Enhancing the fight against doping: using copulae for multivariate testing analysis on ABP profiles

Nina Deliu & Brunero Liseo
Sapienza Università di Roma

SAPIENZA
UNIVERSITÀ DI ROMA

XII Giornate della Ricerca MEMOTEF
31 May – 1 June 2022, Rome

WADA's ABP Program

WADA's Mission

- The campaign to achieve doping-free sport is international and is headed by the *World Anti-doping Angency* (WADA)
- WADA promotes, coordinates and monitors the fight against doping through the *World Anti-Doping Code* – the core document harmonising anti-doping policies with eight international standards, including:
  - The Prohibited List
  - Standard for Testing & Investigations

[1] World Anti-Doping Code 2021, page 21, Comment 9 [WADA, 2021]

WADA's Mission

- The campaign to achieve doping-free sport is international and is headed by the *World Anti-doping Angency* (WADA)
- WADA promotes, coordinates and monitors the fight against doping through the *World Anti-Doping Code* – the core document harmonising anti-doping policies with eight international standards, including:
  - The Prohibited List
  - Standard for Testing & Investigations
- Article 2.2[1]: *Use* or *Attempted Use* by an *Athlete* of a *Prohibited Substance* or *Method* may also be established by other reliable means that include conclusions drawn from longitudinal *Athlete Biological Passport* (ABP) profiling

[1] World Anti-Doping Code 2021, page 21, Comment 9 [WADA, 2021]

The ABP Program

- The fundamental principle of the ABP is to monitor over time athletes' individual profiles – with respect to selected *biomarkers*
- Currently, ABP profiles are established according to two modules:
  1. The **Haematological Module** (2009) – collects *markers of blood doping*
  2. The **Steroidal Module** (2014) – collects *markers of steroid doping*

The ABP Program

- The fundamental principle of the ABP is to monitor over time athletes' individual profiles – with respect to selected *biomarkers*
- Currently, ABP profiles are established according to two modules:
  1. The **Haematological Module** (2009) – collects *markers of blood doping*
  2. The **Steroidal Module** (2014) – collects *markers of steroid doping*
- Standardization and harmonization of ABP programs is achieved through the use of ADAMS: an online database management tool for data entry, storage, sharing, **analyses and decision making**, and reporting
- Doping violations are detected by noting *deviations* **in each individual biomarker** from an athlete's established levels (i.e., *marker* values outside permissible limits), based on the **ADAPTIVE Model** [Sottas et al., 2007]

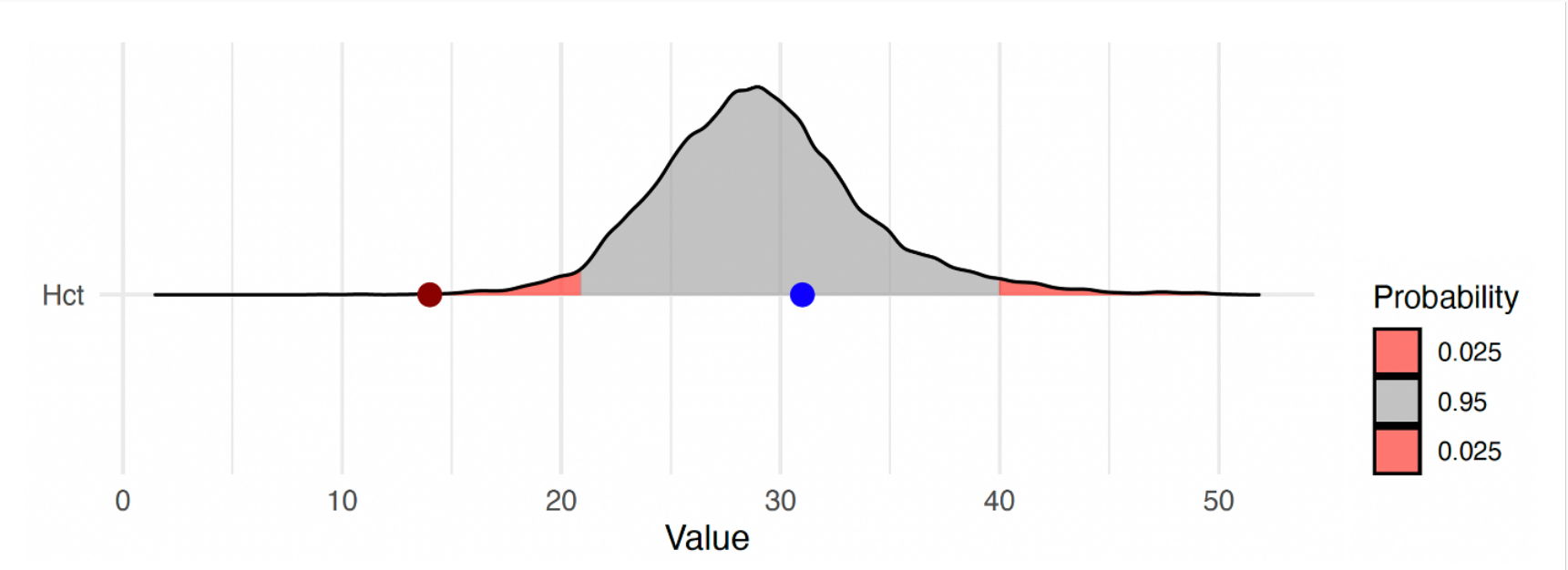
ADAPTIVE Model

The State of the Art: Idea

Given a sequence of measurements of a *biomarker* Y on a single athlete, say $\mathbf{y_t} = (y_1, y_2, \ldots, y_t)$, the *idea* is to produce a predictive distribution for $y = y_{t+1}$

$$p(y|y_1, \ldots, y_t) = \int_\mu \int_\sigma p(y|\mu, \sigma)\, \pi(\mu, \sigma|y_1, \ldots, y_t) d\mu d\sigma$$

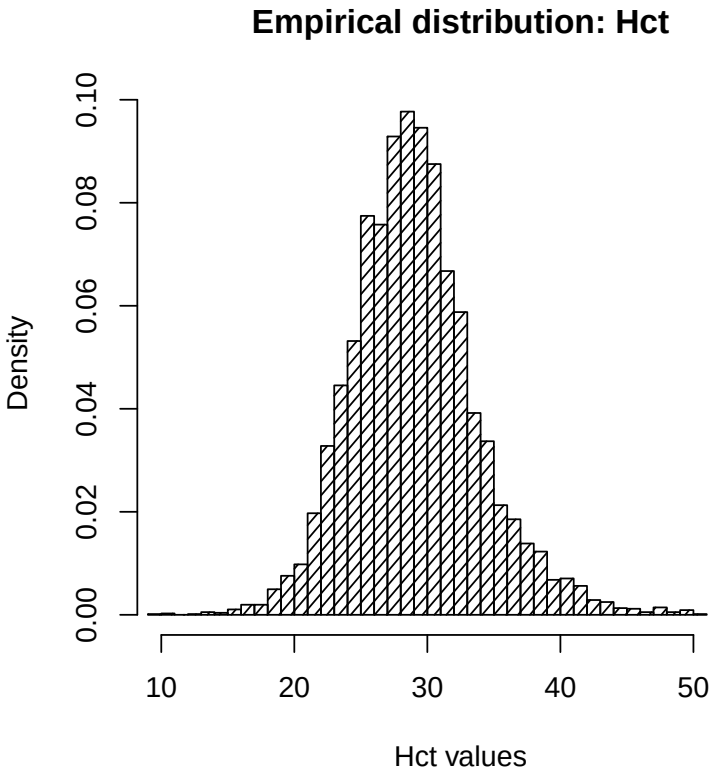
and check whether the observed y_{t+1} falls between certain tolerance limits, pre-specified according to α (e.g., 0.05)

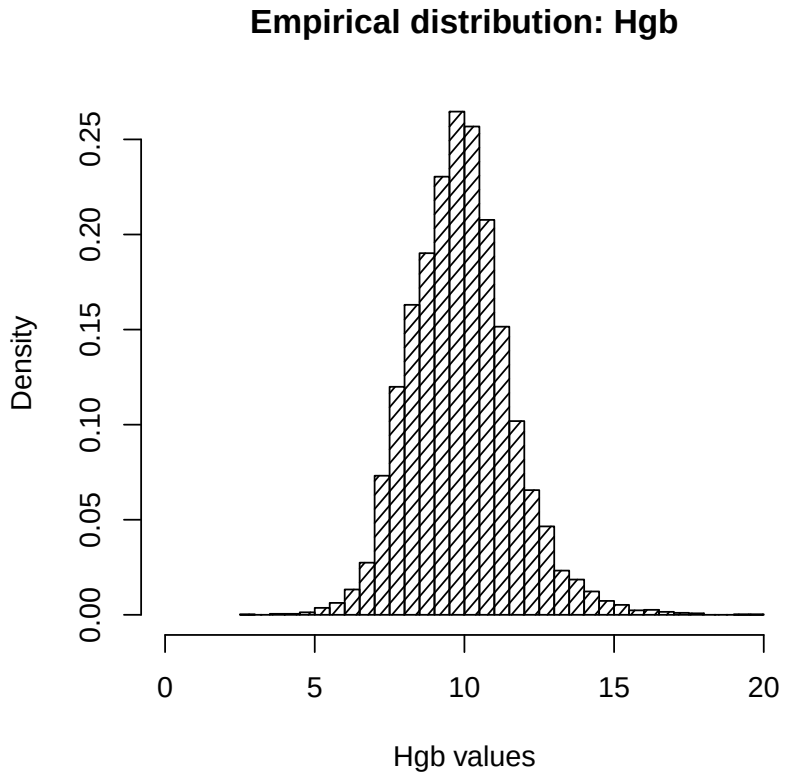


The State of the Art: Assumptions (1)

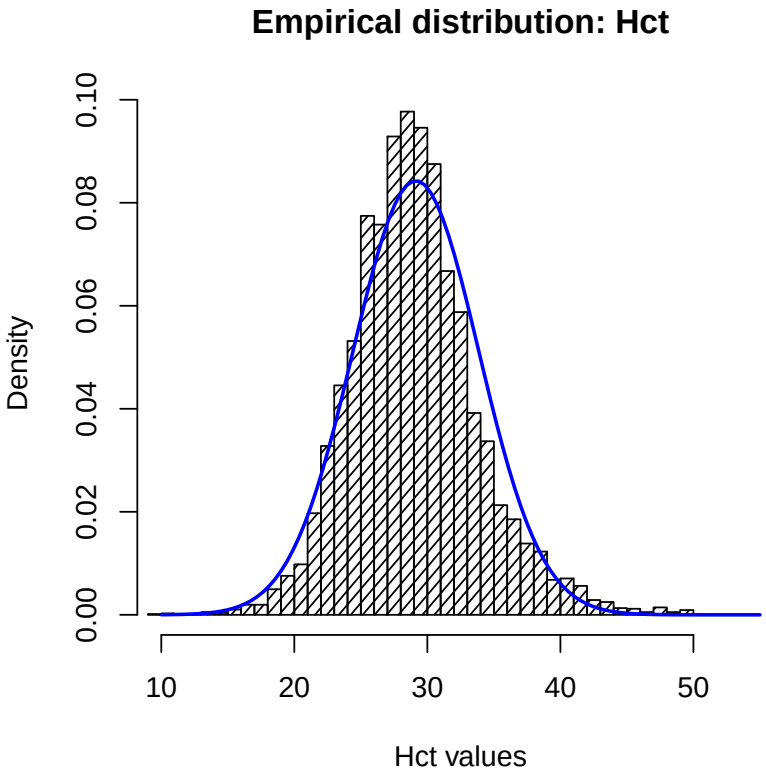
- **Independence** of intra-individual data (over time)
- **Normality**: $Y_1, Y_2, \ldots, Y_t, Y_{t+1}$ i.i.d. $\sim N(\mu, \sigma)$
- (μ, σ) accounts for the inter-individual and intra-individual variation:
  - a prior distribution $\pi(\mu, \sigma)$ can be elicited from a population of "controls": clean athletes or volunteers (non athletes)
  - the tolerance limit(s) are then calibrated based on the individual observed history
- **Numerical integration** is used for deriving the predictive distribution

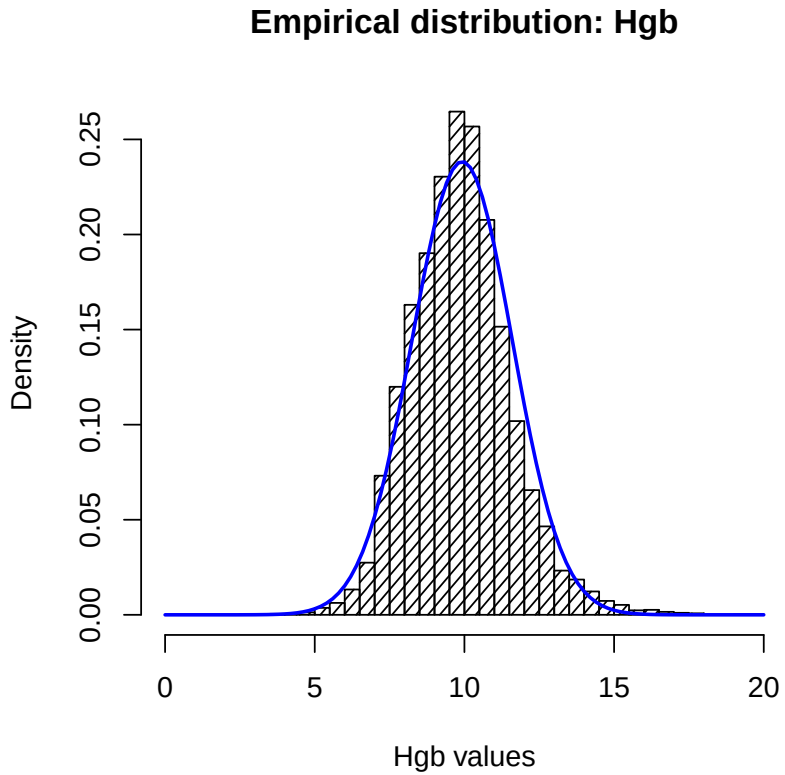
Statistical Issues (1): Does it look normal?





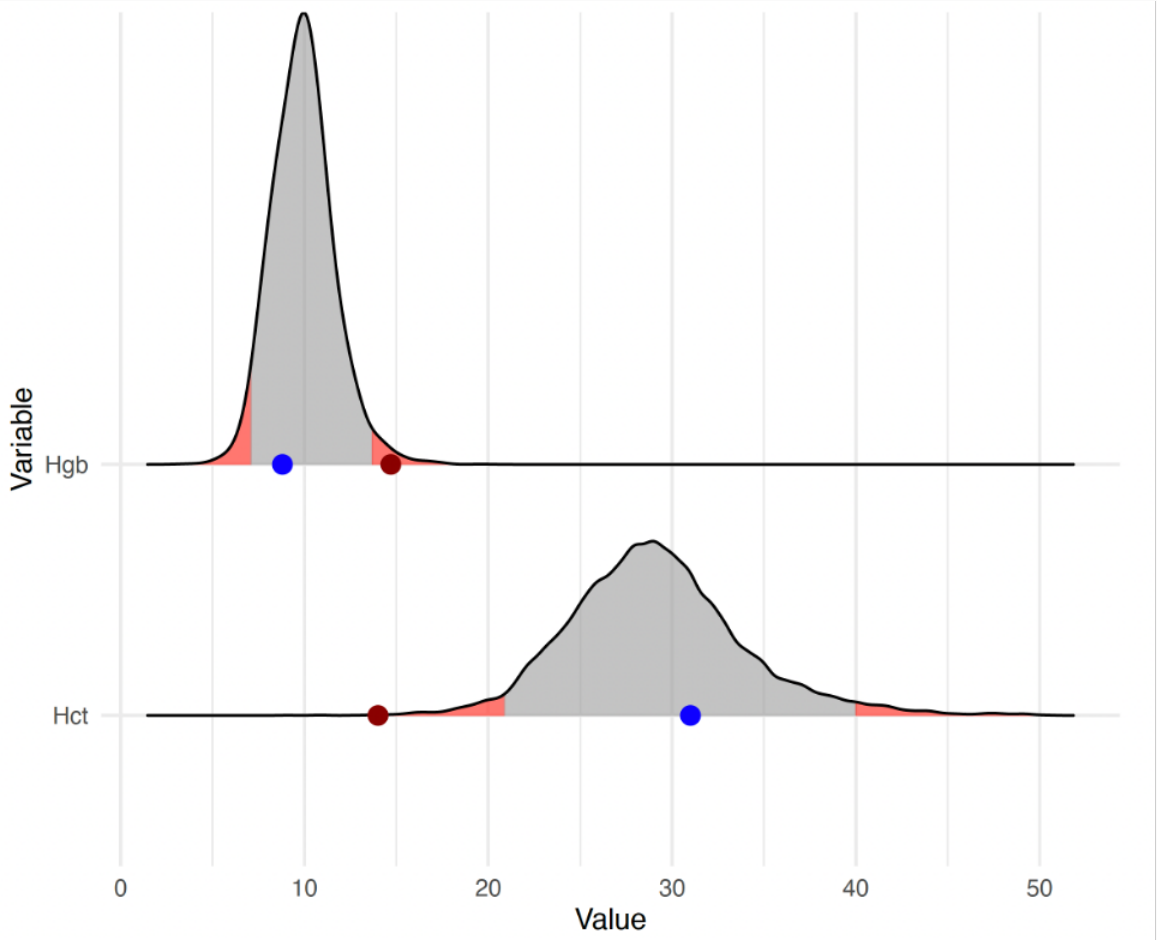
Statistical Issues (1): Does it look normal?



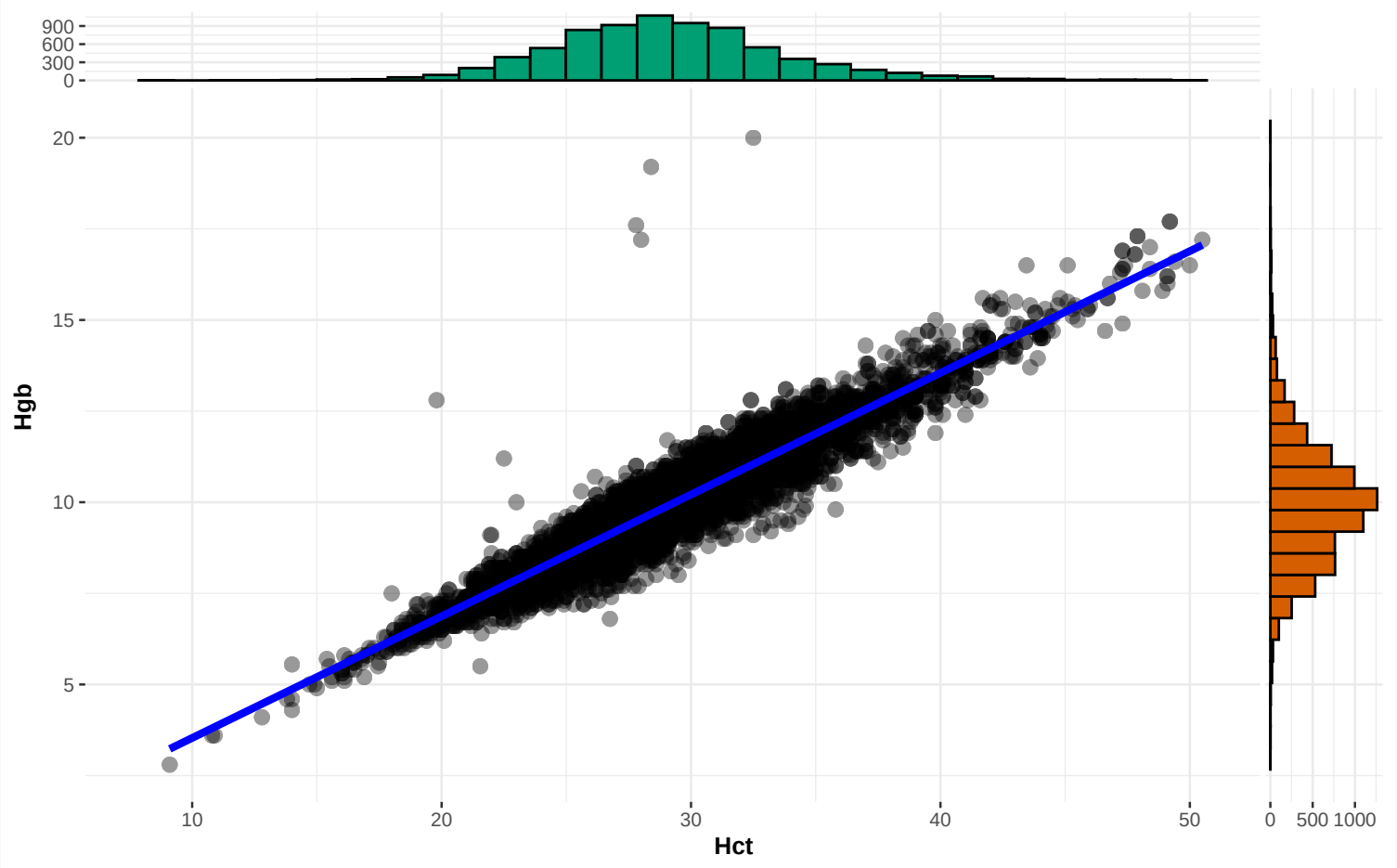


The State of the Art: Assumptions (2)

- **One-dimensional analysis**: biomarkers are analyzed separately, without accounting for their interdependence and with a multiple use of the same prior
  $\Rightarrow$ **Independence** between biomarkers

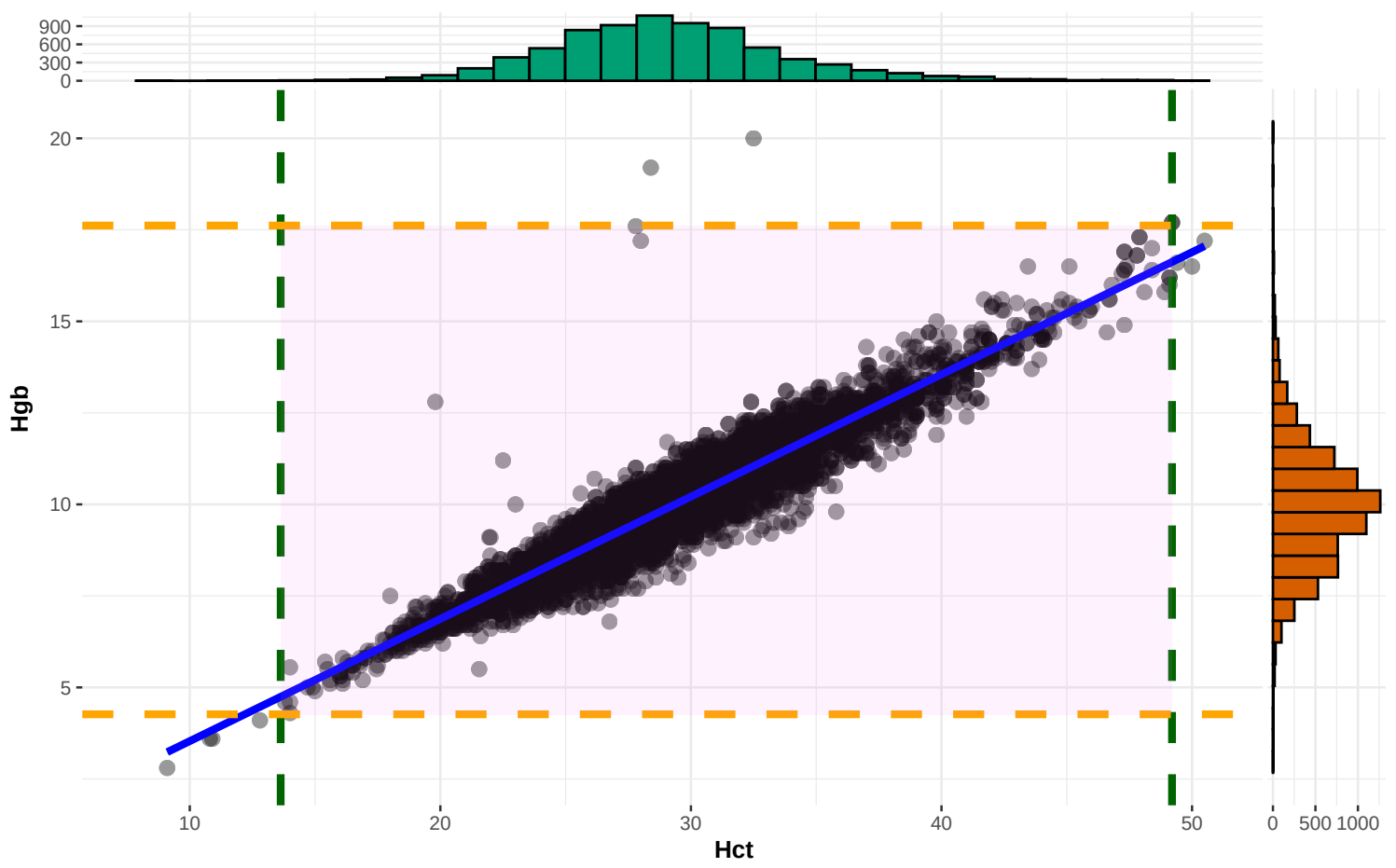


Assumptions (2): Do They Look Independent?



$\hat{\rho} = 0.95$

Assumptions (2): Do They Look Independent?



$\hat{\rho} = 0.95$

Biological Issues

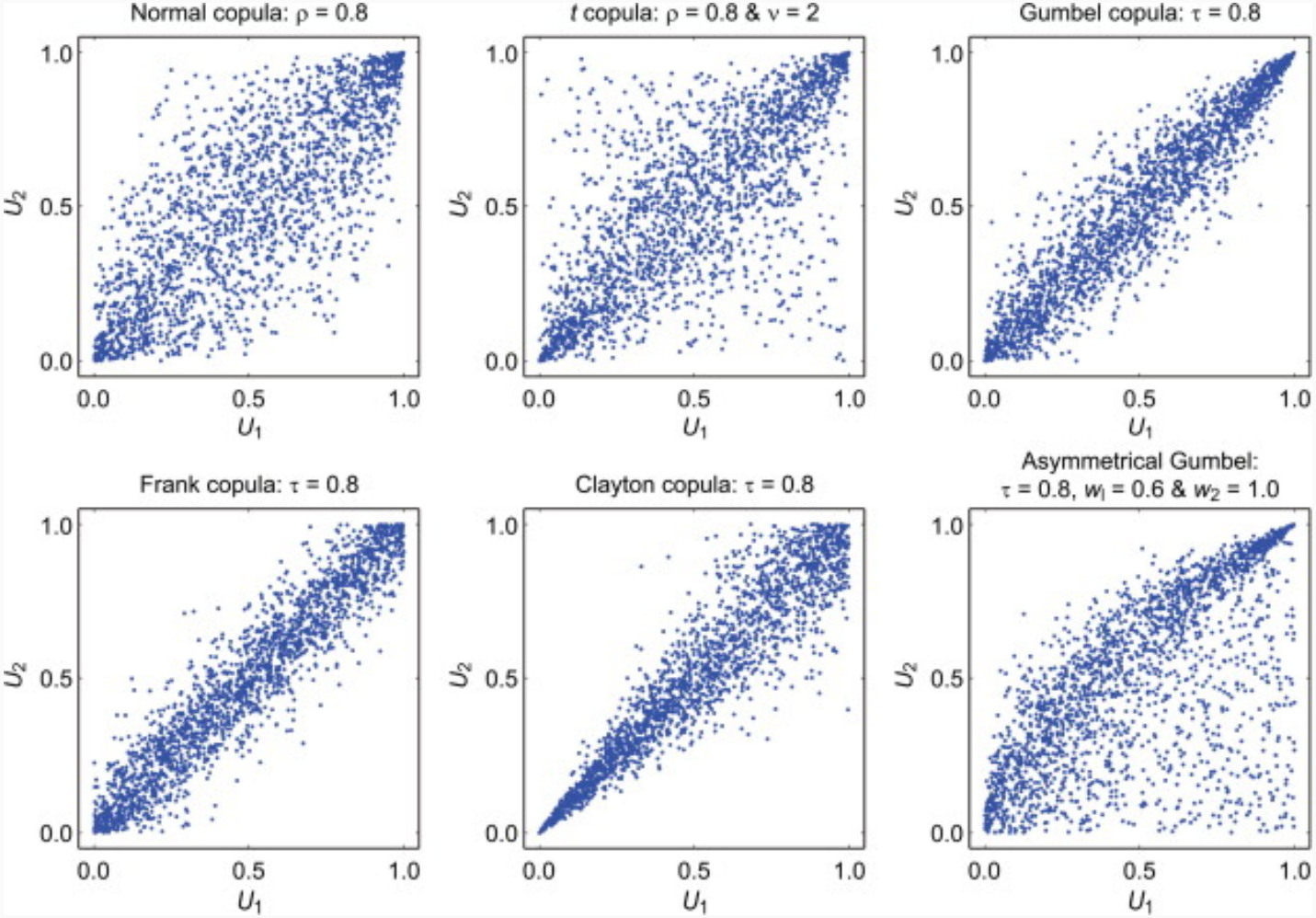
1. In general, biomarkers do not provide *orthogonal* information, either due to their intrinsic characteristics or because these are often derived quantities
2. Whereas in a control population biomarkers might be expected to provide *orthogonal* information, little is known about their (simultaneous) alteration due to the presence of some prohibited substances

Name	Short form	Number of parameters	rHuEPO doping		Blood transfusion	
			Loading phase	Maintenance phase	Removal	Infusion
Hemoglobin	Hgb	1	↑	↑	↓	↑
Hematocrit	Hct	1	↑	↑	↓	↑
Red Blood cells	Rbc	1	↑	↑	↓	↑
Reticulocytes	Ret%	1	↑	↓	↑	↓
OFF-score	OFFS	2	→	↑	↓	↑
Abnormal Blood Profile Score	ABPS	2–12	↑	↑	↓	↑

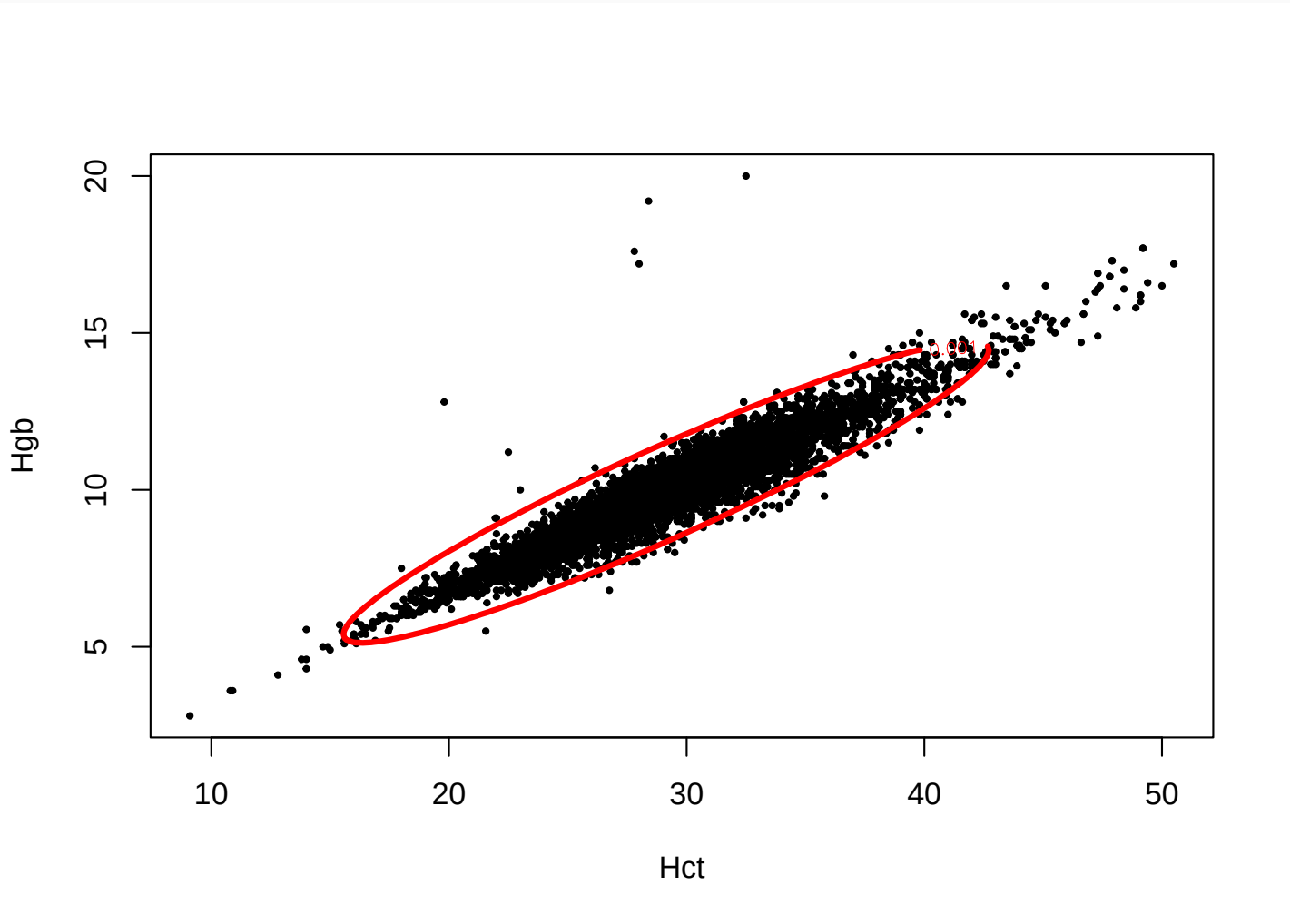
Figure 1: Common markers of blood doping and their response to rHuEPO treatment and blood transfusion. Source: Sottas et al. [2008]

Notably...

Alteration can produce different shapes of potential interaction, well illustrated by different types of copula



Why Copulae?



Could a **Multivariate** Normal distribution capture the (asymmetric) dependency on the tails?

Copulae and Sklar's Theorem [Sklar, 1959]

- Copulae model the dependence (inter-correlation) between random variables
- They provide a flexible way of representing the joint distribution of $\mathbf{Y}$

Given a random vector $\mathbf{Y} = (Y_1, \ldots, Y_d)$ and a d-variate CDF $\mathbf{F}$ which can depend on some parameter λ, Sklar [1959] showed that there always exists a d-variate function $C = C_\theta : [0,1]^d \to [0,1]$, such that:

$$F(y_1, \ldots, y_d; \lambda_1, \cdots, \lambda_d, \theta) = C_\theta(F_1(y_1; \lambda_1), \ldots, F_d(y_d; \lambda_d)),$$

where F_j is the marginal CDF of Y_j.

Therefore, in case that the multivariate distribution has a density $\mathbf{f}$, and this is available, it holds further that

$$\underbrace{\mathbf{f}(y_1, \ldots, y_d)}_{\textit{Multivariate density}} = \underbrace{c_\theta(F_1(y_1), \ldots, F_d(y_d))}_{\textit{Copula density}} \times \underbrace{f_1(y_1) \times \cdots \times f_d(y_d)}_{\textit{Marginal univariate densities}}$$

Copulae and Sklar's Theorem [Sklar, 1959]

- Copulae model the dependence (inter-correlation) between random variables
- They provide a flexible way of representing the joint distribution of $\mathbf{Y}$

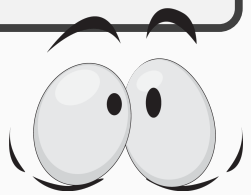
Given a random vector $\mathbf{Y} = (Y_1, \ldots, Y_d)$ and a d-variate CDF $\mathbf{F}$ which can depend on some parameter λ, Sklar [1959] showed that there always exists a d-variate function $C = C_\theta : [0,1]^d \to [0,1]$, such that:

$$F(y_1, \ldots, y_d; \lambda_1, \cdots, \lambda_d, \theta) = C_\theta(F_1(y_1; \lambda_1), \ldots, F_d(y_d; \lambda_d)),$$

where F_j is the marginal CDF of Y_j.

Therefore, in case that the multivariate distribution has a density $\mathbf{f}$, and this is available, it holds further that

$$\underbrace{\mathbf{f}(y_1, \ldots, y_d)}_{\textit{Multivariate density}} = \underbrace{c_\theta(F_1(y_1), \ldots, F_d(y_d))}_{\textit{Copula density}} \times \underbrace{f_1(y_1) \times \cdots \times f_d(y_d)}_{\textit{Marginal univariate densities}}$$



Our Proposal

Our Goal: In a Nutshell...

Given the current state-of-the-art, we aim to extend the ADAPTIVE model building a more sophisticated model framework that:

- Goes **beyond the Gaussianity** assumption of the marginal biomarkers
- Accounts for **multiple response variables simultaneously**, while describing/modeling:
  - the **inter-dependence (inter-correlation) among biomarkers**
  - the **intra-individual temporal dependence**
  - athletes' individual characteristics: fixed and **random effects**
- Replaces numerical integration with a safer and faster **Monte Carlo computation**

Our Ingredients: In a Nutshell...

- We focus on the haematological module: Hgb and Hct
- For illustrative purposes (at this stage), we use real data of a population of controls (non-athletes):

  (I) Model marginal pdfs $\Rightarrow$ **Location-Scale Student's t-distribution**[2]

  (II) Derive a copula model $\Rightarrow$ t-**copula**[3]

  (III) Estimate a linear mixed model for each biomarker[4]:

  $$y_{itr} = \boldsymbol{\beta}_r \mathbf{X}_{it} + u_{ir} + \epsilon_{itr}, \quad i = 1, \ldots, n; t = 1, \ldots, T_n; r = 1, 2$$

  - $\boldsymbol{\beta}_r = (\beta_{0,r}, \ldots, \beta_{p,r})$: model's coefficients related to the fixed effects $\mathbf{X}_{it}$. We consider $p = 3$ covariates: Age, Gender, Diastolic Blood Pressure
  - u_{ir}: unobserved random effect of group i (i.e., the individual itself)

  $\Rightarrow$ Prior elicitation

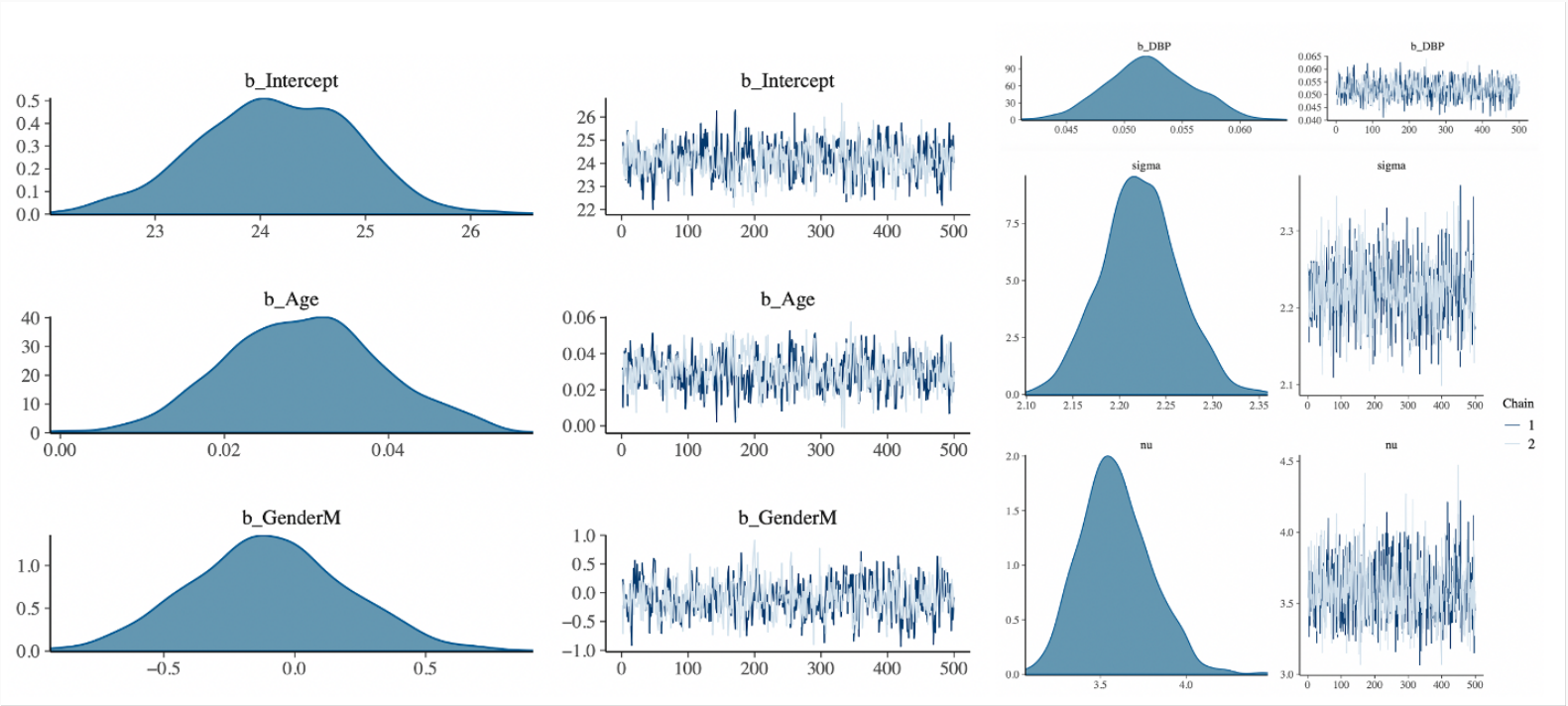
---

[2] See e.g., Jackman [2009]

[3] See e.g., [Demarta and McNeil, 2005]

[4] We use the Bayesian `brms` package [Bürkner, 2017] in `R`

Prior Elicitation



For $t = 1$, posterior parameters' draws define the prior (inter-individual variations), which is then used for deriving the predictive distribution:

$$p(\mathbf{y_1}|\mathbf{y_0} = \emptyset) = \int_{\mu} \int_{\sigma} p(\mathbf{y_1}|\boldsymbol{\mu}, \boldsymbol{\sigma})\, \pi(\boldsymbol{\mu}, \boldsymbol{\sigma}) d\boldsymbol{\mu} d\boldsymbol{\sigma}$$

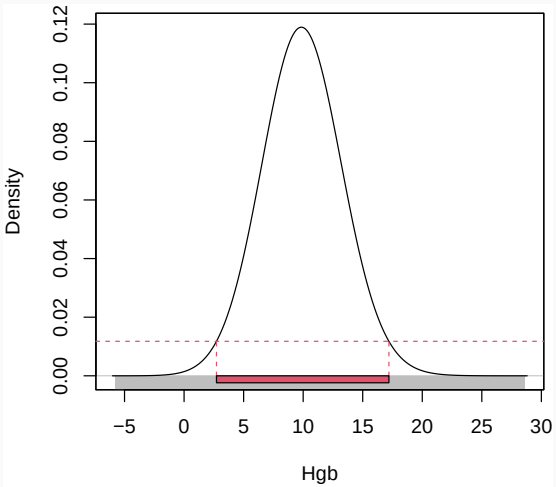
Detection of Abnormal Values

From Intervals to HDR

(a) For univariate distributions, once we compute our predictive distribution (updated at each time point t), a test result y_t is considered abnormal if it falls outside the $(1-\alpha)\%$ percentile range–the tolerance interval, where α is the predefined type-I error or false positive rate

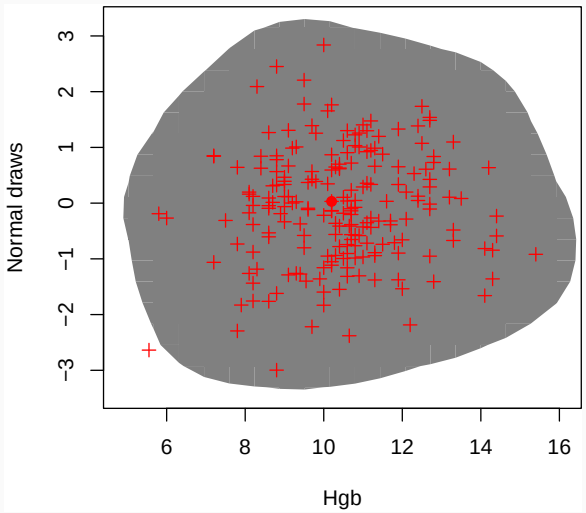
(b) In the multivariate setting, the predictive distribution is multivariate. Thus, given α, we look at identifying a Highest Density Region [HDR; Hyndman, 1996]. Vectors falling outside the HDR are considered abnormal

(a) **Univariate HDR**



⇒

(b) **Bivariate HDR**



Results

HDR for $t = 1$ (prior predictive)

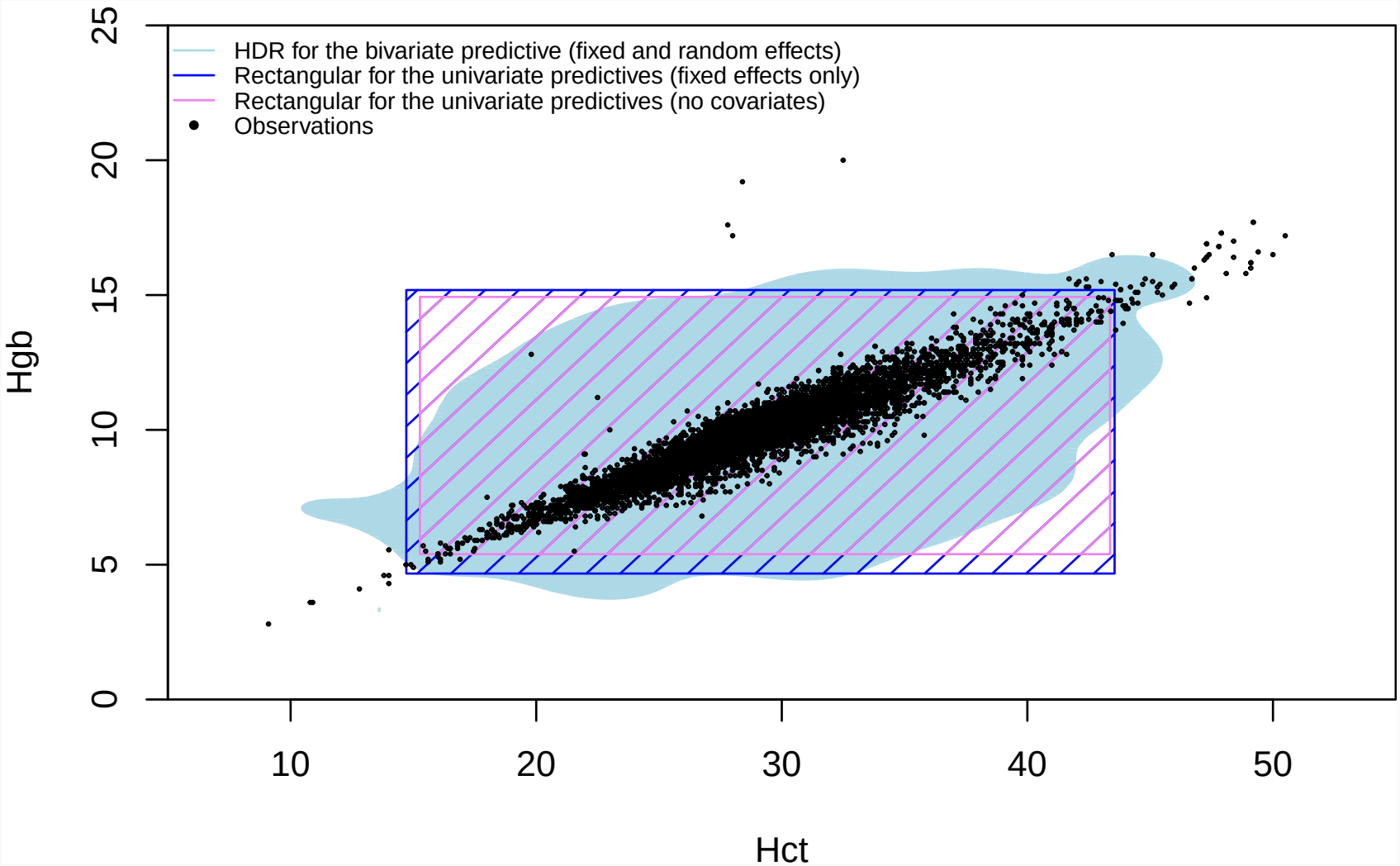


Figure 2: Tolerance regions for a male individual with average characteristics

HDR for $t = 1$ (prior predictive)

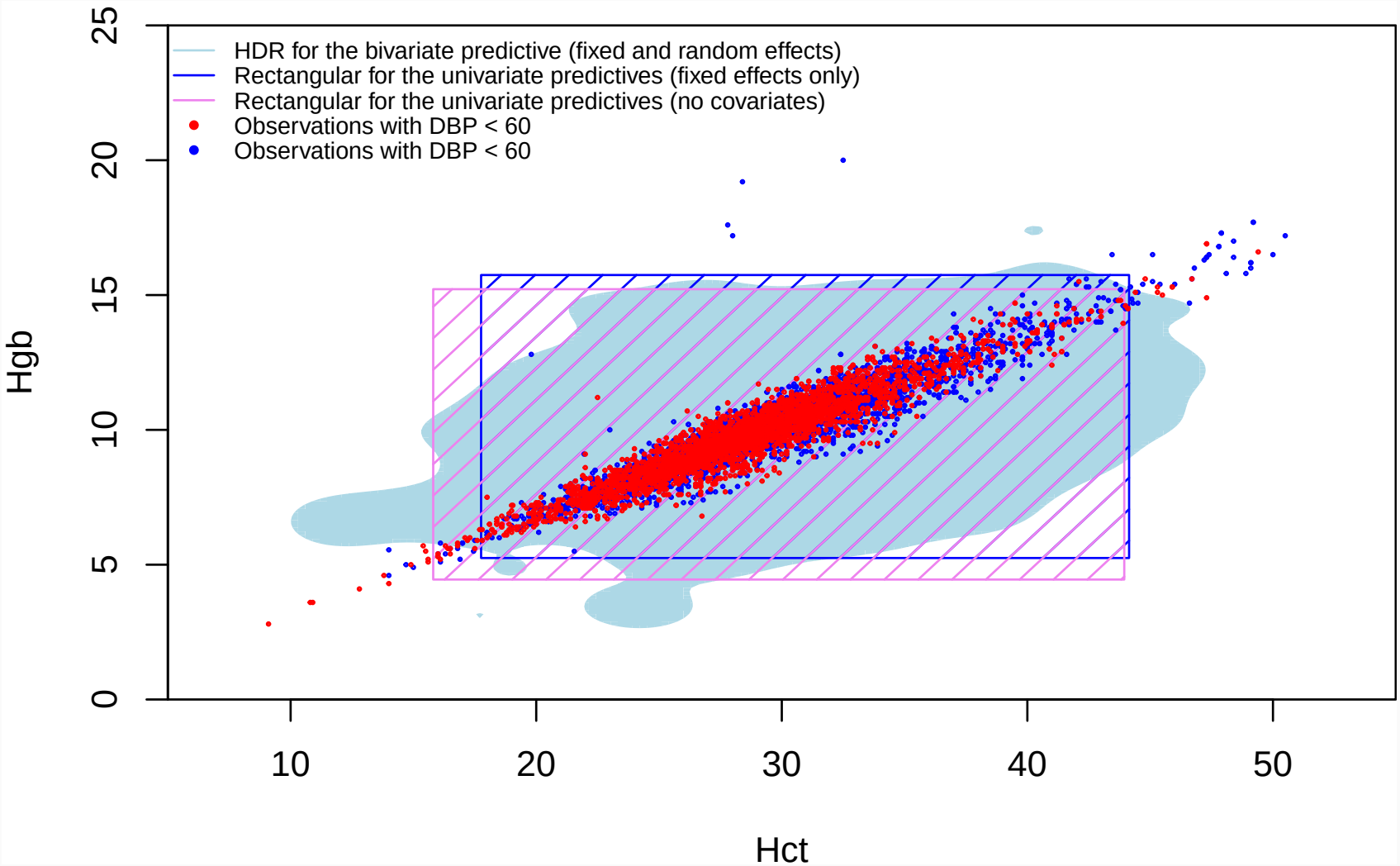


Figure 3: Tolerance regions for a male individual with **higher** DBP

HDR for $t > 1$ (posterior predictive)

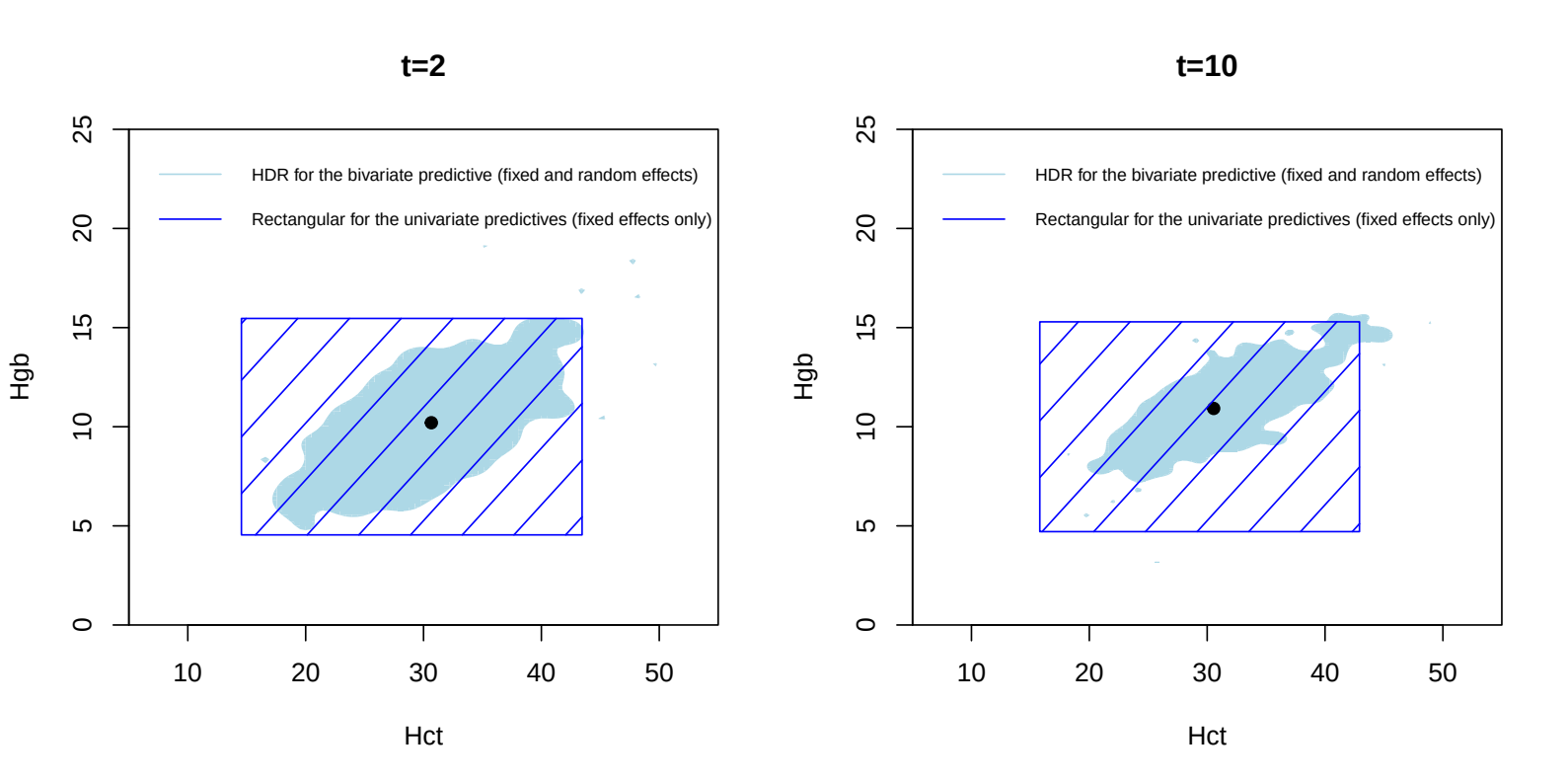


Figure 4: Change of the tolerance region over time for an individual with average characteristics

Conclusion and Future Work

Conclusions and Future work

- Indirect doping detection is become increasingly popular among forensic scientists for assessing evidence of drug abuse
- The current approach rely on certain assumptions, including biomarkers normality and intra-individual (time) independence, which may not be valid
- Furthermore, while multiple biomarkers are part of the longitudinal ABP profiles, their relationship is neither modeled nor assessed
- Our work extends the current state-of-the-art, overcoming existing issues

Conclusions and Future work

- Indirect doping detection is become increasingly popular among forensic scientists for assessing evidence of drug abuse
- The current approach rely on certain assumptions, including biomarkers normality and intra-individual (time) independence, which may not be valid
- Furthermore, while multiple biomarkers are part of the longitudinal ABP profiles, their relationship is neither modeled nor assessed
- Our work extends the current state-of-the-art, overcoming existing issues

Ongoing and future work:

- Account for skewness of the marginal distribution: Skewed Student's t
- Account for covariates in the copula model: Conditional Copula
- Implementation with ADO data

Thank you!

nina.deliu@uniroma1.it

References i

P.-C. Bürkner. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28, 2017.

S. Demarta and A. J. McNeil. The t copula and related copulas. *International statistical review*, 73(1):111–129, 2005.

R. J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126, 1996.

S. Jackman. *Bayesian analysis for the social sciences*. John Wiley & Sons, 2009.

R. B. Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

M. Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

P.-E. Sottas, N. Baume, C. Saudan, C. Schweizer, M. Kamber, and M. Saugy. Bayesian detection of abnormal values in longitudinal biomarkers with an application to t/e ratio. *Biostatistics*, 8(2):285–296, 2007.

P.-E. Sottas, N. Robinson, M. Saugy, and O. Niggli. A forensic approach to the interpretation of blood doping markers. *Law, Probability and Risk*, 7(3):191–210, 2008.

WADA. World Anti-Doping Code, 2021. URL https://www.wada-ama.org/sites/default/files/resources/files/2021_wada_code.pdf.

The ABP Program

- Since its introduction in 2009, the ABP has been established as a complementary and essential pillar in the detection of doping in some specific disciplines, e.g., the haematological (paralympic) ABP involves:

AQUATICS	LONG DISTANCE, OPEN WATER
ATHLETICS	LONG DISTANCE, MIDDLE DISTANCE
BIATHLON	
CANOE/KAYAK	LONG DISTANCE, MARATHON
CYCLING	CYCLOCROSS, TRACK ENDURANCE
	MOUNTAIN BIKE – CROSS COUNTRY, ROAD
ROLLER SPORT	INLINE SPEED SKATING $>$ 1000M
ROWING	
SKATING	SPEED SKATING $>$ 1500M
SKIING	CROSS-COUNTRY, NORDIC COMBINED
SKI MOUNTAINEERING	
TRIATHLON	
UNDERWATER SPORTS	FINSWIMMING OPEN WATER

Sketch of the ADAPTIVE method

Modeling assumptions in Sottas et al. [2007] – steroid module:

- $Y_1, Y_2, \ldots, Y_t, Y_{t+1}$ i.i.d. $\sim N(\mu, \sigma)$
- (μ, σ) accounts for the inter-individual and intra-individual variation:
  - a prior distribution $\pi(\mu, \sigma)$ can be elicited from a population of "controls": clean athletes or volunteers (non athletes)
  - individual tolerance levels are then calibrated based on

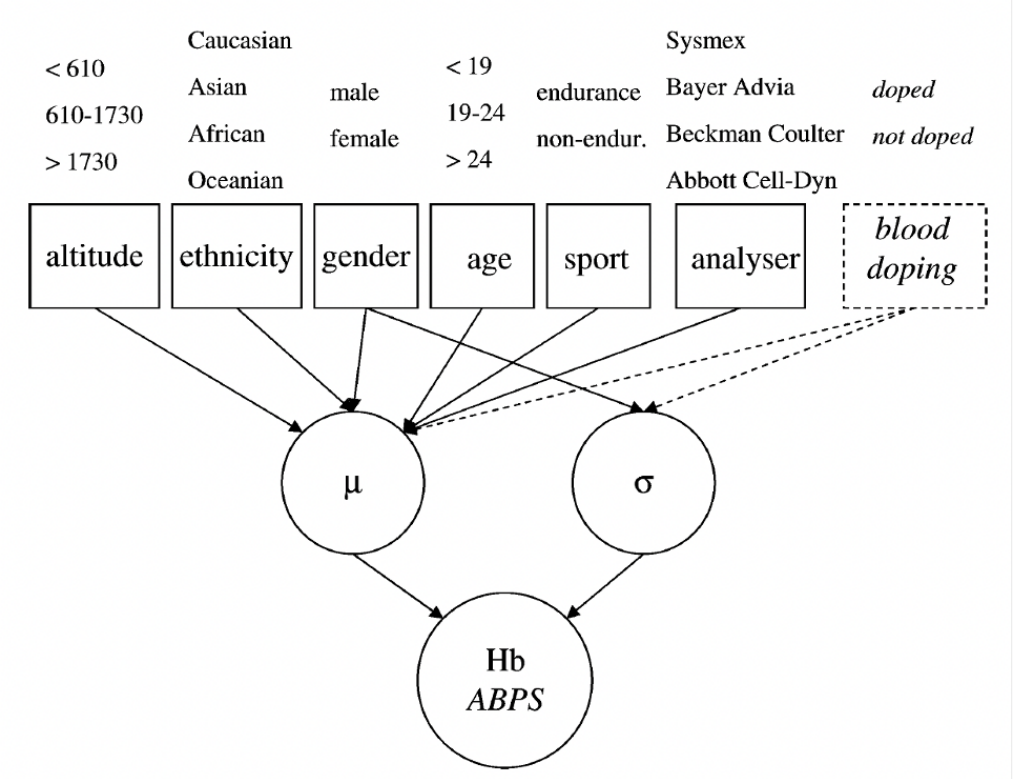
$$p(y|y_1, \ldots, y_t) = \int_\mu \int_\sigma p(y|\mu, \sigma)\, \pi(\mu, \sigma|y_1, \ldots, y_t) d\mu d\sigma$$

  $t = 1$: these are based on the inter-individual variation only (prior predictive)
  $t > 1$: intra-individual variation calibrates the limits (posterior predictive)

  - the integral is calculated by numerical integration

Sketch of the ADAPTIVE method:

The model is subsequently extended to the hematologic module and population heterogeneity is also taken into consideration [Sottas et al., 2008][5]



[5] A forensic approach to the interpretation of blood doping markers. *Law, Probability and Risk* (2008), 7.3: 191-210.

Application of the ADAPTIVE model: an example

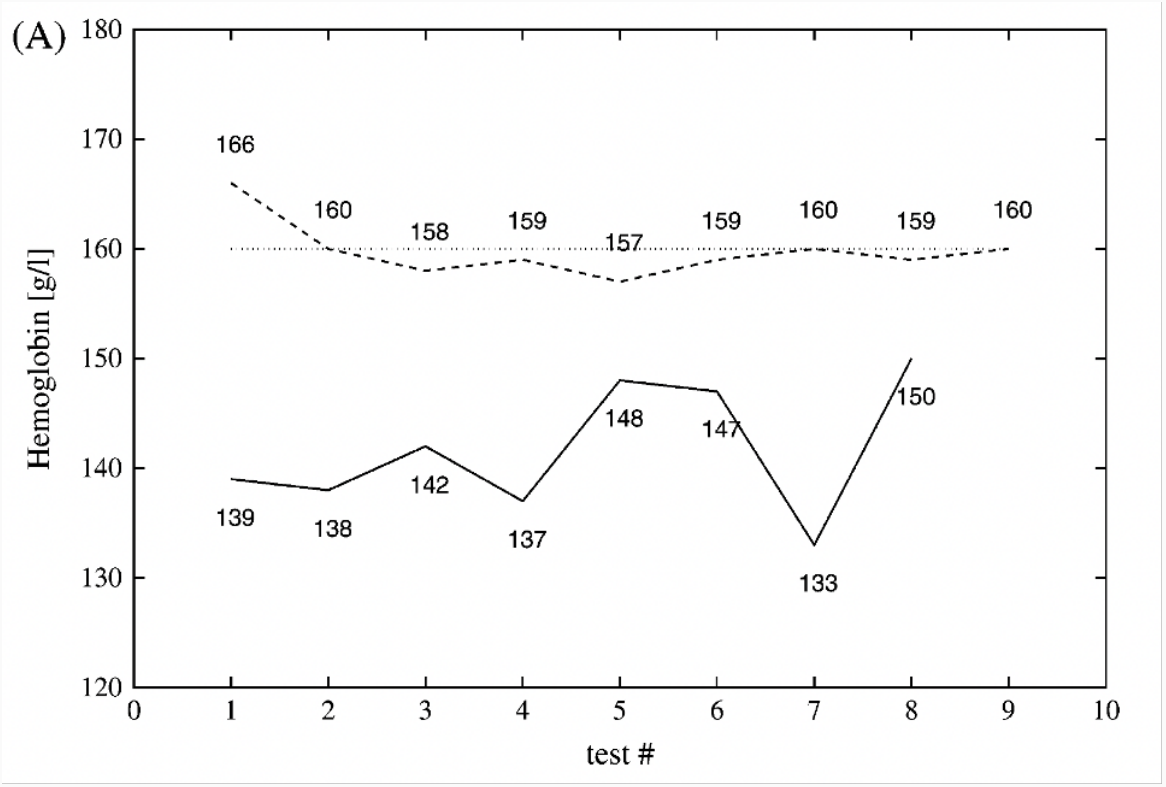


Figure 5: Solid line: longitudinal Hgb data of a female elite athlete. Dashed line: threshold limit values returned by the ADAPTIVE method. Dotted line: traditional population-based limit (here 160 g/l). Source: Sottas et al. [2008]

Copulae: Definition

A function $C : [0,1]^d \to [0,1]$ is a d-dimensional **copula** if it represents a joint cumulative density function (CDF) of a d-dimensional random vector $\mathbf{Y} = (Y_1, \ldots, Y_d)$ with uniform marginals [Nelsen, 2007].

Note that, assuming $\mathbf{Y}$ has continuous marginals, by applying the **probability integral transform** to each component, the random vector

$$(U_1, U_2, \ldots, U_d) = (F_1(Y_1), F_2(Y_2), \ldots, F_d(Y_d))$$

has marginals that are uniformly distributed on the interval $[0,1]$. The copula of $\mathbf{Y}$ can thus be defined as the joint CDF of $(U_1, U_2, \ldots, U_d)$:

$$C(u_1, u_2, \ldots, u_d) = \Pr[U_1 \leq u_1, U_2 \leq u_2, \ldots, U_d \leq u_d].$$

Copulae: Definition

A function $C : [0,1]^d \to [0,1]$ is a d-dimensional **copula** if it represents a joint cumulative density function (CDF) of a d-dimensional random vector $\mathbf{Y} = (Y_1, \ldots, Y_d)$ with uniform marginals [Nelsen, 2007].

Note that, assuming $\mathbf{Y}$ has continuous marginals, by applying the **probability integral transform** to each component, the random vector

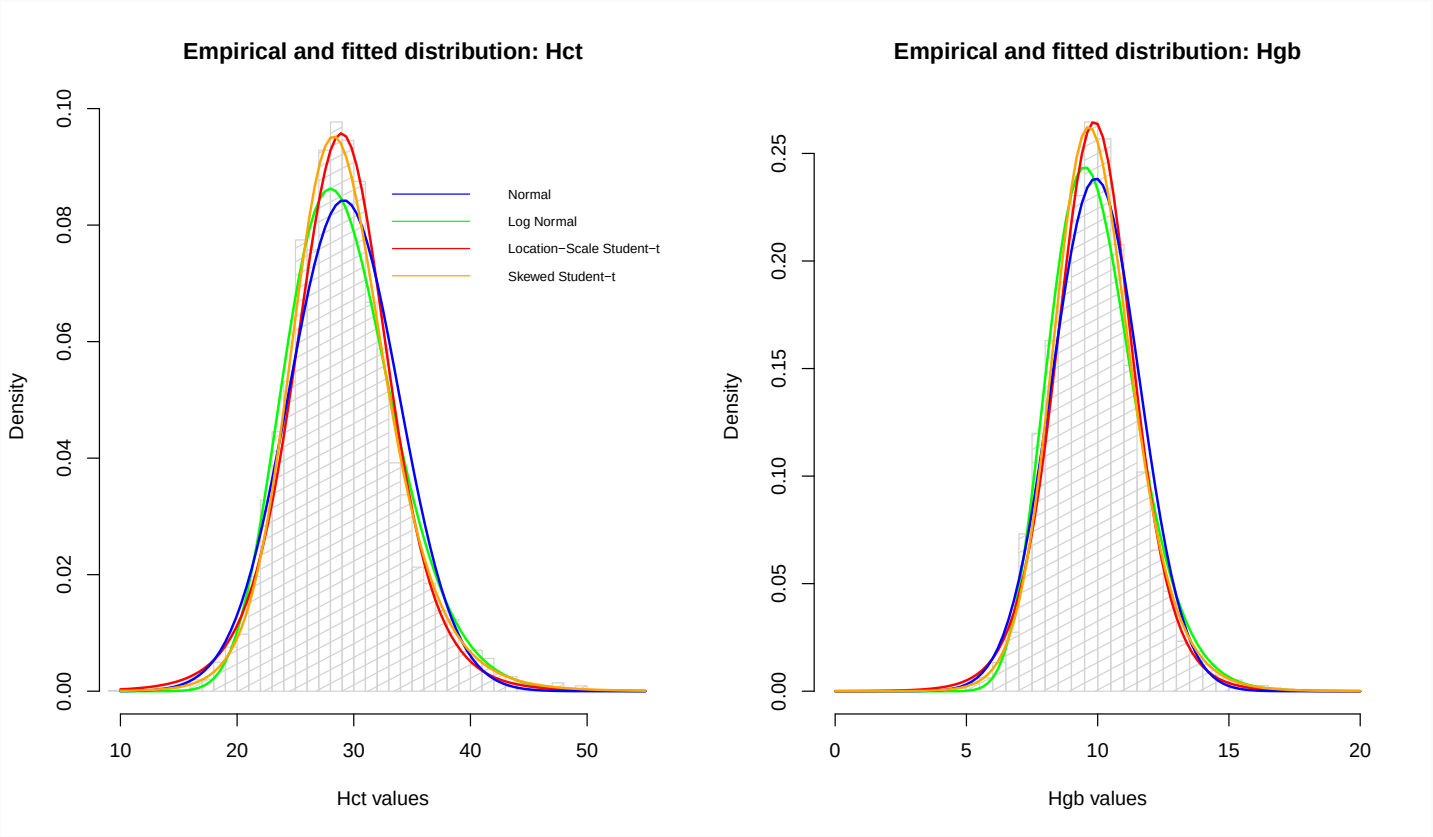
$$(U_1, U_2, \ldots, U_d) = (F_1(Y_1), F_2(Y_2), \ldots, F_d(Y_d))$$

has marginals that are uniformly distributed on the interval $[0,1]$. The copula of $\mathbf{Y}$ can thus be defined as the joint CDF of $(U_1, U_2, \ldots, U_d)$:

$$C(u_1, u_2, \ldots, u_d) = \Pr[U_1 \leq u_1, U_2 \leq u_2, \ldots, U_d \leq u_d].$$

- Copulae model the dependence (inter-correlation) between random variables
- They provide a flexible way of representing the joint distribution of $\mathbf{Y}$

Part I: Marginal distribution



$\Rightarrow$ We choose a **Location-Scale Student's t-distribution**[6]

[6] See e.g., Jackman [2009]

Part II: Modeling inter-dependence with copulae

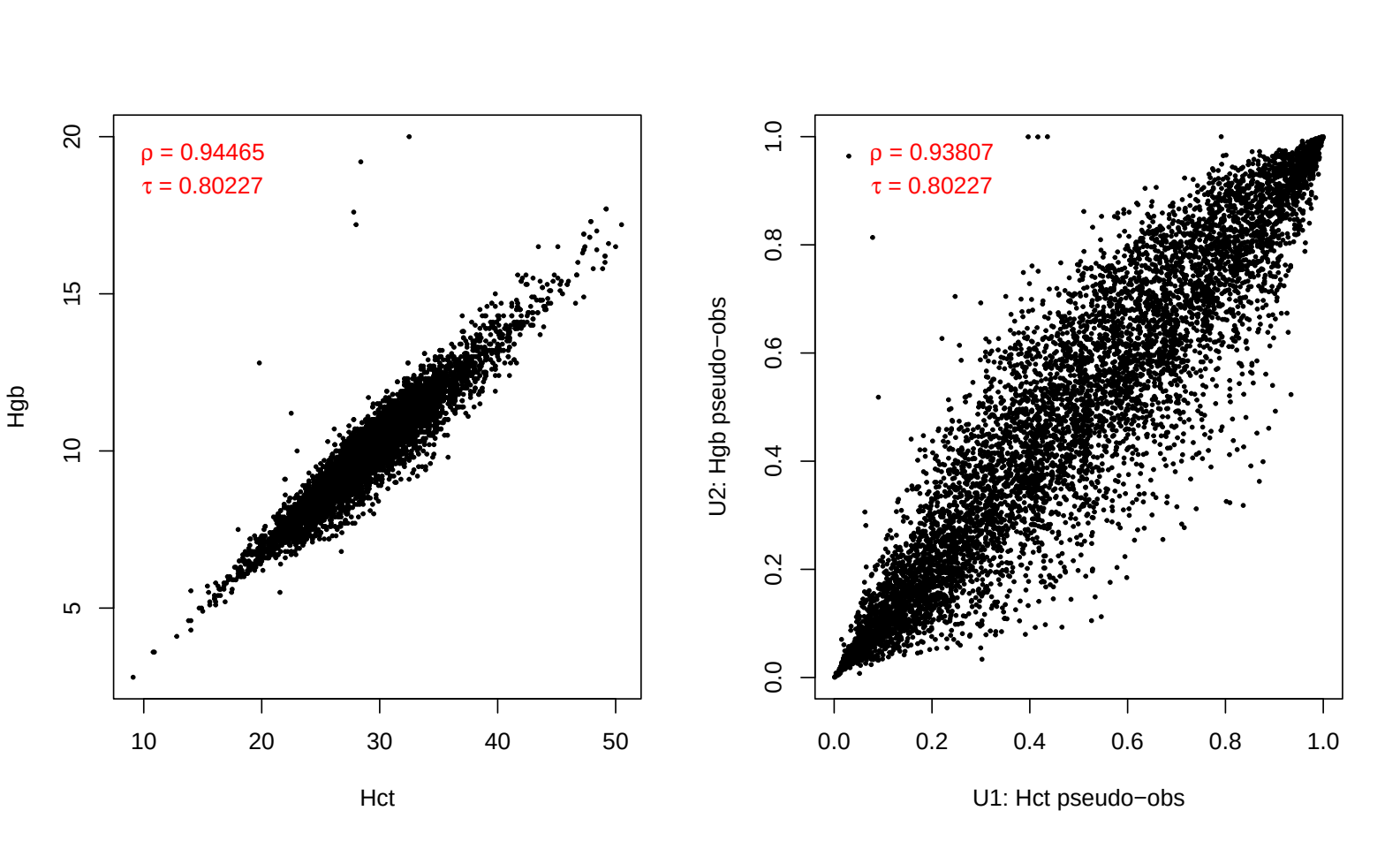
- We choose our copula density among a number of families and selected "our best" one according to the Akaike and Bayesian Information Criteria
- The best one is given by a t-copula [see e.g., Demarta and McNeil, 2005]

In a bi-variate setting, the bi-variate t-copula $C^t_{\rho,\nu}(u_1, u_2)$ with ρ and ν parameters, is given by:

$$C^t_{\rho,\nu}(u_1, u_2) = \int_{-\infty}^{t_\nu^{-1}(u_1)} \int_{-\infty}^{t_\nu^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \left(1 + \frac{s^2 - 2\rho st + t^2}{\nu(1-\rho^2)}\right)^{-(\nu+2)/2} dsdt,$$

where $t_\nu^{-1}(\cdot)$ is the inverse bi-variate t-distribution with ν dof.

Part II: Modeling inter-dependence with copulae



Part III: Linear mixed model

- We assume that $y_{itr} = \boldsymbol{\beta}_r \mathbf{X}_{it} + u_{ir} + \epsilon_{itr}, \quad i = 1, \ldots, n; t = 1, \ldots, T_n; r = 1, 2$
  - $\boldsymbol{\beta}_r = (\beta_{0,r}, \ldots, \beta_{p,r})$ are the model's coefficients related to the fixed effects $\mathbf{X}_{it}$. We consider three covariates ($p = 3$): Age, Gender, Diastolic Blood Pressure (DBP)
  - u_{ir} is the unobserved random effect of group i (i.e., the individual itself) in relation to variable Y_r
- We use the `brms` package [Bürkner, 2017] in `R`:
  - It performs a full Bayesian inference
  - It is based on Stan (MCMC: NUTS)
  - The formula syntax is based on the syntax applied in the `lme4` package

Linear mixed model: Predictive check

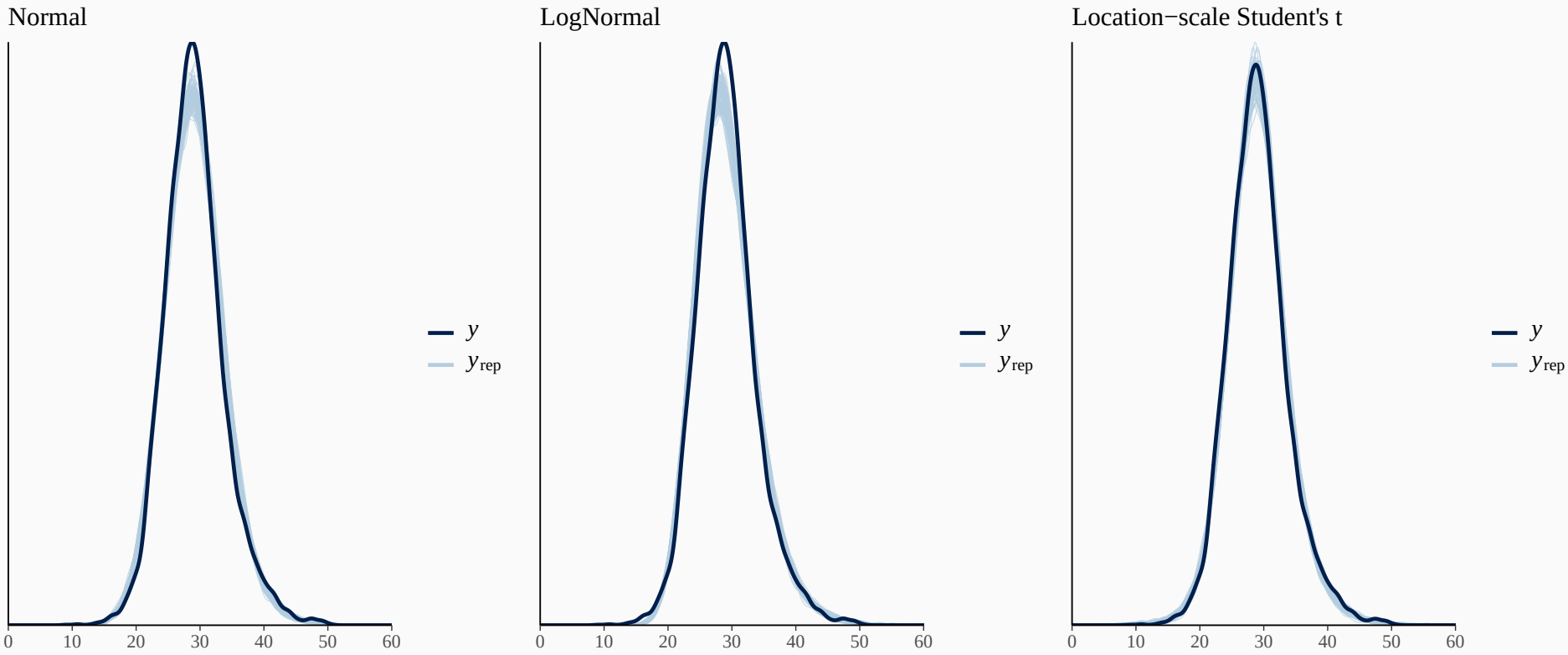


Figure 6: Comparison between the empirical distribution of the observed values and the predictive distribution under an LMM for Hct

Linear mixed model: Role of covariates

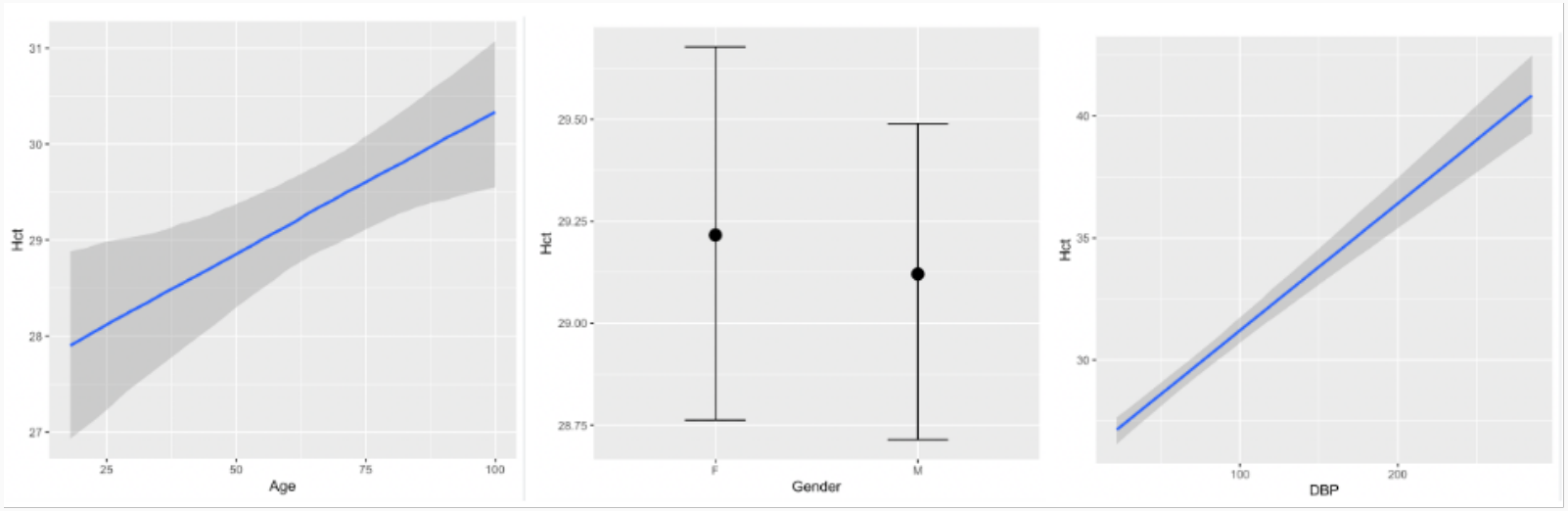
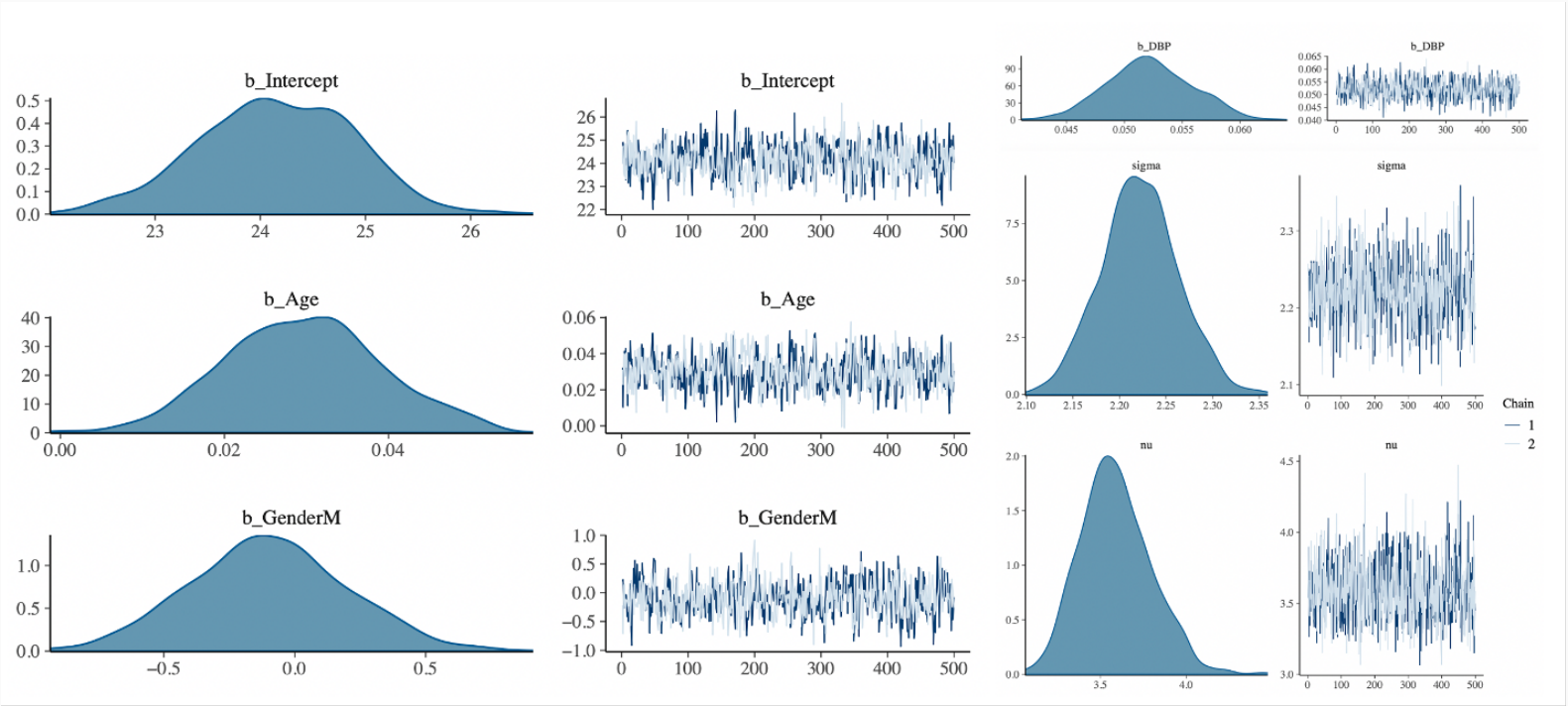


Figure 7: Conditional effects of the three covariates under an LMM for Hct

Prior Elicitation



$\Rightarrow$ Posterior parameters' draws are used as prior (inter-individual variations) for deriving the predictive distribution. Note that the predictive for $t = 1$:

$$p(\mathbf{y_1}|\mathbf{y_0} = \emptyset) = \int_{\boldsymbol{\mu}} \int_{\boldsymbol{\sigma}} p(\mathbf{y_1}|\boldsymbol{\mu}, \boldsymbol{\sigma})\, \pi(\boldsymbol{\mu}, \boldsymbol{\sigma}) d\boldsymbol{\mu} d\boldsymbol{\sigma}$$

$$= \int_{\mu_a} \int_{\mu_b} \int_{\sigma_a} \int_{\sigma_b} p_a(y_{1,a}|\mu_a, \sigma_a)\, p_b(y_{1,b}|\mu_b, \sigma_b)\, c_\theta(F_a(y_{1,a}), F_b(y_{1,b}))$$

$$\times \pi(\mu_b, \sigma_b)\pi(\mu_a, \sigma_a) d\mu_a d\mu_b d\sigma_a d\sigma_b$$