Sparse Variational Inference: Bayesian Coresets from Scratch

THE UNIVERSITY OF BRITISH COLUMBIA

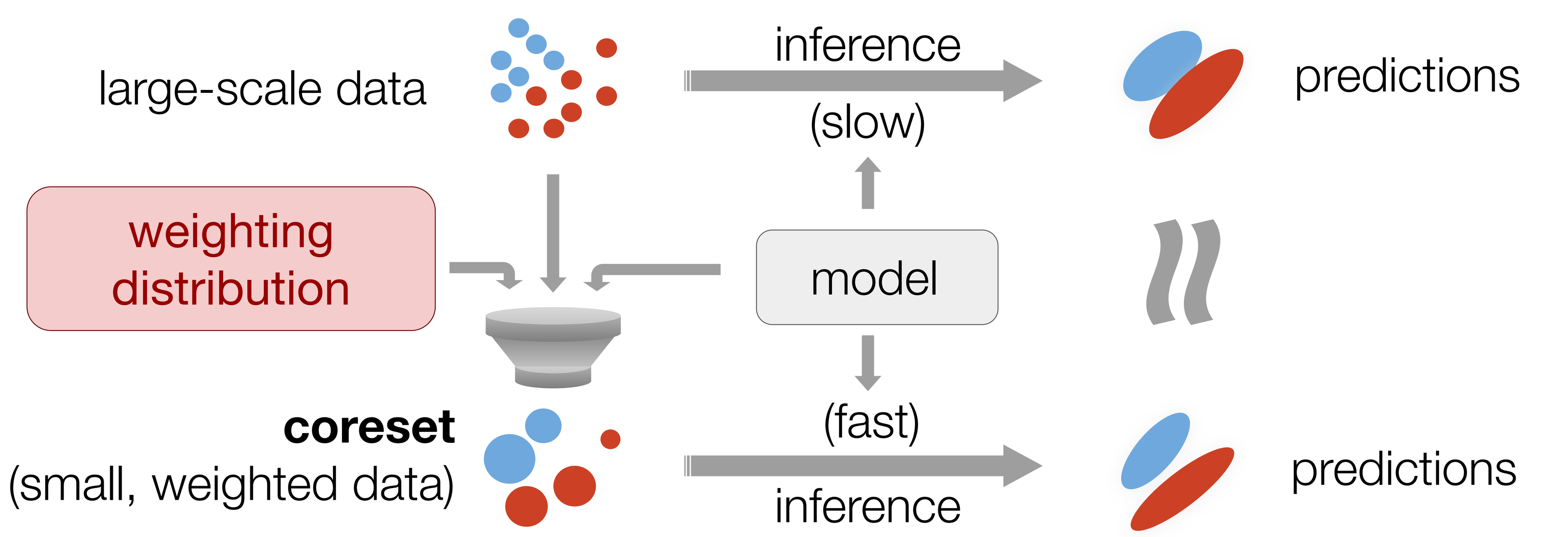
Trevor Campbell, Boyan Beronov • UBC Statistics, Computer Science

Summary

Automated algorithms in Bayesian statistics have provided practitioners access to reproducible data analysis with complex models. But obtaining scalability, guarantees, and automation together remains a challenge.

Bayesian coresets

Scalable inference with **statistical guarantees** via **data summarization**.



coreset posterior · prior · data point n log-likelihood · weight

$$\pi_w(\theta) = \frac{1}{Z_w}\pi_0(\theta)\prod_{n=1}^N \exp(f_n(\theta))^{w_n} = \pi_0(\theta)\exp\left(f(\theta)^T w - \log Z_w\right)$$

π_1 : exact posterior $\qquad 1 \in \mathbb{R}^N_{\geq 0},\ \|1\|_0 = N$

π_w : sparse coreset posterior $\quad w \in \mathbb{R}^N_{\geq 0},\ \|w\|_0 \leq M \ll N$

Contributions

Existing work requires the choice of a fixed weighting distribution. Using a novel information-geometric perspective, we show this fundamentally limits coreset quality, and develop a new, tuning-free construction algorithm with superior accuracy.

Previous State of the Art

Hilbert coresets [CB17,18]: sparse nonnegative least squares

1) discretize log-likelihoods

$$\theta_1, \dots, \theta_S \overset{\text{i.i.d.}}{\sim} \hat{\pi}$$ weighting distribution

$$g_n = \frac{1}{\sqrt{S}}\begin{bmatrix} f_n(\theta_1) - \bar{f}_n \\ \vdots \\ f_n(\theta_S) - \bar{f}_n \end{bmatrix} \qquad \bar{f}_n = \frac{1}{S}\sum_{s=1}^S f_n(\theta_s)$$

2) minimize distance to sum

$$w^\star = \arg\min_{w \in \mathbb{R}^N} \left\| \sum_{n=1}^N g_n - \sum_{n=1}^N w_n g_n \right\|_2^2 \quad \text{s.t.} \quad w \geq 0,\ \|w\|_0 \leq M$$

How should we pick $\hat{\pi}$, and what is its effect?

Sparse Variational Inference

Key Insight 1: Coreset posteriors form an **exponential family** with natural parameter w and sufficient statistic $f(\theta)$.

Hence, the objective of **sparse variational inference,**

$$w^\star = \arg\min_{w \in \mathbb{R}^N} \mathcal{D}_{\mathrm{KL}}(\pi_w \| \pi_1) \quad \text{s.t.} \ \|w\|_0 \leq M, \ , w \geq 0$$

has tractable gradient:

$$\nabla_w \mathcal{D}_{\mathrm{KL}}(\pi_w \| \pi_1) = \mathrm{Cov}_{\pi_w}\left[f, f^T(w-1)\right]$$

Problem: Estimating $\nabla_w \mathcal{D}_{\mathrm{KL}}$ requires sampling from π_w.

Key Insight 2: Sampling from π_w is practical for sparse w.

SparseVI: Iterative Greedy Algorithm

Initialize: weights $w \leftarrow 0 \in \mathbb{R}^N$, active index set $\mathcal{I} \leftarrow \emptyset$

Iterate:

Estimate $\hat{C} := \mathrm{Corr}_{\pi_w}\left[f, f^T(w-1)\right] \in \mathbb{R}^N$ via $\theta_1, \dots, \theta_S \overset{\text{i.i.d.}}{\sim} \pi_w$

Select data point: add $\arg\max_n \hat{C}_n$ to $\mathcal{I}$

Update w: SGD for active set $\mathcal{I}$ using $\nabla_w \mathcal{D}_{\mathrm{KL}}$ estimates from

Information-Geometric Perspective

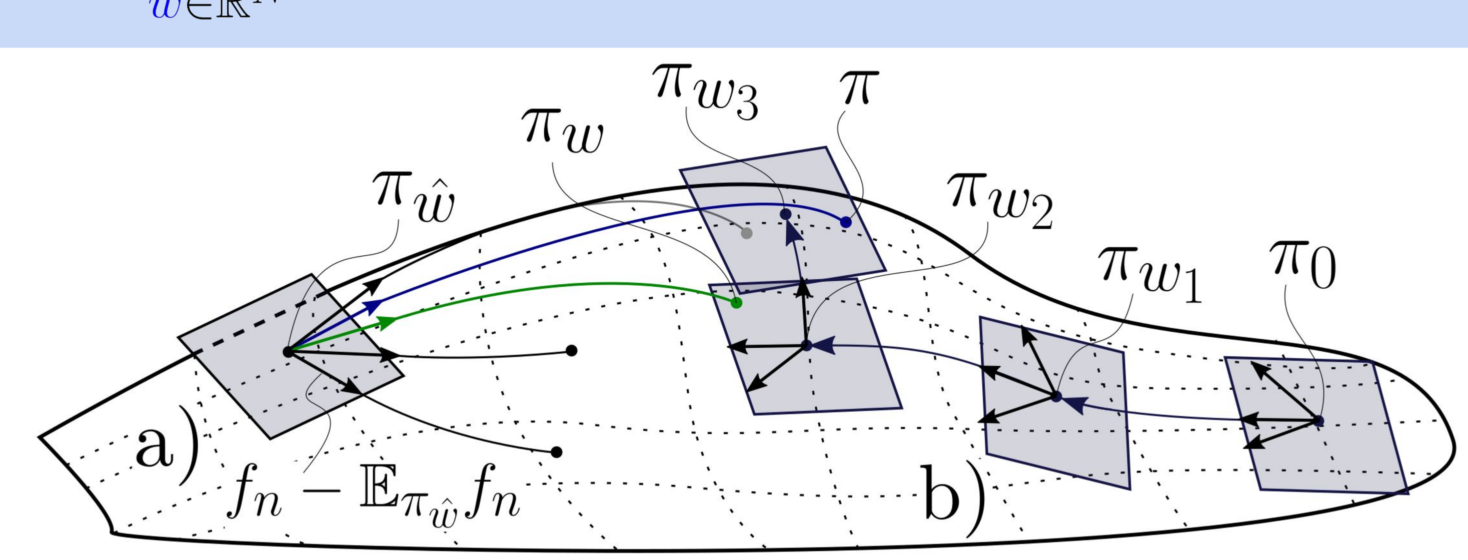
The Fisher information metric on the **coreset posterior manifold** is

$$G(w) = \mathbb{E}_{\pi_w}\left[(\nabla_w \log \pi_w)(\nabla_w \log \pi_w)^T\right] = \nabla_w^2 \log Z_w = \mathrm{Cov}_{\pi_w}[f]$$

We show that past constructions operate on a *fixed* tangent space, whereas SparseVI is a Riemannian optimization algorithm that adapts *iteratively* to the manifold geometry.

Theorem: Both Hilbert coreset construction and *each iteration* of SparseVI are equivalent to local alignment of geodesic initial tangents:

$$w^\star = \arg\min_{w \in \mathbb{R}^N} \|\xi_{\hat{w}\to 1} - \xi_{\hat{w}\to w}\|_{G(\hat{w})} \quad \text{s.t.} \ \|w\|_0 \leq M, \ , w \geq 0$$

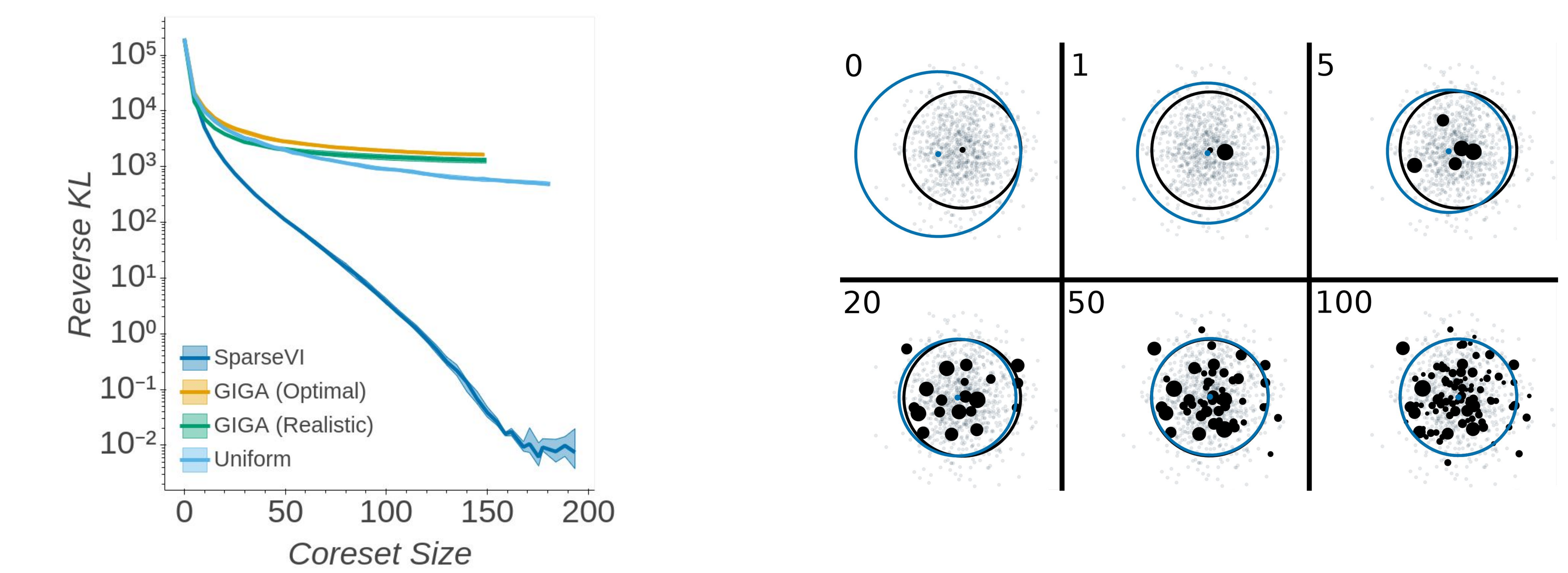


Experimental Results

Coreset post. convergence of SparseVI vs. Uniform subsampling and GIGA [CB18] with weighting distributions: Exact (Optimal) / Noisy (Realistic)

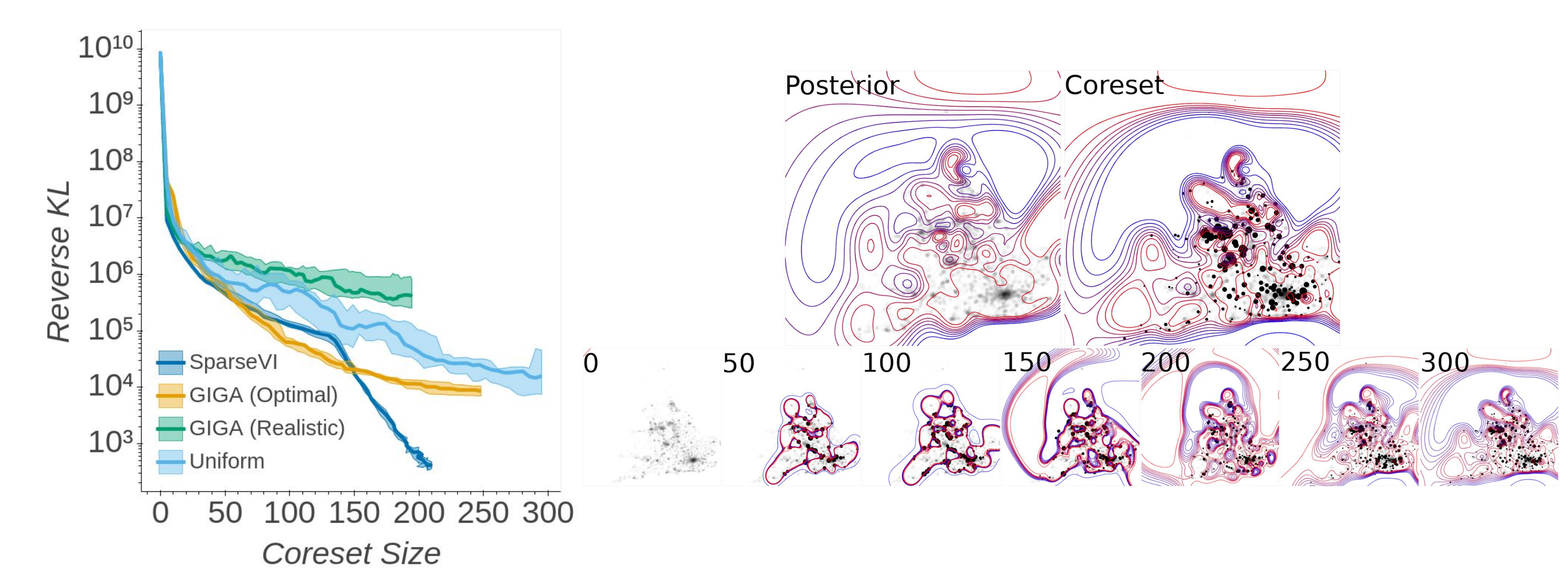
Synthetic Gaussian

200 dimensions, 1K samples



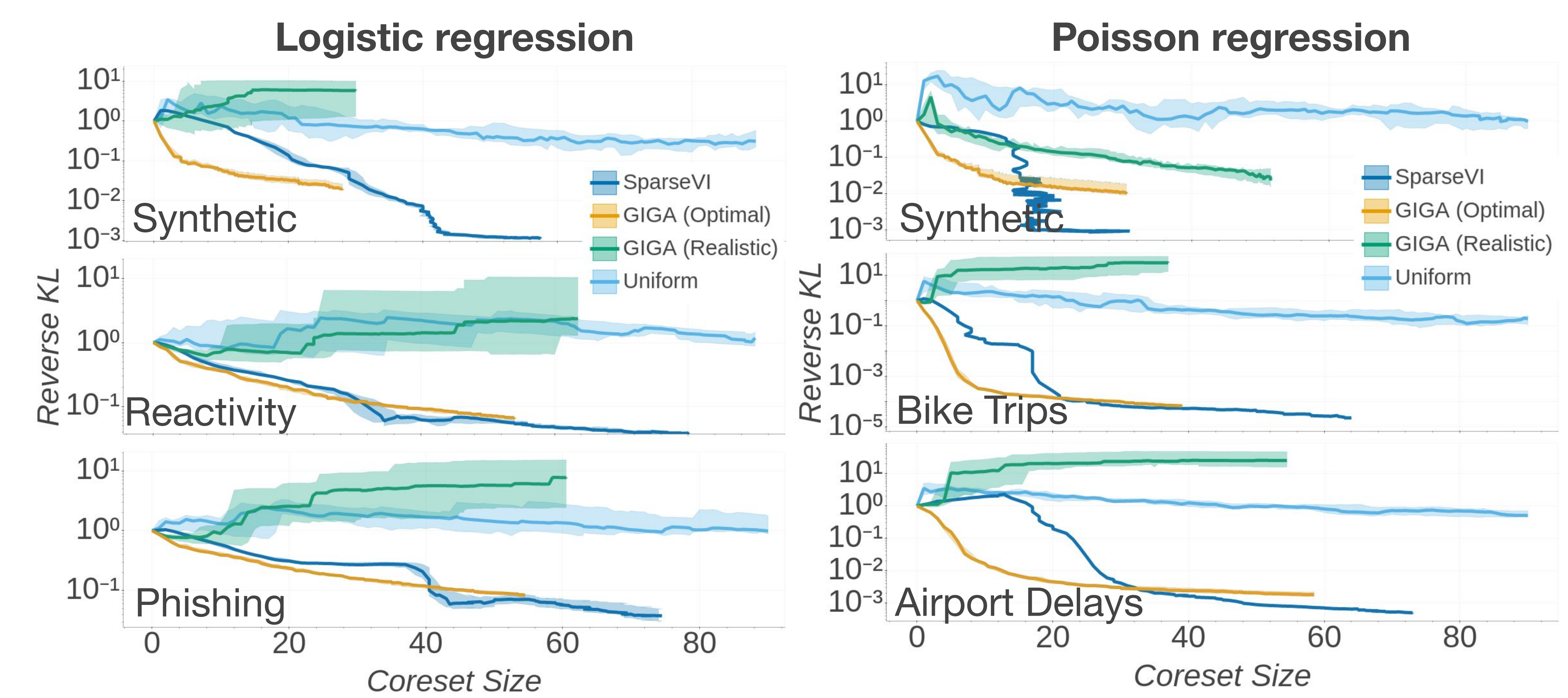
Basis Function Regression

301 dimensions, 10K records (2018 UK Land Registry Dataset)



Logistic & Poisson Regression

6 datasets, 2-15 dimensions, 500 data points



Supported by two NSERC Discovery Grants and an NSERC Discovery Launch Supplement.

E-Mail: trevor@stat.ubc.ca