A factor analysis based method for characterizing the covariance structure of related datasets

Teal Guidici[1], N. Lynn Henry[2], Charles Burant[3], George Michailidis[4]

[1]Statistics, University of Michigan, [2]Internal Medicine, University of Utah, [3] Internal Medicine, University of Michigan, [4]Statistics, University of Florida

Introduction

- Identifying and characterizing patterns of association between biomolecules in omics experiments has played a crucial role in understanding a wide variety of biological phenomena from the dynamics of human disease to condition specific alternations to metabolic pathways.
- Identifying changes in patterns of co-variation/co-expression in genomic data between healthy and diseased states can lead to insight regarding disease-affected regulatory systems. Patterns of co-variation are also particularly important in the metabolome, where they can additionally illuminate flow of metabolites through unobserved metabolic processes.
- Statistically, these patterns of association are commonly described in terms of covariance matrices (marginal associations) or inverse covariance matrices (conditional associations).
- Factor Analysis is an intuitive fit for modeling data in systems biology context – latent factors represent unobserved biological processes (e.g. pathways or similar); these are then projected into the space of observed variables (genes, proteins, metabolites).
- We present a factor analysis based model for capturing the similarities and differences in these patterns of association, as captured by covariance matrices, between multiple datasets from an experiment with factorial design (including those with a single design factor having multiple levels, the classic 2x2 factorial design, or other factorial designs.)

FACTOR ANALYSIS for a single dataset

$$X = \Lambda F + E$$

FACTOR ANALYSIS for multiple datasets

$$X^k = \Lambda^k F + E^k$$
$$\Lambda^k = B^k Q$$

- $k \in \{1, \dots, K\}, F \in \mathbb{R}^{m \times n},\ \Lambda^k \in \mathbb{R}^{p \times m}$
- F, E^k are orthogonal
- $\mathbb{E}(F) = \mathbb{E}(E^k) = 0$
- $Cov(F) = I, Cov(E^k) = \Psi^k$
- F common across all $k \in \{1, \dots, K\}$
- $B^k \in \mathbb{R}^{p \times p}$, diagonal, captures condition specific differences
- $Q \in \mathbb{R}^{p \times m}$, captures common structure

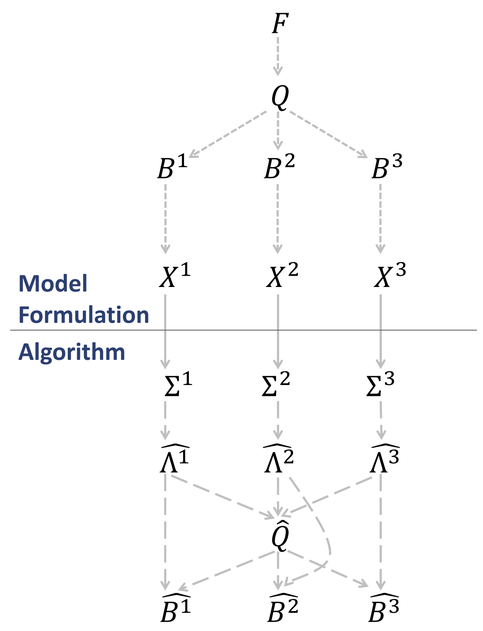
LOW RANK DECOMPOSITION

Each dataset X^k has a condition specific covariance matrix:

$$\Sigma^k = Cov(\Lambda^k F) + Cov(E)$$
$$= B^k Q Q' B^k + \Psi^k$$

- Σ^k is rank m.
- Eckhart-Young-Mirsky theorem: best low rank is given by $U^k_{1:m} D^k_{1:m} U^k_{1:m}{}'$ where UDU' is the eigen-decomposition of Σ^k.
- $B^k Q \approx U^k_{1:m} \sqrt{D^k_{1:m}} = \widehat{\Lambda^k}$
- Identifiability constraints ($B^k \succcurlyeq 0$, $\sum_{k=1}^K B^k = I$, others depending on experimental design structure) allow us to solve for $\hat{Q}$ and $\widehat{B^k}$.

METHOD SCHEMATIC



- Latent variables F are mapped into the space of observed variables by common loading matrix Q.
- B^k modulates this mapping up or down in each individual condition k.
- Observed data X^k is sum of projected latent variables $B^k Q F$ plus error E^k.
- Condition specific covariance matrices Σ^k used to calculate scaled eigen-vectors Λ^k.
- All K scaled eigen-vectors used to estimate

$$\hat{Q} = \sum_{k=1}^{3} \widehat{\Lambda^k}$$

- $\hat{Q}$ and condition specific $\widehat{\Lambda^k}$ used to calculate

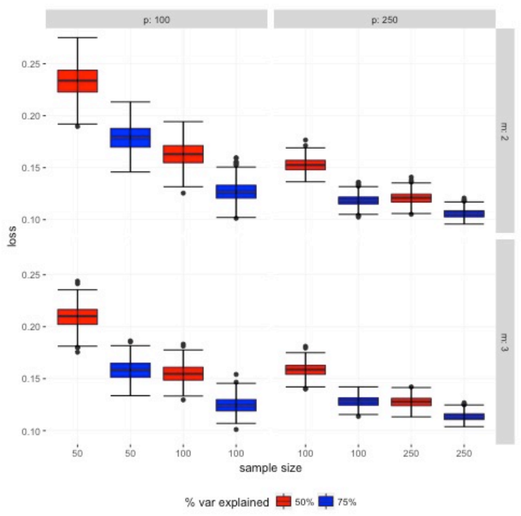
$$\widehat{B^k_{ii}} = \frac{1}{m} \sum_{j=1}^{m} \widehat{\Lambda^k_{ij}} / \hat{Q}_{ij}$$

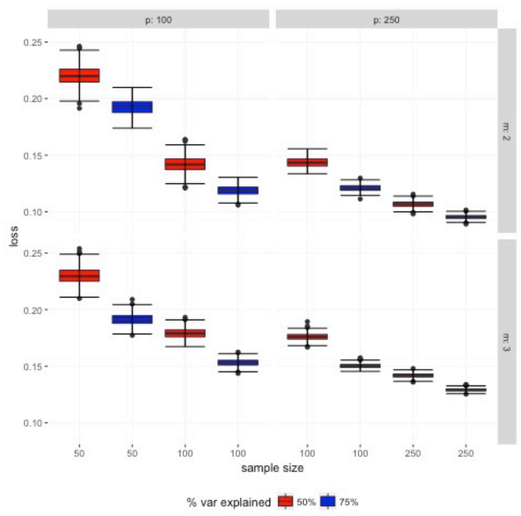
Numerical Results

Our algorithm performs well across a range of simulation settings. In general, loss values decrease as sample size increases, or the percentage of explained variance increases.
Representative numerical results are presented below – data is simulated from a 3×1 experimental design (a single design factor with 3 levels). We consider m=2,3, and Ψ^k such that Λ^k explains 50% or 75% of observed variance. We consider p=100,250 with n = 50,100,250, depending on p. All scenarios are run for 500 error realizations, over which the loss is averaged.

B-loss: $\frac{\sum_{k=1}^K \|B^k - \hat{B}^k\|}{\sum_{k=1}^K \|B^k\|}$

Q-loss: $\frac{\| |\hat{Q}| - |Q| \|_F}{\|Q\|_F}$





VISUALIZATION

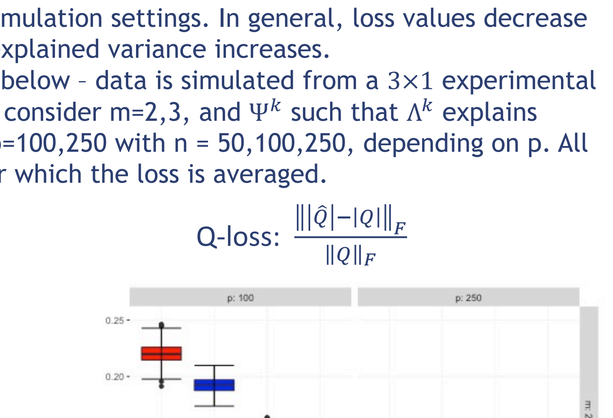
$\hat{Q}$ and $\widehat{B^k}$ can be used to visualize results, illuminating both common structure and condition specific differences, while allowing comparisons across conditions.

Reconstructed covariance:

- $\widehat{\Sigma^k} = \widehat{B^k} \hat{Q} \hat{Q}' \widehat{B^k}$
- Visualized as a heatmap
- Differences in intensity reflect condition specific modulation
- Overall structure reflects common elements.

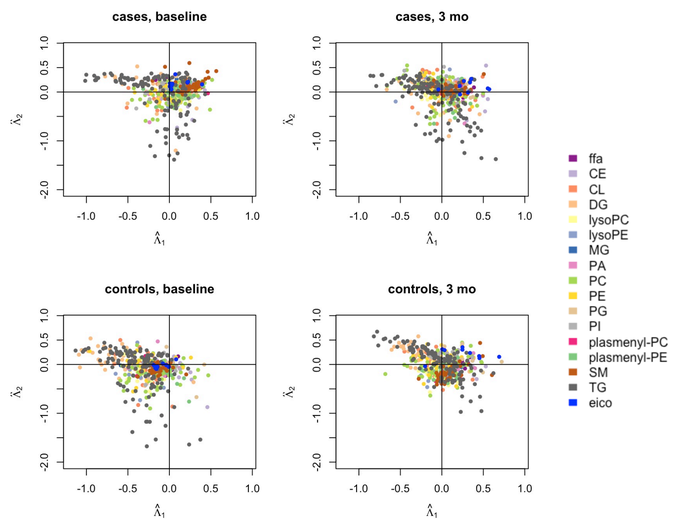
Pseudo-log fold change:

- $r_i^{kk'} = \log(\frac{\widehat{B^k_{ii}}}{\widehat{B^{k'}_{ii}}})$
- Visualized as bar charts
- Bars extend *towards* larger $\widehat{B^k_{ii}}$ values and *away* from smaller values.

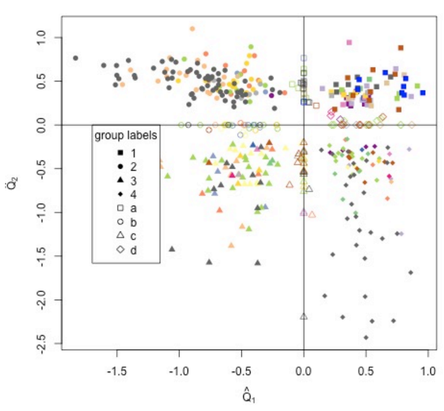


3×1 experimental design where 2 latent factors account for 75% of observed variance. $p = 30, n = 100$

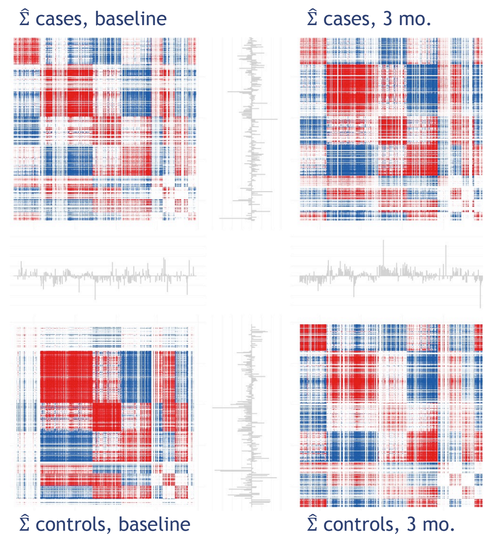
Case Study



- Aromatase Inhibitors (AIs) are an effective class of drugs used to treat breast cancer.
- 50 women with breast cancer, treated with AIs.
  - 25 will remain on drug > 24 months (controls)
  - 25 will remain on drug < 6 months due to bone pain (cases)
- Lipidomes assayed at 0 and 3 months.
- No mean differences between groups.
- Subtle changes in patterns of association can be seen in scatter plots of $\widehat{\Lambda^k_1}$ and $\widehat{\Lambda^k_2}$, the first two scaled eigen-vectors for each dataset.



- Lipids can be partitioned into 8 groups, based how they load onto $\widehat{Q_1}$ and $\widehat{Q_2}$.
- These groups can then be tested for enrichment in lipid classes or saturation levels.
- Groups can also be tested for group level differences of abundance.
- Group 4 is enriched for polyunsaturated lipids *and* is the **only group** to show difference in abundance – with the controls at 3 months being lower than at baseline.
- Results suggest that alterations in dietary intake or metabolism of these unsaturated fatty acids may differ between women with and without adverse effects following AI therapy.



- Lipids arranged by $\hat{Q}$ grouping.
- Large blocks illustrate common structure
- Patterns of light and dark crosshatching, as well as bar charts, indicate condition specific differences.

Conclusions and Continuing Work

- Factor analysis provides a natural structure for modeling omics data.
- Algorithm yields good numerical results, both in non sparse and sparse cases (sparse not shown).
- Method provides data driven approach to grouping variables, yielding biologically meaningful results.
- Method successfully captures weak signal in noisy data.
- Continuing work: time course data; B^k having block or functional form.
- Manuscript in preparation – check arXiv soon!

This work was supported by NIH T32 Proteome Informatics of Cancer Training Program at the University of Michigan (T32-CA-140044) and by the Michigan Regional Comprehensive Metabolomics Resource Core (U24 DK097153).