Scalable Learning of Multivariate Distributions via Coresets

AISTATS 2026

Zeyu Ding, Katja Ickstadt, Nadja Klein, Alexander Munteanu, Simon Omlor

May 2, 2026

TU Dortmund University Lamarr Institute for Machine Learning and AI Karlsruhe Institute of Technology

1. Motivation & Background

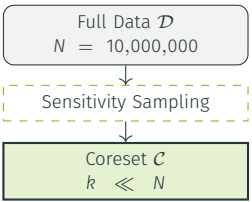
Motivation: Scalable Inference for Large-Scale Data

The core problem

"How do we fit complex multivariate models when N is in the millions?"

The scalability bottleneck

- Maximum likelihood for MCTMs requires iterative optimization over all N points.
- Each gradient step costs $O(N)$ — prohibitive for large datasets.
- Running on a **small, carefully chosen subset** could yield the same result at a fraction of the cost.



Goal: A weighted subset $\mathcal{C}$ such that optimizing on $\mathcal{C}$ gives the same result as on $\mathcal{D}$.

2. Multivariate Conditional Transformation Models

What are Multivariate Conditional Transformation Models (MCTMs)?

The core concept

Instead of assuming $Y \sim$ ComplexDist, we learn a **transformation** h such that:

$$h(Y \mid X) = Z \sim \mathcal{N}(0, \Sigma)$$

Why MCTM?

- **Flexible:** Captures skewness, heavy tails, and non-linear dependencies.
- **Interpretable:** Coefficients ϑ describe the change in distribution.
- **Multivariate:** Models joint distributions via Gaussian copula structure.

MCTM Structure & the Optimization Objective

Model definition

Model the conditional distribution of $Y \in \mathbb{R}^J$ via transformation h:

$$Z = h(Y \mid X) \sim \mathcal{N}_J(0, \Sigma) \implies F_Y(y \mid x) = \Phi_\Sigma(h_1(y_1), \ldots, h_J(y_J))$$

Bernstein parametrization

Flexibility and monotonicity via Bernstein basis polynomials $\mathbf{a}(y)$:

$$h_j(y_j \mid x) = \mathbf{a}(y_j)^\top \boldsymbol{\vartheta}_j(x)$$

Constraint: Monotonicity ($h' > 0$) requires:

$$\frac{\partial h_j}{\partial y_j} = \mathbf{a}'(y_j)^\top \boldsymbol{\vartheta}_j > 0$$

(Constraint enforces monotonicity)

The optimization objective

Log-likelihood splits into two conflicting terms:

$$\ell_i = \underbrace{-\tfrac{1}{2}\|h(\mathbf{y}_i)\|^2_{\Sigma^{-1}}}_{\text{Quadratic term}} + \underbrace{\sum \log(\mathbf{a}'^\top \boldsymbol{\vartheta})}_{\text{Barrier (Log-Jacobian)}}$$

- **Blue:** Standard ℓ_2 term — handled by leverage scores.
- **Red:** Barrier — singular if monotonicity violated. **Requires Convex Hull.**

3. Coreset Construction for MCTMs

Coreset and Sensitivity Sampling

Core mechanism: Sensitivity sampling

Sensitivity of point x_i: its maximum relative contribution to the loss

$$\zeta_i = \sup_{\theta} \frac{f_i(\theta)}{\sum_{j=1}^n f_j(\theta)}$$

Sampling: Draw k points with probability $p_i \propto s_i \geq \zeta_i$, reweight by $w_i = \frac{1}{k \cdot p_i}$.
Goal: $(1 \pm \varepsilon)$-approximation of the objective uniformly over all parameters.

The Challenge: Log-Barrier Singularity

The mathematical barrier

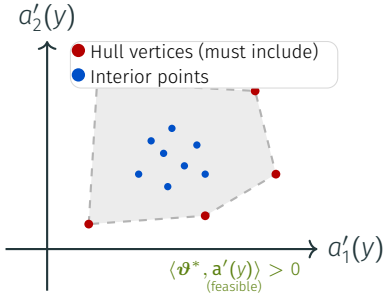
The log-likelihood depends on the **derivative** of the transformation:

$$\mathcal{L} \propto \sum \log(\underbrace{h'(y_i)}_{\text{slope}})$$

Constraint: Monotonicity requires $h'(y) > 0$.

If we naively subsample...

- Interior points are overrepresented; boundary points are missed.
- The estimated slope at boundaries can become negative.
- $\log(\text{negative}) \rightarrow$ **NaN / crash**.



Feasibility: $\langle \vartheta, \mathbf{a}'(y) \rangle > 0$ for all $y \Leftrightarrow$ all points form **acute angles** with ϑ.

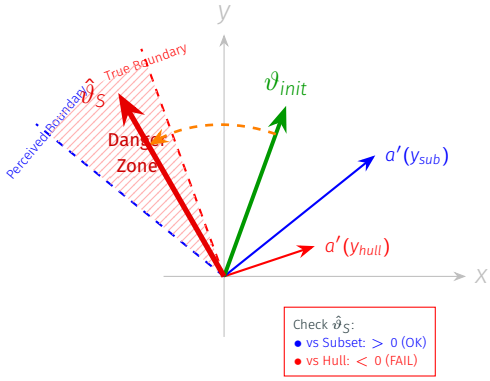
Missing hull vertices allows the optimizer to find ϑ that violates monotonicity at the boundaries.

Why Standard Subsampling Fails

1. **Feasible region constraint**
   - Full data constraints: Red
   - Sampled constraints without hull: Blue

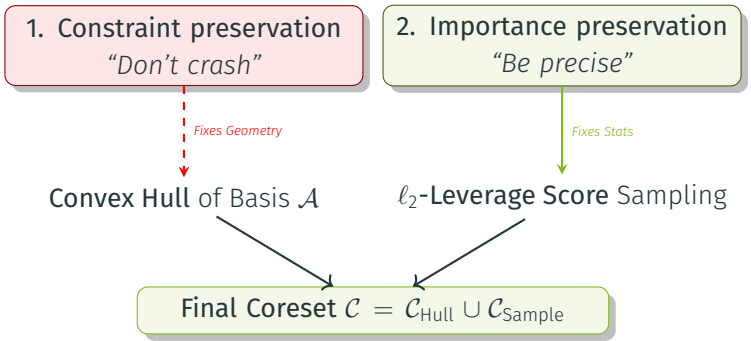
2. **The optimizer**
   - Finds $\hat{\vartheta}_S$ maximizing likelihood on the subset.

3. **The log-likelihood crash**
   - $\hat{\vartheta}_S$ is valid for the sample ($\langle \cdot, \cdot \rangle > 0$).
   - But **invalid** for the hidden hull points ($\langle \cdot, \cdot \rangle \leq 0$).
   - Full-data likelihood $\to -\infty$.



Our Solution: A Hybrid Coreset Strategy

We propose a **Hybrid Coreset** that separates two concerns: **Feasibility** vs. **Accuracy**.



Result: A small subset that guarantees a valid likelihood *and* high distributional accuracy.

4. Theoretical Guarantee

Theoretical Guarantee: The Hybrid Coreset Theorem

Hybrid Construction

The coreset $\mathcal{C}$ is constructed by combining two sets:

- **Geometric Set:** All extreme points of the convex hull of derivatives $\{a'_{ij}\}$.
- **Sampling Set:** Points sampled with probability proportional to:

$$p_i \propto \underbrace{u_i}_{\ell_2\text{-leverage}} + \underbrace{1/n}_{\text{uniform}}$$

Main Theorem (Loss Approximation)

Let $f(A, \vartheta, \lambda)$ be the negative log-likelihood. With high probability, for all parameters in the feasible domain $D(\eta) = \left\{(\vartheta, \lambda) : \ \mathbf{a}'(y_{ij})^\top \vartheta_j > \eta, \ \forall i, j\right\}$:

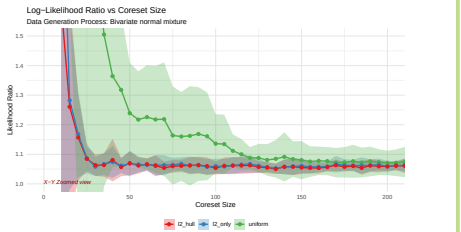
$$\left|f(A, \vartheta, \lambda) - f_{\mathcal{C}}(A(S), \vartheta, \lambda)\right| \leq \varepsilon \cdot f(A, \vartheta, \lambda)$$

- **Result:** A $(1 \pm \varepsilon)$-approximation of the full-data loss.

5. Experiments

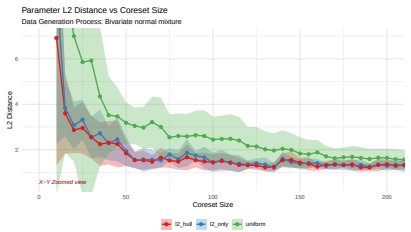
Quantitative Results: Error Convergence

1. Likelihood ratio ($\to 1.0$)



Metric: $\mathcal{L}_{\text{coreset}} / \mathcal{L}_{\text{full}}$

2. Parameter error ($\to 0.0$)



Metric: $\|\hat{\vartheta}_{\mathcal{C}} - \hat{\vartheta}_{\text{full}}\|_2$

Conclusion: The Hybrid method (Red Line) achieves lower variance and faster convergence than Uniform (Green) and pure ℓ_2 methods.

Marginal Density Recovery via Coresets

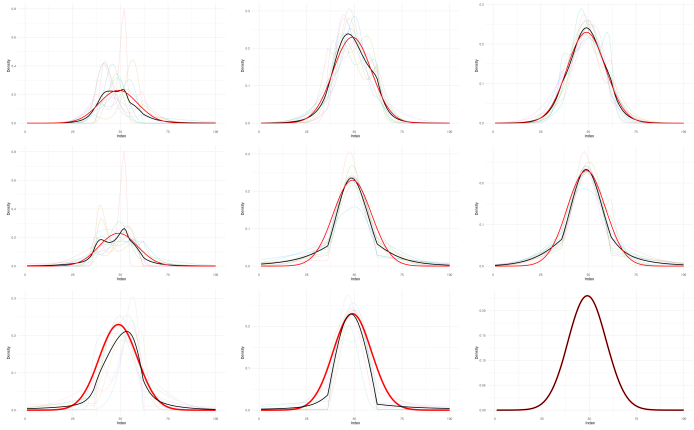
Task

Predict the marginal density of Y using a tiny coreset ($k = 50, 100, 500$) from $N = 10{,}000$ points. **Red line** = True Density. **Black line** = Coreset Estimate (average of 10 runs).

Uniform

ℓ_2-Only

Hybrid (Ours)



$k = 50$ $k = 100$ $k = 500$

6. Conclusion

Summary & Future Directions

Contributions

1. **Identified the core challenge**
   - Log-barrier singularity makes standard sensitivity sampling fail for MCTMs.
2. **Proposed the Hybrid Coreset**
   - Convex Hull guarantees feasibility.
   - ℓ_2-leverage scores guarantee accuracy.
3. **Theoretical guarantee**
   - $(1 \pm \varepsilon)$ loss approximation for all $\vartheta \in D(\eta)$.
4. **Empirical validation**
   - Accurate density recovery with $k \ll N$.

Future Directions

1. **Dynamic / streaming coresets**
   - Handle data insertion and deletion online.
2. **Structural generalizations**
   - Beyond Gaussian copulas.
   - Other basis function families for MCTMs.
3. **Generative modeling**
   - MCTM + Normalizing Flows.
   - Efficient large-scale density estimation.
4. **Bayesian inference**
   - Extend to full posterior approximation via MCMC on the coreset.

Goal: Scalable, theoretically grounded inference for flexible multivariate models.

References

Ding, Z., Ickstadt, K., Klein, N., Munteanu, A., & Omlor, S. (2025). Scalable learning of multivariate distributions via coresets. *Accepted by AISTATS 2026*.

Klein, N., Hothorn, T., Barbanti, L., & Kneib, T. (2022). Multivariate conditional transformation models. *Scandinavian Journal of Statistics*, 49(1), 116–142.

Hothorn, T., Kneib, T., & Bühlmann, P. (2014). Conditional transformation models. *Journal of the Royal Statistical Society Series B*, 76(1), 3–27.

Huggins, J., Campbell, T., & Broderick, T. (2016). Coresets for scalable Bayesian logistic regression. *NeurIPS*, 29, 4080–4088.

Munteanu, A., Omlor, S., & Peters, C. (2022). p-Generalized probit regression and scalable maximum likelihood estimation via sketching and coresets. *AISTATS*, pp. 2073–2100.

Munteanu, A., & Omlor, S. (2024). Optimal bounds for ℓ_p sensitivity sampling via ℓ_2 augmentation. *ICML*. PMLR, pp. 36769–36796.

Campbell, T., & Broderick, T. (2019). Automated scalable Bayesian inference via Hilbert coresets. *JMLR*, 20(1), 551–588.

Thank You!

Questions welcome.

zeyu.ding@tu-dortmund.de

Backup Slides

Necessity of the Convex Hull: Formal Argument

The Goal: ε-Approximation

We seek a subset S such that for all $\theta = (\vartheta, \lambda) \in D(\eta)$: $(1-\varepsilon)\mathcal{L}(\vartheta; \mathcal{D}) \leq \tilde{\mathcal{L}}(\vartheta; S) \leq (1+\varepsilon)\mathcal{L}(\vartheta; \mathcal{D})$, where $D(\eta) = \{\theta : \langle \vartheta_j, a'_{ij} \rangle > \eta \; \forall i, j\}$.

Step 1: Optimize on S

The optimizer finds $\hat{\vartheta}_S$ on subset S (lacking hull vertices):

$$\forall y \in S : \langle \hat{\vartheta}_S, a'(y) \rangle > 0$$

$\tilde{\mathcal{L}}(\hat{\vartheta}_S; S)$ is **Finite** & **High**

Step 2: Evaluate on $\mathcal{D}$

Apply $\hat{\vartheta}_S$ to full data containing **hull vertices**:

$$\exists y \in \mathcal{D} : \langle \hat{\vartheta}_S, a'(y) \rangle \leq 0$$

$$\Downarrow$$

$$\mathcal{L}(\hat{\vartheta}_S; \mathcal{D}) \to \textbf{Undefined} / -\infty$$

Conclusion: Coreset guarantee fails

$$\tilde{\mathcal{L}}(\hat{\vartheta}_S; S) \text{ (Finite)} \quad \not\approx \quad \mathcal{L}(\hat{\vartheta}_S; \mathcal{D}) \; (-\infty) \quad \Longrightarrow \quad \text{No finite } \varepsilon \text{ exists.}$$

Technical Proof: Loss Decomposition Strategy

Problem: The MCTM loss mixes quadratic and logarithmic barrier terms:

$$f(A, \vartheta, \lambda) = \sum_{i,j} \left[\frac{1}{2} \left(\sum \lambda \cdot \vartheta^T a_{ij} \right)^2 - \log(\vartheta^T a'_{ij}) \right]$$

Standard coreset techniques fail because the **log** term is unbounded near the boundary.

Divide and Conquer Strategy

We split the loss into three parts and bound them separately:

1. f_1 **(Squared Part):** The dominant quadratic error term.
   Technique: ℓ_2-Subspace Embedding (Matrix Sketching).
2. f_2 **(Positive Log Part):** Where $\log(\cdot) > 0$.
   Technique: Bounded by f_1 + VC-dimension argument.
3. f_3 **(Negative Log Part):** Where $\log(\cdot) < 0$ (near singularity).
   *Technique: Domain $D(\eta)$ + **Convex Hull** of derivatives.*

Proof Detail: The Squared Part (f_1)

Challenge: The quadratic term involves cross-terms between λ and ϑ.

Solution: Rewrite the sum of squares as a single matrix norm $\|\mathbf{B}\theta\|_2^2$ via a block-diagonal structure.

$$\mathbf{B}_i = \begin{pmatrix} b_i & 0 & \cdots \\ 0 & b_i & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \in \mathbb{R}^{J \times dJ^2}$$

where b_i contains the basis vectors a_{ij}.

Lemma 3.3.1:
Applying ℓ_2-leverage score sampling to $\mathbf{B}$ yields:

$$(1-\varepsilon)f_1 \leq \tilde{f}_1 \leq (1+\varepsilon)f_1$$

with size $\tilde{O}(dJ^2/\varepsilon^2)$.

This reduces the semi-parametric interaction to a standard linear algebra problem.

Proof Detail: The Logarithmic Barrier (f_2, f_3)

Positive Log (f_2)

Idea: Log growth is slower than linear/quadratic.

$$\log(x) \leq x - 1$$

The sensitivity of the log part is bounded by the sensitivity of the squared part (u_i) plus a uniform term ($1/n$).

Result: Sampling $p_i \propto u_i + 1/n$ covers f_2.

Negative Log (f_3)

The Singularity: If $\vartheta^T a' \to 0$, Loss $\to \infty$. Sensitivity is unbounded.

Fix:

1. Restrict domain to $D(\eta) = \{\vartheta \mid \vartheta^T a' > \eta\}$.
2. Include the **Convex Hull** of $\{a'_{ij}\}$.

Geometry: Hull points define the "walls" of the feasible cone.

Main Theorem: Combining these sets gives a $(1 \pm \varepsilon)$-approximation for the total likelihood.

Why Convex Hull Guarantees Feasibility

The Problem: Unbounded Sensitivity

The log-likelihood contains $-\log(\langle \vartheta, a'(y) \rangle)$. When $\langle \vartheta, a'(y) \rangle \to 0$: sensitivity $\to \infty$, so the multiplicative $(1 \pm \varepsilon)$ approximation becomes **undefined**.

Solution: Restricted Domain $D(\eta)$

Restrict to parameters bounded away from singularity:

$$D(\eta) = \{\vartheta : \langle \vartheta, a'(y) \rangle > \eta, \ \forall y\}$$

Choice: $\eta = \Theta(\varepsilon)$
Lemma: Optimal $\vartheta^* \in D(\eta)$ exists with error $O(\varepsilon)$.

Why Convex Hull Suffices

Inner product is **linear**. If $\langle \vartheta, v \rangle > \eta$ for all hull vertices v, then for any convex combination $p = \sum_i \alpha_i v_i$:

$$\langle \vartheta, p \rangle = \sum_i \alpha_i \langle \vartheta, v_i \rangle > \eta$$

$\Rightarrow$ Hull vertices alone enforce $D(\eta)$.

Chain: Convex hull $\to \vartheta \in D(\eta) \to$ bounded sensitivity $\to (1 \pm \varepsilon)$ approx.

Convex Hull in High Dimensions

MCTM Has Special Structure

The design matrix is **block-diagonal**: each block $a(y_{ij}) \in \mathbb{R}^d$ is a Bernstein polynomial of a **single variable**.
$\Rightarrow$ Convex hull is computed **per margin**, not jointly. Effective dimension is just d, not $d \times J$.

User-Controlled Dimension

- Bernstein degree d is a hyperparameter.
- Typical choice: $d = 5 \sim 10$.
- Compute hull per margin and take the union:

$$S_{\text{hull}} = \bigcup_{j=1}^{J} \text{Hull}\left(\{a_j'(y_{ij})\}_{i=1}^{n}\right)$$

Practical Experience

On a standard laptop:

- $d \leq 10$: Exact hull computation is **fast**.
- High J: iterate per margin, union the indices.

Most applications use moderate d, so exact hull is entirely feasible.

Relationship to Normalizing Flows

Shared Mathematical Foundation

Both MCTMs and normalizing flows rely on the probability transformation formula:

$$\log p_Y(y \mid x) = \underbrace{\log p_Z(\mathbf{h}(y \mid x))}_{\text{Base Density}} + \underbrace{\log \left| \det \left(\frac{\partial \mathbf{h}(y \mid x)}{\partial y} \right) \right|}_{\text{Volume Correction}} \tag{1}$$

MCTM

- **Parametrization:** Semi-parametric — Bernstein polynomials.
- **Structure:** Explicit Cholesky-like triangular Jacobian.
- **Focus: Interpretability** & statistical inference.

Normalizing Flows

- **Parametrization:** Deep learning — neural network weights.
- **Structure:** Deep composition $\mathbf{h} = h_L \circ \cdots \circ h_1$.
- **Focus: Expressivity** & generative capabilities.

Key Insight

MCTMs can be viewed as a **statistically interpretable, single-layer autoregressive flow**, balancing flexibility with structural understanding.

Backup: Role of VC Dimension in Coreset Theory

The Core Challenge

Coreset requires **uniform approximation**: the guarantee must hold for **all** θ simultaneously.

- Single-θ: $O(1/\varepsilon^2)$ samples vs All-θ: infinitely many parameters?

VC Dimension: Controlling Function Class Complexity

VC dimension measures the "effective degrees of freedom" of $\{f_\theta : \theta \in \Theta\}$.

Finite VC dim $\Rightarrow$ finitely many "behavior patterns" $\Rightarrow$ **single-θ to all-θ** via union bound.

Uniform convergence bound: With $O\left(\frac{S}{\varepsilon^2} \cdot \Delta \cdot \log S\right)$ samples, we achieve $(1 \pm \varepsilon)$ approximation for **all** θ simultaneously. Here Δ denotes the VC dimension.

VC Dimension in Our Models

Model	VC Dim (Δ)	Reason
MCTM	$O(d+1)$	Monotonicity $\Rightarrow$ linear classifiers