BASIL: Scalable Bayesian Semi-supervised Clustering with Feature Selection and Adaptive Constraint Weighting

Luwei Wang [1] Dagmara Panas [1] Ke Wang [1] Bruce Guthrie [2] Sohan Seth [1]

[1]School of Informatics, University of Edinburgh
[2]Usher Institute, University of Edinburgh

BASIL

Motivation & Contributions

- **Operationalisability.** How we cluster should reflect what the clusters are for. Multimorbidity subgroups in older adults are most useful when they align with healthcare outcomes, which calls for outcome-aware rather than purely unsupervised clustering.
- **Constrained clustering** uses pairwise must-/cannot-link constraints when labels are scarce or K is unknown.
- **Existing methods** are limited in **scalability** (full-dataset passes), **interpretability** (no per-cluster attribution), and **robustness** to noisy constraints.
- We propose BASIL (**Ba**yesian **S**emi-supervIsed c**L**ustering), a unified hierarchical Bayesian model with stochastic variational inference (SVI) jointly addressing all three.

Contributions. BASIL targets **binary** observations and combines HMRF-structured clustering with **SVI for scalable inference** on $N \geq 5 \cdot 10^5$, integrated **per-cluster feature selection** for interpretable cluster-specific features, and **adaptive constraint weighting** via Gamma-prior latent weights, robust to noisy supervision.

The BASIL Generative Model

The hierarchical generative model combines a Dirichlet-multinomial mixture, Gamma-prior constraint weights, and Beta-shrinkage feature selection,

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}_0), \quad w_{nm} \sim \mathrm{Gamma}(\boldsymbol{\lambda}_w), \quad \bar{w}_{nm} \sim \mathrm{Gamma}(\boldsymbol{\lambda}_{\bar{w}}),$$

$$\mathbf{z} \mid \boldsymbol{\pi}, W, \bar{W}, Y \sim p(\mathbf{z} \mid \boldsymbol{\pi}, W, \bar{W}, Y),$$

$$\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}_k), \quad \gamma_{kd} \sim \mathrm{Beta}(\boldsymbol{\lambda}_{\gamma 1}, \boldsymbol{\lambda}_{\gamma 2}),$$

$$\tilde{\theta}^{\mu}_{kd} = \gamma_{kd}\theta^{\mu}_{kd} + (1 - \gamma_{kd})\theta^{\mu}_{0d},$$

$$\mathbf{x}_n \mid z_n = k, \tilde{\boldsymbol{\theta}}_k \sim p(\mathbf{x}_n \mid z_n = k, \tilde{\boldsymbol{\theta}}_k).$$

- Feature relevance $\gamma_{kd} \in (0, 1)$ shrinks each cluster mean toward the population mean θ^{μ}_{0d}; high γ_{kd} marks a discriminative feature for cluster k.
- Constraint weights $w_{nm}, \bar{w}_{nm}$ are latent variables, identifying and down-weighting noisy supervision.
- HMRF-Gibbs prior on $\mathbf{z}$ uses a **cluster-divergence** potential (symmetric KL between cluster-conditional likelihoods) for must-links and a fixed penalty for violated cannot-links.

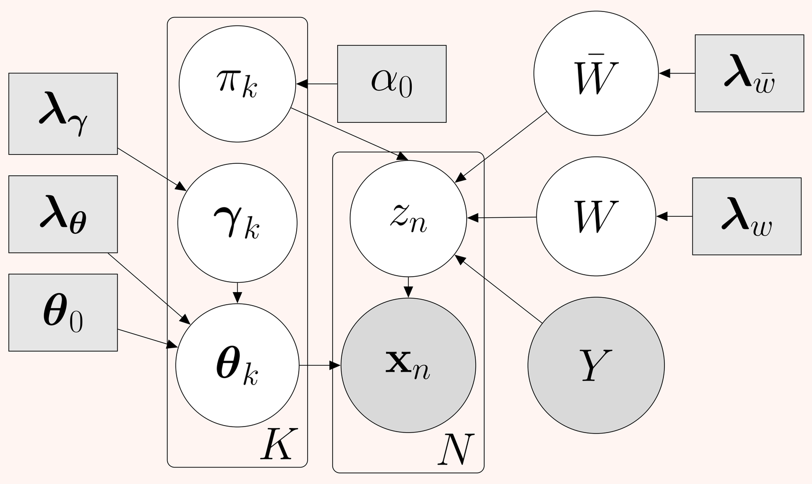


Figure 1. Plate diagram of BASIL. Shaded nodes are observed; clear nodes are latent; rectangles are hyperparameters.

Stochastic Variational Inference and Phased Warmup

We maximise the ELBO $\mathcal{L}(q) = \mathbb{E}_q[\log p(X, \mathbf{z}, \Theta \mid Y)] - \mathbb{E}_q[\log q(\mathbf{z}, \Theta)]$ with $\Theta = \{\boldsymbol{\pi}, W, \bar{W}, \{\boldsymbol{\theta}_k\}, \{\boldsymbol{\gamma}_k\}\}$ via closed-form coordinate ascent on local parameters and stochastic natural-gradient steps on global parameters. Per-batch cost: $O(MKD) + O(Mg_{\max}K)$.

Key SVI updates

Feature importance (CVI-style mini-batch):

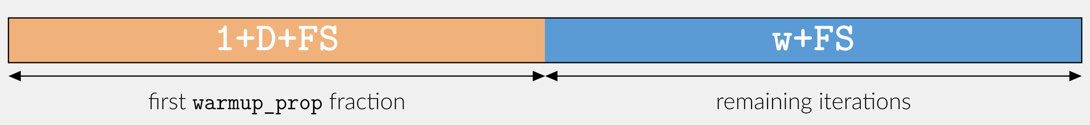
$$\hat{c}_{kd} = \boldsymbol{\lambda}_{\gamma 1} + \tfrac{N}{M}\max(0, \hat{l}_{kd}), \quad \hat{d}_{kd} = \boldsymbol{\lambda}_{\gamma 2} + \tfrac{N}{M}\max(0, -\hat{l}_{kd}).$$

Adaptive constraint weights (Gamma rate posterior):

$$\beta_{nm,2} = \boldsymbol{\lambda}_{w2} + \mathbb{I}(y_{nm} = 1)\sum_{j,k}\Phi_{nj}\Phi_{mk}\mathbb{D}^{\mathrm{sym}}_{\mathrm{KL}}.$$

Phased `warmup_prop` schedule

Joint KL + adaptive weights is unstable; stabilise cluster geometry first, then engage weight inference.



Recommendations. Trusted: fix w=1 + D + FS. Noisy: adaptive w + FS. Unknown noise: phased `warmup_prop`.

AIM CISC
ACRC Advanced Care Research Centre
BiomedAI
Data Science Unit

Benchmark Results

Datasets: Digits (N=1,797, D=64), MNIST (N=70k, D=784). **Baselines:** MML, PCK, MPCK, DCGM. Test ACC mean±SD over 10 seeds.

Data	Model	20% sup. ACC	20% sup. Time	50% sup. ACC	50% sup. Time
Digits	MML	0.61± 0.03	1.1 m	0.71± 0.05	4.2 m
	PCK	**0.80**± 0.01	9.9 s	0.84± 0.00	4.7 m
	MPCK	0.75± 0.02	1.5 m	0.76± 0.07	9.7 m
	DCGM	0.51± 0.30	2.2 m	0.61± 0.38	2.2 m
	BASIL	**0.80**± 0.02	7.0 m	**0.87**± 0.00	44.4 m
MNIST	ML/PCK/MPCK	–	> 2 d	–	> 2 d
	DCGM	0.51± 0.13	1.95 h	0.72± 0.07	2.54 h
	BASIL	**0.68**± 0.02	1.52 h	**0.79**± 0.00	1.57 h

Table 1. Test clustering accuracy and training time. BASIL reduces training time by > 96% on MNIST where classical baselines do not converge within two days.

Feature importance on MNIST

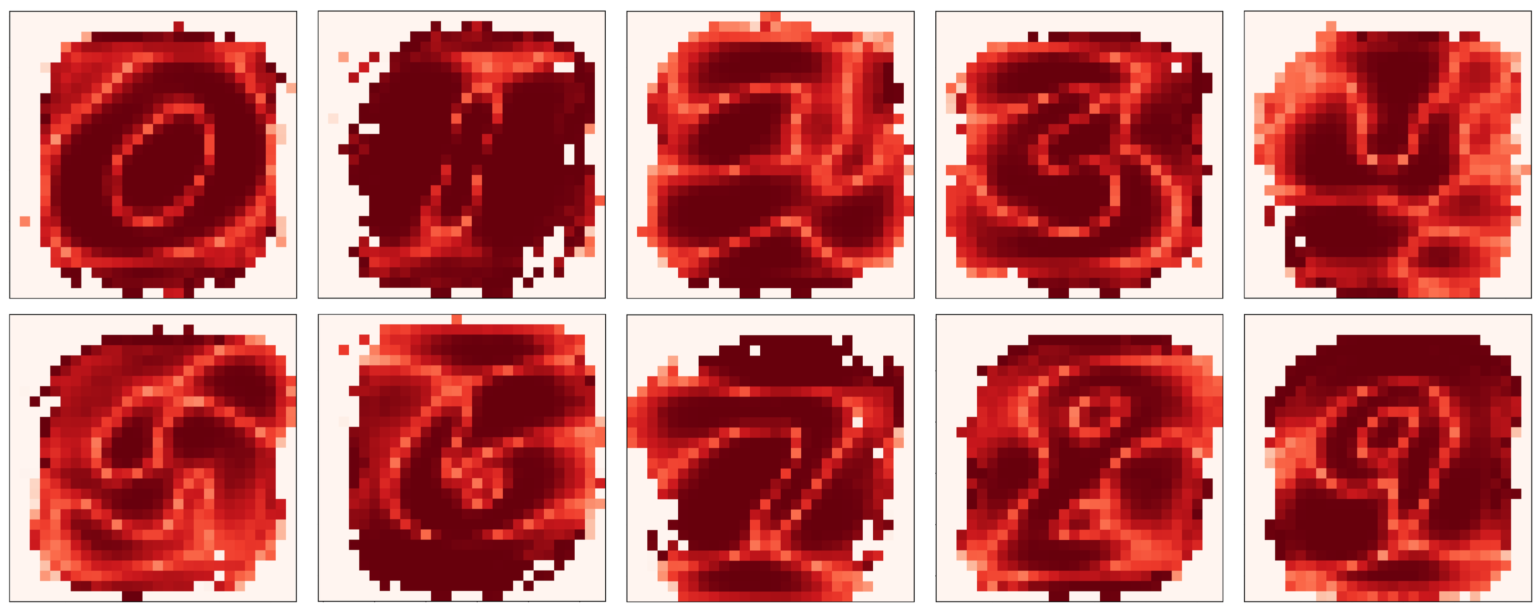


Figure 2. BASIL MNIST feature importance (digits 0–9). Darker pixels are more discriminative. AUROC = 1.00 on the synthetic feature-selection benchmark.

Robustness to noisy constraints

On Digits at 20% supervision, BASIL with adaptive w flags corrupted pairs at **AUROC** $= 0.97 \pm 0.01$ and holds accuracy to 30% constraint noise where deterministic baselines collapse.

Application to Large-scale Health Data

CPRD (UK primary care, N=501,207 patients aged 70–74, 61 binary conditions) provides must/cannot-link constraints from a composite outcome of planned admission, emergency admission, and mortality. BASIL discovers 28 clinically interpretable subgroups with greater outcome alignment and diversity than unsupervised clustering.

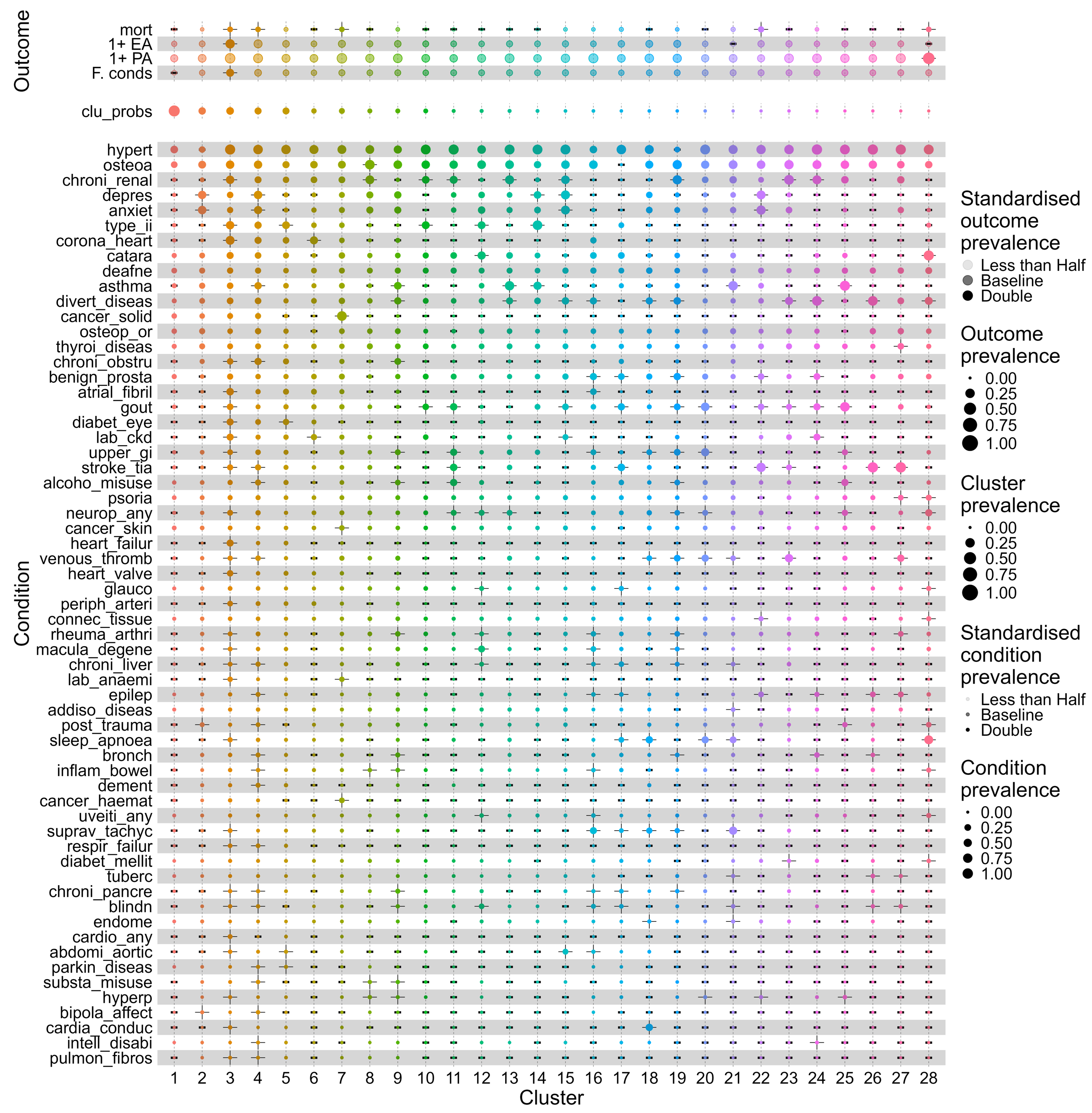


Figure 3. CPRD cluster profiles. Bubble size shows within-cluster condition prevalence. +/– mark conditions or outcomes with relative risk > 2× or < 0.5× the population rate.

Takeaways

Scalability: SVI scales BASIL to $N \geq 5 \cdot 10^5$ with $\sim 58\times$ wall-clock reduction over full-batch Bayesian baselines. **Interpretability:** Per-cluster Beta-shrinkage feature selection recovers ground-truth informative features at AUROC = 1.00 and yields cluster-specific digit-stroke feature maps on MNIST. **Robustness:** Gamma-prior adaptive weights flag noisy supervision at AUROC = 0.97 (20% noise) and hold accuracy to 30% noise.