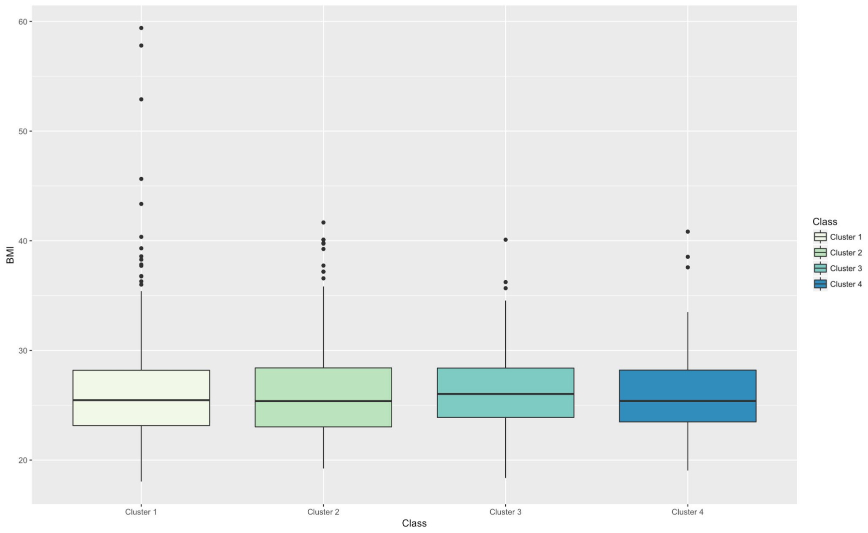
Result – Cluster Characteristics: Sex, Age



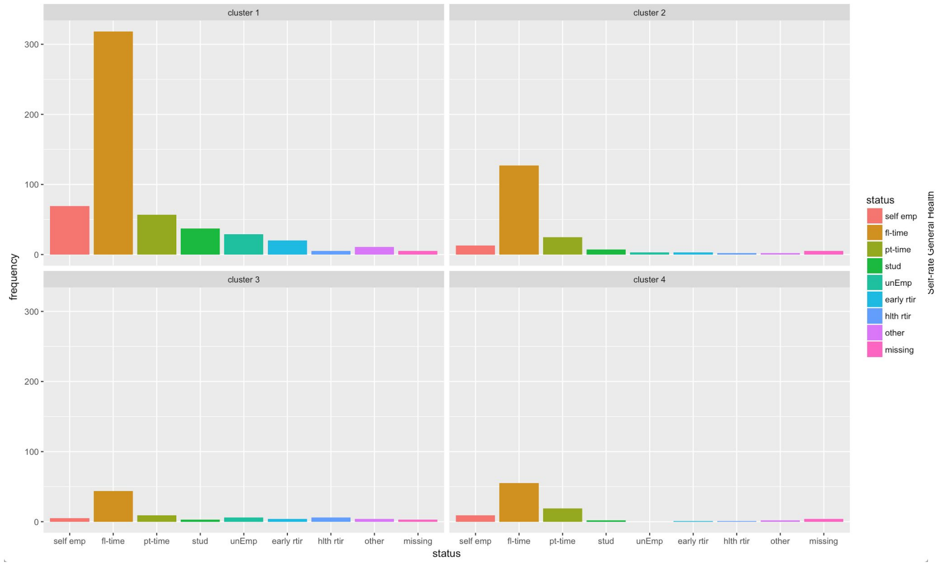


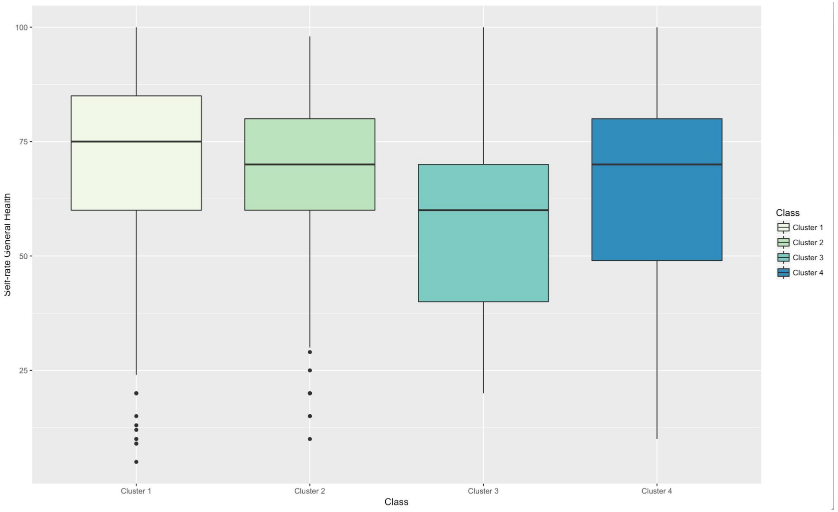
Result – Cluster Characteristics: Height, BMI





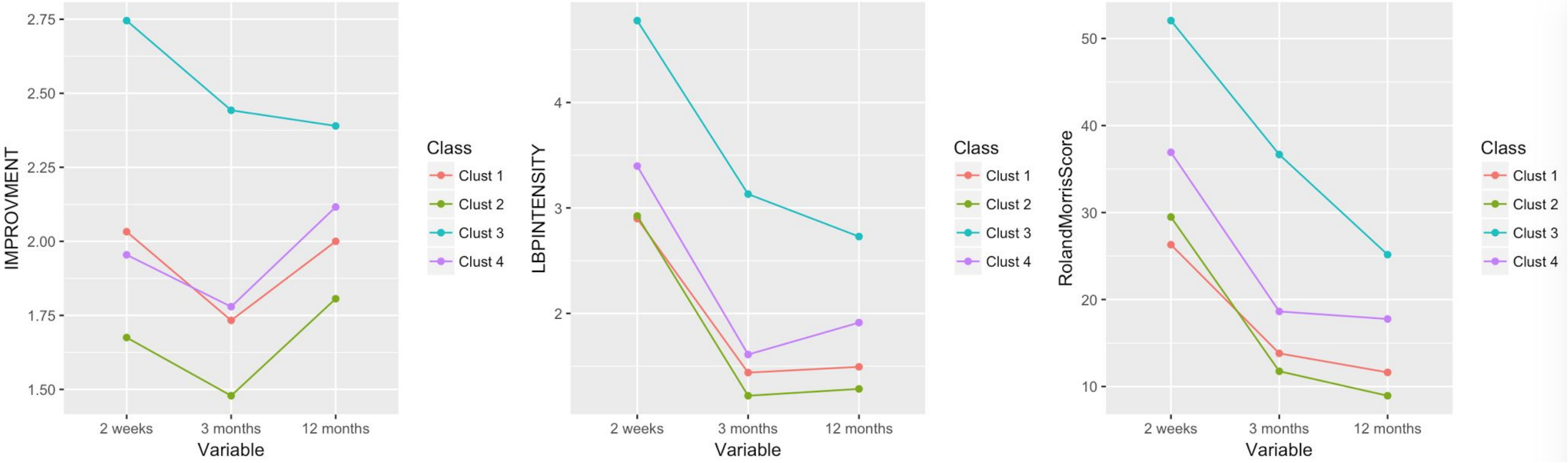
Cluster Characteristics – Employment, Self-Rated Health





Result – Validation Variables

- Three outcomes with 9 variables
  - Global perceived improvements: 2-week, 3-month, 12-month
  - LBP intensity: 2-week, 3-month, 12-month
  - Roland-Morris score: 2-week, 3-month, 12-month



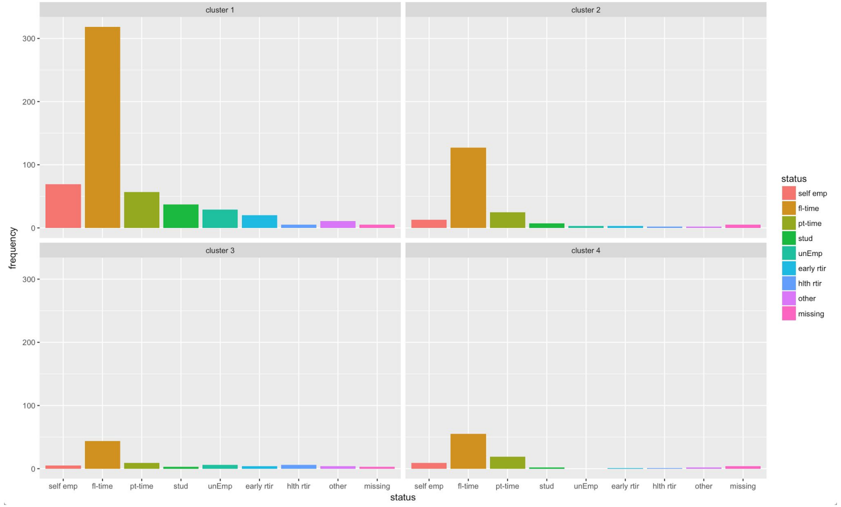
Interpreting Validation Variables

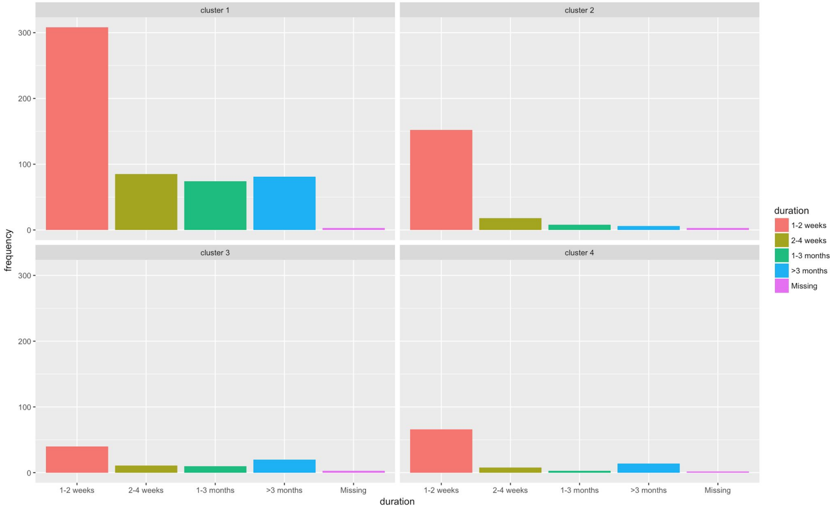
- Perceived improvements
  - Older patients (cluster 4) exhibit greater incremental perceived improvement from 3-month to 12-month after clinical consultation in all time period.
  - Patients who has greater LBP and Roland-Morris score experienced greater improvement.
- LBP correlated with RMS: highest LBP patients exhibit highest Roland-Morris score.
- Patient **height** could be a predictor of LBP and Roland-Morris score
  - Patients in cluster 2: tallest, least LBP intensity, lowest Roland-Morris score
  - Patients in cluster 3: shortest, greatest LBP, highest Roland-Morris score
-

Interpreting – Validation Variables (cont.)

- LBP intensity is negatively correlated with employment situation
    - Patients with high LBP intensity also report “non-working” status
- Patients who self-rated “poor general health” are more pessimistic than others
    - As their LBP and RMS improve, their perceived improvement continue to decline
- Patients who work full-time (i.e., cluster one) have short-term (1-2 week) LBP at the time of clinical consultation

Short-term LBP (1-2 week) doesn't affect working status





Interpreting – Validation Variables (cont.)

LBP intensity impact working status