# Risk-Based Stratification Accounting for Group Membership Using the HazClust R Package

Alessandra Ragni[1(✉)], Lara Cavinato[1], and Francesca Ieva[1,2]

[1] MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
{alessandra.ragni,lara.cavinato,francesca.ieva}@polimi.it

[2] Health Data Science Centre, Human Technopole, Viale Rita Levi–Montalcini 1, 20157 Milan, Italy

**Abstract.** For time-to-event data in which units are nested within groups, the `HazClust` R package provides a likelihood-based framework that combines a parametric shared frailty model – suitable for hierarchical data – with spectral clustering through a penalization term based on a graph adjacency matrix. Building on the parametric shared frailty models implemented in the `parfm` R package, `HazClust` employs an iterative algorithm to jointly estimate survival parameters and risk-driven clusters, extending existing methods by integrating clustering directly into the modelling phase. By stratifying units within groups according to their risk of experiencing an event while accounting for group-level effects, `HazClust` supports more informed decision-making in settings with heterogeneous risk patterns – for example, clustering patients within hospitals based on their time-to-event risk of adverse outcomes while adjusting for hospital-level frailty.

**Keywords:** shared frailty models · clustering · hierarchical data · penalization

## 1 Introduction

Time-to-event data often involve units organized into groups, inducing dependence in survival outcomes. While classical frailty models account for such group-level correlation [1], they do not reveal latent subgroups; conversely, survival-informed clustering methods capture heterogeneous risks but often ignore hierarchical dependence.

In this work, we present `HazClust`, an R package publicly available on GitHub[1], built on the `parfm` R package for parametric shared frailty models [2]. In our package, we simultaneously model group dependence and risk-based

[1] Available at https://github.com/alessandragni/HazClust.


F. Martella et al. (Eds.): SIS-FENStatS 2026, ISSSAS, pp. 348–353, 2026.
https://doi.org/10.1007/978-3-032-30665-4_57

clustering within a survival framework, and parameters and clustering structure are learned jointly.

This is specifically addressed by integrating the negative marginal log-likelihood of a parametric shared frailty model with an adaptive spectral clustering penalty. Indeed, rather than specifying the clustering structure a priori, it is inferred from the data through a learned similarity matrix that depends on the model parameters. Joint estimation of survival parameters and clustering structure is achieved by solving iteratively an optimization problem. In this way, clustering is driven by estimated log-hazards, allowing the similarity structure to adapt to the underlying risk patterns.

The remainder of the paper is organized as follows. In Sect. 2, we introduce the proposed model; Sect. 3 describes implementation details and design choices; and Sect. 5 concludes with a discussion and directions for future research.

## 2 Methodology

### 2.1 Notation and Data

With focus on right-censored[2] survival outcomes, let units $i = 1, \ldots, n_g$ belong to groups $g = 1, \ldots, M$. The overall sample size is $N := \sum_{g=1}^{M} n_g$. Denote the observed dataset by $\mathcal{Z} = \{(y_{gi}, \delta_{gi}, \boldsymbol{x}_{gi}) : i = 1, \ldots, n_g,\ g = 1, \ldots, M\}$, where $y_{gi}$ represents the observed time (either event or censoring), $\delta_{gi}$ is the event indicator, and $\boldsymbol{x}_{gi} \in \mathbb{R}^{d \times 1}$ denotes the vector of covariates.

### 2.2 Model Formulation

We propose a framework that jointly estimates model parameters and clustering structure by combining the negative marginal log-likelihood of a parametric shared frailty survival model with an adaptive spectral clustering penalty, where the similarity graph is learned from the data, resulting in the following optimization problem:

$$
\begin{aligned}
\min_{\boldsymbol{\beta}, \boldsymbol{\psi}, \theta, \mathbf{S}, \mathbf{F}} \quad & -\ell(\boldsymbol{\psi}, \boldsymbol{\beta}, \theta \mid \mathcal{Z}) \\
& + \gamma \Big\{ \sum_{g,h=1}^{M} \sum_{i=1}^{n_g} \sum_{j=1}^{n_h} \big( d_{gi,hj}(\boldsymbol{\beta}, \theta, \mathcal{Z})\, s_{gi,hj} + \mu\, s_{gi,hj}^2 \big) + 2\lambda Tr(\mathbf{F}'\mathbf{L}_{\mathbf{S}}\mathbf{F}) \Big\} \\
\text{s.t.} \quad & \sum_{h=1}^{M} \sum_{j=1}^{n_h} s_{gi,hj} = 1, \quad \mathbf{s}_{gi} \geq 0, \quad \forall\, i = 1, \ldots, n_g,\ g = 1, \ldots, M, \\
& \mathbf{F}'\mathbf{F} = \mathbf{I},\ \mathbf{F} \in \mathbb{R}^{N \times C}
\end{aligned}
\tag{1}
$$

[2] Independent and noninformative censoring is assumed.

**Likelihood $\boldsymbol{\ell(\psi, \beta, \theta \mid \mathcal{Z})}$.** The marginal loglikelihood $\ell(\boldsymbol{\psi}, \boldsymbol{\beta}, \theta \mid \mathcal{Z})$ in Eq. (1) is given by

$$\begin{aligned}\ell(\boldsymbol{\psi}, \boldsymbol{\beta}, \theta \mid \mathcal{Z}) = \sum_{g=1}^{M} \Bigg[ \Bigg[ \sum_{i=1}^{n_g} \delta_{gi} \left( \log(h_0(y_{gi}; \boldsymbol{\psi})) + \boldsymbol{x}'_{gi}\boldsymbol{\beta} \right) \Bigg] \\ + \log \left[ (-1)^{d_g} \mathcal{L}^{(d_g)} \left( \sum_{i=1}^{n_g} H_0(y_{gi}; \boldsymbol{\psi}) \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta}) \right) \right] \Bigg] \end{aligned} \quad (2)$$

where $d_g := \sum_{i=1}^{n_g} \delta_{gi}$ is the number of events in group $g$. This likelihood corresponds to a *shared frailty model* [1] defined as

$$h_{gi}(t \mid u_g) = h_0(t; \boldsymbol{\psi})\, u_g \, \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta}), \quad (3)$$

where $\boldsymbol{\beta} \in \mathbb{R}^{d \times 1}$ denotes the vector of fixed-effect coefficients, $u_g \sim f_U(\cdot \mid \theta)$ is the frailty term associated with group $g$, $h_0(t; \boldsymbol{\psi})$ is a parametric baseline hazard function, and $H_0(t; \boldsymbol{\psi}) = \int_0^t h_0(s; \boldsymbol{\psi})\, ds$ is the corresponding cumulative baseline hazard. The marginal log-likelihood in Eq. (2) is further obtained by integrating out the frailty distribution over its support, employing the Laplace transform $\mathcal{L}(s)$ $(s \geq 0)$ such that $(-1)^k \mathcal{L}^{(k)}(s) := (-1)^k \frac{d^k}{ds^k} \mathcal{L}(s)$, see [3].

**Penalization Depending on S.** The similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ represents a probabilistic neighborhood structure, where each entry $s_{gi,hj}$ measures the affinity between two units and is constrained to lie on the probability simplex. The quadratic penalty term $\mu\, s^2_{gi,hj}$ regularizes the similarity matrix and avoids degenerate neighborhood assignments, as discussed in [4]. The tuning parameter $\gamma$ balances the contribution of the clustering component and the survival estimation, following the idea proposed in [5].

**Frailty-Adjusted Distance $\boldsymbol{d_{gi,hj}}(\boldsymbol{\beta}, \theta, \mathcal{Z})$.** Similarity between two units $i$ and $j$, respectively, in groups $g$ and $h$, namely $(g, i)$ and $(h, j)$, is quantified through the frailty-adjusted distance

$$d_{gi,hj}(\boldsymbol{\beta}, \theta, \mathcal{Z}) = \left\| (\boldsymbol{x}'_{gi}\boldsymbol{\beta} + \log u_g) - (\boldsymbol{x}'_{hj}\boldsymbol{\beta} + \log u_h) \right\|^2,$$

which measures squared differences in log-hazard contributions, where $\log h_0(t; \boldsymbol{\psi})$ cancels out in pairwise comparisons. Here, the frailty terms $\log u_g$ and $\log u_h$ are computed by their conditional expectations given the observed data.

**Partition into $\boldsymbol{C}$ Clusters.** Given that we interpret $\mathbf{S}$ as the adjacency matrix of a weighted graph with Laplacian $\mathbf{L_S}$, the trace term $\mathrm{Tr}(\mathbf{F}'\mathbf{L_S}\mathbf{F})$, weighted by $\lambda$, corresponds to a spectral relaxation that encourages a graph structure with $C$ well-separated connected components.

### 2.3 Estimation

To solve Eq. (1), given the large number of unknown parameters, we adopt an iterative scheme that alternates the following three steps until convergence:

- Update of $\mathbf{F}$: With $\mathbf{S}$, $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, and $\theta$ fixed, this step reduces to a spectral decomposition problem. The solution is obtained by selecting the $C$ eigenvectors of the Laplacian matrix $\mathbf{L_S}$ associated with the smallest eigenvalues.
- Update of $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, and $\theta$: Keeping $\mathbf{F}$ and $\mathbf{S}$ fixed, we perform penalized frailty survival estimation following the approach of [2], augmented with an additional similarity-based penalty.
- Update of $\mathbf{S}$: With $\mathbf{F}$, $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, and $\theta$ held constant, $\mathbf{S}$ is updated by solving a constrained quadratic optimization problem. This problem separates across units and admits a closed-form solution for each row; sparsity is enforced by retaining only the $k$ nearest neighbors.

Further implementation details can be found in [6].

## 3 Implementation Details

**Initialization.** The similarity matrix $\mathbf{S}$ is initialized uniformly (or supplied by the user), and the penalty $\lambda$ starts large and is adaptively updated to enforce $C$ connected components. Regression and frailty parameters are set from user input or defaults as in `parfm`, with support for various baseline hazards and frailty distributions.

**Convergence.** The algorithm stops when $\mathbf{S}$ stabilizes, the log-likelihood improvement is minimal, and the number of connected components equals $C$, or when maximum iterations are reached.

**Parameters Tuning.** As $k$ sets the sparsity of $\mathbf{S}$ and influences graph connectivity, small $k$ can over-fragment the graph, while large $k$ can over-connect it; thus, $k$ should be selected by the user over a grid, in accordance with the target number of connected components $C$. The penalization parameter $\gamma$, the number of clusters $C$, and the parametric specifications for the frailty distribution and baseline hazard must all be provided by the user; practical guidance on their selection is discussed in Sect. 4. The parameter $\lambda$, instead, is adaptively tuned within the algorithm to enforce exactly $C$ connected components, while $\mu$ is determined analytically from the update of $\mathbf{S}$ (see Eq. (10) in [6] for further details).

## 4 Experiments and Guidance on the Model Fit

**Simulation Study.** To assess the sensitivity of cluster recovery and parameter estimation to $k$ and $\gamma$, we refer to the simulation study in [6], where survival times are generated from three well-separated latent clusters across 10 groups, using a Gamma frailty (with variance $\theta$) and Weibull baseline hazard (with shape $\rho$ and scale $\xi^\rho$ parameters), subject to administrative censoring. The model recovers the three clusters with at least 95% frequency across all tested values of $\gamma$, achieving 100% for $\gamma \leq 0.1$. As shown in Fig. 1, bias generally increases with $\gamma$, while parameter variability remains stable, with the exception of $\xi$.

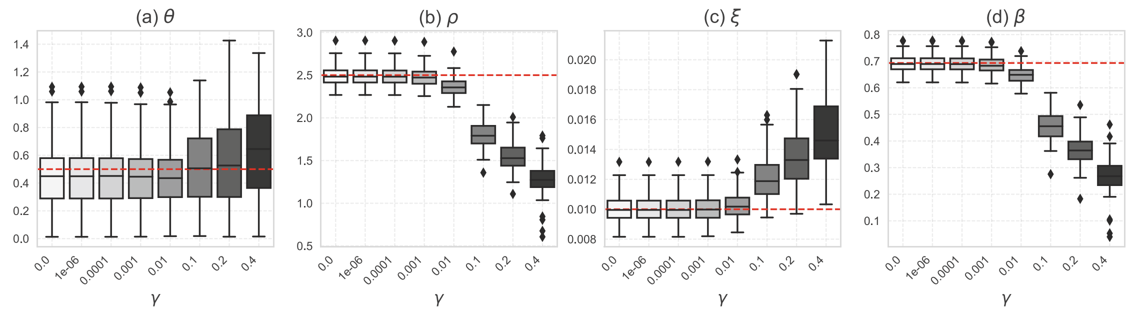


**Fig. 1.** Boxplots of parameter estimates $(\theta, \rho, \xi, \beta)$ across penalization levels $\gamma$ (x-axis) $k = 20$ and $C = 3$. The true parameter values are shown as a piecewise-constant horizontal line.

**Model and Hyperparameter Selection.** The selection of $C$, $k$, $\gamma$, and the parametric specifications for the frailty distribution and baseline hazard require careful evaluation. Clustering quality can be assessed via the mean silhouette coefficient, computed using survival-informed distances, or through the penalized log-likelihood; additional criteria such as stability measures or the gap statistic may further inform the choice, though these are not currently implemented and their use is left to the user. In [6], we select the optimal configuration by maximizing the mean silhouette coefficient over a grid of candidate values for $C$, $k$, $\gamma$, and distributions for baseline hazard and frailty.

**Computational Aspects.** The algorithm combines spectral decomposition, likelihood maximization, and constrained quadratic optimization. As sparsity enforced via $k$-nearest neighbors, in practice, the most computationally demanding step is the estimation of $(\boldsymbol{\beta}, \boldsymbol{\psi}, \theta)$, as it requires numerical maximization of the marginal likelihood (penalized), while eigendecomposition of the graph Laplacian may also become costly for large $N$. Convergence is typically achieved in a limited number of iterations. If the maximum number of iterations is reached before convergence criteria are satisfied, this is reported in the output. Memory requirements may also become relevant for large $N$, due to the storage of the similarity matrix.

## 5 Discussion and Future Work

In this work, we present the `HazClust` R package, which jointly performs unit-level clustering and survival modeling for hierarchical time-to-event data, incorporating shared frailty to account for within-group dependence. Limitations of the current approach include sensitivity to the choice of neighborhood size $k$, which affects cluster formation, as well as the need to pre-specify the number of clusters $C$. Misspecifying the number of clusters $C$ affects both estimation and interpretation. If $C$ is too small, heterogeneous risk profiles are merged, masking meaningful differences. If $C$ is too large, the model may produce unstable or weakly separated clusters, increasing estimation variability. Overall, selecting $C$

involves a bias–variance trade-off, and data-driven criteria are recommended to ensure a reliable clustering structure.

Future developments will focus on more flexible, potentially nonparametric formulations – such as cluster-specific baseline hazards and alternative similarity measures – to better capture heterogeneous risk patterns across patient subgroups, as well as improved criteria for selecting $C$ and supporting model selection.

**Acknowledgments.** The authors acknowledge the financial support from the Department of Mathematics of Politecnico di Milano through the grant 'Dipartimento di Eccellenza 2023–2027' by MUR, Italy. A. Ragni and L. Cavinato are funded by the National Plan for NRRP Complementary Investments 'Advanced Technologies for Human-centred Medicine' (PNC0000003).

## References

1. Duchateau, L., Janssen, P.: The Frailty Model. Springer, New York (2008)
2. Munda, M., Rotolo, F., Legrand, C.: parfm: parametric frailty models in R. J. Stat. Softw. **51**, 1–20 (2012)
3. Munda, M., Legrand, C.: Adjusting for centre heterogeneity in multicentre clinical trials with a time–to–event outcome. Pharm. Stat. **13**(2), 145–152 (2014)
4. Nie, F., Wang, X., Huang, H.: Clustering and projected clustering with adaptive neighbors. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 977–986 (2014)
5. Liu, C., et al.: Supervised graph clustering for cancer subtyping based on survival analysis and integration of multi-omic tumor data. IEEE/ACM Trans. Comput. Biol. Bioinf. **19**(2), 1193–1202 (2020)
6. Ragni, A., Cavinato, L., Ieva, F.: Penalized likelihood optimization for adaptive neighborhood clustering in time-to-event data with group-level heterogeneity. arXiv preprint arXiv:2601.07446 (2026)