



HAL
open science

Support of temporal structure in the statistical analysis of high-throughput proteomic data

Wilfried Heyse, Vincent Vandewalle, Philippe Amouyel, Guillemette Marot,
Christophe Bauters, Florence Pinet

► **To cite this version:**

Wilfried Heyse, Vincent Vandewalle, Philippe Amouyel, Guillemette Marot, Christophe Bauters, et al.. Support of temporal structure in the statistical analysis of high-throughput proteomic data. Journées de Statistique 2021, Jun 2021, Nice, France. hal-03525345

HAL Id: hal-03525345

<https://inria.hal.science/hal-03525345>

Submitted on 13 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

SUPPORT OF TEMPORAL STRUCTURE IN THE STATISTICAL ANALYSIS OF HIGH-THROUGHPUT PROTEOMIC DATA

HEYSE WILFRIED (UMR 1167)

Thesis supervisor : Pr. BAUTERS CHRISTOPHE (UMR 1167)

Co-supervisor : Dr. MAROT GUILLEMETTE (ULR 2694)

Dr. VANDEWALLE VINCENT (ULR 2694)

Team supervisor : Dr. PINET FLORENCE (UMR 1167)



Scientific Context

Survival predictive models

Taking into account temporal structure

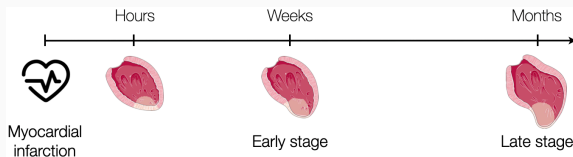
k-means

Mixture model

SCIENTIFIC CONTEXT

Heart failure (HF) : multi-factor disease resulting in the incapacity of the heart to pump enough blood to supply all organs.

→ In France, **70 000** persons die from heart failure each year.



Left Ventricular Remodeling (LVR) : Progressive dilatation of the LV leading to a growth of the LV that occurs in response to myocardial infarction (MI).

LVR quantification is given by :

$$\text{LVR} = \frac{\text{LV volume 1 year after MI} - \text{LV volume after MI}}{\text{LV volume after MI}}$$

→ LVR is an indicator of a high risk of HF or death after MI. (St John Sutton et al., 1994)

REVE-1 AND REVE-2 COHORTS

In order to study LVR and longterm survival, two cohorts were designed by Pr. Bauters where included patients were :

- Affected by a **first myocardial infarction**
- Monitored with blood samples (1 to 4 samples)
- **Followed for up to 13 years** for heart failure or death for cardiovascular reasons

REVE-1 : (2002-2004)

255 patients

77 events observed



REVE-2 : (2006-2008)

238 patients

41 events observed



→ **Over 5000 plasma proteins were measured** out of each plasma sample collected during the two studies

QUESTIONS

1. Can we predict longterm survival using baseline measurements of both cohorts ?
2. Can we build a selection method that support the temporal structure and the high dimension of the data in order to select proteins we could use for prediction based on REVE-2 measurements ?

Scientific Context

Survival predictive models

Taking into account temporal structure

k-means

Mixture model

Objective : Selecting proteins in order to predict a clinical outcome using baseline measurements of the two cohorts.

Statistical framework

- Construction of a predictive model for a variable Y with a set of variables X_1, \dots, X_p measured over n individuals.

Clinical framework

- Y : Time before the occurrence of a event (survival analysis).
- X_1, \dots, X_p : Clinical variables (~ 10) + Proteomic variables (~ 5000)

Statistical difficulties : High dimension of the data ($n \ll p$)

- Considered solution : variable selection and individual clustering \rightarrow identification of patients subtypes

Used model : Cox proportional hazards regression model

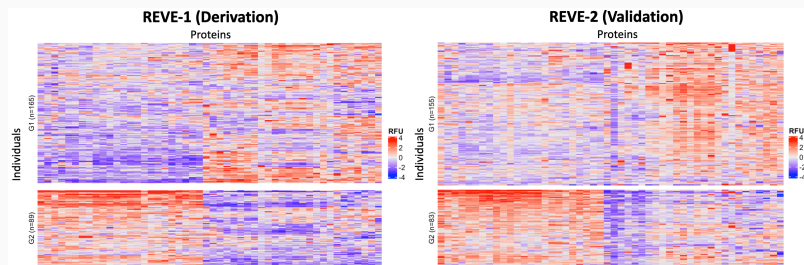
$$\underbrace{h(t, X)}_{\text{Risk function}} = \underbrace{h_0(t)}_{\text{Baseline risk}} \underbrace{\exp(\beta^T X)}_{\text{Variables effect}}$$

Variable Selection : Univariate analysis

Univariate Cox models were fitted for each protein and the **50 proteins** significantly associated to longterm survival were selected.

Clustering of patients : *k*-means clustering was used in order to identify groups of patients based on protein expression related to longterm survival.

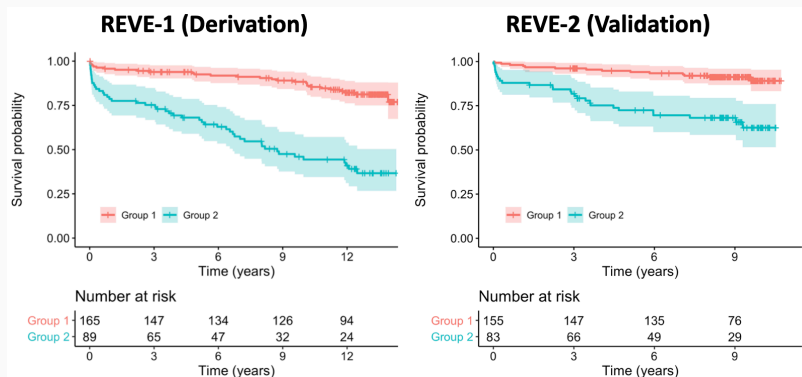
2 clusters of patients were identified on REVE-1 using k-means and were applied to REVE-2 showing **opposite proteomic expression profiles** and **distinct clinical profiles**.



→ Significant differences on Age, Diabetes, Hypertension and WMSI (Wall Motion Systolic Index) between patients of the 2 clusters for both cohorts.

SURVIVAL

Clusters were used as a new variable to predict longterm survival :



→ Cluster effect was significant on both cohorts even when adjusted on clinical variables with significant **hazard ratio of 2.6** for both cohorts after adjustment on clinical variables.

Scientific Context

Survival predictive models

Taking into account temporal structure

- k -means

- Mixture model

Classic goal :

Prediction of Y with baseline measurement

Particularity of the data

Proteomic measures are available at baseline (for both cohorts) and at 3 other times for REVE-2.

Offered outlook :

- Study of the temporal evolution of the protein measurements
- **Proteins clustering** with measures at all times → more meaningful groups

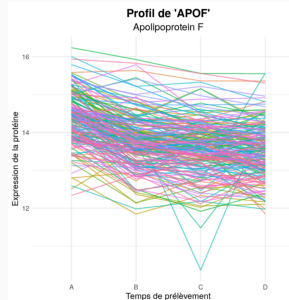
Usefulness for the prediction of Y :

- Construction of a penalty taking in charge the temporal structure with the previously created groups
- **Selected proteins potentially more relevant**

PROTEINS CLUSTERING

Notation

- x_{ijt} : measurement of protein j for patient i at time t
- $\mathbf{x}_j = (x_{ijt})_{i,t}$: all measurements of protein j stored as a $n(= 238) \times T(= 4)$ matrix.



Considered clustering approaches

- *k-means* algorithm
 - on all the data \mathbf{x}_j which could be vectorized as a $n \times T$ vector.
 - on a summary of \mathbf{x}_j (mean slope for example)
- Use of a *mixture model* based approach

$$g(\mathbf{x}) = \sum_{k=1}^G \pi_k f_k(\mathbf{x})$$

Idea : Create groups of proteins varying together

→ Slope formula to sum-up \mathbf{x}_j to a vector of size $T - 1$:

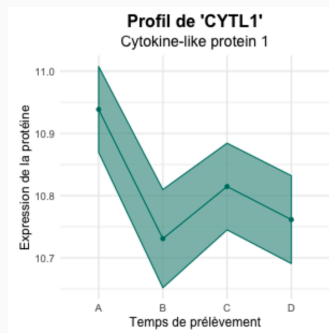
$$d_{jl} = \frac{\bar{x}_{jt_{l+1}} - \bar{x}_{jt_l}}{\sqrt{\frac{\sigma_{jt_l}^2}{n} + \frac{\sigma_{jt_{l+1}}^2}{n}}},$$

where $\bar{x}_{jt_l} = \frac{1}{n} \sum_{i=1}^n x_{ij_{t_l}}$ and $\sigma_{jt_l}^2 = \frac{1}{n} \sum_{i=1}^n (x_{ij_{t_l}} - \bar{x}_{jt_l})^2$.

k-means criteria :

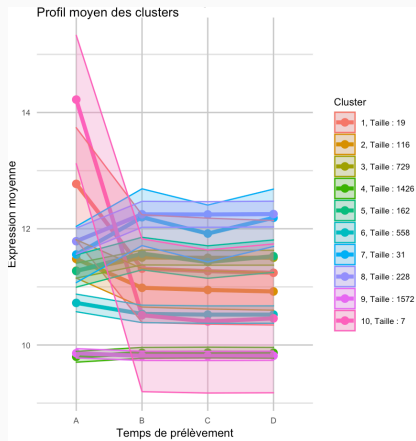
$$\arg \min_{\mu_1, \dots, \mu_G} \sum_{k=1}^G \sum_{\mathbf{d}_j \in S_k} \|\mathbf{d}_j - \mu_k\|^2$$

where $\mathbf{d}_j = (d_{jt_1}, \dots, d_{jt_{T-1}})$.



Number of groups : inertia criteria.

Groups obtained with k -means algorithm are stable (specially for groups with distinctive profiles).



Huge sensitivity to data pre-treatment (slope modelization) and difficulties to interpret groups.

Mixture model for G groups :

$$g(\mathbf{x}) = \sum_{k=1}^G \pi_k f_k(\mathbf{x})$$

where g is the law of a model to model proteins.

Using linear mixed models (Celeux, 2005) we could :

- Use objective statistical criteria to compare models (and groups)
- Control more accurately the temporal structure
- Adapt the model to the data by adding multiple effects

MIXTURE MODEL OF LINEAR MIXED MODELS

Using a mixture model, for each protein \mathbf{x}_j we have :

- $\mathbf{z}_j = (z_{j1}, \dots, z_{jG}) \sim \mathcal{M}(\pi_1, \dots, \pi_G)$ the class of the variable \mathbf{x}_j
- Knowing $\mathbf{x}_j | z_{jk} = 1 \sim MM(\theta_k)$, the protein \mathbf{x}_j will be modeled by :

$$x_{ijt} = \mu_k + b_{ij} + \beta_{kt} + \varepsilon_{ijt}$$

Where :

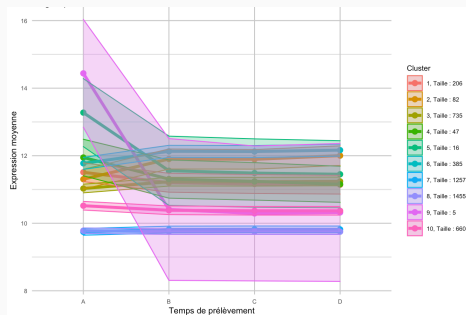
- μ_k the fixed effect of protein of class k
- $b_{ij} \sim \mathcal{N}(0, \sigma_{1,k}^2)$ random effect of individual i for proteins of class k
- β_{kt} the fixed effect of time t for proteins of class k
- $\varepsilon_{ijt} \sim \mathcal{N}(0, \sigma_{2,k}^2)$ the error term for proteins of class k

MIXTURE MODEL OF LINEAR MIXED MODELS

Using this model we obtain classification like this one :

Number of group was decided a priory

High specificity of the groups depending on the effects of the linear mixed model.



Unlike k -means clusters, groups can be interpreted by interpreting the parameters of their linear mixed model.

Prediction of longterm survival

- Selection of proteins
- Meaningful clusters of individuals
- Prediction of longterm survival

Temporal Structure

- Mixture model of mixed model
- Very flexible modelisation of temporal structure
- Use the created groups for the prediction of the outcome

THANK YOU FOR YOUR ATTENTION !