



Finanziato  
dall'Unione europea  
NextGenerationEU



Ministero  
dell'Università  
e della Ricerca



# A COVARIATE-ENHANCED HIDDEN MARKOV MODEL FOR RISK PREDICTION IN HEART FAILURE

---

LUCA BRUSA<sup>1</sup>, FULVIA PENNONI<sup>1</sup>,  
FRANCESCO BARTOLUCCI<sup>2</sup>  
([luca.brusa@unimib.it](mailto:luca.brusa@unimib.it))

<sup>1</sup>University of Milano-Bicocca, Department of Statistics and  
Quantitative Methods (Milan, Italy)

<sup>2</sup>University of Perugia, Department of Economics (Perugia, Italy)

# Outline

- 1 The model
- 2 The data
- 3 Results
- 4 Conclusions
- 5 References

# Hidden Markov model: notation and formulation

- **Univariate categorical response variables**  $Y_i^{(t)}$  with  $c$  categories, observed for unit  $i$  at time occasion  $t$
- **Hidden process**  $U_i^{(t)}$ , following a first-order Markov chain with state-space  $\{1, \dots, k\}$
- **Covariates**  $\mathbf{x}_i^{(t)}$ , representing the vector of observed individual covariates for unit  $i$  at time  $t$
- Composed of two sub-models:
  - 1 **Measurement model** (affected by covariates): conditional distribution of the response variable  $Y_i^{(t)}$  given the latent variable  $U_i^{(t)}$  and the covariates  $\mathbf{x}_i^{(t)}$
  - 2 **Latent model** (not affected by covariates): non-parametric distribution of the latent process; accounts for the so-called **unobserved heterogeneity** between individuals

# Hidden Markov model: parameters

- **Initial probabilities:**  $\pi_u = p(U_i^{(1)} = u)$
- **Transition probabilities:**  $\pi_{u|\bar{u}} = p(U_i^{(t)} = u | U_i^{(t-1)} = \bar{u})$
- **Conditional response probabilities:**

$$\phi_{y|ux}^{(t)} = p(Y_i^{(t)} = y | U_i^{(t)} = u, \mathbf{X}_i^{(t)} = \mathbf{x}),$$

relying on the following logit parameterization:

$$\log \frac{\phi_{y|ux}^{(t)} + \dots + \phi_{c-1|ux}^{(t)}}{\phi_{0|ux}^{(t)} + \dots + \phi_{y-1|ux}^{(t)}} = \mu_y = \alpha_u + \mathbf{x}'\boldsymbol{\beta},$$

with:

- $\mu_1, \dots, \mu_{c-1}$ : intercepts specific for each category of the response
- $\alpha_1, \dots, \alpha_k$ : support points corresponding to the latent states
- $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ : regression parameters for the covariates

# Maximum likelihood estimation

- The **expectation-maximization** (EM) algorithm is often employed to perform **maximum likelihood estimation**
- It maximizes the observed-data log-likelihood function  $\ell(\boldsymbol{\theta})$  relying on the **complete-data log-likelihood** function  $\ell^*(\boldsymbol{\theta})$
- It alternates the following steps until convergence:
  - **E-step: compute the conditional expected value** of  $\ell^*(\boldsymbol{\theta})$  given the value of the parameters at the previous step and the observed data
  - **M-step: update the model parameters** by maximizing the expected value of  $\ell^*(\boldsymbol{\theta})$ :
    - explicit solutions are available for  $\pi_u$  and  $\pi_{u|\bar{u}}$
    - a Newton-Raphson algorithm is used for updating  $\alpha$  and  $\beta$

# Forecasting

- **Forecasting probability** of category  $y$  for unit  $i$  at a future time occasion  $t^*$ :

$$\hat{p}_{iy}^{(t^*)} = \sum_{u=1}^k \hat{q}^{(t^*)}(u|\mathbf{x}_i, \mathbf{y}_i) \hat{\phi}_{y|ux}^{(t^*)}$$

- $\hat{\phi}_{y|ux}^{(t^*)}$ : estimated conditional response probabilities at time  $t^*$
- $\hat{q}^{(t^*)}(u|\mathbf{x}_i, \mathbf{y}_i)$ : estimated posterior distribution of the latent variable  $U_i^{(t^*)}$  given the observed responses  $\mathbf{y}_i$  and covariates  $\mathbf{x}_i$
- Event  $y$  is then forecast if  $\hat{p}_{iy}^{(t^*)} > c$ , with  $c \in [0, 1)$  a suitable cutoff chosen according to:
  - either the **Yuoden's J statistics**
  - or the **F1 score**

# Choice of the Cutoff

- Selection through a ***k*-fold cross-validation** procedure:
  - Data are split into training ( $k - 1$  folds) and validation (1 fold) sets
  - The model is estimated on the training folds and forecasting probabilities are computed on the test fold
  - Performance is evaluated using both the **F1 score** and the **Youden's J statistics**, and the (fold-specific) cutoffs are selected by maximizing these quantities
  - The final (overall) cutoffs are obtained by averaging the fold-specific values
- This strategy ensures that the cutoff is **data-driven**, accounts for **class imbalance**, and **avoids overfitting**

# Outline

- 1 The model
- 2 The data
- 3 Results
- 4 Conclusions
- 5 References

# Heart failure

- **Heart failure (HF)** is a complex clinical syndrome in which the heart is unable to pump blood at a rate sufficient to meet the body's needs, or only at the cost of elevated filling pressures
- It is influenced by multiple heterogeneous and interacting **factors**:
  - **Non-modifiable**, like age, sex, genetic predispositions
  - **Modifiable**, like obesity, diabetes, hypertension, smoking, sedentary lifestyle, excessive alcohol consumption
- **Global burden**:
  - Leading cause of morbidity and mortality, affecting ~55.5 million people worldwide (2021)
  - Associated with high hospital readmission rates and substantial healthcare costs (estimated at \$108 billion/year in 2012)
- **Early stages** of HF can be **asymptomatic**; timely identification of cardiac changes is crucial to prevent progression and reduce mortality

# The data

- Retrospective study of **2008 patients** hospitalized with HF at Zigong Fourth People's Hospital (Sichuan, China), observed between **December 2016 and June 2019**
- **Outcomes:** hospital readmission or death, measured at three follow-up times (28 days, 3 months, 6 months after discharge)
- A wide set of patient-level information, measured only at baseline
  - Demographic information
  - Clinical characteristics and comorbidities
  - Laboratory tests and treatments
  - Therapy data for 18 drugs (diuretics, inotropes, vasodilators, others)

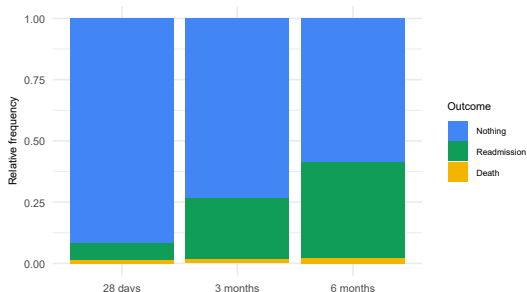
# Data preparation and variable selection

- Preliminary **removal of some variables**:
  - Variables with fewer than two categories (non-informative)
  - Variables with more than 10% missing values
  - Binary variables with one category occurring in less than 0.5% of cases
  - Redundant variables (e.g., weight and height removed, keeping BMI)
- Missing data imputation performed using the **Spectral Regularization Algorithm**
- Removal of variables with excessively high Variance Inflation Factor (VIF) to assess **multicollinearity**
- Final selection of variables using the Akaike Information Criterion (AIC) to balance model fit and complexity

# Response variable

- **Categorical response variable**  $Y_{it}$  with **three categories**:

$$Y_i^{(t)} = \begin{cases} 2 & \text{if patient } i \text{ died at time } t \\ 1 & \text{if patient } i \text{ was readmitted at time } t \\ 0 & \text{otherwise} \end{cases}$$



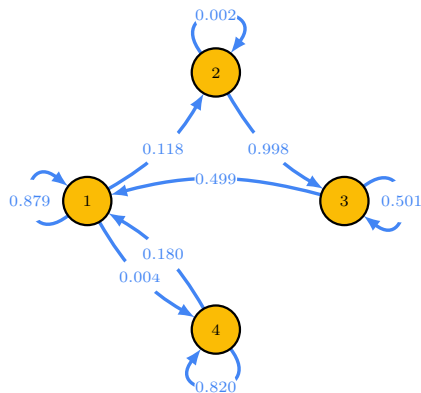
# Outline

- 1 The model
- 2 The data
- 3 Results**
- 4 Conclusions
- 5 References

# HM model estimation

- The final dataset consists of **1981 patients**, observed over **3 time points**, including 29 covariates and one categorical response variable
- The HM model is estimated with a number of latent states  $k$  ranging from 1 to 5
- For each  $k$ , the model is estimated  **$20 \cdot k$  times**, using both deterministic and random initializations to reduce the risk of local maxima
- The optimal number of states is selected using the **Akaike Information Criterion (AIC)**, which identifies **4 latent states**
- Latent states are interpreted as **patient risk profiles**

# Estimated model parameters



State	$\hat{\pi}_u$	$\hat{\alpha}_u$
1	0.728	-5.265
2	0.086	12.750
3	0.172	13.597
4	0.015	27.233

Table: Estimated initial probabilities ( $\hat{\pi}_u$ ) and support points ( $\hat{\alpha}_u$ )

Figure: Estimated transition probabilities ( $\hat{\pi}_{u|\bar{u}}$ )

# Estimated regression coefficients

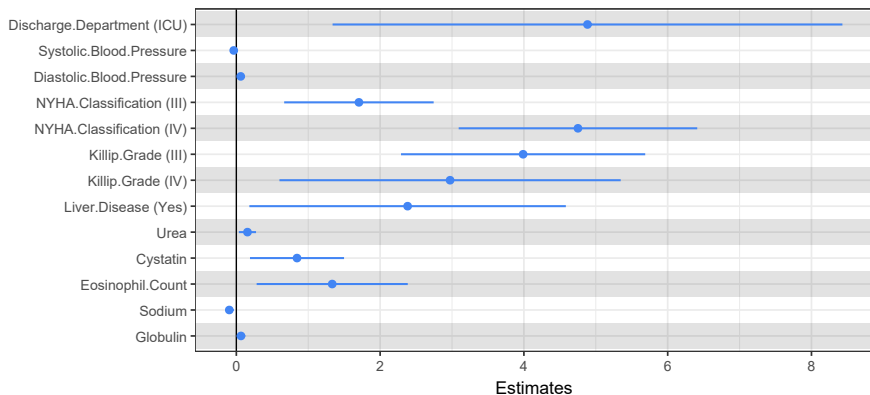
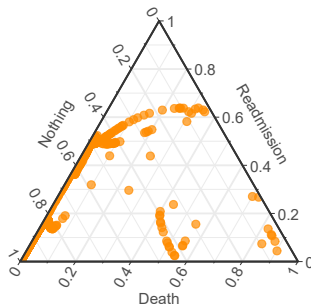


Figure: Point estimate and corresponding confidence interval (confidence level: 95%) for some of the considered covariates

# Forecasting

- The HM model (estimated using  $k = 4$  latent states) is trained on data from **times 1 and 2** and used to forecast responses at **time 3**
- For each subject a vector of forecasting probabilities is defined:

$$\hat{\mathbf{p}}_i^{(t^*)} = (\hat{p}_{\text{NOTHING}}^{(t^*)}, \hat{p}_{\text{READMISSION}}^{(t^*)}, \hat{p}_{\text{DEATH}}^{(t^*)})$$



# Forecasting results - Metrics

Cutoff selected through **F1 score**

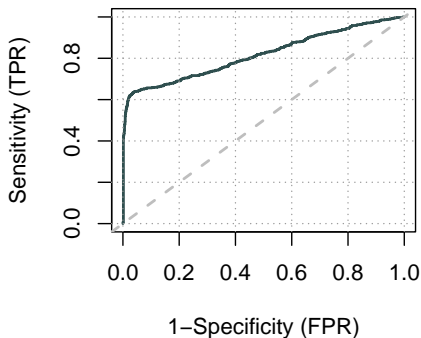
	Readmission	Death
Cutoff ( $c$ )	0.231	0.518
Specificity	0.983	1.000
Sensitivity/Recall	0.587	0.500
Precision	0.958	1.000
Accuracy	0.829	0.988
F1 score	0.728	0.667

Cutoff selected through **Youden's J statistics**

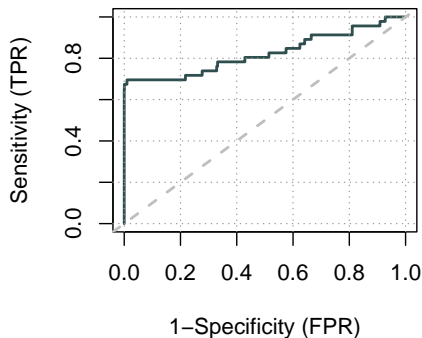
	Readmission	Death
Cutoff ( $c$ )	0.163	0.448
Specificity	0.942	1.000
Sensitivity/Recall	0.614	0.630
Precision	0.942	1.000
Accuracy	0.835	0.991
F1 score	0.744	0.773

# Forecasting results - ROC curve

**Readmission (AUC = 0.817)**



**Death (AUC = 0.824)**



# Outline

- 1 The model
- 2 The data
- 3 Results
- 4 Conclusions**
- 5 References

# Conclusions

- **Main contributions:**

- Application of a **hidden Markov model** for forecasting with a categorical response variable with more than two categories
- Data-driven **selection of the cutoff** through cross-validation, explicitly accounting for class imbalance

- **Future work:**

- Exploration of alternative strategies for cutoff selection, such as **cost-sensitive approaches**
- Simulation study to **better assess** the performance and robustness of the proposed method
- Extension of the model to **longer follow-up times** and possibly **time-varying covariates**

# Outline

- 1 The model
- 2 The data
- 3 Results
- 4 Conclusions
- 5 References**

# References

- BARTOLUCCI, F. AND FARCOMENI, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent markov heterogeneity structure. *J. Am. Stat. Assoc.*, **104**, 816–831.
- BARTOLUCCI, F., FARCOMENI, A., AND PENNONI, F. (2013). *Latent Markov models for longitudinal data*. Chapman & Hall/CRC, Boca Raton.
- BRAUNWALD, E., ZIPES, D. P., AND LIBBY, P. (2001). *Heart disease: a textbook of cardiovascular medicine (6th ed.)*. Saunders.
- DEMPSTER, A., LAIRD, N., AND RUBIN, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. Series B Stat. Methodol.*, **39**, 1–38.
- MAZUMDER, R., HASTIE, T., AND TIBSHIRANI, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, **11**, 2287–2322.
- WELCH, L. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Inform. Theory Soc. Newsl.*, **50**, 10–13.
- ZHANG, Z., CAO, L., CHEN, R., ZHAO, Y., LV, L., XU, Z., AND XU, P. (2021). Electronic healthcare records and external outcome data for hospitalized patients with heart failure. *Scientific Data*, **8**, 1–6.

# Acknowledgments

**Acknowledgment:** L. Brusa, F. Pennoni, and F. Bartolucci acknowledge the financial support from the grant “Hidden Markov Models for Early Warning Systems” of Ministero dell’Università e della Ricerca (PRIN 2022TZEXKF) funded by European Union - Next Generation EU, Mission 4, Component 2, CUP J53D23004990006.