Impact of joint Dimension Reduction methods for survival prediction - Extension of a multi-omics benchmark study

Vincent Le Goff, Vincent Guillemot, Cathy Philippe, Gwendoline Mendes, Jean-François Deleuze, Edith Le Floch, Arnaud Gloaguen

Impact of joint Dimension Reduction methods for survival prediction
Extension of a multi-omics benchmark study

Vincent Le Goff[1,@], Vincent Guillemot[2], Cathy Philippe[3], Gwendoline Mendes[4], Jean-François Deleuze[1], Edith Le Floch[1], Arnaud Gloaguen[1]

[1] Centre National de Recherche en Génomique Humaine (CNRGH), CEA, Université Paris-Saclay, Evry, France
[2] Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France
[3] NeuroSpin, CEA Saclay, Université Paris-Saclay, France [4] CentraleSupélec, Gif-sur-Yvette, France
@ vincent.legoff@cnrgh.fr

Introduction

This work aims to expand on a previous study [1] that compares the **survival predictions** of **13 different supervised multi-omics methods** on **datasets from 18 cancer types** from The Cancer Genome Atlas (TCGA).
This comparison is based on the analysis of **4 types of molecular data** (mRNA, miRNA, CNV, Mutations) and **clinical data**. Methods are divided into 3 groups: **reference models** (Kaplan-Meier and a Cox regression on clinical data only), **"naive" models**, unable to distinguish between each omics data, and **"structured" models.**

Objectives:
- Building on the conclusion in [1] that structured methods perform better than naive ones, we include methods not only capable of exploiting the group structure of omics data but also of extracting links between them: **Joint Dimension Reduction** (JDR) methods. We test these methods in **unsupervised and supervised** settings whenever possible.
- Then, to go further on the evaluation of the importance of clinical data **we compare all of these methods (naive/structured/JDR) without clinical data** to see if molecular data only can lead to similar prediction performance as with the addition of clinical data.

JDR methods

RGCCA/SGCCA
Regularized Generalized Canonical Correlation Analysis [2]

$$\underset{\mathbf{w}_1,\dots,\mathbf{w}_L}{\operatorname{argmax}} \sum_{k,l=1}^{L} c_{kl}\, g\left(\operatorname{cov}(\mathbf{X}_k\mathbf{w}_k, \mathbf{X}_l\mathbf{w}_l)\right)$$

$$\text{s.t.} \begin{cases} (1-\tau_l)\operatorname{var}(\mathbf{X}_l\mathbf{w}_l) + \tau_l\|\mathbf{w}_l\|_2^2 = 1 \\ \|\mathbf{w}_l\|_2^2 = 1 \text{ and } \|\mathbf{w}_l\|_1 \le s_l \end{cases}, l = 1, \dots, L$$

JIVE
Joint and Individual Variation Explained [3]

$$\mathbf{X}_l^\top \approx \mathbf{U}_l\mathbf{S}_\mathbf{J} + \mathbf{W}_l\mathbf{S}_l \text{ s.t} \begin{cases} \mathbf{S}_\mathbf{J}\mathbf{S}_l^\top = 0 \\ \operatorname{rank}(\mathbf{S}_\mathbf{J}) = r_j \\ \operatorname{rank}(\mathbf{S}_l) = r_a \end{cases}$$

IntNMF
Integrative Non-Negative Matrix Factorization [4]

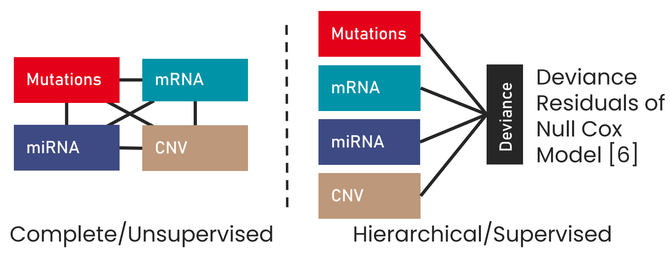
$$\mathbf{X}_l \approx \mathbf{W}\mathbf{H}_l^\top \text{ s.t. } \mathbf{W}, \mathbf{H}_l \ge 0, l = 1, \dots, L$$

MOFA
Multi-Omics Factor Analysis [5]

$$\mathbf{X}_l \approx \mathbf{Z}\mathbf{W}_l^\top$$

fit through Bayesian methods

RGCCA Unsupervised & Supervised



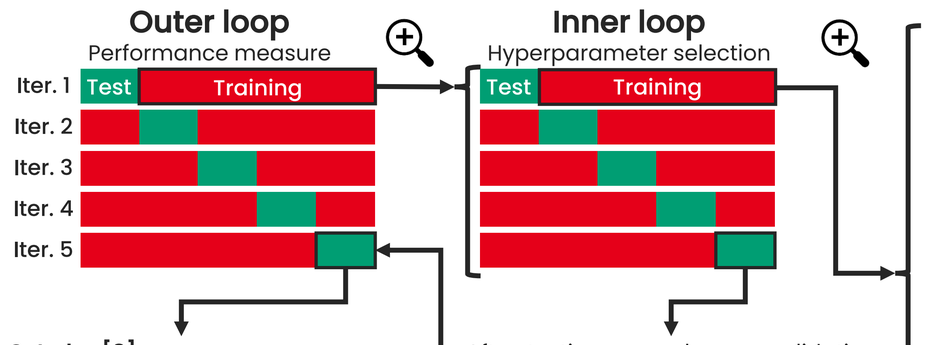
Supervised JIVE [7]

$$\underset{\substack{\mathbf{U},\mathbf{S}_\mathbf{J},\boldsymbol{\theta}_1 \\ \mathbf{W}_1,\cdots,\mathbf{W}_L \\ \mathbf{S}_1,\cdots,\mathbf{S}_L \\ \boldsymbol{\theta}_{21},\cdots,\boldsymbol{\theta}_{2L}}}{\operatorname{argmin}} (1-\eta) \left\| \begin{bmatrix} \mathbf{X}_1^\top \\ \vdots \\ \mathbf{X}_L^\top \end{bmatrix} - \mathbf{U}\mathbf{S}_\mathbf{J} - \begin{bmatrix} \mathbf{W}_1\mathbf{S}_1 \\ \vdots \\ \mathbf{W}_L\mathbf{S}_L \end{bmatrix} \right\|_F^2 + \eta \left\| \mathbf{y}^\top - \mathbf{S}_\mathbf{J}^\top\boldsymbol{\theta}_1 - \sum_{l=1}^{L} \mathbf{S}_l^\top\boldsymbol{\theta}_{2l} \right\|_2^2$$

$\mathbf{y}$: Deviance Residuals of Null Cox Model

Survival Prediction with JDR

Nested Cross-Validation



Uno's C-Index [8]

$$\hat{C} = \frac{\sum_{i,j=1}^{n} \Delta_i \{\hat{G}(T_i)\}^{-2} I(T_i < T_j, T_i < t_{\max}) I(\hat{\eta}_i > \hat{\eta}_j)}{\sum_{i,j=1}^{n} \Delta_i \{\hat{G}(T_i)\}^{-2} I(T_i < T_j, T_i < t_{\max})}$$

It corresponds to how well the model predicts the order of deaths.

Integrated Brier Score [8]

$$\hat{L} = \frac{1}{n}\sum_{i=1}^{n} \int_0^{t_{\max}} \Delta_i I(t \ge T_i) \frac{\hat{S}(t)^2}{\hat{G}(T_i)} + I(t < T_i) \frac{\left(1-\hat{S}(t)\right)^2}{\hat{G}(T_i)} dt$$

It corresponds to how close the survival curve is to the « true » curve.
$t_{\max}$ is chosen s.t. we keep 90% of the data to avoid score inflation

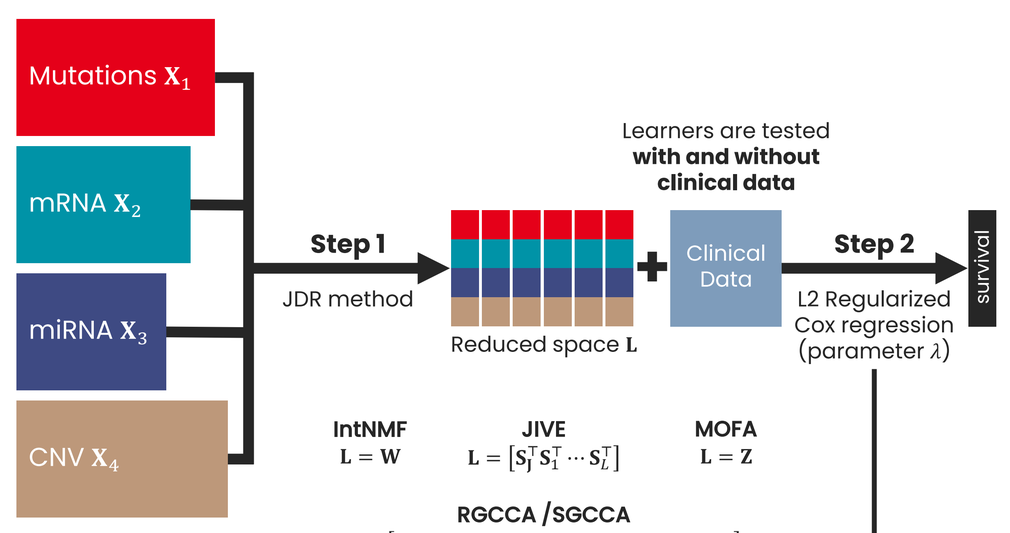
After testing several cross-validation metrics, we use Verweij and Van Houwelingen Cross-Validation Error (CVE) [9]:

$$CVE(\lambda) = -2\sum_{k=1}^{K} \ell\left(\hat{\beta}_{-k}(\lambda)\right) - \ell_{-k}\left(\hat{\beta}_{-k}(\lambda)\right)$$

The best λ is the one that minimizes the CVE:

$$\lambda_{best} = \underset{\lambda}{\operatorname{argmin}}\, CVE(\lambda)$$

Sequential Modeling



Tested Parameters

RGCCA:
$\tau_l \in \{0,1; 1\}, R_l \in \{1; 3\}, g(x) = x^2$
The 2 supervised/unsupervised connection graphs described above.

SGCCA:
$s_l \in \{0,5; 0,75\}, R_l \in \{1; 3\}, \ g(x) = x^2$
The 2 supervised/unsupervised connection graphs described above.

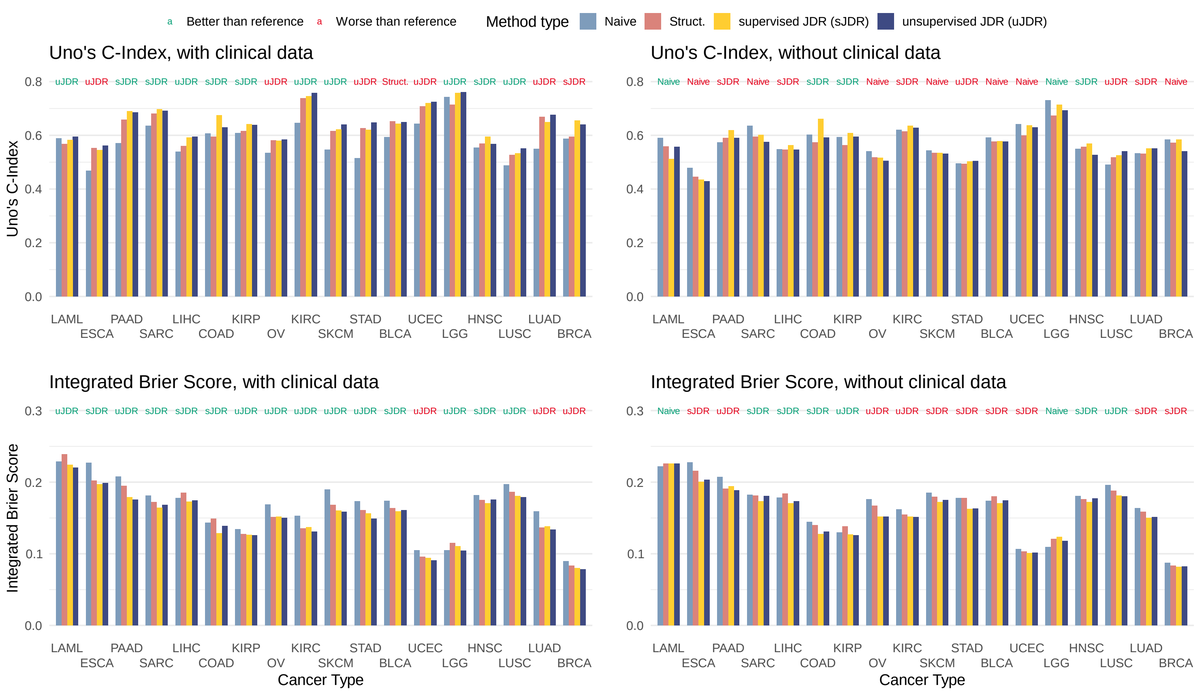
JIVE:
$r_j \in \{0; 1; 3\}$ and $r_a \in \{0; 1; 3\}$ except $\{0; 0\}$

sJIVE:
$r_j \in \{0; 1; 3\}$ and $r_a \in \{0; 1; 3\}$ except $\{0; 0\}$
$\eta \in \{0,1; 0,5; 0,9\}$

IntNMF:
$R \in \{2; 3; 5\}$

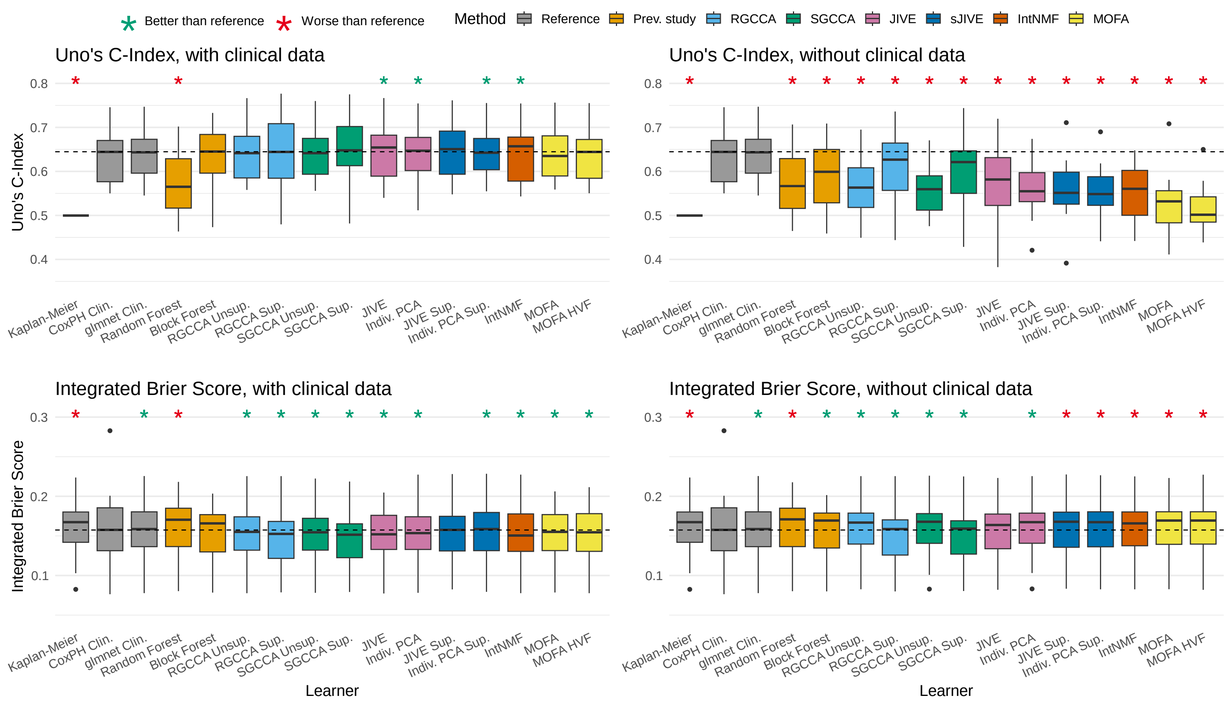
MOFA:
$R = 3$
With and without restriction to Highly Variable Features (HVFs): we keep the top 10% most variable features of each block.
With and without scaling blocks to unit variance.

Results by cancer



Barplots for each class of methods and each cancer type. All tested parameter combinations were averaged. The best class of method for each dataset is written above, and colored green if the average performance of the winning class is better than the mean performance of CoxPH Clin. We consider individual PCA (JIVE with $r_j = 0$) as structured and joint PCA (JIVE with $r_a = 0$) as a JDR method. Dataset are ordered by number of observations.

Results by method



Boxplots of each tested method. Several parameter combinations were tested for each JDR methods, only the best performing ones are represented here. P-values were computed using a paired Wilcoxon test between the folds of the considered methods and those of CoxPH Clin. Combinations kept: RGCCA Sup./Unsup.: $R_l = 1$, $\tau_l = 0.1$, SGCCA Sup./Unsup.: $R_l = 1$, $s_l = 0,5$, JIVE: $r_j = 3$, $r_a = 1$, PCA: $r_j = 0$, $r_a = 1$, JIVE Sup.: $r_j = 3$, $r_a = 1$, $\eta = 0,9$ PCA Sup.: $r_j = 0$, $r_a = 1$, $\eta = 0,9$ IntNMF: $R = 3$, MOFA: 3 factors, scaled MOFA HVF: 3 factors, scaled.

	Best method with Clinical Data		Best method without Clinical Data	
Dataset	C-Index	IBS	C-Index	IBS
BRCA	**RGCCA Sup.**	CoxPH Clin.	CoxPH Clin.	CoxPH Clin.
LUAD	**sJIVE**	CoxPH Clin.	CoxPH Clin.	CoxPH Clin.
LUSC	**Joint PCA ***	**sJIVE**	**Joint PCA**	Priority LASSO *
HNSC	**RGCCA Sup. ***	**SGCCA Sup. ***	**RGCCA Sup. ***	**SGCCA Sup. ***
LGG	LASSO *	**MOFA HVF ***	LASSO	**MOFA HVF ***
UCEC	CoxPH Clin.	CoxPH Clin.	CoxPH Clin.	CoxPH Clin.
BLCA	**IntNMF ***	**SGCCA Sup. ***	**RGCCA Sup.**	**SGCCA Sup. ***
STAD	**sJIVE ***	**IntNMF***	Glmnet Clin. *	CoxPH Clin.
SKCM	**JIVE ***	**sJIVE ***	Glmnet Clin.	Glmnet Clin.
KIRC	**Joint PCA ***	**Joint PCA ***	Glmnet Clin.	CoxPH Clin.
OV	**MOFA HVF**	**MOFA HVF**	**IntNMF**	**IntNMF**
KIRP	**IntNMF ***	Block Forest	**RGCCA Unsup.**	Individual PCA
COAD	**RGCCA Sup. ***	**RGCCA Sup. ***	**RGCCA Sup. ***	**RGCCA Sup *.**
LIHC	Individual PCA *	Ranger *	**RGCCA Sup.**	Ranger *
SARC	**RGCCA Sup. ***	**SGCCA Sup. ***	LASSO	**RGCCA Sup. ***
PAAD	**sJIVE ***	**SGCCA Unsup. ***	Glmnet Clin. *	**SGCCA Unsup.**
ESCA	**SGCCA Unsup.**	**RGCCA Sup.**	CoxPH Clin.	**RGCCA Sup.**
LAML	Block Forest *	**MOFA HVF ***	Ranger *	Ranger *

Conclusion

With clinical data
- **JDR methods** (supervised and unsupervised) **improve both C-Index and IBS.**
- **IntNMF has the best median performance** across datasets.

Without clinical data
- C-index and IBS are inferior across all methods.
- JDR are the best performing non reference methods.
- **Supervised RGCCA and SGCCA** have the highest performance.
- **Supervising JIVE does not seem to improve performance,** as is the case with RGCCA/SGCCA.

Perspectives

- Add **more JDR methods** in the benchmark (MCIA, iCluster, Scikit-Fusion).
- Use an **automatic procedure** to select the **best performing parameters** for each method.
- Include methods capable of extracting **Common, Local and Distinct** (CLD) components, in line with the good performance of JIVE.
- **Design a new method** based on the best performing ones of the benchmark to exploit biological *a priori* information.

References
[1] Herrmann, M., Probst, P., Hornung, R., Jurinovic, V., & Boulesteix, A. L. (2021). Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in bioinformatics*, *22*(3), bbaa167. https://doi.org/10.1093/bib/bbaa167
[2] Tenenhaus, M., Tenenhaus, A., & Groenen, P. J. F. (2017). Regularized Generalized Canonical Correlation Analysis: A Framework for Sequential Multiblock Component Methods. *Psychometrika*, 10.1007/s11336-017-9573-x. Advance online publication. https://doi.org/10.1007/s11336-017-9573-ᵛ
[3] Lock, E. F., Hoadley, K. A., Marron, J. S., & Nobel, A. B. (2013). Joint and Individual Variation Explained (JIVE) for Integrated Analysis of Multiple Data Types. *The annals of applied statistics*, *7*(1), 523–542. https://doi.org/10.1214/12-AOAS597
[4] Chalise, P., & Fridley, B. L. (2017). Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PloS one*, *12*(5), e0176278. https://doi.org/10.1371/journal.pone.0176278
[5] Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., & Stegle, O. (2018). Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, *14*(6), e8124. https://doi.org/10.15252/msb.20178124
[6] Bastien, P., Bertrand, F., Meyer, N., & Maumy-Bertrand, M. (2015). Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data. *Bioinformatics (Oxford, England)*, *31*(3), 397–404. https://doi.org/10.1093/bioinformatics/btu660
[7] Palzer, E. F., Wendt, C. H., Bowler, R. P., Hersh, C. P., Safo, S. E., & Lock, E. F. (2022). sJIVE: Supervised Joint and Individual Variation Explained. *Computational statistics & data analysis*, *175*, 107547. https://doi.org/10.1016/j.csda.2022.107547
[8] Rahman, M. S., Ambler, G., Choodari-Oskooei, B., & Omar, R. Z. (2017). Review and evaluation of performance measures for survival prediction models in external validation settings. *BMC medical research methodology*, *17*(1), 60. https://doi.org/10.1186/s12874-017-0336-2
[9] Verweij, P. J., & Van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Statistics in medicine*, *12*(24), 2305–2314. https://doi.org/10.1002/sim.4780122407

JOBIM 2024 25>28 JUIN Toulouse