Machine Learning for Omics Integration - Day 2 Notes

Unsupervised Multi-Omics Integration

Course Notes

`r Sys.Date()`

Convert to PDF: `pandoc Notes_day2.Rmd -o Notes_day2.pdf --pdf-engine=xelatex -V mainfont="Helvetica Neue" -V monofont="Monaco" -V mathfont="TeX Gyre Termes Math" -V fontsize=12pt -V geometry:margin=1in -V linestretch=1.2`

Continuation of Day 1 laboratory session (lab 2)

Key Challenges with Binary Data

Problem with Binary/Sparse Data

- **Binary data** (e.g., mutation data with mostly 0s and 1s) presents significant challenges
- **DIABLO** has known difficulties handling such sparse binary datasets
- **Recommendation**: Check mixOmics discussion forum for best practices and workarounds

Day 2 lecture ## Pre-analysis Strategy: MOFA

Why Run MOFA First?

- **Purpose**: Understanding relationships between different omics layers before integration
- **Benefit**: Provides insights into data structure and inter-omics correlations
- **Strategy**: Use MOFA as exploratory analysis prior to supervised methods

Unsupervised Machine Learning Philosophy

The "Fishing Expedition" Approach

- **Concept**: Unsupervised ML is like a "fishing expedition"
- **No prior hypothesis**: We don't understand the biological hypothesis beforehand
- **Discovery-driven**: Let the data reveal patterns and relationships

  Key Reference: *"A hypothesis is a liability"* - article published in Genome Biology
  [Link to be added]

MOFA: Multi-Omics Factor Analysis

Overview

- **MOFA & MOFA+**: Leading examples of unsupervised multi-omics integration
- **Methodological approach**: Hybrid of PLS/CCA and Bayesian methods
- **Applications**: Widely used for discovering latent factors across omics layers

Core Concept: Factor Analysis

What is Factor Analysis?

- **Central idea**: All observed data (gene expression, methylation, mutations) are generated by a few **latent variables** (factors/vectors)
- **Goal**: Learn these hidden factors from observed data
- **Process**: Start from observed data infer hidden factors that explain the patterns

Mathematical Foundation: Matrix Factorization **Concept**: Decomposing the original data matrix into multiple component matrices

$$X_{ij} = U_{ik} \times V_{kj}$$

Where: - **X**: Original data matrix (samples × features) - **U**: Factor loadings matrix (samples × factors)
- **V**: Factor weights matrix (factors × features) - **k**: Number of latent factors

MOFA Mathematical Framework

Multi-View Factor Model For multiple omics datasets, MOFA extends the basic factor model:

$$X_{ij}^{(m)} = \sum_{k=1}^{K} Z_{ik} \cdot W_{kj}^{(m)} + \epsilon_{ij}^{(m)}$$

Where: - **X^(m)**: Data matrix for omics type m (samples × features) - **Z**: Shared latent factor matrix (samples × factors) - **W^(m)**: Factor loadings for omics m (factors × features) - **K**: Number of latent factors - **ε^(m)**: Noise term for omics m

Bayesian Formulation MOFA uses a **Bayesian approach** with prior distributions:

Factor Prior:

$$Z_{ik} \sim \mathcal{N}(0, 1)$$

Loading Prior (with sparsity):

$$W_{kj}^{(m)} \sim \mathcal{N}(0, (\alpha_k^{(m)})^{-1})$$

Precision Prior (Automatic Relevance Determination):

$$\alpha_k^{(m)} \sim \text{Gamma}(a_0, b_0)$$

Likelihood Functions **For continuous data (e.g., gene expression):**

$$X_{ij}^{(m)} | Z, W^{(m)} \sim \mathcal{N}\left(\sum_{k=1}^{K} Z_{ik} W_{kj}^{(m)}, (\tau^{(m)})^{-1}\right)$$

For count data (e.g., RNA-seq):

$$X_{ij}^{(m)} | Z, W^{(m)} \sim \text{Poisson}\left(\exp\left(\sum_{k=1}^{K} Z_{ik} W_{kj}^{(m)}\right)\right)$$

For binary data (e.g., mutations):

$$X_{ij}^{(m)}|Z, W^{(m)} \sim \text{Bernoulli}\left(\sigma\left(\sum_{k=1}^{K} Z_{ik} W_{kj}^{(m)}\right)\right)$$

Where σ is the sigmoid function: $\sigma(x) = \frac{1}{1+e^{-x}}$

Variational Inference MOFA uses **variational Bayes** to approximate the posterior distribution:

Objective Function (ELBO):

$$\mathcal{L} = \mathbb{E}_q[\log p(X, Z, W, \alpha)] - \mathbb{E}_q[\log q(Z, W, \alpha)]$$

Where: - **p(X, Z, W, α)**: Joint probability of data and parameters - **q(Z, W, α)**: Variational approximation to posterior - **ELBO**: Evidence Lower BOund (maximized during training)

Key Advantages of MOFA

1. **Missing Value Compensation**: Values missing in one omics layer can be compensated by information from other omics layers
2. **Variance Explanation**: Uses R^2 to quantify how much variance is explained by the model
3. **Cross-omics Discovery**: Identifies shared and unique factors across different data types

Applications

scNMT Study

- **Example application**: Single-cell multi-omics integration
- **Reference**: scNMT paper [Link to be added]
- **Demonstrates**: Practical utility in real biological datasets

MOFA vs MOFA+: Detailed Comparison

MOFA (Multi-Omics Factor Analysis)

Core Methodology

- **Statistical Framework**: Bayesian factor analysis with group sparsity
- **Key Innovation**: Handles multiple omics datasets simultaneously
- **Mathematical Foundation**:
  - Assumes data is generated from a low-dimensional latent space
  - Uses variational inference for model fitting
  - Incorporates automatic relevance determination (ARD) for factor selection

MOFA Architecture

```
Input: Multiple omics matrices (RNA-seq, ATAC-seq, Methylation, etc.)
↓
Latent Factor Model: Z (samples × factors)
↓
Factor Loadings: W_m (factors × features) for each omics m
↓
Reconstruction: X_m = Z × W_m + noise
```

Key Features of MOFA

1. **Factor Interpretability**: Each factor can be interpreted biologically
2. **Sparsity**: Automatically selects relevant features and factors
3. **Uncertainty Quantification**: Provides confidence intervals for estimates
4. **Missing Data Handling**: Naturally accommodates missing observations

MOFA+ (MOFA Plus)

Major Improvements Over MOFA

- **Scalability**: Handles much larger datasets (>10,000 samples)
- **GPU Acceleration**: Faster computation using GPU implementations
- **Enhanced Flexibility**: Better handling of different data types and structures
- **Improved Convergence**: More robust optimization algorithms

New Features in MOFA+

1. **Stochastic Variational Inference**: Enables mini-batch processing
2. **Non-Gaussian Likelihoods**: Better modeling of count data, binary data
3. **Smoothness Constraints**: For spatial/temporal data integration
4. **Transfer Learning**: Pre-trained models can be applied to new datasets

When to Use MOFA vs MOFA+

Aspect	MOFA	MOFA+
Dataset Size	< 5,000 samples	> 5,000 samples
Data Types	Continuous/Gaussian	Mixed data types
Computational Resources	CPU-friendly	Requires GPU for large data
Interpretability	High (simpler model)	High (with more complexity)
Development Status	Mature, stable	Active development

MOFA/MOFA+ Workflow

1. Data Preprocessing

```
# Example preprocessing steps
- Log-transformation for count data
- Feature filtering (highly variable features)
- Normalization across samples
- Quality control checks
```

2. Model Training

```
# Basic MOFA model setup
MOFAobject <- create_mofa(data_list)
model_opts <- get_default_model_options(MOFAobject)
model_opts$num_factors <- 10
train_opts <- get_default_training_options(MOFAobject)
MOFAmodel <- run_mofa(MOFAobject, model_opts, train_opts)
```

3. Model Analysis

- **Factor inspection**: Which factors explain most variance?
- **Feature loadings**: Which genes/features drive each factor?
- **Sample scores**: How do samples project onto factors?
- **Variance decomposition**: How much variance per omics layer?

Biological Interpretation of MOFA Results

Factor Types

1. **Shared Factors**: Active across multiple omics layers
    - Often represent fundamental biological processes
    - Examples: Cell cycle, differentiation states, stress responses
2. **Specific Factors**: Active in single omics layer
    - Capture omics-specific technical or biological variation
    - Examples: RNA processing effects, chromatin accessibility patterns

Downstream Analysis

- **Pathway Enrichment**: Gene set analysis on factor loadings
- **Cell Type Identification**: Factor scores as features for clustering
- **Temporal Analysis**: Factor dynamics across time points
- **Clinical Association**: Correlate factors with phenotypes

Advantages of MOFA Approach

Over Traditional Methods

1. **vs PCA**: Handles multiple data types simultaneously
2. **vs CCA**: No requirement for paired samples across all omics
3. **vs Concatenation**: Accounts for different scales and noise levels
4. **vs Individual Analysis**: Identifies shared regulatory mechanisms

Statistical Benefits

- **Dimensionality Reduction**: From thousands of features to ~10-50 factors
- **Noise Reduction**: Separates signal from technical noise
- **Integration**: Leverages complementary information across omics
- **Flexibility**: Accommodates different experimental designs

Limitations and Considerations

MOFA Limitations

1. **Linear Assumptions**: May miss non-linear relationships
2. **Factor Interpretation**: Requires biological expertise
3. **Hyperparameter Tuning**: Number of factors needs careful selection
4. **Computational Complexity**: Can be slow for very large datasets

Best Practices

- **Factor Number Selection**: Use model selection criteria (ELBO, cross-validation)
- **Feature Selection**: Pre-filter for highly variable features
- **Batch Effects**: Address technical confounders before analysis
- **Validation**: Replicate findings in independent cohorts

Case Study Applications

Single-Cell Multi-Omics (scNMT-seq)

- **Data**: Single-cell RNA, DNA methylation, chromatin accessibility
- **Findings**: Identified developmental trajectories and regulatory relationships
- **Impact**: Revealed cell-type-specific regulatory mechanisms

Cancer Multi-Omics

- **Data**: Gene expression, copy number, methylation, mutation
- **Findings**: Discovered pan-cancer molecular subtypes
- **Clinical Relevance**: Biomarkers for treatment stratification

Population Studies

- **Data**: Multi-omics across large population cohorts
- **Findings**: Environmental and genetic factors affecting molecular profiles
- **Applications**: Precision medicine and risk prediction

Model Evaluation

R^2 (R-squared)

- **Definition**: Proportion of variance in the data explained by the model
- **Range**: 0-1 (or 0-100%)
- **Interpretation**: Higher R^2 indicates better model fit and factor explanatory power

Mathematical Formulation of R^2 For each omics layer m and factor k, the variance explained is:

$$R^2_{mk} = \frac{\text{Var}(\text{Predicted}_{mk})}{\text{Var}(\text{Observed}_m)}$$

Total variance explained across all factors for omics m:

$$R^2_m = \sum_{k=1}^{K} R^2_{mk}$$

Overall model R² (across all omics):

$$R^2_{\text{total}} = \frac{1}{M} \sum_{m=1}^{M} R^2_m$$

MOFA-Specific Evaluation Metrics

Variance Decomposition The total variance in the data can be decomposed as:

$$\text{Var}(X^{(m)}) = \text{Var}_{\text{explained}} + \text{Var}_{\text{noise}}$$

Where:

$$\text{Var}_{\text{explained}} = \text{Var}\left(\sum_{k=1}^{K} Z_{ik} W^{(m)}_{kj} \right)$$

$$\text{Var}_{\text{noise}} = \text{Var}(\epsilon^{(m)}_{ij})$$

Factor-Specific Variance Contribution For factor k in omics m:

$$\text{Contribution}_{mk} = \frac{\text{Var}(Z_{ik} W^{(m)}_{kj})}{\text{Var}(X^{(m)}_{ij})}$$

Evidence Lower Bound (ELBO) The ELBO objective function can be decomposed as:

$$\mathcal{L} = \underbrace{\mathbb{E}_q[\log p(X|Z,W)]}_{\text{Reconstruction}} - \underbrace{D_{KL}(q(Z)||p(Z))}_{\text{Factor Regularization}} - \underbrace{D_{KL}(q(W)||p(W))}_{\text{Loading Regularization}}$$

Where: - **Reconstruction term**: How well the model reconstructs the original data - **KL divergence terms**: Regularization preventing overfitting - **D_KL**: Kullback-Leibler divergence measuring difference between distributions

Model Selection Criteria

ELBO-Based Model Selection Compare models with different numbers of factors K:

$$\text{Best K} = \arg\max_K \mathcal{L}(K)$$

Cross-Validation for Factor Selection **K-fold CV procedure:** 1. Split data into K folds 2. For each fold i: Train MOFA on remaining folds, test on fold i 3. Compute CV error:

$$\text{CV Error} = \frac{1}{K}\sum_{i=1}^{K} ||\hat{X}^{(i)} - X^{(i)}||^2$$

4. Select number of factors that minimize CV error

Factor Activity Measure Sparsity of factor k in omics m:

$$\text{Activity}_{mk} = \frac{\text{Number of non-zero } W_{kj}^{(m)}}{\text{Total number of features in omics m}}$$

Biological Validation Metrics

Gene Set Enrichment For factor k, compute enrichment p-value:

$$p_{\text{enrichment}} = P(\text{overlap} \geq \text{observed}|\text{random})$$

Using hypergeometric distribution:

$$P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

Where: - **N**: Total genes in background - **K**: Genes in pathway
- **n**: Genes associated with factor - **x**: Overlap between factor and pathway

Factor Reproducibility Correlation between factors across independent datasets:

$$\text{Reproducibility} = \text{cor}(Z_{\text{dataset1}}, Z_{\text{dataset2}})$$

Good reproducibility: correlation > 0.7