

Bayesian Profile Regression for Multimorbidity Clustering at Population Scale

Dr Jim Rafferty

Swansea University Medical School

Acknowledgements

Reference: <https://arxiv.org/abs/2602.24038>



HDRUK
Health Data Research UK

Project aims

- Find clusters of MLTC in EHRs held in SAIL
- Write a BPR model that can be fit using SVI
- Validate model performance using simulation studies

Bayesian Profile Regression

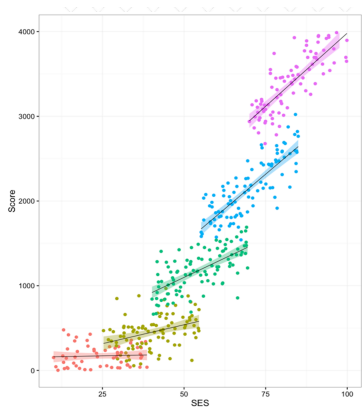


Image taken from <https://www.r-bloggers.com/2016/10/multilevel-modeling-of-educational-data-using-r-part-1/> accessed 17th April 2026

Bayesian Profile Regression

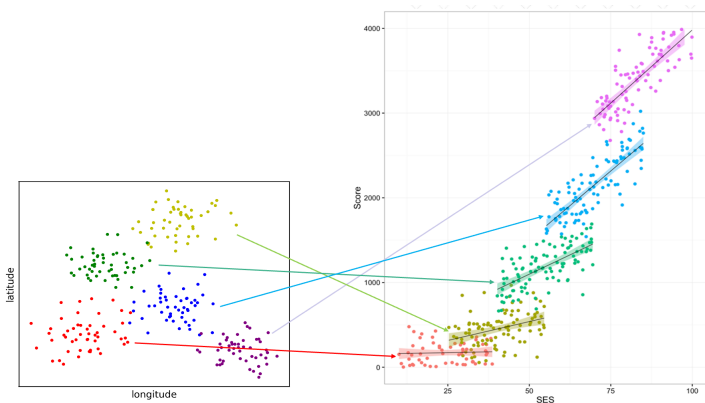


Image taken from <https://www.r-bloggers.com/2016/10/multilevel-modeling-of-educational-data-using-r-part-1/> accessed 17th April 2026

Aside - Dirichlet Process Mixture models

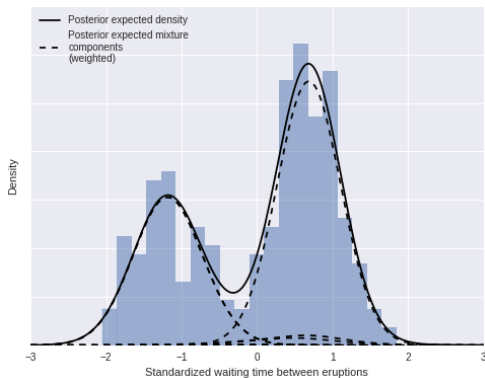


Image taken from <https://austinrochford.com/posts/2016-02-25-density-estimation-dpm.html> accessed 17th April 2026

Bayesian Profile Regression

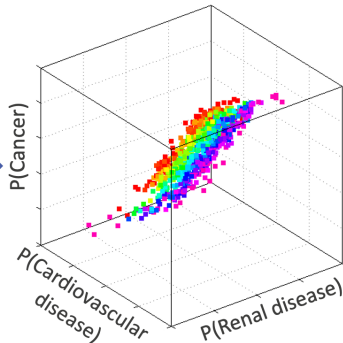
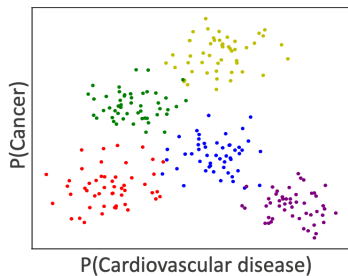


Image courtesy of Rhiannon Owen

References

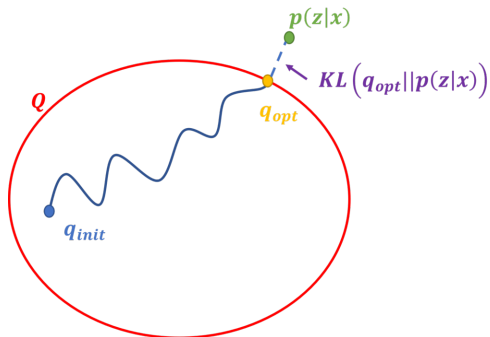
- Molitor J et al. Bayesian profile regression with an application to the National Survey of Children's Health. *Biostatistics*. 2010 Jul 1;11(3):484-98.
- Liverani S et al. PReMiuM: An R package for profile regression mixture models using Dirichlet processes. *Journal of statistical software*. 2015 Mar 20;64:1-30.

Standard fitting methods

- In a Bayesian framework we calculate the full posterior distribution rather than just point estimates
- Normally done with MCMC sampling.
- MCMC is computationally quite expensive.

Stochastic Variational Inference

Stochastic Variational Inference



from <https://meerkatstatistics.com/courses/variational-inference-in-r/> accessed 17th April 2026

Advantages & Disadvantages

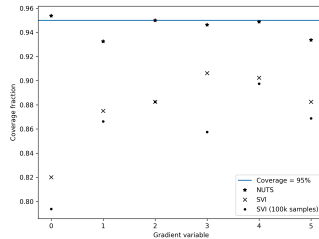
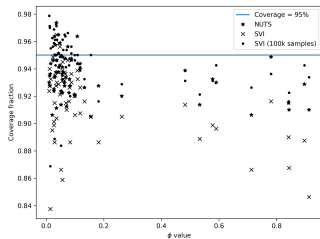
No-U-Turn Sampler

- ✓ Convergence guarantees
- ✓ Convergence diagnostics
- ✗✗ Computationally expensive

Stochastic Variational Inference

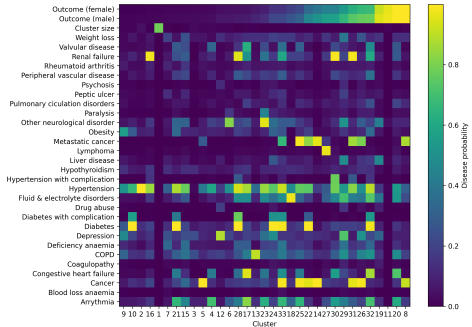
- ✗ No convergence guarantees
- Convergence diagnostics
- ✓✓ Computationally efficient

Results - Simulation study

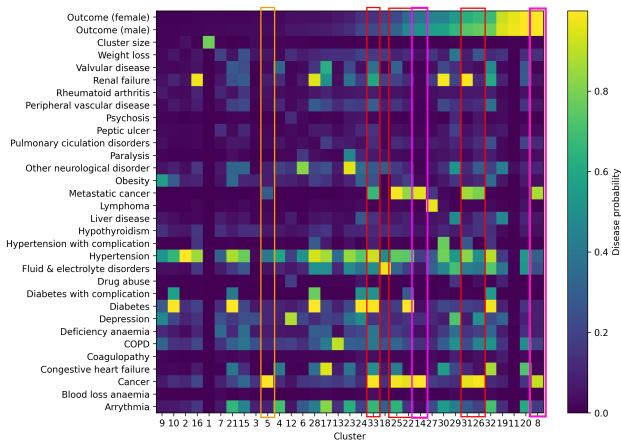


Results - SAIL analysis

Variable	N(%) / Mean(SD)
Population N	1296463
Male Sex N(%)	631453 (48.70%)
Age Mean (SD) (years)	42.51 (15.95)
Died N(%)	160709 (12.40%)
Total Follow up (years)	20949059



Results - SAIL analysis



Summary & Future work

- BPR allows for granular clustering, conditioning on covariates.
- SVI allows Bayesian model fitting even at population-scale
- Future work
 - Time to event response model
 - More exotic mixture sector structures

Thank you

Questions?



j.m.rafferty@swansea.ac.uk

See https://github.com/jim-rafferty/HDR_midlands_data_grand_rounds_20290421 for slides and notes

References:

- JR et al. 2026 Bayesian Profile Regression using Variational Inference to Identify Clusters of Multiple Long-Term Conditions Conditioning on Mortality in Population-Scale Data
<https://arxiv.org/abs/2602.24038>. Submitted to *Biostatistics*
- J Molitor et al. 2010 Bayesian profile regression with an application to the National Survey of Children's Health. *Biostatistics*
- S Liverani et al. 2015 PReMiuM: An R package for profile regression mixture models using Dirichlet processes. *Journal of statistical software*
- J Lyons et al. 2021 Protocol for the development of the Wales Multimorbidity e-Cohort (WMC): data sources and methods to construct a population-based research platform to investigate multimorbidity *BMJ Open*

SVI - losses

