

# Bayesian Modelling of Metabolic Syndrome Risk in AVIS Blood Donors

An analysis of routine laboratory data

Noemi Bongiorno, Martina Caliendo, Davide Marchesi,  
Greta Minazzi, Elisa Nordera, Matteo Zanetti

**Tutor:** Simone Colombara

Bayesian Statistics 2025/2026

February 19, 2026



**POLITECNICO**  
MILANO 1863

- ▶ 100,203 laboratory measurements.
- ▶ 4,300 donors - longitudinal data.
- ▶ 5 target variables: Blood pressure, BMI, Glucose, HDL Cholesterol, Triglycerides.

## Research question:

Is it feasible to implement a **Metabolic Syndrome screening**, based on AVIS laboratory measurements?  
Is this able to identify the predictive biomarkers for high-risk patients and support clinicians in treatment strategies?

Hierarchical multivariate Bayesian model:

$$\mathbf{y}_{ij} \mid \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma} \stackrel{\text{ind}}{\sim} \mathcal{N}_K(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}), \quad i = 1, \dots, I, j = 1, \dots, n_i$$

$$\boldsymbol{\mu}_{ij} = \mathbf{x}_{ij} \boldsymbol{\beta} + \mathbf{b}_i$$

$K$  : number of target variables (5)

$I$  : number of donors (35)

$n_i$ : number of visits per donor

$\boldsymbol{\beta}$ : fixed effects

$\mathbf{b}_i$ : donor specific random intercept

## Residual covariance: LKJ

$$\Sigma = \mathbf{L}_\Sigma \mathbf{L}_\Sigma^\top, \quad \mathbf{L}_\Sigma = \text{diag}(\boldsymbol{\tau}) \mathbf{L}_\Omega, \quad \mathbf{L}_\Omega \sim \text{LKJ-Cholesky}(4)$$

$$\tau_k \stackrel{\text{iid}}{\sim} \mathcal{N}^+(0, 0.5^2)$$

## Random effects: Gaussian

$$\mathbf{b}_i \mid \Sigma_b \sim \mathcal{N}(\mathbf{0}, \Sigma_b)$$

$$\Sigma_b = \mathbf{L}_{\Sigma_b} \mathbf{L}_{\Sigma_b}^\top, \quad \mathbf{L}_{\Sigma_b} = \text{diag}(\boldsymbol{\tau}) \mathbf{L}_{\Omega_b}$$

$$\mathbf{L}_{\Omega_b} \sim \text{LKJ-Cholesky}(4)$$

$$\tau_k \stackrel{\text{iid}}{\sim} \mathcal{N}^+(0, 0.5^2)$$

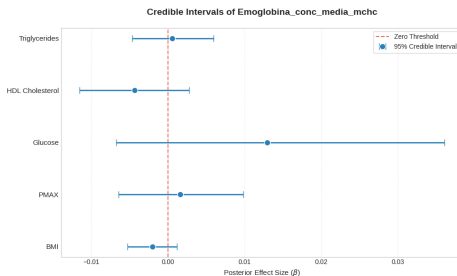
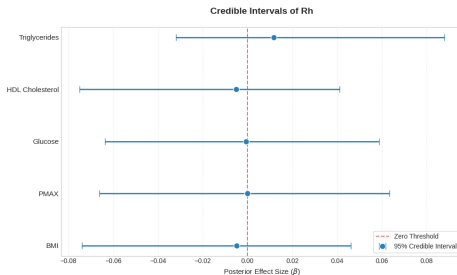
## Fixed effects: Bayesian Lasso

$$\beta_{pk} \mid \tau_{pk} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \tau_{pk})$$

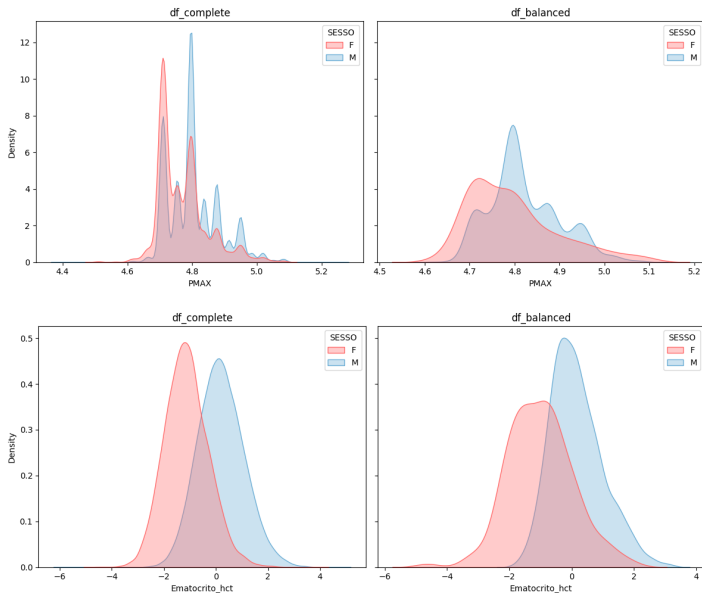
$$\tau_{pk} \stackrel{\text{iid}}{\sim} \text{Exponential}\left(\frac{\lambda^2}{2}\right)$$

$$\lambda \sim \text{Gamma}(1, 1)$$

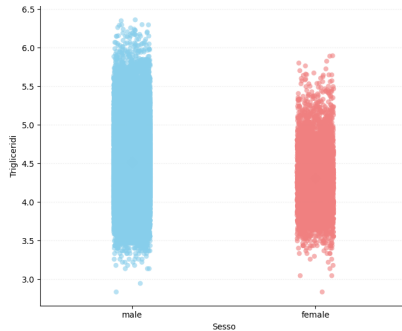
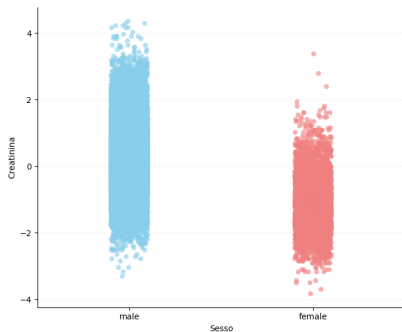
$i=1, \dots, I$  (35): donors' index  
 $p=1, \dots, P$  (23): covariate index  
 $k=1, \dots, K$  (5): target index



**Non-informative:**  
Basophili\_perc  
Creatinina  
Emoglobina  
Ferritina  
Ferro\_totale  
Data  
Data\_nascita  
Rh  
ABO  
SESSO.



## Creatinina and Trigliceridi values of males and females



## Fixed Effects:

$$\beta_{pk} \stackrel{iid}{\sim} \mathcal{N}(0, 2^2)$$

## Random Effects:

Idea (DPMM):

$$\mathbf{b}_i \mid \boldsymbol{\theta}_i \stackrel{ind}{\sim} k(\cdot; \boldsymbol{\theta}_i)$$

$$\boldsymbol{\theta}_i \mid P \stackrel{iid}{\sim} P$$

$$P \sim DP(\alpha, P_0)$$

$k(\cdot; \boldsymbol{\theta}_i)$  is a multivariate Gaussian density

Actually:

$$P = \sum_{m=1}^M \pi_m \delta_{\boldsymbol{\theta}_m^*}, \quad \pi_1, \dots, \pi_M : \text{stick-breaking weights}$$

$$\boldsymbol{\theta}_m^* = (\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

$$\boldsymbol{\mu}_m \sim \mathcal{N}_K(\mathbf{0}, 2^2 \mathbf{I})$$

$$\boldsymbol{\Sigma}_m = \mathbf{L}_m \mathbf{L}_m^\top, \quad \mathbf{L}_m = \text{diag}(\boldsymbol{\tau}_m) \mathbf{L}_{\Omega_m}$$

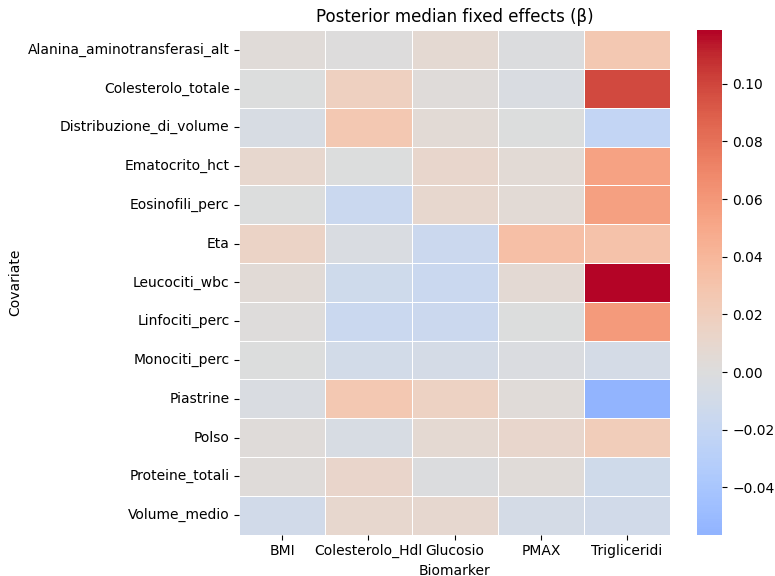
$$\mathbf{L}_{\Omega_m} \sim \text{LKJ-Cholesky}(4), \quad \tau_{mk} \sim \mathcal{N}^+(0, 0.5^2)$$

$p = 1, \dots, P$  (23): covariate index

$k = 1, \dots, K$  (5): target index

$i = 1, \dots, I$  (35): donor index

$m = 1, \dots, M$  (20): cluster index



## Cluster-level mean random effects

$$\mu_m =$$

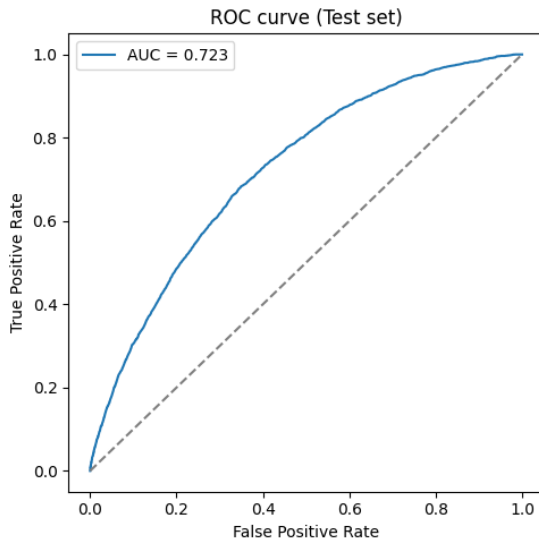
	PMAX	Glucosio	Trigliceridi	HDL	BMI
Cluster 1	4.7929	4.5373	4.6824	3.9904	3.2908

## Covariance matrix of random effects

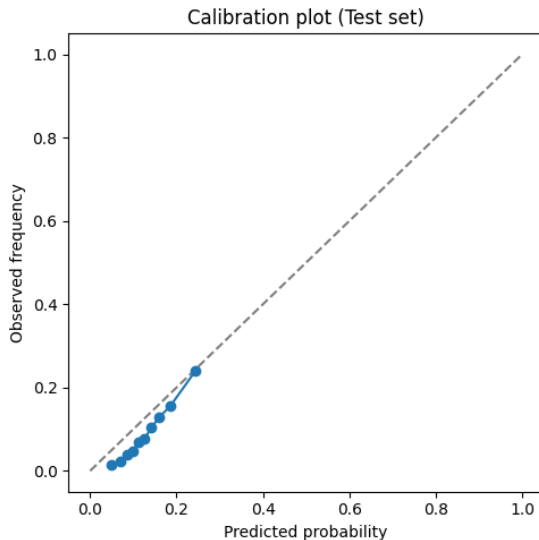
$$\Sigma_m =$$

	<i>P</i>	<i>G</i>	<i>T</i>	<i>H</i>	<i>B</i>
<i>P</i>	0.0020	0.0008	0.0014	-0.0010	0.0016
<i>G</i>	0.0008	0.0044	0.0015	-0.0030	0.0021
<i>T</i>	0.0014	0.0015	0.0856	-0.0279	0.0131
<i>H</i>	-0.0010	-0.0030	-0.0279	0.0303	-0.0051
<i>B</i>	0.0016	0.0021	0.0131	-0.0051	0.0114

# Predictive performance: ROC curve



# Predictive performance: calibration plot



## Key Findings:

- ▶ Strong predictive performance
- ▶ Interpretable associations at the population level
- ▶ No evidence of latent subgroups: population appears continuous
- ▶ Flexible hierarchical Bayesian model providing an early risk flag and interpretable biomarkers

---

## Further Developments:

- ▶ Employ a larger dataset
- ▶ Validate interpretation findings with a field specialist

- ▶ Park, T., & Casella, G. (2008). *The Bayesian Lasso*. Journal of the American Statistical Association, 103(482), 681–686.  
<https://doi.org/10.1198/016214508000000337>
- ▶ Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). *The horseshoe estimator for sparse signals*. Biometrika, 97(2), 465–480.  
<https://doi.org/10.1093/biomet/asq017>
- ▶ Piironen, J., & Vehtari, A. (2017). *Sparsity information and regularization in the horseshoe and other shrinkage priors*. Electronic Journal of Statistics, 11(2), 5018–5051. <https://doi.org/10.1214/17-EJS1337SI>
- ▶ Yang, S., Rouder, J. N. Assessing Two Common Priors of Covariance in Hierarchical Designs. *OFS*. [https://doi.org/10.31234/osf.io/jen65\\_v2](https://doi.org/10.31234/osf.io/jen65_v2)
- ▶ Guglielmi, A. (2025). [Bayesian Statistics Slides]. WeBeep, Politecnico di Milano.

**The complete source code and the scripts for this project are available on GitHub at the following repository:**

`https://github.com/elisanordera/  
Bayesian-Statistics-Project`.