Imperial College London

CorrMapper: an online research tool for the integration and visualization of multi-omics studies

Daniel Homola, Elaine Holmes

Problem

Modern omics datasets are extremely feature rich and in multi-omics studies this complexity is compounded by a second or even third dataset. Many of these features however, might be completely irrelevant to the studied biological problem, or redundant in the context of others (**multicollinearity**). Learning from such feature rich datasets inevitably incurs an **increased computational cost**. It also increases the **chance of overfitting** the noise in our data, while reducing the predictive power of our models.

Finally, the correlation networks arising from these high-throughput datasets are often hard to interpret and explore due to their density and lack of interactive tools.

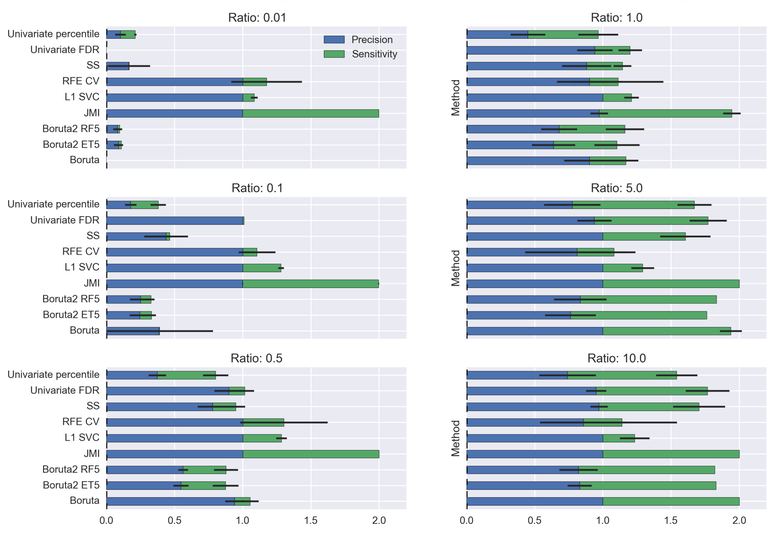
Contribution

CorrMapper brings together powerful feature selection, covariance estimation and visualisation algorithms. It reduces the dimensionality of complex biological datasets and exposes their core correlation structures in an interactive web-based interface. It can integrate any two omics datasets (or work with just one), and will be made available soon at www.corrmapper.com.

CorrMapper can help with the interpretation of complex multi-omics datasets, while pilot and exploratory studies can use it as an **excellent hypothesis-generating tool**.

Feature Selection

Based on our benchmarking results, four different Feature Selection (FS) methods are provided in CorrMapper:



Boruta[1] is an all-relevant FS method, based on Random Forests and artificial contrast. It tries to find all features that have any information about the outcome variable.

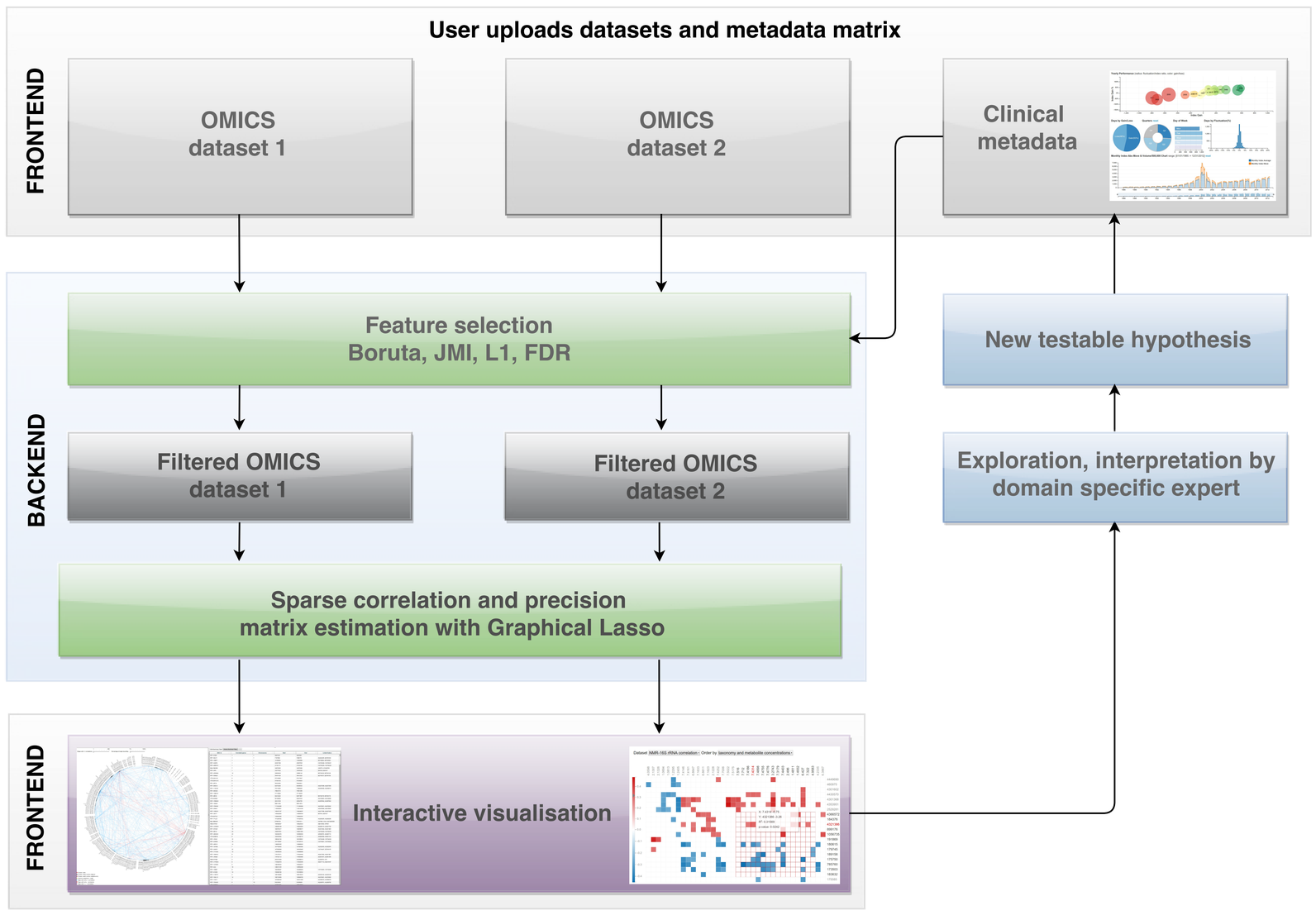
Joint Mutual Information based FS method[2] relies on information theoretic concepts and a greedy search algorithm to identify a subset of predictive features.

L1 regularised SVM[3] and **LASSO regression**[4] are based on L1 norm penalisation, which makes it possible in linear models to shrink certain coefficients to zero and consequently perform FS at training time.

Univariate tests coupled with FDR correction are also offered as they were found to perform well on $n << p$ datasets[5].

Boruta & JMI: github.com/danielhomola

Overview

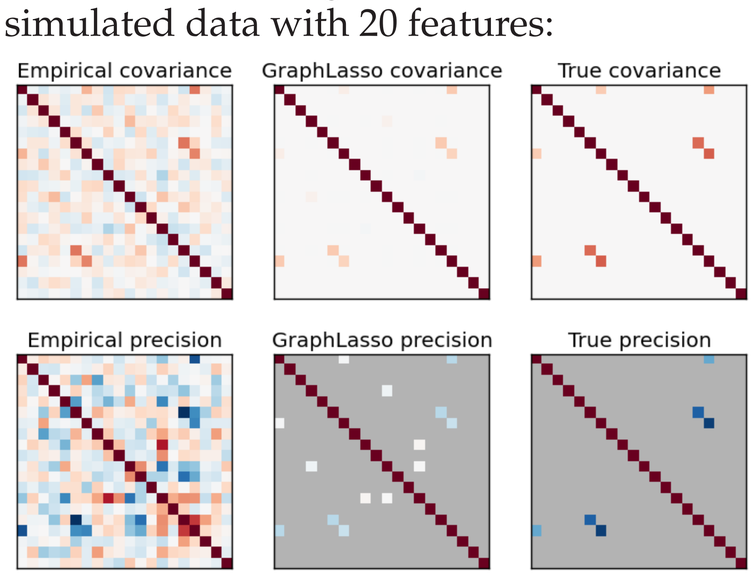


Covariance Estimation

Given a normally distributed random vector $X = (X_1, ..., X_p) \sim \mathcal{N}(\mu, \Sigma)$, with covariance matrix Σ and precision matrix $\Omega = \Sigma^{-1}$, X_i and X_j are conditionally independent given all other variables if and only if $\Omega_{ij} = 0$. Therefore Ω can reveal the core dependence structures in our data, and also help us to find confounding variables.

When $n < p$, the empirical covariance estimate $S_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)^T$ becomes very unstable which could result in S becoming singular. As a consequence S cannot be inverted and the partial correlations cannot be recovered.

Therefore we use the sparse inverse covariance estimator called **Graphical Lasso**[6], which is known to perform well in such $n < p$ scenarios. The figure below shows this on simulated data with 20 features:



The sparse precision and covariance matrix estimation is done on the selected features from both datasets to ensure we only explore the dependence structures which are relevant to the biological outcome variable.

The matrices are visualised as bipartite graphs, where the two different sets of nodes represent the two omics datasets.

Visualisation modules

CorrMapper has two interactive visualisation modules (please ask for a demo):

Genomic data is binned into 300 buckets that make up the whole genome. These bins are visualised in a circular graph, which is linked with sortable and searchable tables.

Omics datasets without genomic locations are visualised using interlinked bipartite graphs and a sortable heatmap that could be used to filter the networks dynamically.

Technology

Frontend: **HTML5**, **CSS** (Bootstrap), **Javascript**: AJAX calls, file uploads, form validation, d3.js for visualisation.

Backend: **Flask** (lightweight micro web framework), **Celery** (asynchronous job queue), **Redis** (in-memory message broker), **scikit-learn, numpy, scipy, pandas, SQLite** could be changed to SQL or PostgreSQL easily thanks to SQLAlchemy. This stack could be run on AWS or any university cluster.

ScienceFlask: The source-code of the CorrMapper website will be made available on GitHub and turned into an open project to serve as a template and make the deployment of scientific tools easy and quick.

References

1. Kursa, M. B. & Rudnicki, W. R. Feature Selection with the Boruta Package. *Journal Of Statistical Software* **36,** 1–13 (2010).
2. Howard Hua Yang, J. M. Feature Selection Based on Joint Mutual Information.
3. Vapnik, V. *Statistical learning theory [...] [...]* (Wiley, 1998).
4. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58,** 267–288 (1996).
5. Haury, A. C., Gestraud, P. & Vert, J. P. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE* **6,** 1–12 (2011).
6. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. en. *Biostatistics (Oxford, England)* **9,** 432–41 (July 2008).