BAYESIAN PARTIAL LEAST SQUARES (BPLS)

Szymon Urbaś

Maynooth University, Ireland

:: THANKS AND ACKNOWLEDGEMENTS ::

This joint work under VistaMilk SFI Research Centre.

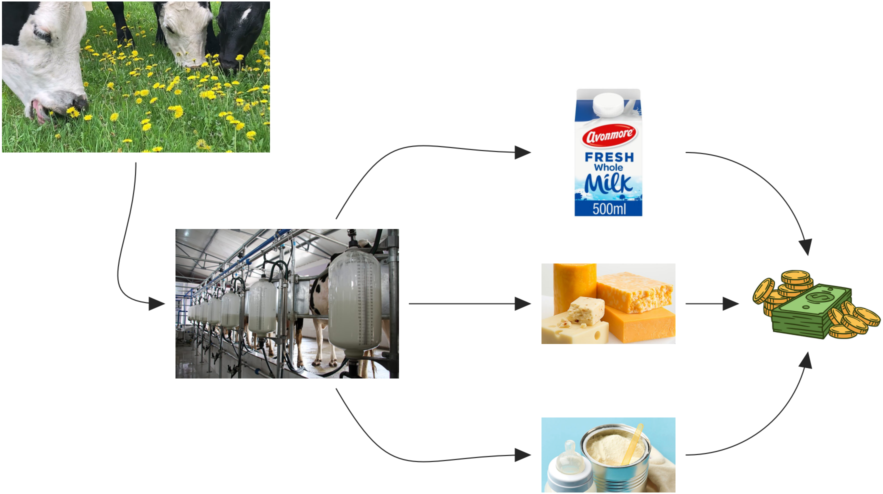
- Prof Claire Gormley *(UCD)*
- Prof Donagh Berry *(Teagasc)*
- Dr Pierre Lovera, Dr Rob Daly and Prof Alan O'Riordan *(Tyndall Institute, Cork)*

VistaMilk

A World Leading SFI Research Centre

Science Foundation Ireland SFI For what's next

MOTIVATING PROBLEM—PREDICTING TRAITS OF MILK



MOTIVATING PROBLEM—PREDICTING TRAITS OF MILK

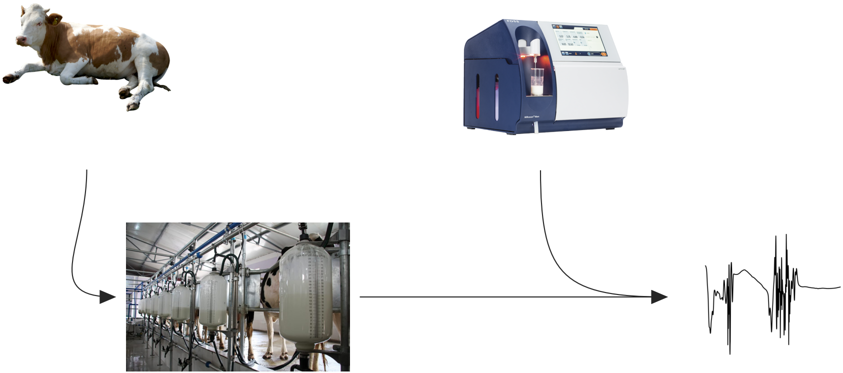
The quality and the value of a food product (e.g. milk) are determined by its chemical and technological traits (properties).

The process of measuring each of these in a lab can be costly and time-consuming.

Spectrometry: examining how light at different wavelengths passes through the substance (cheaper, quicker, non-destructive)

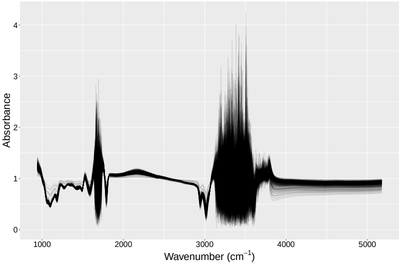
Aim: Devise prediction models which can reliably predict the traits from a given spectral reading.

MOTIVATING PROBLEM—PREDICTING TRAITS OF MILK



MOTIVATING PROBLEM — MID-INFRARED SPECTRA OF MILK

$\mathtt{P}$ = 538 wavenumbers



Traits ($\mathtt{R}$ = 3) are measured in a lab:

- **heat stability**
  — for making infant formula;
- **rennet coagulation time** (RCT)
  — for making cheese;
- **casein content**
  — major protein.

We have N = 363 complete observations. (large $\mathtt{P}$, small N problem)

PARTIAL LEAST SQUARES

Predictors: $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^\top = \boldsymbol{X} \in \mathbb{R}^{N \times \mathrm{P}}$ **Responses**: $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)^\top = \boldsymbol{Y} \in \mathbb{R}^{N \times \mathrm{R}}$

Suppose there is a common projection onto a Q-dimensional latent space with variables $\boldsymbol{Z} \in \mathbb{R}^{N \times \mathtt{Q}}$,

$$
\begin{aligned}
\boldsymbol{X} &= \boldsymbol{Z} W^\top + \boldsymbol{E}_x, \\
\boldsymbol{Y} &= \boldsymbol{Z} C^\top + \boldsymbol{E}_y,
\end{aligned}
$$

where W and C are loading matrices, and $\boldsymbol{E}_x$ and $\boldsymbol{E}_y$ are residuals.

PLS: Iteratively find $\operatorname{argmax} \|\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})\|$.

- Use W and C estimates to predict traits from new sample spectra.

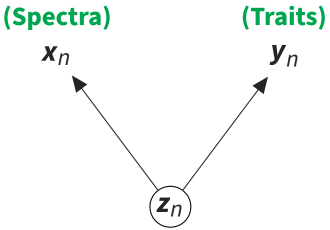
PARTIAL LEAST SQUARES

+ Fast, easy, accurate*, easy to modify the “regression” part

− Not a statistical model $\implies$ no uncertainty, difficult to introduce sample correlations

~ Choosing `Q` can be tricky, strongly depends on the quality of data

BAYESIAN PARTIAL LEAST SQUARES



Each sample is generated through:

$$\boldsymbol{x}_n = W\boldsymbol{z}_n + \varepsilon_n, \quad n = 1, \ldots N,$$

$$\boldsymbol{y}_n = C\boldsymbol{z}_n + \eta_n,$$

where $\boldsymbol{z}_n \overset{iid}{\sim} \mathrm{N}(\mathbf{0}, I_Q)$, $\varepsilon_n \overset{iid}{\sim} \mathrm{N}(\mathbf{0}, \Sigma)$ and $\eta_n \overset{iid}{\sim} \mathrm{N}(\mathbf{0}, \Psi)$.

This allows for likelihood-based inference of model parameters:

$$\mathrm{p}(\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{\Theta}) = \prod_{n=1}^{N} f(\boldsymbol{x}_n, \boldsymbol{y}_n | \boldsymbol{\Theta}),$$

where $\boldsymbol{\Theta} = (\boldsymbol{Z}, W, C, \Sigma, \Psi, \ldots)$ are all (unknown) model parameters.

BAYESIAN PARTIAL LEAST SQUARES

This allows for likelihood-based inference of the model:

- Frequentist methods for these types of models typically require $\mathtt{Q} \leq \mathtt{R}$ which greatly limits prediction utility.
- Bayesian inference with appropriate priors bypasses this constraint.
  → Bayesian partial least squares (**BPLS**)

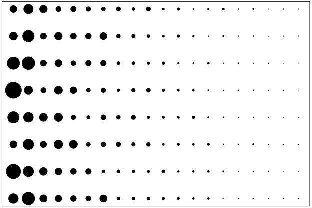
Priors: Here we assume vague(ish) priors: apart from regularising model fitting, they won't heavily contribute to the final predictions.

*Identifiability of Θ is not required for predictions.

SHRINKAGE ON THE LATENT VARIABLES

Structure of loading matrix:

$W =$



$\Longrightarrow$ Subsequent latent components contribute less to the signal.

$\rightarrow$ Suppose `Q` is infinite:

- Take a **stochastically increasing** sequence $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots)$.
- Elements of W are given conjugate normal priors such that

$$\mathbb{V}[w_{pq}] \propto \frac{1}{\tau_q}, \quad p=1,\ldots,P,\ q=1,2,\ldots\,.$$

- **Multiplicative gamma process** prior:

$$\tau_q = \prod_{k=1}^{q} \delta_k, \quad \text{where} \quad \delta_k \sim \text{Gamma}(\alpha, \beta).$$

SPARSITY, PRIORS AND INFERENCE

Large parts of the spectrum may tell you nothing about the traits.

Response part of BPLS: $\boldsymbol{y}_n = C\boldsymbol{z}_n + \eta_n, \;\; n = 1, \ldots, N$

We consider two sparse variants:

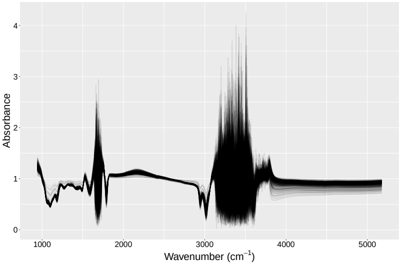
- **Spike-and-slab** — sets some columns of C to be exactly zero (**ss-BPLS**)
  →Corresponds to Two-Way Orthogonal PLS (O2-PLS)
- **Bayesian LASSO** — emulates the ℓ_1-penalty on elements of C (**L-BPLS**)
  →Corresponds to sparse PLS (sPLS)

Can assign conjugate priors everywhere → Gibbs sampling

Output: Posterior predictive distribution of $\boldsymbol{Y}^{\text{new}} | \boldsymbol{X}^{\text{new}}, \mathcal{D}$
(marginalised over parameter posterior)

MOTIVATING PROBLEM — MID-INFRARED SPECTRA OF MILK

$\mathtt{P} = 538$ wavenumbers



Traits ($\mathtt{R} = 3$) are measured in a lab:

- **heat stability**

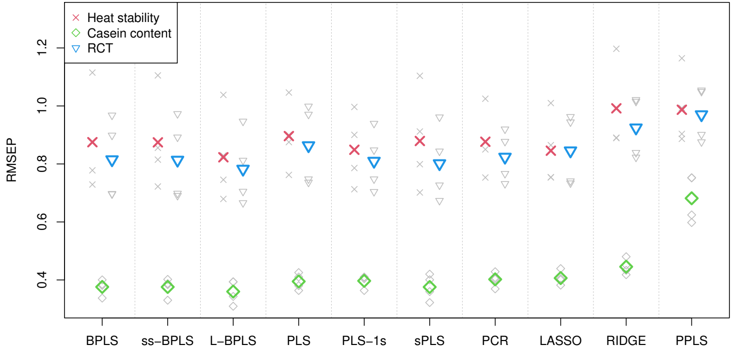
  — for making infant formula;
- **rennet coagulation time** (RCT)

  — for making cheese;
- **casein content**

  — major protein.

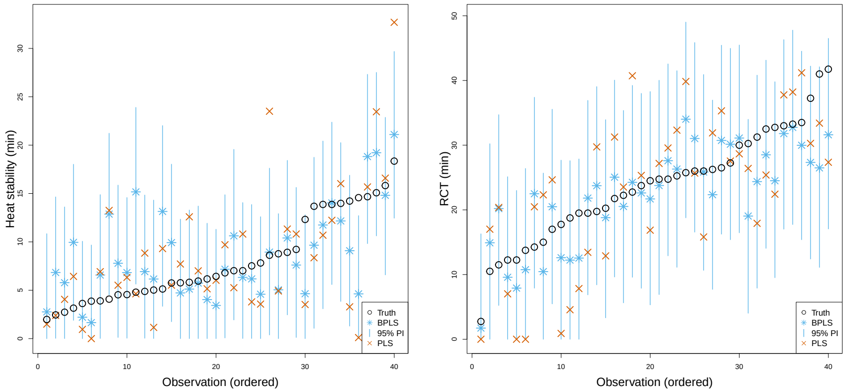
We have $N = 363$ complete observations. (large $\mathtt{P}$, small N problem)

MILK MIR SPECTRAL DATA RESULTS



- **L-BPLS** is consistently the most accurate method.
- Statistical model outputs have a lot more utility.

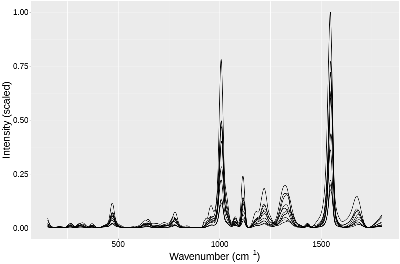
MILK MIR SPECTRAL DATA RESULTS



MOTIVATING PROBLEM — SERS-BASED PH SENSORS

pH sensors using surface-enhanced Raman spectroscopy
→ detect product spoilage or indicate mastitis.

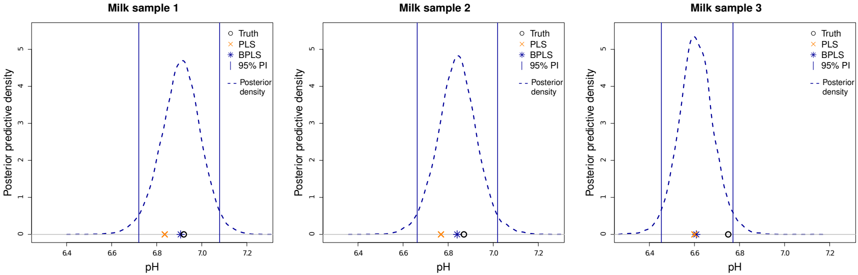
$\mathtt{P}$ = 1733 wavenumbers



The pH of two cartons of milk is measured in a lab over 6 days:

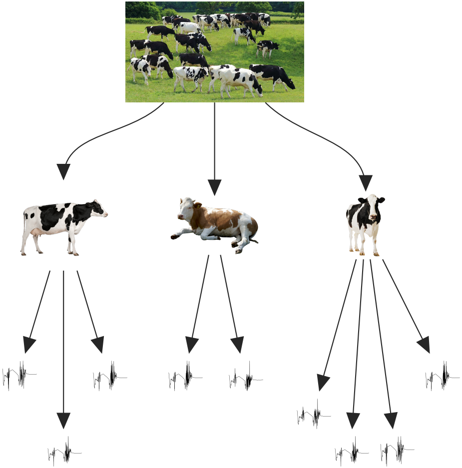
- Very small dataset (N = 11) — can we recover any signal?

MILK SERS DATA RESULTS



- Standard PLS methods fail to identify signal + cross-validation methods used for finding `Q` really struggle here.
- BPLS methods manage to produce reasonable point predictions but highlight the uncertainty due to the small dataset.

TOWARDS HIERACHICAL MODELLING



Paper:

Urbas, S., Lovera, P., Daly, R., O'Riordan, A., Berry, D., & Gormley, I. C. (2024). Predicting milk traits from spectral data using Bayesian probabilistic partial least squares regression. *The Annals of Applied Statistics*, 18(4), 3486–3506.

Code:

- `bplsr` package available on CRAN.

SPARSITY — SPIKE-AND-SLAB

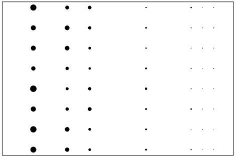
- **Idea**: large components explaining variation in $\boldsymbol{X}$ may play no part in explaining the variation in $\boldsymbol{Y}$
- In the model, this implies that some columns of C may be zero.
- **Spike and slab approach:** For each column introduce Bernoulli variables $b_q,\ q = 1, 2, \ldots$ which can "switch" the latent variable on and off.

With $B = \mathrm{diag}(b_1, b_2, \ldots)$,the response part of the model becomes

$$\boldsymbol{y}_n = CB\boldsymbol{z}_n + \varepsilon_n.$$

Elements of B can be inferred via posterior Gibbs sampling.

$C =$



SPARSITY — BAYESIAN LASSO

- Explicit prior of the form $\pi_0(C) \propto \exp\left(-\lambda \sum_{r,q} |c_{rq}|\right)$, $\lambda > 0$ results in an intractable and non-differentiable posterior
- Park and Casella (2009) uses a scale mixture of normal distributions to get a similar form:

$$\frac{\lambda}{2} e^{-\lambda|c|} = \int_0^\infty \frac{1}{\sqrt{2\pi v}} e^{c^2/(2v)} \times \frac{\lambda^2}{2} e^{\lambda^2 v/2}\, \mathrm{d}v.$$

- So

$$c_{rq}|\tau_q, v_{rq} \sim \mathrm{N}(0, v_{rq}/\tau_q), \;\; v_{rq}|\lambda \sim \mathrm{Exp}(\lambda^2/2), \;\; r = 1, \ldots, \mathtt{R}, \; q = 1, 2, \ldots$$

  gives us what we want.
- Posterior conditionals of v_{rq} are inverse-Gaussian so can still use Gibbs