STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–19.
Published online in International Academic Press (www.IAPress.org)

Accelerating Parametric Inference in Healthcare Data Streams Using Symbolic Histogram Likelihoods

Ahmad Hakiim Jamaluddin[1,2,*] Syaiful Anam[3], Muhammad Jaffri Mohd Nasir[4], Andrea Tri Rian Dani[5], Nur Chamidah[6]

[1] *Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia*

[2] *Statistical Data Science Group, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia*

[3] *Department of Mathematics, Faculty of Science, Technology and Mathematics, Universitas Brawijaya, Malang 65145, East Java, Indonesia*

[4] *Faculty of Entrepreneurship and Business, Universiti Malaysia Kelantan, Kota Bharu, 16100, Kelantan, Malaysia*

[5] *Statistics Study Programme, Department of Mathematics, Faculty of Mathematics and Natural Sciences, Mulawarman University, Samarinda 75123, Indonesia*

[6] *Department of Mathematics, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia*

Abstract Maximum likelihood estimation (MLE) is a fundamental tool for parametric inference, but likelihood evaluations become computationally prohibitive for very large samples and for moving-window analyses of streaming data. We develop a symbolic maximum likelihood framework for univariate continuous data in which the raw sample is replaced by a histogram-valued symbol. The resulting objective is a multinomial log-likelihood whose cost per evaluation depends on the number of histogram boundary points B, rather than directly on the sample size n. Utilising the symbolic data analysis (SDA) framework, we demonstrate the substantial computational efficiency improvements of MLE via histogram-valued symbols where the raw sample is replaced by a compressed symbolic representation. The evaluation cost of the resulting multinomial log-likelihood objective depends on the number of histogram boundary points B, rather than directly on the sample size n. We show that the symbolic likelihood is exact for the aggregated histogram counts and that, under increasingly fine partitions, it approximates the classical log-likelihood based on the individual observations up to an additive bin-width term that does not depend on the unknown parameter. We present a numerically stable implementation, compare equal-width and equal-probability binning, and describe a sliding-window workflow with constant-time count updates and $\mathcal{O}(B)$ likelihood evaluation per optimisation step. The simulation benchmark covers Exponential, Normal, Gamma, Weibull, and Lognormal models for sample sizes up to 10^6 and histogram resolutions from $B-1=10$ to $B-1=500$ bins. In the largest-sample setting, the equal-width symbolic estimator achieves speed-ups ranging from 2.64-fold to 90.05-fold across the reported configurations. Accuracy improves steadily with finer partitions: at $n=10^6$ and $B-1=500$, the relative RMSE of the equal-width estimator ranges from 1.000 to 1.042 across the five families, while equal-probability bins are especially helpful for strongly skewed models at coarse resolution. A public-data illustration on the MIMIC-III heart-rate stream shows that the symbolic approach remains practical in window-wise fitting. For the preferred model, the median per-window speed-up is 5-fold, with close agreement between classical and symbolic parameter estimates. These findings position symbolic MLE as a practical likelihood-based compression strategy for large-scale simulation, estimation, and near-real-time monitoring.

Keywords Symbolic Data Analysis, Maximum Likelihood Estimation, Histogram-Valued Data, Streaming Inference, Computational Statistics, Healthcare Analytics

AMS 2010 subject classifications 62F10, 62P10, 62M10

*Correspondence to: Ahmad Hakiim Jamaluddin (Email: ahmadhakiim@upm.edu.my). Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia.

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 202x International Academic Press

DOI: 10.19139/soic-2310-5070-4242

1. Introduction

Maximum likelihood estimation (MLE) remains one of the central tools of modern statistical inference because it supports parametric estimation, uncertainty quantification, likelihood-based model comparison, and principled model selection. In many contemporary applications, however, the dominant computational burden is no longer the optimiser itself but the evaluation of the likelihood across very large samples. This issue becomes acute in streaming, telemetry, digital health, and high-frequency monitoring settings, where the same model must be re-estimated over thousands of overlapping windows. The central motivation of this paper is that this repeated re-estimation makes the per-evaluation likelihood cost, rather than the optimiser itself, the binding computational constraint, and that histogram-based symbolic compression can reduce this cost by orders of magnitude while preserving the likelihood-based interpretation of the resulting inference.

A natural response is to compress the raw sample before inference. Symbolic Data Analysis (SDA) provides a coherent framework for this purpose by replacing individual observations with aggregated objects such as intervals, histograms, or more general distribution-valued summaries [1, 2, 3, 4, 5, 6, 8]. Histogram-valued symbols are particularly attractive because they retain much more distributional structure than a single summary statistic while sharply reducing the dimension of the object entering estimation. Recent likelihood-based SDA research has shown that symbolic likelihoods can be constructed directly from an underlying measurement-level model, thereby avoiding ad hoc assumptions and preserving a principled inferential connection between the grouped representation and the original micro-data model [4, 5, 6, 8]. Histogram-valued and related symbolic representations also continue to motivate recent methodological work, including symbolic classification rules for histogram-valued data [17] and the maximum-likelihood development of flexible parametric families in the statistics, optimisation and information computing literature [15, 16].

This paper studies a computationally focused symbolic MLE strategy for *univariate continuous data* aggregated into histograms. The goal is not to replace classical MLE based on the individual observations when n is modest and the model is fitted once, but to provide a principled and implementation-ready surrogate when n is very large. Once a sample is compressed into a histogram defined by B boundary points, hence $B-1$ bins, the optimiser no longer traverses all n observations at each iteration. Instead, each objective evaluation depends only on the bin counts and the model-implied probabilities of those bins.

While the foundational mathematics and theoretical properties of this framework were established by Beranger (2023) [6], applied users still need a univariate formulation that is fully explicit, notationally transparent, computationally stable, and benchmarked in the terms that matter for statistical computing: approximation accuracy and runtime reduction. The present manuscript addresses that gap by combining a boundary-based notation, a direct grouped-likelihood derivation, a finite-partition approximation argument, and a large empirical benchmark. The computational emphasis also places the work in a broader class of reduced-likelihood and compressed-inference strategies, akin in spirit to the trade-offs considered in composite likelihood and binned approximation methods [11, 12].

There are three contributions of the paper. We present the paper as a comprehensive computational evaluation and a practical implementation guide that builds directly on the theoretical framework of Beranger (2023), rather than as a claim of new statistical theory. First, we present a finite-partition approximation result showing that, under shrinking bin widths, the symbolic log-likelihood approaches the classical log-likelihood based on the individual observations up to an additive term independent of the unknown parameter. Second, we present its practical computational performance by detailing implementation aspects like unconstrained reparameterisation, stable handling of boundary zones, and warm-started streaming updates. Third, we report a broad simulation benchmark across five standard

continuous families together with a public physiological-data application based on the MIMIC-III heart-rate stream, demonstrating its viability for near-real-time monitoring.

The remainder of the paper is organized as follows. Section 2 presents the symbolic likelihood, its theoretical approximation link to classical MLE, and our computational considerations. Section 3 reports the simulation design and the empirical accuracy–runtime trade-off. Section 4 presents the MIMIC-III public-data illustration. Section 5 discusses practical recommendations, limitations, and extensions. Section 6 concludes.

2. Methodology

2.1. Notation and problem setup

Let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables from a continuous model with density $f(y \mid \boldsymbol{\theta})$ and cumulative distribution function (CDF)

$$F(y \mid \boldsymbol{\theta}) = \Pr_{\boldsymbol{\theta}}(Y \leq y),$$

where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ is an unknown parameter vector. Throughout the manuscript, $F(y \mid \boldsymbol{\theta})$ always denotes the *model CDF* associated with the assumed parametric family; it does not denote the likelihood and it does not denote the grouped-data distribution of the histogram counts. Throughout the manuscript, we adopt the convention that B denotes the number of histogram *boundary points*, so that the number of histogram bins is $B - 1$. Table 1 presents the convention for all notations used consistently in all formulas, tables, algorithms, and figure captions.

Table 1. Core notation used throughout the manuscript.

Symbol	Meaning
n	Raw sample size
$Y_1, \ldots, Y_n$	Independent observations from the target continuous model
$\boldsymbol{\theta} \in \Theta$	Unknown parameter vector
$f(y \mid \boldsymbol{\theta})$	Model density
$F(y \mid \boldsymbol{\theta})$	Model cumulative distribution function (CDF), $F(y \mid \boldsymbol{\theta}) = \Pr_{\boldsymbol{\theta}}(Y \leq y)$
B	Number of histogram boundary points
$B - 1$	Number of histogram bins
$s_1 < \cdots < s_B$	Ordered boundary points defining the histogram partition
$I_b = [s_b, s_{b+1})$	Bin b, for $b = 1, \ldots, B - 1$
c_b	Count of observations falling in bin I_b
$\boldsymbol{c} = (c_1, \ldots, c_{B-1})$	Histogram count vector
$\pi_b(\boldsymbol{\theta})$	Model probability of bin I_b
$\ell_n(\boldsymbol{\theta})$	Classical log-likelihood based on the ungrouped observations
$\ell_{\mathrm{sym}}(\boldsymbol{\theta})$	Symbolic log-likelihood based on grouped counts
Δ_B	Maximum bin width, $\max_{1 \leq b \leq B-1}(s_{b+1} - s_b)$
w	Moving-window width in streaming analysis
r	Moving-window step size

2.2. Classical likelihood-based inference and its connection to grouped data

The classical log-likelihood based on the ungrouped sample is

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(Y_i \mid \boldsymbol{\theta}), \tag{1}$$

where $f(\cdot \mid \boldsymbol{\theta})$ represents any underlying continuous probability density function (PDF) assumed to generate the data. The classical maximum likelihood estimator (classical MLE) is therefore

$$\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell_n(\boldsymbol{\theta}). \tag{2}$$

A single evaluation of (1) requires $\mathcal{O}(n)$ operations, and the total fitting cost depends on how many objective and gradient evaluations are required by the optimiser. When n is large, or when the same model is re-fitted across many windows, traversal of all observations becomes the dominant computational bottleneck.

To bypass this bottleneck, symbolic MLE coarsens the exact point evaluations of the PDF $f(Y_i \mid \boldsymbol{\theta})$ by replacing them with integrals of $f(\cdot \mid \boldsymbol{\theta})$ over histogram bins. Crucially, as formally established in Section 2.5, substituting these aggregated bin probabilities into a symbolic likelihood objective creates a direct analytical link to the classical log-likelihood in (1). By applying the Mean Value Theorem to the bin integrals of the PDF $f(\cdot \mid \boldsymbol{\theta})$, the working symbolic log-likelihood $\tilde{\ell}_{\mathrm{sym}}(\boldsymbol{\theta})$ approximates the classical log-likelihood $\ell_n(\boldsymbol{\theta})$ up to an additive discretization term, K_B. Because this shift K_B depends strictly on the empirical bin counts and bin widths rather than the unknown parameter $\boldsymbol{\theta}$, the symbolic approximation structurally preserves the geometry of the classical log-likelihood objective.

2.3. Histogram symbols and symbolic compression

Let

$$-\infty \le s_1 < s_2 < \cdots < s_B \le \infty$$

define a partition of the support through the $B-1$ bins

$$I_b = [s_b, s_{b+1}), \qquad b = 1, \ldots, B-1,$$

with the convention that the last interval may be closed on the right if needed. The histogram-valued symbol associated with the sample is

$$S = (\boldsymbol{s}, \boldsymbol{c}),$$

where $\boldsymbol{s} = (s_1, \ldots, s_B)^\top$ and $\boldsymbol{c} = (c_1, \ldots, c_{B-1})^\top$ with

$$c_b = \sum_{i=1}^{n} \mathbf{1}\{Y_i \in I_b\}, \qquad b = 1, \ldots, B-1,$$

$$\sum_{b=1}^{B-1} c_b = n.$$

Constructing the histogram has one-time cost $\mathcal{O}(n)$. The computational gain appears after compression: subsequent likelihood evaluations depend only on the grouped counts $\boldsymbol{c}$ and the boundary vector $\boldsymbol{s}$.

2.4. Symbolic likelihood for histogram-valued data

For a given parameter value $\boldsymbol{\theta}$, the probability that one observation falls in bin I_b is

$$\pi_b(\boldsymbol{\theta}) = \Pr_{\boldsymbol{\theta}}(Y \in I_b) = F(s_{b+1} \mid \boldsymbol{\theta}) - F(s_b \mid \boldsymbol{\theta}), \qquad b = 1, \ldots, B-1. \tag{3}$$

Equation (3) makes the role of $F(y \mid \boldsymbol{\theta})$ explicit: it is the model CDF, and each symbolic cell probability is obtained by subtracting the model CDF at adjacent histogram boundaries.

The likelihood induced by grouped continuous observations is standard in the grouped-data literature: once the raw sample is coarsened into non-overlapping bins with fixed boundaries, the resulting count

vector is analysed through cell probabilities obtained by integrating the underlying continuous model over the bins; see, for example, [7, 3, 9, 5, 6]. For histogram-valued symbols in the SDA context, this mathematical formulation was formalized by Beranger (2023) [6].

Proposition 1 (Beranger, 2023[6])
Conditional on the fixed boundary vector $\boldsymbol{s}$, the histogram count vector $\boldsymbol{c}$ has multinomial likelihood

$$L_S(\boldsymbol{\theta}; \boldsymbol{c}, \boldsymbol{s}) = \frac{n!}{\prod_{b=1}^{B-1} c_b!} \prod_{b=1}^{B-1} \pi_b(\boldsymbol{\theta})^{c_b}, \tag{4}$$

where

$$\pi_b(\boldsymbol{\theta}) = F(s_{b+1} \mid \boldsymbol{\theta}) - F(s_b \mid \boldsymbol{\theta}), \qquad b = 1, \ldots, B-1.$$

Hence the symbolic log-likelihood is

$$\ell_{\mathrm{sym}}(\boldsymbol{\theta}) = \sum_{b=1}^{B-1} c_b \log \pi_b(\boldsymbol{\theta}) + C(\boldsymbol{c}), \tag{5}$$

with $C(\boldsymbol{c}) = \log n! - \sum_{b=1}^{B-1} \log c_b!$, which is constant in $\boldsymbol{\theta}$. This is the standard grouped-data likelihood representation used in both classical grouped continuous-data inference and modern symbolic-likelihood modelling [7, 3, 9, 5].

Because the combinatorial term $C(\boldsymbol{c})$ is independent of the unknown parameter $\boldsymbol{\theta}$, the symbolic likelihood is proportional to the parameter-dependent terms as the partition is refined [6]. Therefore, for estimation purposes and to facilitate the asymptotic approximation, we drop $C(\boldsymbol{c})$ and define the working symbolic log-likelihood:

$$\tilde{\ell}_{\mathrm{sym}}(\boldsymbol{\theta}) = \sum_{b=1}^{B-1} c_b \log \pi_b(\boldsymbol{\theta}). \tag{6}$$

The symbolic estimator is thus defined by maximizing this working objective:

$$\hat{\boldsymbol{\theta}}_S = \arg\max_{\boldsymbol{\theta} \in \Theta} \tilde{\ell}_{\mathrm{sym}}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta} \in \Theta} \sum_{b=1}^{B-1} c_b \log[F(s_{b+1} \mid \boldsymbol{\theta}) - F(s_b \mid \boldsymbol{\theta})]\,. \tag{7}$$

Once the histogram has been constructed, each objective evaluation requires only $\mathcal{O}(B)$ operations. If the boundary points are data-driven, for example through empirical quantiles, the same likelihood is interpreted conditionally on the realised partition. This conditional viewpoint is the relevant inferential object once the symbolic representation has been formed [4, 5, 6].

Two implementation aspects deserve explicit comment. First, empty or sparse bins: when a bin contains no observations ($c_b = 0$) it contributes nothing to the working log-likelihood in (6); under equal-probability binning on small samples, however, an empty interior bin can leave the model-implied probability weakly identified near that region. The probability floor introduced in Section 2.6 stabilises such boundary cases, and the multinomial weighting automatically down-weights unoccupied bins. Second, data-driven boundaries: when the boundary points are computed from the sample, for example through empirical quantiles, the likelihood is interpreted conditionally on the realised partition, and the approximation of Section 2.5 is understood at the realised Δ_B; resampling or perturbing the boundary construction is a natural robustness extension that we leave for future work.

2.5. Approximation to the full-data log-likelihood

The symbolic likelihood is exact for the grouped data, but the main computational question is how well it approximates the classical objective. The key driver is the maximal bin width

$$\Delta_B = \max_{1\le b\le B-1}(s_{b+1} - s_b),$$

assuming finite effective widths over the region occupied by the data.

The next result formalizes, for the present symbolic-MLE setting, the standard grouped-data approximation principle: when the partition is sufficiently fine, the grouped likelihood differs from the classical likelihood mainly through a discretization term induced by the bin widths. The following finite-partition approximation result, established by Beranger (2023) [6], formalizes the standard grouped-data approximation principle: when the partition is sufficiently fine, the grouped likelihood differs from the classical likelihood mainly through a discretization term induced by the bin widths. Related grouped-data and binned-likelihood arguments appear in [7, 11, 9, 10, 5].

Proposition 2
Suppose that for every $\boldsymbol{\theta} \in \Theta$, the density $f(\cdot \mid \boldsymbol{\theta})$ is positive and continuous on a compact interval containing the sample, and that $\log f(\cdot \mid \boldsymbol{\theta})$ is uniformly continuous in y, uniformly over $\boldsymbol{\theta} \in \Theta$. Then there exists a function $\omega(\Delta_B) \to 0$ as $\Delta_B \to 0$ such that

$$\sup_{\boldsymbol{\theta}\in\Theta} \left| \tilde{\ell}_{\mathrm{sym}}(\boldsymbol{\theta}) - \sum_{i=1}^{n} \log f(Y_i \mid \boldsymbol{\theta}) - K_B \right| \le n\,\omega(\Delta_B), \tag{8}$$

where

$$K_B = \sum_{b=1}^{B-1} c_b \log(s_{b+1} - s_b)$$

does not depend on $\boldsymbol{\theta}$.

Proof sketch. By the mean value theorem for integrals, for each bin I_b and each $\boldsymbol{\theta}$ there exists $\xi_b(\boldsymbol{\theta}) \in I_b$ such that

$$\pi_b(\boldsymbol{\theta}) = \int_{I_b} f(y \mid \boldsymbol{\theta})\, dy = f(\xi_b(\boldsymbol{\theta}) \mid \boldsymbol{\theta})\,(s_{b+1} - s_b).$$

Substituting this into the working symbolic log-likelihood yields

$$\tilde{\ell}_{\mathrm{sym}}(\boldsymbol{\theta}) - K_B = \sum_{b=1}^{B-1} c_b \log f(\xi_b(\boldsymbol{\theta}) \mid \boldsymbol{\theta}).$$

For each observation Y_i lying in I_b, the difference between $\log f(Y_i \mid \boldsymbol{\theta})$ and $\log f(\xi_b(\boldsymbol{\theta}) \mid \boldsymbol{\theta})$ is bounded by the modulus of continuity of $\log f$ over intervals of length at most Δ_B. Summing over all observations yields (8). This argument formalizes, in the present histogram-symbol setting, the usual grouped-data approximation idea that discretization error is controlled by within-bin variation and therefore shrinks as the partition is refined [7, 11, 9, 10]. □

Remark 1
Proposition 2 shows that maximizing the symbolic criterion is asymptotically equivalent to maximizing the classical log-likelihood, up to an additive constant independent of the parameter, whenever the partition is refined so that $\Delta_B \to 0$. This conclusion is consistent with the broader grouped-data and symbolic-data literature, where increasingly fine partitions recover the inferential content of the underlying continuous model [7, 11, 9, 10, 5, 6].

Remark 2
The bound in (8) also indicates how statistical efficiency is lost relative to the classical MLE: the discrepancy is controlled by $n\,\omega(\Delta_B)$, so the symbolic estimator approaches the classical estimator as the maximal bin width shrinks, while for a fixed partition the loss depends on the shape of the density through its modulus of continuity. This is the standard grouped-data efficiency mechanism [7]: refining

the partition (larger $B-1$) purchases back efficiency at a computational cost of $\mathcal{O}(B)$ per evaluation. The relative-RMSE tables in Section 3 quantify exactly this efficiency–resolution trade-off empirically; deriving a closed-form finite-sample efficiency bound for specific families is left as future work.

2.6. Computational complexity and implementation

The symbolic approach has two computational phases. The first is histogram construction, which is a one-time $\mathcal{O}(n)$ preprocessing step. The second is likelihood evaluation during optimisation, which becomes $\mathcal{O}(B)$ per evaluation. This separation matters most when the same model is fitted many times: under multiple initializations, across competing model families, throughout Monte Carlo studies, or along sliding windows in a stream.

Numerical stability requires care when a cell probability is extremely small. To avoid ambiguity with the stopping tolerance used later in Algorithm 1, let ε_π denote the probability floor. In implementation, we evaluate

$$\widetilde{\pi}_b(\boldsymbol{\theta}) = \max\{\pi_b(\boldsymbol{\theta}), \varepsilon_\pi\},$$

with $\varepsilon_\pi = 10^{-12}$ in all simulations and in the public-data illustration, and optimize the stabilized objective

$$\widetilde{\ell}_{\mathrm{sym}}(\boldsymbol{\theta}) = \sum_{b=1}^{B-1} c_b \log \widetilde{\pi}_b(\boldsymbol{\theta}).$$

The value 10^{-12} was chosen as a fixed machine-safe threshold because it is small enough to be negligible for practically relevant bin probabilities, yet large enough to prevent numerical issues caused by catastrophic cancellation or floating-point underflow in $F(s_{b+1} \mid \boldsymbol{\theta}) - F(s_b \mid \boldsymbol{\theta})$, which can yield exactly zero and result in the evaluation of $\log(0)$. In other words, it acts as a conservative numerical safeguard rather than a substantive tuning parameter. We therefore keep it fixed for reproducibility in the reported results. In other applications, users may change this floor after sensitivity checks, but the replacement should remain sufficiently small so that it does not materially alter the likelihood surface away from the boundary of the parameter space. We emphasise that the floor interacts with the empty-bin discussion of Section 2.4: because it only binds when a model-implied bin probability underflows, it does not affect the optimiser in the interior of the likelihood surface, and in the reported configurations no fitted parameters lay at the floor boundary. A systematic sensitivity study over alternative floors (for example, 10^{-14} versus 10^{-10}) remains a worthwhile robustness check for applications with extreme bin sparsity.

For constrained parameters, optimisation is conveniently carried out on an unconstrained scale. For example, a positive scale parameter σ or a rate parameter λ may be represented through an exponential transform of the form $\theta = \exp(\eta)$, where $\eta \in \mathbb{R}$. In window-wise applications, warm starts from the previous fitted value can materially reduce the number of iterations and improve numerical stability.

The choice of binning strategy affects both approximation quality and operational simplicity. Equal-width (EW) bins are easy to interpret and simple to update online. Equal-probability (QU) bins devote more resolution to high-density regions and can substantially reduce approximation error for skewed or heavy-tailed models. In that sense, the binning problem is not merely a display choice but a statistical design choice for the symbolic representation itself [5, 6, 8]. The simulation benchmark in Section 3 therefore compares both strategies directly.

2.7. Streaming update mechanism

Suppose that a moving window of width w is used over a stream. If the boundary points are fixed in advance, then when one observation leaves the window and another enters, only two bin counts need to be updated. Thus the count update is $\mathcal{O}(1)$, while the remaining cost lies in the symbolic optimisation. This makes the approach attractive for near-real-time telemetry and online physiological monitoring.

Algorithm 1 Sliding-Window Symbolic Maximum Likelihood

Require: Parametric family $f(\cdot \mid \boldsymbol{\theta})$, window width w, step size r, a binning scheme with B boundary points (hence $B-1$ bins), and stopping tolerance ϵ_{stop}
Ensure: Sequence of runtimes, fitted parameters, and monitoring scores
Build the initial histogram for the first window of width w
Compute the initial symbolic MLE $\boldsymbol{\theta}^{(0)}$ by maximizing Eq. (7)
Record runtime, fitted parameters, and monitoring scores for the initial window
Set window index $t \leftarrow 1$
while new observations arrive in the stream **do**
    Move the window forward by r observations
    Update the histogram counts $\boldsymbol{c}$
    Recompute the symbolic MLE $\boldsymbol{\theta}^{(t)}$ by maximizing Eq. (7) ▷ Optionally use $\boldsymbol{\theta}^{(t-1)}$ as a warm start
    Compute parameter discrepancy metric $\delta \leftarrow \|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\|$
    if $\delta \le \epsilon_{\text{stop}}$ **then** ▷ Check if the stopping tolerance is achieved
        break ▷ Terminate the stream estimation loop early
    end if
    Record runtime, fitted parameters, and monitoring score of interest
    $t \leftarrow t+1$
end while

3. Simulation study

3.1. Design

The simulation study quantifies the trade-off between statistical fidelity and computational gain across five continuous families: the Exponential distribution $Y \sim \text{Exp}(\lambda)$ with true value $\lambda = 1.2$; the Normal distribution $Y \sim \mathcal{N}(\mu, \sigma^2)$ with $(\mu, \sigma) = (0, 1)$; the Gamma distribution $Y \sim \text{Gamma}(\alpha, \beta)$ with $(\alpha, \beta) = (2, 1)$, where β denotes the rate; the Weibull distribution $Y \sim \text{Weibull}(k, \lambda)$ with $(k, \lambda) = (1.5, 1)$; and the Lognormal distribution $Y \sim \text{Lognormal}(\mu, \sigma^2)$ with $(\mu, \sigma) = (0, 0.75)$. For each family we consider $n \in \{10^4, 10^5, 10^6\}$ and $B - 1 \in \{10, 20, 50, 100, 200, 500\}$. Equivalently, the number of boundary points is $B \in \{11, 21, 51, 101, 201, 501\}$.

To avoid ambiguity, the model CDF $F(y \mid \boldsymbol{\theta})$ used in Eq. (3) is, for each fitted family, the corresponding parametric CDF: for the Exponential model, $F(y \mid \lambda) = 1 - \exp(-\lambda y)$ for $y \ge 0$; for the Normal model, $F(y \mid \mu, \sigma) = \Phi((y-\mu)/\sigma)$; for the Gamma model, $F(y \mid \alpha, \beta) = \gamma(\alpha, \beta y)/\Gamma(\alpha)$ for $y \ge 0$, where $\gamma(\cdot, \cdot)$ is the lower incomplete gamma function; for the Weibull model, $F(y \mid k, \lambda) = 1 - \exp\{-(y/\lambda)^k\}$ for $y \ge 0$; and for the Lognormal model, $F(y \mid \mu, \sigma) = \Phi((\log y - \mu)/\sigma)$ for $y > 0$.

Two binning strategies are benchmarked: equal-width (EW) bins and equal-probability (QU) bins based on empirical quantiles from each sample. For every configuration, $R = 1000$ Monte Carlo replications are run. The design intentionally spans very coarse to moderately fine symbolic resolutions so that the paper can identify where the accuracy–runtime trade-off becomes practically favourable.

The simulation settings follow the replication script exactly. All experiments are run in `R` with a fixed seed of 20260101. For each replication, a fresh sample is generated independently from the target family. Equal-width histograms are constructed over the sample range using evenly spaced cut points, whereas equal-probability histograms use empirical quantiles of the same sample. To ensure that all observations are captured at the extremes, the first and last boundaries are expanded by 10^{-8}. Classical and symbolic likelihoods are optimized with the same numerical backend, namely unconstrained BFGS optimization with exponential reparameterizations for positive parameters and a maximum iteration cap of 1000. For the symbolic estimator, the recorded runtime includes both histogram construction and optimization, so

the reported speed-up reflects the total computational cost of the symbolic workflow rather than only the grouped-likelihood evaluation.

3.2. Estimators compared and performance criteria

For each replication and each model family, we compute (1) the classical MLE, obtained by maximizing (1); (2) the symbolic MLE with equal-width bins (Sym-EW); and (3) the symbolic MLE with equal-probability bins (Sym-QU). All optimisations use the same numerical backend, the same unconstrained parameterisations, and the same maximum-iteration setting. The comparison is therefore between data representations rather than between different optimisers.

The simulation records the following metrics for each parameter component: (1) root mean squared error (RMSE); (2) relative RMSE with respect to the classical MLE; (3) average runtime; (4) speed-up factor defined as

$$\text{Speed-up} = \frac{\text{classical runtime}}{\text{symbolic runtime}};$$

and (5) optimiser convergence frequency (Conv.), defined as the percentage of replications for which the optimisation routine reports successful convergence. Because the negative log-likelihood gap was not central to the inferential or computational comparison, it is not reported in the revised manuscript.

3.3. Main empirical findings

Tables 2–5 reveal four consistent patterns. First, as illustrated in Figure 1, the symbolic MLE becomes more accurate as the number of bins $B-1$ increases. This behaviour is the finite-sample manifestation of Proposition 2: as the maximal bin width decreases, the grouped criterion becomes a closer surrogate to the classical log-likelihood. At $n = 10^6$ and $B - 1 = 500$, the relative RMSE of the equal-width symbolic estimator is 1.000 for the Exponential model, 1.000 for the Normal model, 1.001 for the Weibull model, 1.002 for the Lognormal model, and 1.042 for the Gamma model. The pattern is especially compelling because it is monotone or near-monotone across most families and resolutions, which is exactly the behaviour expected from likelihoods built on increasingly refined histogram symbols [4, 5, 6, 8].

Second, equal-probability binning is especially valuable for skewed models when the symbolic resolution is coarse. The clearest example is the Lognormal model at $n = 10^6$ and $B - 1 = 10$, where the relative RMSE drops from 9.834 under equal-width binning to 1.065 under equal-probability binning. A similar, though more moderate, gain appears for the Gamma family at coarse resolution. Statistically, this is intuitive: equal-probability partitions allocate more effective resolution to regions with high mass, thereby reducing information loss where the likelihood is most sensitive to local shape. This interpretation is consistent with the symbolic-data literature, which emphasizes that the inferential quality of a symbol depends materially on how the underlying observations are aggregated [5, 6, 8].

Third, as illustrated by the runtime scaling profiles in Figure 2, the computational gains become substantial as the sample size grows. For $n = 10^6$, the equal-width symbolic estimator yields speed-ups between 2.64-fold and 90.05-fold across the reported settings. The strongest gains occur for Gamma and Weibull models, where classical optimisation is intrinsically more expensive than for the Exponential, Normal, and Lognormal cases. This ordering reflects model-specific computational burden, not sample size alone. In the classical likelihood, Gamma and Weibull fitting involves repeated evaluation of more complex likelihood components and a more demanding optimisation landscape for multiple positive parameters, whereas the Exponential model is especially simple and the Normal model is comparatively well behaved. Consequently, the per-evaluation and per-iteration costs are larger for Gamma and Weibull, which inflates *Time (Classical)* in Tables 4–5. This is also why the symbolic speed-up is most pronounced for those families: once the data have been compressed, the symbolic objective avoids traversing all n observations while retaining the same model-specific structure at a much lower dimensional cost.

Fourth, numerical stability is strong across the full design. Across all 180 factorial configurations (five families, three sample sizes, and six resolutions), the classical MLE, equal-width symbolic MLE, and equal-probability symbolic MLE all achieved a 100% convergence rate. This uniform convergence is expected

by construction: the working symbolic objective is smooth and well-behaved on the unconstrained reparameterised scale, it is maximised by BFGS with finite-difference gradients under a generous iteration cap, and warm starting is used in the streaming setting. Importantly, convergence should not be confused with accuracy: even in the most challenging configurations (e.g., coarse equal-width bins on the Lognormal model, relative RMSE up to 9.834 at $B-1=10$), the optimiser successfully converged to the maximiser of the grouped objective. The error lies in the objective's discretisation, not in the optimisation. We therefore advise readers to interpret this perfect convergence record jointly with the accuracy tables, and we note that more adversarial settings (e.g., extremely sparse bins or near-boundary optima) could expose convergence failures and are recommended as stress tests for future work. Taken together, these results suggest a practical rule of thumb. For light-tailed or roughly symmetric data, equal-width binning with $B-1=100$ often provides an effective default. For strongly skewed models, either a finer equal-width histogram or an equal-probability partition is preferable when only a small number of bins is operationally feasible. Regarding the uniformly reported 100% convergence, two points merit emphasis. First, convergence is expected by construction: the working symbolic objective is smooth and well behaved on the unconstrained reparameterised scale, it is maximised by BFGS with finite-difference gradients under a generous iteration cap, and warm starting is used in the streaming setting. Second, convergence should not be confused with accuracy: even in the most challenging configuration (coarse equal-width bins on the Lognormal model, relative RMSE up to 9.834 at $B-1=10$), the optimiser converged to the maximiser of the grouped objective, the error lies in the objective itself, not in the optimisation. The convergence tables should therefore be read jointly with the accuracy tables, and more adversarial settings (for example, extremely sparse bins or near-boundary optima) are recommended as stress tests for future work.

3.4. Simulation outputs

Taken together, Tables 2–5 show a clear and practically useful speed–accuracy trade-off. In the largest-sample setting, equal-width symbolic MLE yields speed-ups ranging from 2.64-fold to 90.05-fold. Accuracy improves rapidly as the number of bins $B-1$ increases, although the rate of improvement depends on the model family. For the Exponential, Normal, Lognormal, and Weibull models, the equal-width symbolic estimator is already very close to the classical MLE by approximately $B-1=100$ or $B-1=200$. The Gamma model is more demanding, but the approximation error still declines steadily as the symbolic resolution increases. Equal-probability binning is especially helpful for strongly skewed models at coarse resolutions, whereas for the Normal model the difference between the two symbolic schemes becomes negligible once $B-1\geq 50$.

These findings support a practical recommendation. If the target distribution is approximately symmetric and the primary goal is acceleration, equal-width binning with $B-1=100$ provides a sensible default. If the data are strongly right-skewed and only a small or moderate number of bins is feasible, equal-probability binning is preferable because it preserves substantially more inferential accuracy at coarse symbolic resolutions.

4. Public-data illustration

This section illustrates the practical use of symbolic MLE for window-wise estimation on a public physiological data stream derived from the MIMIC-III Waveform Database. The aim is computational validation in a realistic monitoring setting rather than substantive clinical interpretation. Briefly, the observed heart-rate numerics span 52–227 beats per minute (median 161), a range consistent with bedside-monitor recordings from intensive-care patients that includes resting values as well as tachycardic episodes; this wide dynamic range, together with occasional missing values and monitoring artefacts, is precisely the setting in which robust streaming refitting is operationally relevant. All analyses use de-identified

Table 2. Simulation accuracy summary for classical MLE and symbolic MLE with equal-width (Sym-EW) and equal-probability (Sym-QU) bins (Exponential, Lognormal, and Normal).

Model	n	Bins ($B-1$)	RMSE (Classical)	RMSE (Sym-EW)	RMSE (Sym-QU)	Rel. RMSE (EW)	Rel. RMSE (QU)
exp	10,000	10	0.0092	0.0097	0.0099	1.054	1.078
exp	10,000	20	0.0102	0.0103	0.0105	1.015	1.031
exp	10,000	50	0.0094	0.0094	0.0096	1.001	1.020
exp	10,000	100	0.0092	0.0092	0.0093	1.002	1.004
exp	10,000	200	0.0097	0.0097	0.0098	1.000	1.005
exp	10,000	500	0.0095	0.0095	0.0096	1.000	1.002
exp	100,000	10	0.0030	0.0032	0.0032	1.075	1.071
exp	100,000	20	0.0030	0.0031	0.0031	1.029	1.033
exp	100,000	50	0.0029	0.0029	0.0030	1.001	1.006
exp	100,000	100	0.0031	0.0031	0.0031	1.000	1.005
exp	100,000	200	0.0031	0.0031	0.0031	1.000	1.001
exp	100,000	500	0.0029	0.0029	0.0029	1.000	1.001
exp	1,000,000	10	0.0010	0.0011	0.0010	1.074	1.052
exp	1,000,000	20	0.0010	0.0010	0.0010	1.012	1.014
exp	1,000,000	50	0.0009	0.0009	0.0009	1.001	1.002
exp	1,000,000	100	0.0010	0.0010	0.0010	1.001	1.000
exp	1,000,000	200	0.0010	0.0010	0.0010	1.000	1.000
exp	1,000,000	500	0.0010	0.0010	0.0010	1.000	1.001
lognormal	10,000	10	0.0057	0.0139	0.0061	2.434	1.065
lognormal	10,000	20	0.0057	0.0078	0.0060	1.361	1.049
lognormal	10,000	50	0.0058	0.0062	0.0059	1.072	1.021
lognormal	10,000	100	0.0056	0.0057	0.0057	1.017	1.005
lognormal	10,000	200	0.0058	0.0058	0.0058	0.999	1.002
lognormal	10,000	500	0.0059	0.0059	0.0059	1.001	1.001
lognormal	100,000	10	0.0018	0.0084	0.0020	4.694	1.101
lognormal	100,000	20	0.0018	0.0033	0.0018	1.854	1.035
lognormal	100,000	50	0.0018	0.0021	0.0019	1.136	1.007
lognormal	100,000	100	0.0018	0.0019	0.0018	1.042	1.005
lognormal	100,000	200	0.0018	0.0018	0.0018	1.006	1.000
lognormal	100,000	500	0.0018	0.0018	0.0018	1.001	1.002
lognormal	1,000,000	10	0.0006	0.0057	0.0006	9.834	1.065
lognormal	1,000,000	20	0.0006	0.0016	0.0006	2.830	1.030
lognormal	1,000,000	50	0.0006	0.0007	0.0006	1.216	1.005
lognormal	1,000,000	100	0.0006	0.0006	0.0006	1.059	1.008
lognormal	1,000,000	200	0.0006	0.0006	0.0006	1.014	1.001
lognormal	1,000,000	500	0.0006	0.0006	0.0006	1.002	1.003
normal	10,000	10	0.0078	0.0081	0.0085	1.033	1.090
normal	10,000	20	0.0076	0.0076	0.0079	1.011	1.041
normal	10,000	50	0.0076	0.0076	0.0077	1.002	1.013
normal	10,000	100	0.0074	0.0074	0.0075	1.000	1.013
normal	10,000	200	0.0074	0.0074	0.0075	1.000	1.004
normal	10,000	500	0.0076	0.0076	0.0076	1.000	1.002
normal	100,000	10	0.0024	0.0025	0.0026	1.050	1.086
normal	100,000	20	0.0024	0.0025	0.0025	1.007	1.035
normal	100,000	50	0.0023	0.0024	0.0024	1.001	1.009
normal	100,000	100	0.0025	0.0025	0.0025	1.000	1.011
normal	100,000	200	0.0024	0.0024	0.0025	1.001	1.002
normal	100,000	500	0.0023	0.0023	0.0023	1.000	1.000
normal	1,000,000	10	0.0007	0.0008	0.0008	1.050	1.090
normal	1,000,000	20	0.0008	0.0008	0.0008	1.009	1.042
normal	1,000,000	50	0.0007	0.0008	0.0008	1.002	1.011
normal	1,000,000	100	0.0008	0.0008	0.0008	1.000	1.004
normal	1,000,000	200	0.0008	0.0008	0.0008	1.000	1.004
normal	1,000,000	500	0.0008	0.0008	0.0008	1.000	1.001

public data, and the replication workflow operates on a dataset extracted from the original public source [13, 14].

For each moving window, candidate parametric families are fitted by both classical and symbolic MLE, and the workflow records fitted parameters, likelihood values, model preference, and runtime. In this univariate illustration, the principal comparison is between Normal and Gamma working models.

4.1. Data source and preprocessing

The empirical illustration uses heart-rate numerics derived from the MIMIC-III Waveform Database, a PhysioNet resource containing waveform and bedside-monitor numerics collected in intensive-care settings [13, 14]. The analysis assumes that the user has prepared a univariate heart-rate series dataset as published on GitHub repository at https://github.com/HakiimJ/fast-scalable-symbolic-mle. Missing observations are removed, timestamps are ordered, and only the measurement column is used for the present univariate analysis. Because the purpose of this section is methodological rather than clinical, preprocessing is intentionally kept simple and transparent.

Table 3. Simulation accuracy summary for classical MLE and symbolic MLE with equal-width (Sym-EW) and equal-probability (Sym-QU) bins (Gamma and Weibull).

Model	n	Bins ($B-1$)	RMSE (Classical)	RMSE (Sym-EW)	RMSE (Sym-QU)	Rel. RMSE (EW)	Rel. RMSE (QU)
gamma	10,000	10	0.0173	0.0245	0.0205	1.419	1.187
gamma	10,000	20	0.0180	0.0202	0.0197	1.124	1.095
gamma	10,000	50	0.0175	0.0179	0.0180	1.021	1.028
gamma	10,000	100	0.0174	0.0176	0.0177	1.017	1.020
gamma	10,000	200	0.0174	0.0175	0.0175	1.004	1.005
gamma	10,000	500	0.0177	0.0178	0.0179	1.002	1.010
gamma	100,000	10	0.0057	0.0088	0.0065	1.545	1.150
gamma	100,000	20	0.0055	0.0065	0.0060	1.191	1.089
gamma	100,000	50	0.0055	0.0058	0.0057	1.040	1.036
gamma	100,000	100	0.0056	0.0057	0.0058	1.004	1.019
gamma	100,000	200	0.0056	0.0056	0.0057	1.001	1.013
gamma	100,000	500	0.0056	0.0056	0.0056	1.001	1.005
gamma	1,000,000	10	0.0017	0.0030	0.0020	1.717	1.122
gamma	1,000,000	20	0.0018	0.0022	0.0019	1.225	1.079
gamma	1,000,000	50	0.0018	0.0019	0.0018	1.065	1.040
gamma	1,000,000	100	0.0018	0.0018	0.0018	1.037	1.032
gamma	1,000,000	200	0.0018	0.0018	0.0018	1.040	1.052
gamma	1,000,000	500	0.0018	0.0019	0.0019	1.042	1.054
weibull	10,000	10	0.0083	0.0093	0.0095	1.120	1.145
weibull	10,000	20	0.0085	0.0088	0.0090	1.042	1.059
weibull	10,000	50	0.0084	0.0085	0.0087	1.011	1.036
weibull	10,000	100	0.0084	0.0084	0.0085	1.002	1.014
weibull	10,000	200	0.0087	0.0088	0.0088	1.004	1.007
weibull	10,000	500	0.0084	0.0084	0.0084	1.001	1.002
weibull	100,000	10	0.0026	0.0030	0.0031	1.148	1.160
weibull	100,000	20	0.0026	0.0027	0.0028	1.036	1.073
weibull	100,000	50	0.0027	0.0027	0.0027	1.006	1.022
weibull	100,000	100	0.0027	0.0027	0.0027	1.001	1.016
weibull	100,000	200	0.0026	0.0026	0.0026	1.002	0.998
weibull	100,000	500	0.0026	0.0026	0.0027	1.000	1.005
weibull	1,000,000	10	0.0008	0.0010	0.0009	1.202	1.137
weibull	1,000,000	20	0.0009	0.0009	0.0009	1.062	1.049
weibull	1,000,000	50	0.0008	0.0008	0.0009	1.018	1.021
weibull	1,000,000	100	0.0008	0.0009	0.0009	1.004	1.013
weibull	1,000,000	200	0.0008	0.0008	0.0008	1.001	1.000
weibull	1,000,000	500	0.0009	0.0009	0.0009	1.001	1.007

4.2. Window-wise symbolic estimation pipeline

To align the real-data analysis with the likelihood-based objective of the paper, each moving window is analysed by fitting candidate parametric families, using both classical and symbolic MLE. For each window, the workflow records the fitted parameters, the attained negative log-likelihood, the selected working family, and the runtime. This enables a direct comparison between inferential agreement and computational gain in an estimation setting.

4.3. Heart-rate numerics from MIMIC-III

The filtered MIMIC-III heart-rate stream contains 1,912,844 observations (Min: 52, Q1: 154, Median: 161, Q3: 168, Max: 227). For the empirical illustration, the replication script uses a moving-window length of $w = 10,000$ observations, step size $r = 500$, and a default symbolic resolution of $B - 1 = 20$ bins, corresponding to $B = 21$ boundary points. This setup represents short-horizon refitting along the stream. The primary estimands are the local location and scale parameters under the selected working model, together with the runtime comparison between classical and symbolic MLE.

For the MIMIC-III stream, the Normal model is preferred in 2,383 of the 3,806 windows under the classical negative log-likelihood and in 2,440 windows under the symbolic criterion. To summarise the per-window computational cost and agreement for the MIMIC-III stream, the classical MLE required a median of 0.005 seconds per window, whereas the symbolic MLE required a median of 0.001 seconds. This yields a median speed-up of 5.000-fold (with an interquartile range of 1.000) across the 3,806 evaluated windows. Despite this substantial acceleration, the agreement between the estimators remains exceptionally tight, with a median absolute parameter difference of 0.193 and a maximum absolute difference of 3.874. As illustrated in Figure 3, the agreement between classical and symbolic parameter estimates is exceptionally close over the entire sequence of windows: the median absolute parameter difference is 0.193, the maximum absolute difference is 3.874, and the trajectories reflect strong linear associations, with correlations between classical and symbolic estimates reaching 0.9998 for μ and 0.9719 for σ. These findings indicate that the

Table 4. Average runtime and speed-up for classical MLE and symbolic MLE with equal-width (Sym-EW) and equal-probability (Sym-QU) bins (Exponential, Lognormal, and Normal).

Model	n	Bins ($B-1$)	Time (Classical)	Time (Sym-EW)	Time (Sym-QU)	Speed-up (EW)	Speed-up (QU)
exp	10,000	10	0.0023	0.0014	0.0023	1.64	1.05
exp	10,000	20	0.0024	0.0015	0.0025	1.60	1.06
exp	10,000	50	0.0021	0.0015	0.0026	1.40	0.89
exp	10,000	100	0.0025	0.0017	0.0028	1.97	0.96
exp	10,000	200	0.0022	0.0018	0.0033	1.51	0.73
exp	10,000	500	0.0024	0.0022	0.0043	1.20	0.60
exp	100,000	10	0.0221	0.0059	0.0183	4.06	1.25
exp	100,000	20	0.0223	0.0070	0.0187	3.61	1.21
exp	100,000	50	0.0227	0.0086	0.0191	2.79	1.21
exp	100,000	100	0.0219	0.0091	0.0199	2.51	1.12
exp	100,000	200	0.0218	0.0102	0.0211	2.27	1.06
exp	100,000	500	0.0221	0.0105	0.0219	2.32	1.03
exp	1,000,000	10	0.2468	0.0556	0.1988	4.61	1.26
exp	1,000,000	20	0.2434	0.0613	0.2045	4.08	1.21
exp	1,000,000	50	0.2430	0.0821	0.2102	3.02	1.18
exp	1,000,000	100	0.2511	0.0865	0.2178	2.94	1.18
exp	1,000,000	200	0.2429	0.0937	0.2272	2.64	1.09
exp	1,000,000	500	0.2454	0.0905	0.2280	2.78	1.10
lognormal	10,000	10	0.0038	0.0016	0.0028	2.38	1.56
lognormal	10,000	20	0.0039	0.0019	0.0028	2.05	1.57
lognormal	10,000	50	0.0037	0.0020	0.0032	1.85	1.35
lognormal	10,000	100	0.0037	0.0023	0.0038	1.84	1.23
lognormal	10,000	200	0.0039	0.0035	0.0040	1.54	1.09
lognormal	10,000	500	0.0037	0.0041	0.0052	1.05	0.77
lognormal	100,000	10	0.0414	0.0063	0.0189	7.26	2.25
lognormal	100,000	20	0.0417	0.0068	0.0194	6.62	2.20
lognormal	100,000	50	0.0417	0.0094	0.0195	4.65	2.17
lognormal	100,000	100	0.0415	0.0102	0.0203	4.30	2.08
lognormal	100,000	200	0.0416	0.0109	0.0214	3.98	1.96
lognormal	100,000	500	0.0415	0.0111	0.0220	3.96	1.90
lognormal	1,000,000	10	0.4223	0.0561	0.1985	7.83	2.17
lognormal	1,000,000	20	0.4199	0.0612	0.2042	7.15	2.09
lognormal	1,000,000	50	0.4223	0.0843	0.2075	5.14	2.07
lognormal	1,000,000	100	0.4193	0.0879	0.2142	4.91	1.99
lognormal	1,000,000	200	0.4226	0.0930	0.2227	4.61	1.93
lognormal	1,000,000	500	0.4245	0.0902	0.2252	4.83	1.92
normal	10,000	10	0.0028	0.0017	0.0027	1.65	1.14
normal	10,000	20	0.0027	0.0018	0.0031	1.50	1.06
normal	10,000	50	0.0027	0.0018	0.0034	1.50	0.95
normal	10,000	100	0.0026	0.0018	0.0036	1.74	0.87
normal	10,000	200	0.0027	0.0023	0.0042	1.59	0.75
normal	10,000	500	0.0026	0.0028	0.0055	1.11	0.54
normal	100,000	10	0.0277	0.0078	0.0189	3.84	1.50
normal	100,000	20	0.0287	0.0086	0.0191	3.72	1.53
normal	100,000	50	0.0289	0.0082	0.0192	3.83	1.52
normal	100,000	100	0.0293	0.0089	0.0204	3.45	1.46
normal	100,000	200	0.0266	0.0098	0.0214	2.84	1.27
normal	100,000	500	0.0264	0.0106	0.0226	2.62	1.19
normal	1,000,000	10	0.2704	0.0697	0.2018	3.96	1.37
normal	1,000,000	20	0.2701	0.0749	0.2087	3.70	1.32
normal	1,000,000	50	0.2716	0.0754	0.2156	3.68	1.30
normal	1,000,000	100	0.2726	0.0837	0.2227	3.29	1.25
normal	1,000,000	200	0.2691	0.0916	0.2306	2.99	1.20
normal	1,000,000	500	0.2715	0.0893	0.2295	3.09	1.21

symbolic approximation tracks local distributional changes closely while delivering a clear computational advantage.

4.4. Real-data runtime summary

Sensitivity checks over the number of bins and window size support the same overall pattern seen in the simulation study (Table 6). The preferred working model remains Normal across the sensitivity grid. As the moving-window size increases, the computational gain of symbolic MLE becomes larger because the classical likelihood incurs a higher evaluation cost, whereas the symbolic likelihood depends primarily on the histogram resolution. At the same time, increasing the number of bins reduces approximation discrepancy by preserving more local distributional detail. The empirical illustration therefore reinforces the main message of the paper: symbolic MLE is most attractive when likelihood-based fitting must be executed many times over large or overlapping windows. The sensitivity grid also offers practical guidance on the hyperparameters of the streaming pipeline. The window width w should be large enough for the classical cost $\mathcal{O}(w)$ per evaluation to dominate the symbolic cost $\mathcal{O}(B)$: in Table 6 the median speed-up

Table 5. Average runtime and speed-up for classical MLE and symbolic MLE with equal-width (Sym-EW) and equal-probability (Sym-QU) bins (Gamma and Weibull).

Model	n	Bins ($B-1$)	Time (Classical)	Time (Sym-EW)	Time (Sym-QU)	Speed-up (EW)	Speed-up (QU)
gamma	10,000	10	0.0610	0.0022	0.0032	27.73	21.42
gamma	10,000	20	0.0599	0.0026	0.0038	26.26	17.86
gamma	10,000	50	0.0611	0.0036	0.0053	18.85	13.27
gamma	10,000	100	0.0595	0.0051	0.0068	12.68	9.13
gamma	10,000	200	0.0598	0.0080	0.0105	7.95	5.95
gamma	10,000	500	0.0599	0.0166	0.0211	3.99	2.99
gamma	100,000	10	0.7062	0.0066	0.0185	115.47	38.72
gamma	100,000	20	0.6996	0.0083	0.0195	91.23	36.40
gamma	100,000	50	0.7149	0.0104	0.0214	72.51	34.21
gamma	100,000	100	0.7015	0.0125	0.0244	58.59	29.62
gamma	100,000	200	0.7096	0.0165	0.0282	44.75	25.23
gamma	100,000	500	0.7091	0.0241	0.0388	30.64	18.51
gamma	1,000,000	10	4.9675	0.0569	0.2003	90.05	25.34
gamma	1,000,000	20	5.0517	0.0678	0.2081	76.77	24.79
gamma	1,000,000	50	4.9221	0.0806	0.2109	62.03	23.84
gamma	1,000,000	100	5.1230	0.0881	0.2217	58.94	23.56
gamma	1,000,000	200	5.0018	0.0962	0.2314	52.36	21.95
gamma	1,000,000	500	5.1224	0.0981	0.2360	52.70	21.98
weibull	10,000	10	0.0290	0.0021	0.0028	13.81	11.48
weibull	10,000	20	0.0296	0.0022	0.0031	13.45	11.32
weibull	10,000	50	0.0291	0.0021	0.0030	10.41	9.70
weibull	10,000	100	0.0294	0.0028	0.0037	13.04	8.84
weibull	10,000	200	0.0287	0.0034	0.0047	9.92	6.77
weibull	10,000	500	0.0297	0.0059	0.0069	6.11	4.67
weibull	100,000	10	0.5237	0.0077	0.0186	76.12	29.03
weibull	100,000	20	0.5239	0.0083	0.0194	66.27	27.44
weibull	100,000	50	0.5243	0.0096	0.0202	57.28	26.58
weibull	100,000	100	0.5217	0.0106	0.0217	51.82	24.45
weibull	100,000	200	0.5250	0.0140	0.0236	43.29	22.82
weibull	100,000	500	0.5247	0.0152	0.0273	35.82	19.65
weibull	1,000,000	10	4.9659	0.0588	0.2007	87.35	25.29
weibull	1,000,000	20	4.9668	0.0708	0.2046	71.99	24.71
weibull	1,000,000	50	4.9694	0.0800	0.2106	63.31	24.10
weibull	1,000,000	100	4.9682	0.0856	0.2177	58.95	23.26
weibull	1,000,000	200	4.9667	0.0938	0.2292	53.58	22.19
weibull	1,000,000	500	4.9704	0.0929	0.2332	53.96	21.82

grows from 2-fold at $w = 5{,}000$ to 9-fold at $w = 20{,}000$ for most binning settings, and up to 14-fold at the finest resolution ($B = 20$ boundary points), whereas with too few boundary points the discrepancy is large regardless of w. The step size r controls responsiveness rather than accuracy: smaller steps re-estimate more frequently at proportionally higher total cost, while the per-window fit itself is unaffected. A principled default is therefore to choose the histogram resolution primarily for statistical fidelity (small Δ_B), and to set w and r from the monitoring latency budget.

Table 6. Sensitivity of runtime and estimation discrepancy to the number of bins and window size in the MIMIC-III analysis.

Bins ($B-1$)	Window Size	Avg. Sym. Runtime (s)	Median Speed-up	Median Discrepancy
5	5,000	0.00035	2.00	6.00078
5	10,000	0.00050	5.00	2.80217
5	20,000	0.00046	9.00	2.17438
10	5,000	0.00051	2.00	0.45270
10	10,000	0.00049	5.00	0.38464
10	20,000	0.00046	9.00	0.33318
15	5,000	0.00049	2.00	0.25190
15	10,000	0.00050	5.00	0.22342
15	20,000	0.00047	9.00	0.20203
20	5,000	0.00052	2.00	0.20473
20	10,000	0.00051	5.00	0.19261
20	20,000	0.00076	14.00	0.18419

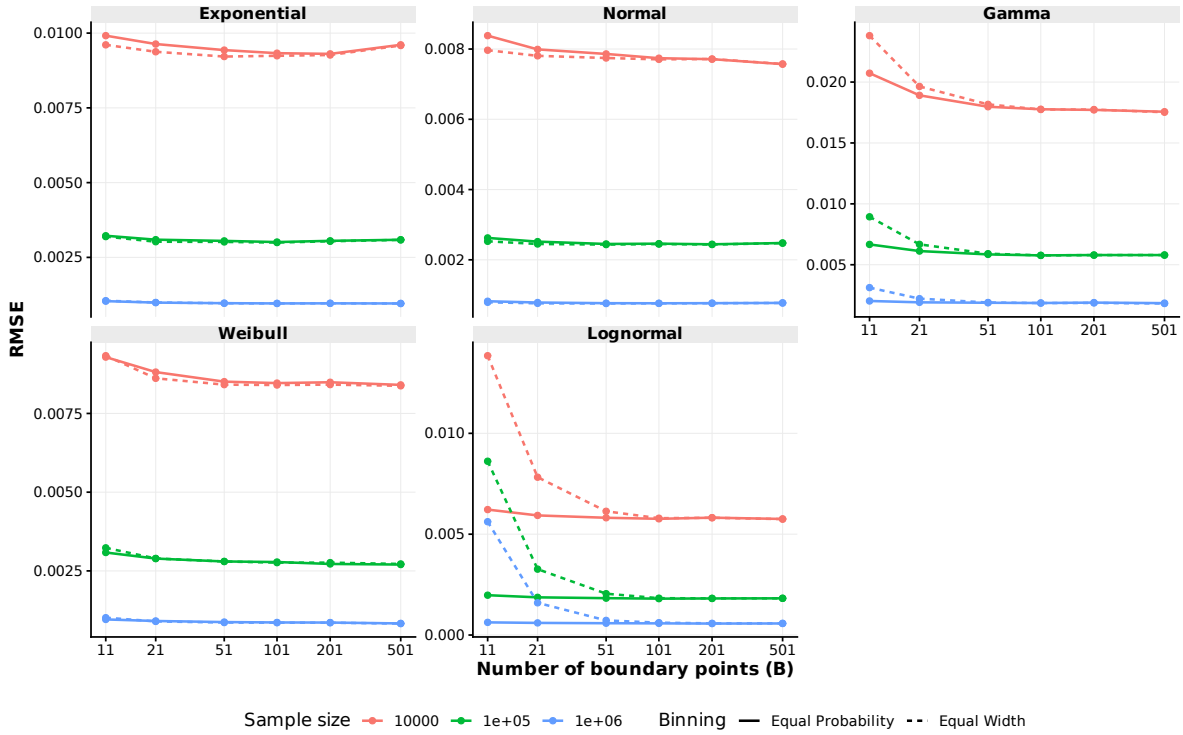


Figure 1. RMSE of symbolic MLE as a function of the number of bin boundaries B, stratified by distribution and sample size.

5. Discussion

The proposed symbolic MLE is best viewed as a likelihood-based compression device for repeated parametric estimation (across different windows, for example). It is not designed to replace classical MLE in every setting. When the sample is small and the model is fitted only once, the one-time cost of building the histogram may offer little practical benefit. The computational gain becomes compelling when n is very large, when several candidate families must be fitted, or when the same optimisation problem must be solved in Monte Carlo studies, resampling exercises, or moving-window monitoring.

The theoretical approximation result in Proposition 2 explains why the method works. The underlying theoretical approximation results formalized by Beranger (2023) [6] explain why the framework works. Under shrinking bin widths, the symbolic objective differs from the classical log-likelihood only by an additive discretisation term that does not depend on the unknown parameter, together with an error that vanishes as the partition is refined. The simulation results show that this asymptotic argument is already useful at finite symbolic resolutions. In the largest-sample setting, the equal-width symbolic estimator attains speed-ups from 2.64-fold to 90.05-fold while preserving near-classical accuracy for sufficiently fine histograms. The empirical evidence also shows that the required symbolic resolution depends materially on the shape of the underlying distribution: light-tailed and approximately symmetric families are well approximated with moderate values of $B - 1$, whereas the Gamma model requires a finer partition and the Lognormal model benefits strongly from equal-probability binning when the number of bins is small.

The results also demonstrate that bin design is not a minor implementation detail. Equal-width bins are operationally attractive because they are simple, interpretable, and convenient for online maintenance. Equal-probability bins, however, can be markedly superior at coarse resolutions for skewed distributions because they allocate more effective resolution to regions where the density is concentrated. In practice,

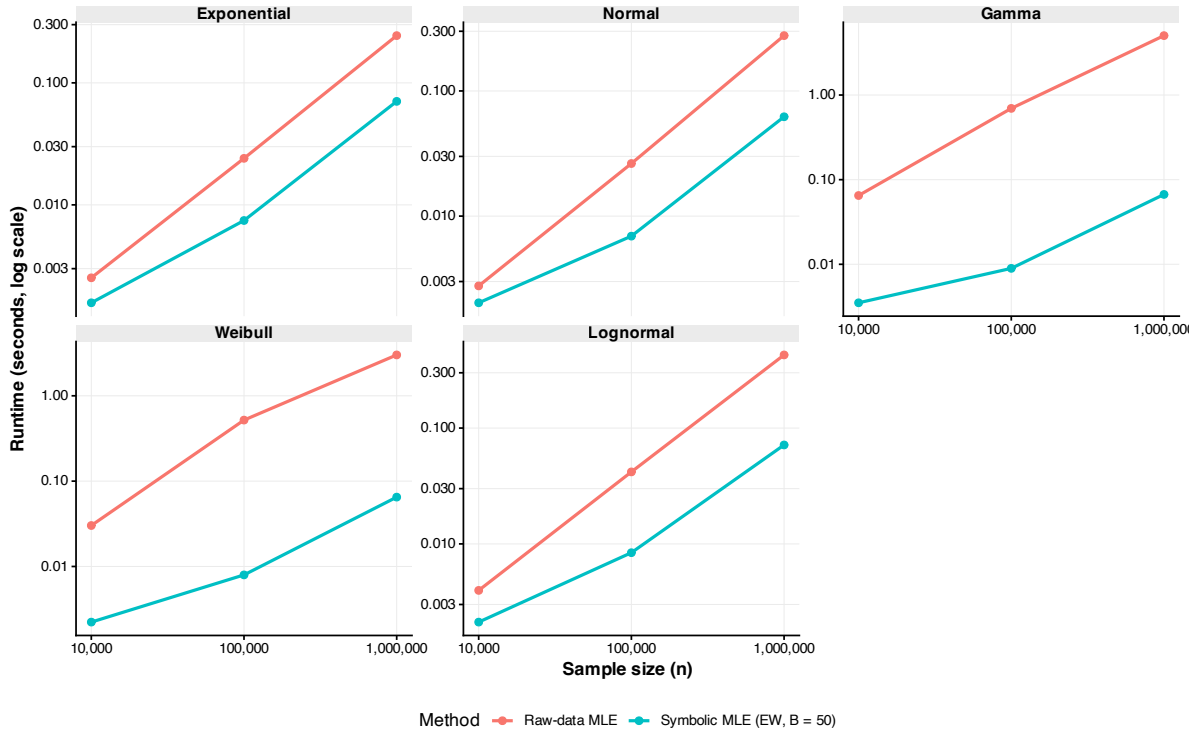


Figure 2. Runtime scaling of classical and symbolic MLE across sample sizes, stratified by distribution.

the choice of partition should therefore be guided jointly by the expected signal shape and the intended deployment context. If the boundary points must remain fixed and cheaply updateable online, equal-width partitions are often preferable. If the analysis is offline and inferential fidelity under skewness is the dominant objective, equal-probability binning deserves priority.

Several limitations should be acknowledged. First, the present manuscript treats only univariate continuous models. Second, the symbolic estimator remains model-based and therefore inherits sensitivity to distributional misspecification. Concretely, the approach is least reliable when the histogram is coarse relative to the complexity of the density: strongly skewed or heavy-tailed families (for example, the Lognormal model, whose equal-width relative RMSE reaches 9.834 at $B - 1 = 10$) need either fine partitions or equal-probability binning; genuinely multimodal densities require enough bins to resolve the modes, since between-mode structure is invisible to a partition that is too coarse; and the one-dimensional construction does not extend trivially to discrete or mixed data, where bin probabilities are not obtained from a smooth CDF. The efficiency argument of Section 2.5 also presumes enough data per bin, so very small samples are a poor setting for symbolic MLE. Third, when boundary points are recomputed too frequently in a truly online setting, maintenance costs may erode part of the computational benefit. Fourth, the paper focuses primarily on point estimation and runtime; standard errors and confidence intervals can nevertheless be pursued within the symbolic framework: the observed information can be approximated by the numerical Hessian of the working symbolic log-likelihood in (6) at the optimum, and profile-likelihood confidence intervals can be constructed directly from (6), since it is a genuine log-likelihood for the histogram counts [7]. The main caveat is that approximation error in the bin probabilities $\pi_b(\boldsymbol{\theta})$ propagates to the curvature, so coverage probabilities should be validated against the classical fit before deployment; likelihood-ratio procedures, model-selection calibration, and a full treatment of uncertainty quantification under symbolic compression remain important topics for future work.

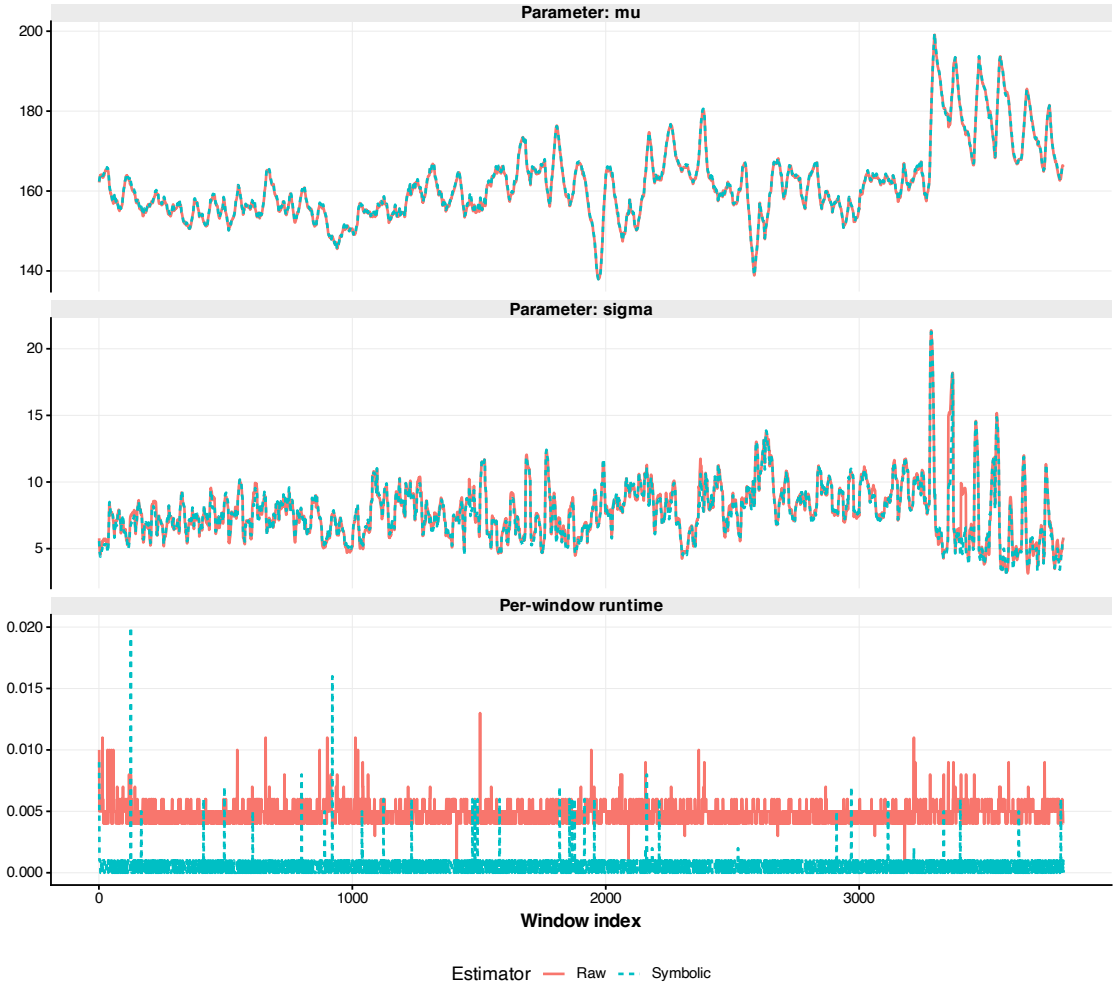


Figure 3. Window-wise parameter trajectories and runtime comparison for the MIMIC-III heart-rate stream.

These limitations suggest several extensions. One natural direction is the development of multivariate histogram-valued symbolic likelihoods, where the computational gains may be even more consequential but the symbolic design problem becomes substantially richer. For multivariate extensions the main challenge is combinatorial: a full multidimensional partition grows exponentially in the number of variables, so practical schemes will likely require low-dimensional marginals combined through composite or copula-style constructions [5, 12]. Extensions involving latent variables require marginalising the latent density over each bin, which generally renders the cell probabilities analytically intractable and would require numerical integration inside the optimisation loop. Finally, for non-stationary streams the sliding-window framework of Algorithm 1 already tracks time-varying parameters through warm-started re-estimation, and the step size governs the responsiveness of the tracker; a fully adaptive scheme that also adjusts the partition online, for example when Δ_B becomes coarse relative to local density variation, is an interesting open problem. Another is adaptive bin construction targeted directly at estimation loss under a computational budget. A third is the integration of symbolic compression with broader approximate-inference strategies, including composite likelihood, surrogate likelihood, and online optimisation frameworks.

It is also useful to position symbolic MLE against the broader landscape of accelerated likelihood inference. Subsampling and mini-batch stochastic-gradient methods reduce cost by processing fewer likelihood terms per update, but they introduce sampling or step-size noise and still target the same n-term objective. Binned and grouped-data methods, including the present approach, instead replace the objective itself with an approximation whose evaluation is $\mathcal{O}(B)$, trading a deterministic and controllable discretisation error for the speed-up [11, 7, 9, 10]. Composite likelihood methods reduce cost by decomposing the joint likelihood into lower-dimensional components [12], which addresses multivariate dependency structures, whereas the present method targets the univariate marginal fit; the histogram-composite construction of Whitaker et al. (2020) [5] is the natural bridge between the two. Modern data-efficient and variational approaches similarly replace exact likelihood evaluations with cheaper surrogates, but often introduce optimiser-dependent bias. Symbolic MLE is therefore best seen as the deterministic, likelihood-preserving member of this family, most attractive when the same univariate fit must be repeated across windows, replicates, or candidate families.

6. Conclusion

This paper presents a computational evaluation of symbolic maximum likelihood estimation for massive univariate continuous data aggregated into histograms, building upon the theoretical framework established by Beranger (2023) [6]. After a one-time compression step, likelihood evaluations depend on the number of histogram boundary points B, rather than on the raw sample size n. The resulting grouped likelihood is exact for the symbolic data and admits a clear approximation link to the classical log-likelihood as the partition is refined.

The simulation benchmark shows that symbolic MLE can provide substantial acceleration with modest accuracy loss once the histogram is sufficiently fine. In the largest-sample setting, equal-width symbolic MLE achieves speed-ups from 2.64-fold to 90.05-fold across the reported configurations. Accuracy improves steadily as the number of bins increases, and the most severe coarse-bin (or wide-bin) errors are concentrated in skewed models, where equal-probability binning is often considerably more reliable. All reported simulation configurations converge successfully, supporting the numerical stability of the proposed implementation.

The public-data illustration further demonstrates that the method is practically useful in window-wise fitting. On the MIMIC-III heart-rate stream, the preferred Normal model achieves a median 5-fold speed-up per window while maintaining close agreement between classical and symbolic estimates. Overall, symbolic MLE offers a practical and principled tool for large-scale simulation, likelihood fitting, and moving-window analysis of data streams.

Author Contributions

Ahmad Hakiim Jamaluddin: Conceptualization, Methodology, Formal analysis, Software, Investigation, Writing – original draft, Writing – review and editing.
Syaiful Anam: Conceptualization, Validation, Writing – review and editing.
Muhammad Jaffri Mohd Nasir: Methodology, Formal analysis, Mathematical derivations, Writing – review and editing.
Andrea Tri Rian Dani: Writing – review and editing.
Nur Chamidah: Writing – review and editing.

Data and Code Availability

All R scripts, generated figures, and tables used to reproduce the simulation study and the public-data illustration are available in the public GitHub repository at https://github.com/HakiimJ/fast-scalable-symbolic-mle. The real-data illustration uses the de-identified MIMIC-III public resource hosted by PhysioNet, subject to the relevant access and credentialing requirements of the data provider [13, 14].

Funding

This research was supported by the Universitas Brawijaya Visiting Lecturer Programme Fund (Grant number: 01236.37/DST/UN10.A0101/B/PJ.00.05.1/2026).

Ethics Statement

This study uses de-identified public data from the MIMIC-III database and does not involve the collection of new identifiable human-participant data by the authors.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgement

The authors acknowledge PhysioNet and the investigators who developed and maintained the MIMIC-III databases. The authors also acknowledge institutional research support provided through the Universitas Brawijaya Visiting Lecturer Programme.

REFERENCES

1. L. Billard and E. Diday, *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Wiley, Chichester, 2006.
2. P. Brito, *Symbolic data analysis: another look at the interaction of data mining and statistics*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 4, no. 4, pp. 281–295, 2014.
3. J. Le-Rademacher and L. Billard, *Likelihood functions and some maximum likelihood estimators for symbolic data*, Journal of Statistical Planning and Inference, vol. 141, no. 4, pp. 1593–1602, 2011.
4. T. Zhang, B. Béranger, and S. A. Sisson, *Likelihood-based inference for interval-valued data in symbolic data analysis*, Scandinavian Journal of Statistics, vol. 47, no. 3, pp. 820–847, 2020.
5. T. Whitaker, B. Béranger, and S. A. Sisson, *Composite likelihood methods for histogram-valued random variables*, Statistics and Computing, vol. 30, no. 5, pp. 1459–1477, 2020.
6. B. Béranger, H. Lin, and S. A. Sisson, *New models for symbolic data analysis*, Advances in Data Analysis and Classification, vol. 17, no. 3, pp. 659–699, 2023.
7. D. F. Heitjan, *Inference from grouped continuous data: A review*, Statistical Science, vol. 4, no. 2, pp. 164–179, 1989.
8. A. H. Jamaluddin, *Optimal Design of Symbolic Data for Statistical Inference in Large Datasets* (Ph.D. Thesis), UNSW Sydney, Sydney, 2025. Available at: https://unsworks.unsw.edu.au/entities/publication/a346f58f-261f-4b30-b383-99dd78ea0aca.
9. P. Lambert and P. H. C. Eilers, *Bayesian density estimation from grouped continuous data*, Computational Statistics & Data Analysis, vol. 53, no. 4, pp. 1388–1399, 2009.
10. S. Rizzi, J. Gampe, and P. H. C. Eilers, *Efficient estimation of smooth distributions from coarsely grouped data*, American Journal of Epidemiology, vol. 182, no. 2, pp. 138–147, 2015.
11. D. W. Scott and S. J. Sheather, *Kernel density estimation with binned data*, Communications in Statistics – Theory and Methods, vol. 14, no. 6, pp. 1353–1359, 1985.

12. C. Varin, N. Reid, and D. Firth, *An overview of composite likelihood methods*, Statistica Sinica, vol. 21, no. 1, pp. 5–42, 2011.
13. B. Moody, G. Moody, M. Villarroel, G. D. Clifford, and I. Silva, *MIMIC-III Waveform Database* (version 1.0), PhysioNet, 2020. doi:10.13026/c2607m.
14. A. E. W. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, *MIMIC-III, a freely accessible critical care database*, Scientific Data, vol. 3, article 160035, 2016.
15. M. Alizadeh, M. N. Khan, M. Rasekhi, and G. H. Hamedani, *A new generalized modified Weibull distribution*, Statistics, Optimization and Information Computing, vol. 9, no. 1, pp. 17–34, 2021. doi:10.19139/soic-2310-5070-1014.
16. L. Benkhelifa, *The Weibull Birnbaum-Saunders distribution and its applications*, Statistics, Optimization and Information Computing, vol. 9, pp. 61–81, 2021. doi:10.19139/soic-2310-5070-887.
17. M. Razmkhah, F. Al-Ma'shumah, and S. Effati, *The k-nearest neighbor classification of histogram- and trapezoid-valued data*, Statistics, Optimization and Information Computing, 2022. doi:10.19139/soic-2310-5070-1451.