Inferring food intake from multiple biomarkers using a latent variable model

Silvia D'Angelo

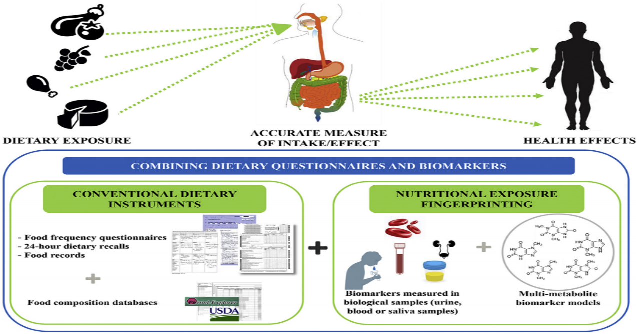
TCD - 15th March 2023

Joint work with...

Prof. Claire Gormley

Prof. Lorraine Brennan

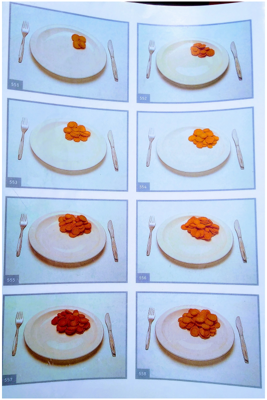
Assessing nutrients and food intake

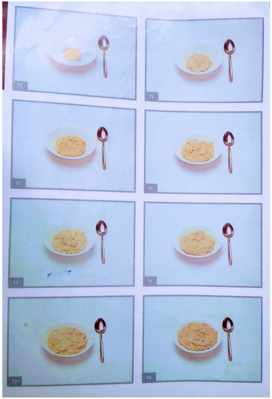

DIETARY EXPOSURE
ACCURATE MEASURE OF INTAKE/EFFECT
HEALTH EFFECTS
COMBINING DIETARY QUESTIONNAIRES AND BIOMARKERS
CONVENTIONAL DIETARY INSTRUMENTS
- Food frequency questionnaires
- 24-hour dietary recalls
- Food records
+
Food composition databases
USDA
+
NUTRITIONAL EXPOSURE FINGERPRINTING
Biomarkers measured in biological samples (urine, blood or saliva samples)
+
Multi-metabolite biomarker models

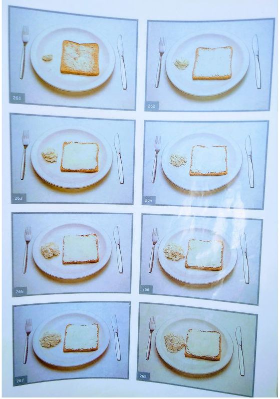
- Intake data are necessary in various fields (nutrition, epidemiology,...)
- Often the only intake data available are self-reported

- **Self-reported** data are **subjective** and possibly biased
- Dietary **biomarkers** can provide more **objective measurements** of intake

Portions assessment







Discovering biomarker candidates

- Experimental studies are designed to investigate the potential of compounds as specific nutrient/food biomarkers
- Participants are fed the food of interest, either directly or indirectly
- Different biological samples are collected and analysed from the participants
- Multiple steps are required to candidate compounds as biomarkers

Experimental studies only investigate the relationship between candidate biomarkers and portions of intake. This should be generalized, to allow biomarkers usage outside of experimental study contexts.

How to generalize experimental studies results?

Previous works on food biomarkers

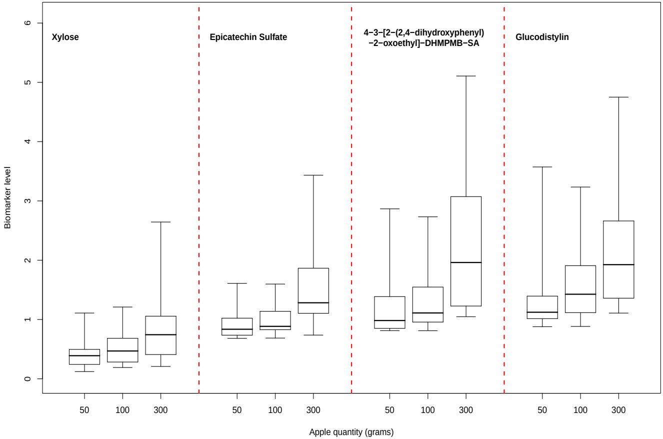
Gürdeniz et al. 2016.

- Consumers
- Non-consumers

Vázquez Manjarrez et al. 2019.

- **Low** intake
- Medium intake
- **High** intake

Apple intake data



- Feeding study, $N = 32$ participants
- $D = 3$ apple quantities: 50, 100 and 300 grams
- Collection: 4 days diet + 5th day urine + wash-out periods
- Data: $n = 86$ observations and $P = 4$ biomarkers

The model

$$y_{ip} = \alpha_p + \beta_p z_i + \epsilon_{ip},$$

$$z_i \sim \sum_{d=1}^{D} \pi_{id} \mathcal{N}_{[0,\infty)}\left(X_d, \theta_d^2\right), \quad \pi_i = \{\pi_{i1}, \dots, \pi_{iD}\}$$

with

- P biomarkers: $\{y_1, \dots, y_P\}$
- n latent intakes: $\{z_1, \dots, z_n\}$
- D food quantities: $\{X_1, \dots, X_D\}$

and

- $\alpha_p \sim \mathcal{N}_{[0,\infty)}\left(\mu_\alpha, \sigma_\alpha^2\right)$
- $\beta_p \sim \mathcal{N}_{(0,\infty)}\left(\mu_\beta, \sigma_\beta^2\right)$
- $\epsilon_{ip} \sim \mathcal{N}\left(0, \sigma_p^2\right)$
- $y_{ip} \sim \mathcal{N}_{[0,\infty)}(\alpha_p + \beta_p z_i, \sigma_p^2)$

Modelling the latent intakes

$$z_i \mid c_i \sim \prod_{d=1}^{D} \Big[\mathcal{N}_{[0,\infty)}(X_d, \theta_d^2)\Big]^{[c_i=d]}$$

$\mathbf{c} = \{c_1, \ldots, c_i, \ldots, c_n\}$ are observation-specific allocation labels.

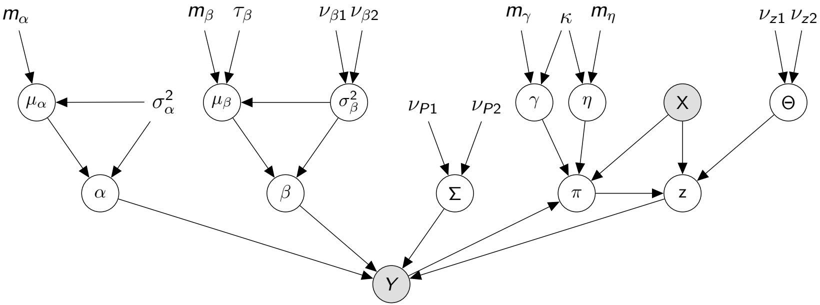
Modelling the weights

$$p(\pi \mid \gamma, \eta, \mathrm{c}, \mathrm{Y}, \mathrm{X}) = \prod_{i=1}^{n}\prod_{d=1}^{D} \pi_{id}^{[c_i=d]} = \prod_{i=1}^{n}\prod_{d=1}^{D} \pi_{id}(y_i|\gamma_d, \eta)^{[c_i=d]}$$

$$= \prod_{i=1}^{n}\prod_{d=1}^{D} \Big[Pr(x_i \leq X_d \mid \gamma_d, \eta, y_i) - Pr(x_i \leq X_{d-1} \mid \gamma_{d-1}, \eta, y_i)\Big]^{[c_i=d]}$$

where $Pr(x_i \leq X_d \mid \gamma_d, \eta, y_i) = \frac{1}{\pi}\Big[\text{arctan}\left(\frac{1}{2}(\gamma_d + \eta y_i)\right) + \frac{\pi}{2}\Big]$.

Model structure



- $\mu_\alpha \sim \mathcal{N}_{[0,\infty)}(m_\alpha, \tau_\alpha \sigma_\alpha^2)$
- $\mu_\beta \sim \mathcal{N}_{[0,\infty)}(m_\beta, \tau_\beta \sigma_\beta^2)$
- $\sigma_\beta^2 \sim Inv\Gamma(\nu_{\beta 1}, \nu_{\beta 2})$
- $\theta_d^2 \sim Inv\Gamma(\nu_{z1}, \nu_{z2})$
- $\Sigma = \{\sigma_1^2, \ldots, \sigma_P^2\}$
- $\Theta = \{\theta_1^2, \ldots, \theta_P^2\}$
- $\gamma_d \sim \mathcal{N}_{(m_{\gamma_{d-1}}, m_{\gamma_{d+1}})}(m_{\gamma_d}, \kappa)$
- $\eta_p \sim \mathcal{N}(m_{\eta_p}, \kappa)$

Estimation

- Metropolis within Gibbs MCMC algorithm -

Outline

1. Initialize all parameters
2. Gibbs steps: update α, β, μ_α, μ_β, σ^2_β and Σ from their full conditional distributions
3. MH steps: update γ and η (random walk)
4. Update labels c
5. Gibbs steps: update Θ and z from their full conditional distributions

$$z_i \mid c_i = d, \cdots \sim \mathcal{N}_{[0,\infty)}\left(\mu^*_{id}, \theta^{2*}_{id}\right)$$

$$\theta^{2*}_{id} = \left(\sum_{p=1}^{P} \frac{\beta_p^2\theta_d^2+\sigma_p^2/P}{\theta_d^2\sigma_p^2}\right)^{-1}, \mu^*_{id} = \sigma^{2*}_{id}\left[\sum_{p=1}^{P} \frac{\beta_p\left(y_{ip}-\alpha_p\right)}{\sigma_p^2} + \frac{X_d}{\theta_d^2}\right]$$

Intake quantification / prediction

AIM: Infer intakes for a new group of n^* observations, when only biomarker measurements are available.

- Sampling distribution for a new intake z_j^*:

  $$p(z_j^* \mid y_j^*, c_j^*, \Omega) = \prod_{d=1}^{D} \left[\mathcal{N}_{[0,\infty)} \left(\frac{\mu_{zj}\theta_d^2 + X_d \sigma_{zj}^2}{\sigma_{zj}^2 + \theta_d^2}, \left(\frac{1}{\theta_d^2} + \frac{1}{\sigma_{zj}^2} \right)^{-1} \right) \right]^{[c_j^* = d]}$$

  where:

  - $\sigma_{zj}^2 = \left(\sum_{p=1}^{P} \frac{\beta_p^2}{\sigma_p^2} \right)^{-1}$
  - $\mu_{zj} = \sigma_{zj}^2 \left(\sum_{p=1}^{P} \frac{\beta_p (y_{jp}^* - \alpha_p)}{\sigma_p^2} \right)$
  - $\Omega = \{\alpha, \beta, \Sigma, \mathrm{X}, \Theta, \eta, \gamma\}$

- Posterior predictive distribution for a new intake z_j^*:

  $$p(z_j^* \mid y_j^*, c_j^*) \propto \int p(z_j^* \mid y_j^*, c_j^*, \Omega) p(\Omega \mid \mathrm{Y}) \, d\Omega$$

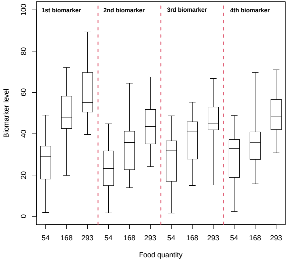
Simulations - details

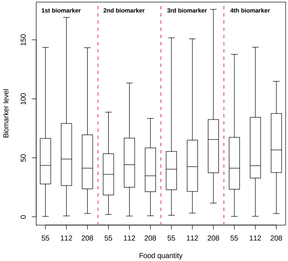
Settings

- $P = 4$ biomarkers
- $n = \{30, 60, 99, 150\}$, $n^* = \lfloor 0.4 \times n \rfloor$
- (α,β): small, medium and large biomarkers range
- Σ: small, mixed and large variability
- X: stable, increasing, and decreasing increments
- Θ: low and high variability
- c: sampled at random in $\{1, \ldots, D\}$

Outline

- 20 datasets per each settings combination
- 30000 MCMC iterations (burn: 6000)

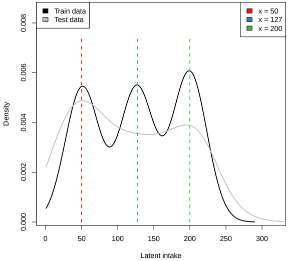


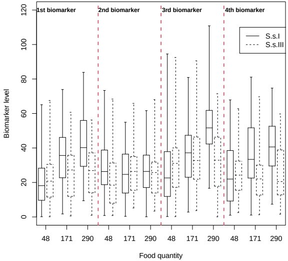


Simulations - intake quantification with...

- **Study I:** varying biomarker variability (σ_p^2)

- **Study II:** discrepancies between training and test data generation





- **Study III:** model misspecification

$$y_{ip} = \alpha_p + \beta_p z_i^2 + \epsilon_{ip}$$

Simulations - results

Absolute error values (in grams)

between true and estimated/predicted intakes

		Estimates			**Predictions**		
Sim. study	Σ	MM	BLR	PLS	MM	BLR	PLS
I	Small	3(7)	77(200)	10(24)	4(9)	76(216)	10(26)
	Mixed	4(8)	62(136)	20(39)	4(33)	111(263)	21(41)
	Large	6(18)	62(136)	35(60)	9(59)	64(137)	37(69)
II	Small	3(8)	61(181)	9(30)	5(25)	62(184)	10(37)
	Mixed	4(8)	113(222)	22(33)	7(22)	112(224)	23(43)
	Large	7(62)	80(118)	39(56)	22(77)	88(108)	41(67)
III	Small	6(35)	66(227)	26(72)	8(49)	67(231)	31(77)
	Mixed	7(44)	98(269)	43(80)	9(56)	112(298)	46(87)
	Large	11(62)	87(197)	64(65)	27(74)	101(138)	70(80)

Models:

- MM:

  multiMarker model
- BLR:

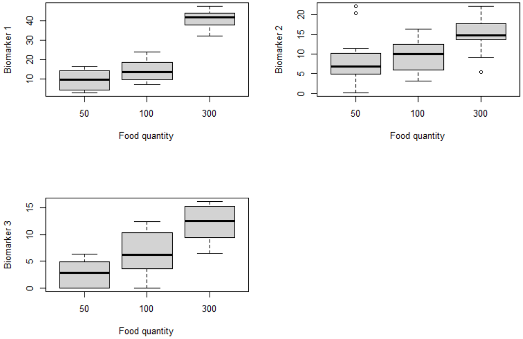
  Bayesian linear regression
- PLS:

  PLS regression

Simulations - Multinomial weights

$$\pi_{id} = \frac{\exp(\gamma_d + \sum_{p=1}^{P} \eta_{dp} y_{ip})}{\sum_{d'=1}^{D} \exp(\gamma_{d'} + \sum_{p=1}^{P} \eta_{d'p} y_{ip})}$$

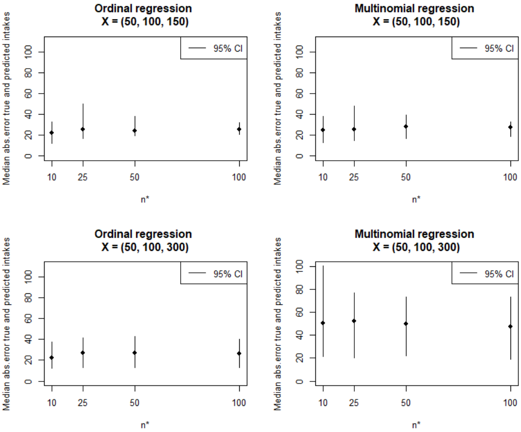
- $n = 50,\ P = D = 3,$
  $\alpha_p \sim \mathcal{N}_{[0,\infty)}(4, 1),$
  $\beta_p \sim \mathcal{N}_{(0,\infty)}(0.001, 0.1),$
  $\sigma_p^2 = 5^2,\ \theta_d^2 = 8^2$
- Two scenarios: "stable increments", $\mathrm{X} = \{50, 100, 150\}$, and "increasing increments", $\mathrm{X} = \{50, 100, 300\}$
- $n^* = \{10, 25, 50, 100\}$



Simulations - Multinomial weights

$$\pi_{id} = \frac{\exp(\gamma_d + \sum_{p=1}^{P} \eta_{dp} y_{ip})}{\sum_{d'=1}^{D} \exp(\gamma_{d'} + \sum_{p=1}^{P} \eta_{d'p} y_{ip})}$$

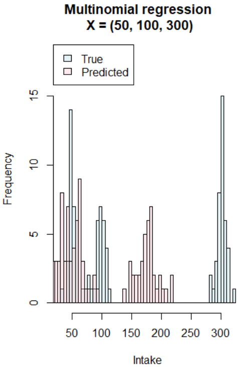
- $n = 50,\ P = D = 3,$
  $\alpha_p \sim \mathcal{N}_{[0,\infty)}(4, 1),$
  $\beta_p \sim \mathcal{N}_{(0,\infty)}(0.001, 0.1),$
  $\sigma_p^2 = 5^2,\ \theta_d^2 = 8^2$
- Two scenarios: “stable increments”, $\mathrm{X} = \{50, 100, 150\}$, and “increasing increments”, $\mathrm{X} = \{50, 100, 300\}$
- $n^* = \{10, 25, 50, 100\}$

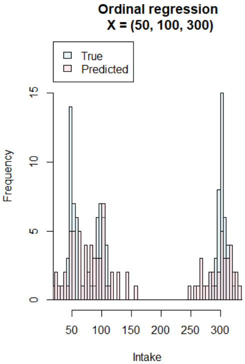


Simulations - Multinomial weights

$$\pi_{id} = \frac{\exp\left(\gamma_d + \sum_{p=1}^{P} \eta_{dp} y_{ip}\right)}{\sum_{d'=1}^{D} \exp\left(\gamma_{d'} + \sum_{p=1}^{P} \eta_{d'p} y_{ip}\right)}$$

- $n = 50,\ P = D = 3,$
  $\alpha_p \sim \mathcal{N}_{[0,\infty)}(4, 1),$
  $\beta_p \sim \mathcal{N}_{(0,\infty)}(0.001, 0.1),$
  $\sigma_p^2 = 5^2,\ \theta_d^2 = 8^2$
- Two scenarios: "stable increments", $\mathrm{X} = \{50, 100, 150\}$, and "increasing increments", $\mathrm{X} = \{50, 100, 300\}$
- $n^* = \{10, 25, 50, 100\}$





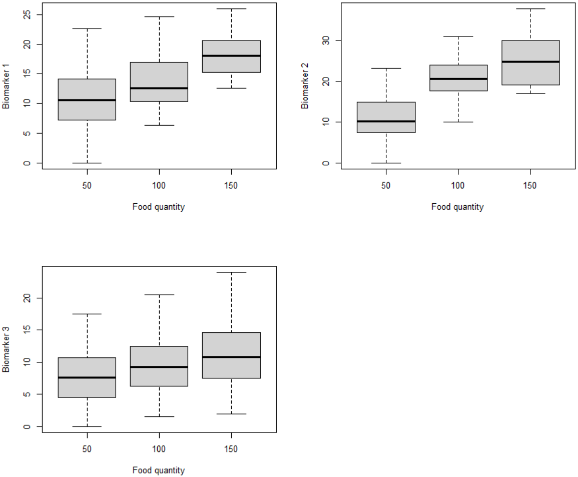
Simulations - extras

General settings:

- $n = 50,\ P = D = 3,$
  $\alpha_p \sim \mathcal{N}_{[0,\infty)}(4, 1),$
  $\beta_p \sim$
  $\mathcal{N}_{(0,\infty)}(0.001, 0.1),$
  $\sigma_p^2 = 5^2,\ \theta_d^2 = 8^2,$
  $\mathrm{X} = \{50, 100, 150\}$

Test data:

- **Uniform intakes**:
  $z_i^* \sim \mathcal{U}(0, 200)$
- **Unbalanced components**: Last component contains most of the observations



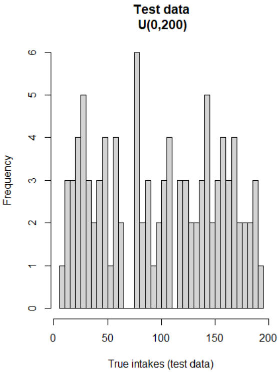
Simulations - extras

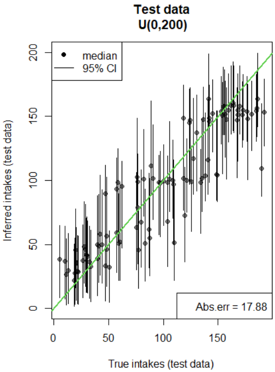
General settings:

- $n = 50,\ P = D = 3,$ $\alpha_p \sim \mathcal{N}_{[0,\infty)}(4, 1),$ $\beta_p \sim$ $\mathcal{N}_{(0,\infty)}(0.001, 0.1),$ $\sigma_p^2 = 5^2,\ \theta_d^2 = 8^2,$ $\mathrm{X} = \{50, 100, 150\}$

Test data:

- **Uniform intakes**: $z_i^* \sim \mathcal{U}(0, 200)$
- **Unbalanced components**: Last component contains most of the observations





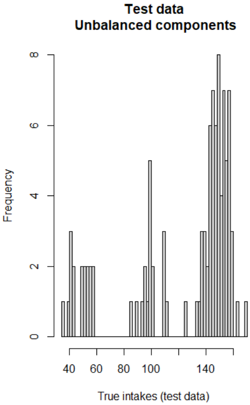
Simulations - extras

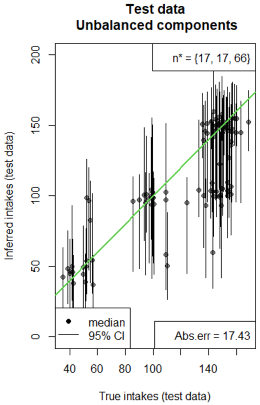
General settings:

- $n = 50,\ P = D = 3,$ $\alpha_p \sim \mathcal{N}_{[0,\infty)}(4, 1),$ $\beta_p \sim$ $\mathcal{N}_{(0,\infty)}(0.001, 0.1),$ $\sigma_p^2 = 5^2,\ \theta_d^2 = 8^2,$ $\mathrm{X} = \{50, 100, 150\}$

Test data:

- **Uniform intakes**: $z_i^* \sim \mathcal{U}(0, 200)$
- **Unbalanced components**: Last component contains most of the observations

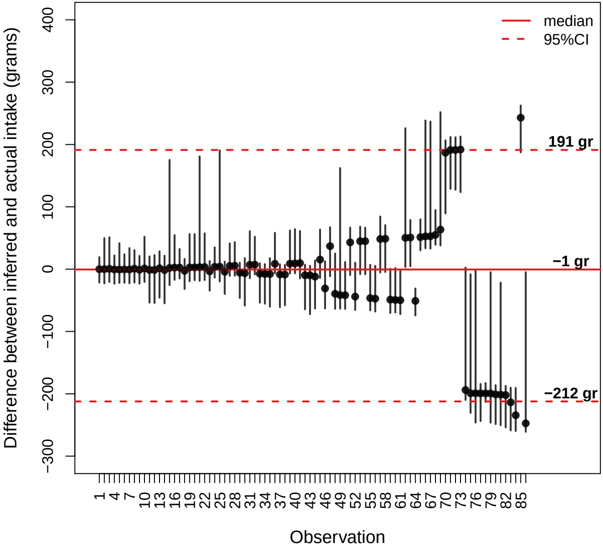

Test data
Unbalanced components


Test data
Unbalanced components

Apples - predicted intakes

- Leave-one-out CV

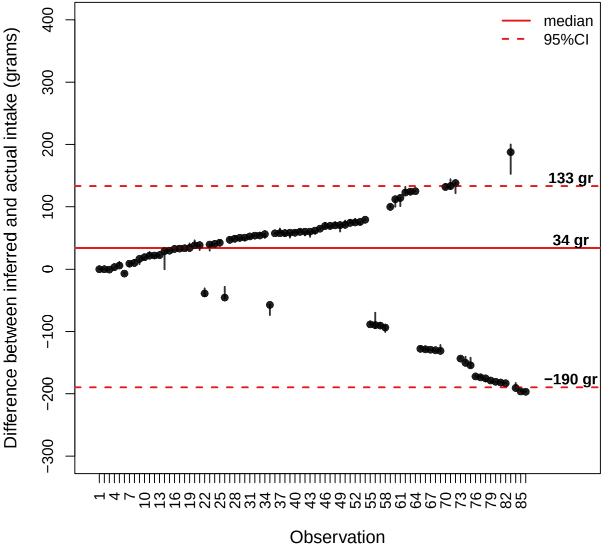
- **Models**:
  - MM
  - BLR
  - PLS



Apples - predicted intakes

- Leave-one-out CV

- **Models**:
  - MM
  - BLR
  - PLS



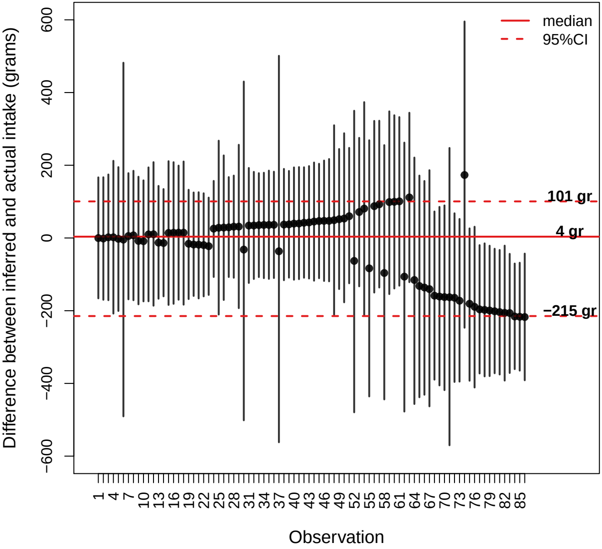
Apples - predicted intakes

- Leave-one-out

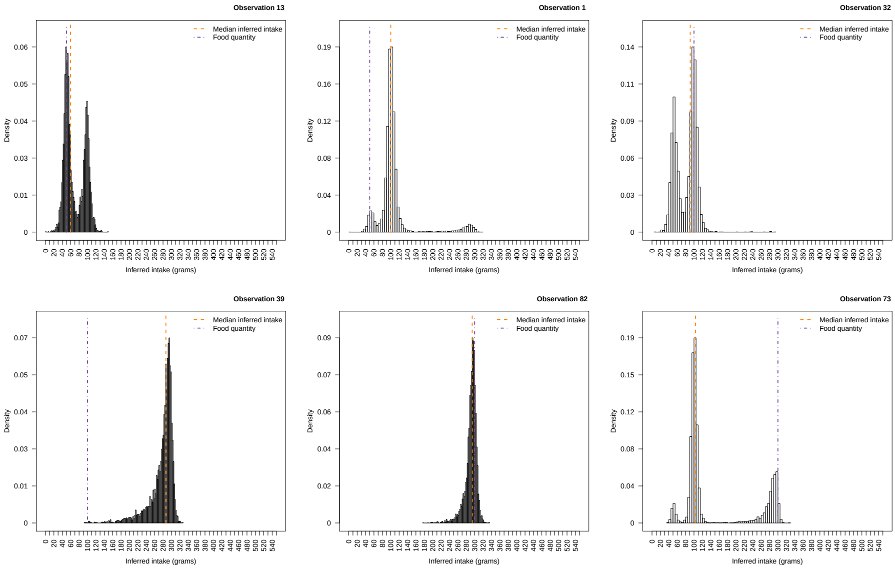
  CV

- **Models**:

  - MM
  - BLR
  - PLS



Apples - posterior predictive distributions



Apples - repeated measures?

		Parameter			
Data	**Dimension**	α_p	β_p	σ_p	θ_d
Original	1	0.206 (0.267)	0.003 (0.002)	0.353 (0.117)	5.546 (6.514)
	2	0.489 (0.215)	0.005 (0.002)	0.271 (0.104)	7.546 (14.038)
	3	0.612 (0.514)	0.007 (0.004)	0.677 (0.234)	98.989 (71.595)
	4	0.614 (0.323)	0.008 (0.004)	0.382 (0.174)	-
Modified	1	0.214 (0.433)	0.003 (0.003)	0.597 (0.157)	1.806 (1.176)
	2	0.515 (0.416)	0.005 (0.004)	0.600 (0.264)	2.032 (1.773)
	3	0.517 (0.755)	0.009 (0.007)	0.746 (0.258)	8.407 (10.358)
	4	0.662 (0.635)	0.008 (0.006)	0.676 (0.226)	-

Original data:

- $n = 86$ observations, treated as independent

Modified data:

- Each one of the 32 participants appears only once

Conclusions

- Flexible framework to infer intake from multiple biomarkers
- Uncertainty quantification
- multiMarker: R package and Shiny app
- Easily extendable to other applied contexts, when multiple outcomes are associated with an unobserved variable of interest
- EXTRA: introduction of covariates
- EXTRA: repeated measurements

Inferring food intake from multiple biomarkers using a latent variable model. (2021)

D'Angelo, Brennan, Gormley. Annals of Applied Statistics.