Bayesian Modelling of Metabolic Syndrome Risk in AVIS Blood Donors

A longitudinal analysis of routine laboratory data

Noemi Bongiorni, Martina Caliandro, Davide Marchesi, Greta Minazzi, Elisa Nordera, Matteo Zanetti

Tutor: Simone Colombara

Bayesian Statistics 2025/2026

28th November 2025

POLITECNICO
MILANO 1863

Context and dataset

POLITECNICO MILANO 1863

- 100,203 laboratory measurements.
- 4,300 donors - longitudinal data.
- 42 variables per visit.

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?

Definition of metabolic syndrome

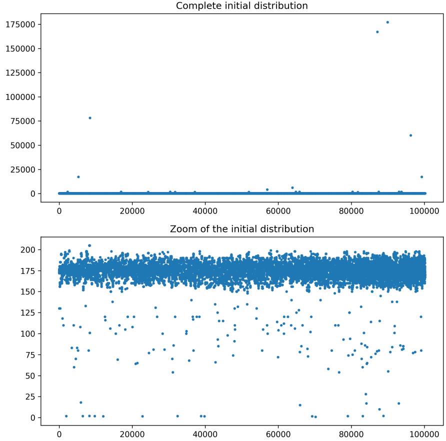
- Metabolic Syndrome (MetS) is a **cluster of conditions** that increase the risk of type 2 diabetes, myocardial infarction, stroke, and other cardiovascular diseases.
- Diagnosis requires at least **three** of the five diagnostic components out of range.

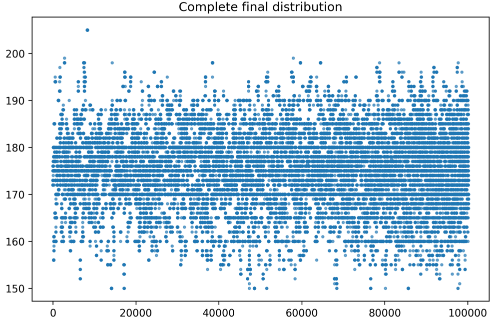
Component	**Men**	**Women**
Waist circumference (cm)	> 102	> 88
Blood pressure (mmHg)	$\geq 130 / \geq 85$	SAME
Fasting Glucose (mg/dL)	≥ 100	SAME
Triglycerides (mg/dL)	≥ 150	SAME
HDL cholesterol (mg/dL)	< 40	< 50

Component	**Men**	**Women**
BMI (kg/m^2)	≥ 30	≥ 30
Blood pressure (mmHg)	$\geq 130 / \geq 85$	SAME
Fasting Glucose (mg/dL)	≥ 100	SAME
Triglycerides (mg/dL)	≥ 150	SAME
HDL cholesterol (mg/dL)	< 40	< 50

Initial Exploration

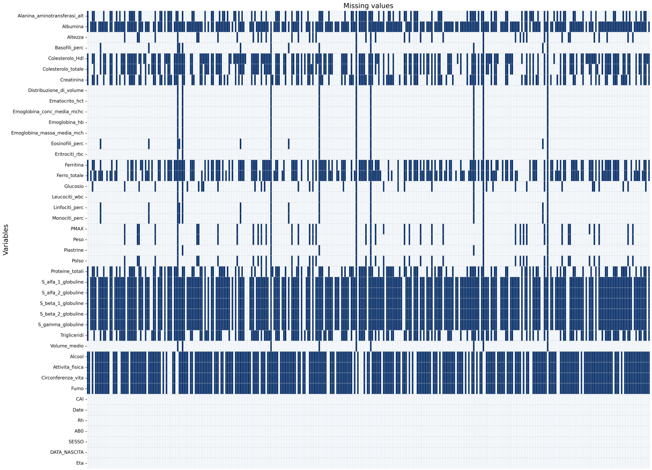
Covariate 'Height' [cm]





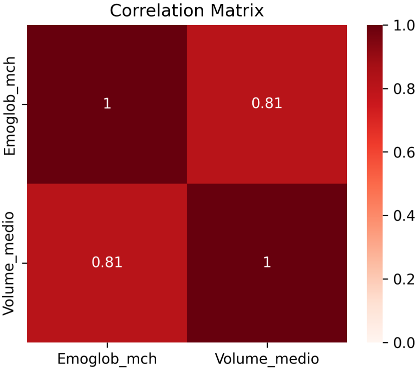
The Missing Data Problem

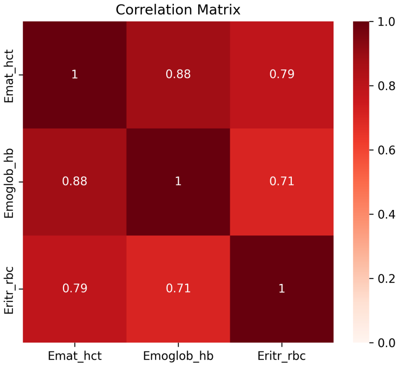
POLITECNICO
MILANO 1863



Solution: **MICE** - Multiple Imputation by Chained Equations (*Van Buuren, et al. 2011*)

Transformations & Variable Selection





- A reduced dataset of about $n \approx 1000$ observations was obtained by retaining all donors with more than 25 visits.

The Bayesian Model: Structure

Hierarchical multivariate Bayesian model:

$$\boldsymbol{y}_{ij} \mid \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma} \overset{\text{ind}}{\sim} \mathcal{N}_K(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}), \qquad i = 1, \ldots, I, \; j = 1, \ldots, n_i$$

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

$$\boldsymbol{\Sigma} \sim \text{Inv-Wishart}(a_\Sigma, \; b_\Sigma \boldsymbol{I}_K)$$

K : number of target variables (5)

$a_\Sigma = 10, \; b_\Sigma = 1$

I : number of donors (34)

n_i: number of visits per donor

The Bayesian Model: Priors

Random effect:

$$\boldsymbol{b}_i \overset{\text{iid}}{\sim} \mathcal{N}_K(\boldsymbol{\mu_b}, \boldsymbol{\Sigma_b}) \qquad i = 1, \dots, I$$

$$\boldsymbol{\mu_b} \sim \mathcal{N}_K(\boldsymbol{0}, \boldsymbol{I_K}) \qquad \boldsymbol{\Sigma_b} \sim \text{Inv-Wishart}(a_{\Sigma_b},\ b_{\Sigma_b}\, \boldsymbol{I_K})$$

Fixed effects - Regularized Horseshoe (*Tew, et al. 2022*):

$$\tilde{\beta}_{pk} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \qquad p = 1, \dots, P+1, \qquad k = 1, \dots, K$$

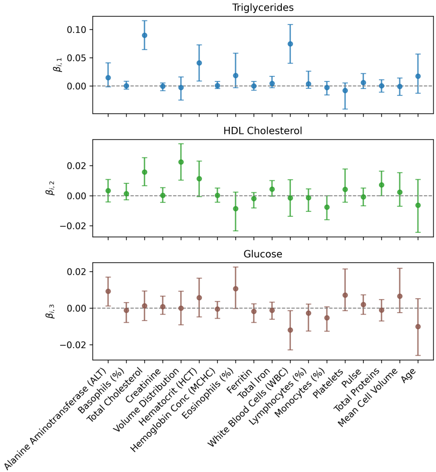
$$\beta_{pk} = \tilde{\beta}_{pk}\, \tau\, \lambda_p \frac{c}{\sqrt{c^2 + \tau^2 \lambda_p^2}}$$

$$\lambda_p \overset{\text{iid}}{\sim} \mathrm{HC}(0, 1), \quad \tau \sim \mathrm{HC}(0, \tau_0), \quad c^2 \sim \mathrm{Inv\text{-}Gamma}(a_c, b_c)$$

$a_{\Sigma_b} = 10,\ b_{\Sigma_b} = 0.5,\ \tau_0 = 0.002,\ a_c = 2,\ b_c = 8$

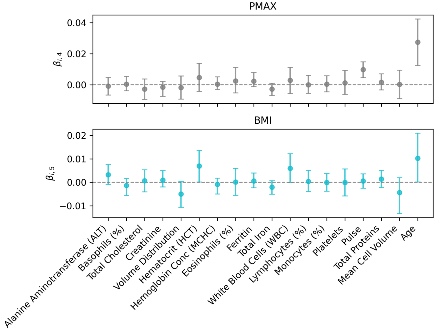
P : number of covariates (23)

K : number of target variables (5)



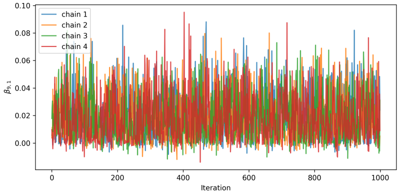
95% Credible Intervals

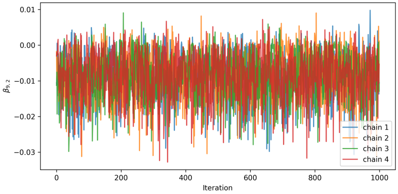
of fixed effect

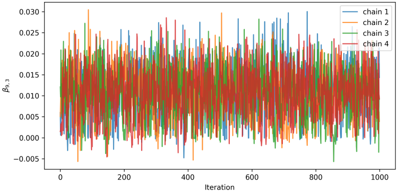


POLITECNICO
MILANO 1863

Trace plots of fixed effects of *Eosinophils* on *Triglycerides*, *HDL Cholesterol* and *Glucose*







Next Steps

POLITECNICO MILANO 1863

- Investigate a **new prior for** Σ, e.g. LKJ (*Wang, et al. 2018*).
- Experiment with **different priors for the random effects**, including Dirichlet process priors.
- Use the **complete dataset** on the relevant covariates identified.
- Fine-tune the **model hyperparameters**.
- Better exploit the **longitudinal structure** of the data.

References

- Samson, S. L., Garber, A. J. (2014). *Metabolic syndrome. Endocrinology and Metabolism Clinics*, 43(1), 1-23.

- Van Buuren, S., Groothuis-Oudshoorn, K. (2011). *mice: Multivariate imputation by chained equations in R. Journal of statistical software*, 45, 1-67.

- Tew, S. Y., Schmidt, D. F., Makalic, E. (2022, September). *Sparse horseshoe estimation via expectation-maximisation. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 123-139)*. Cham: Springer Nature Switzerland.

- Wang, Z., Wu, Y., Chu, H. (2018). *On equivalence of the LKJ distribution and the restricted Wishart distribution.*