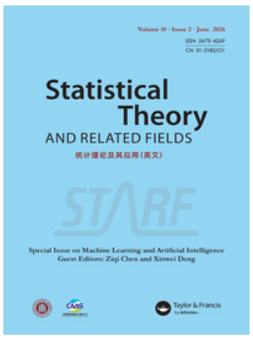

Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: www.tandfonline.com/journals/tstf20

Statistical properties of mean relative entropy and its applications

Lei Huang, Danting Wang & Lanpeng Ji

To cite this article: Lei Huang, Danting Wang & Lanpeng Ji (29 Jun 2026): Statistical properties of mean relative entropy and its applications, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2026.2689052

To link to this article: https://doi.org/10.1080/24754269.2026.2689052

Published online: 29 Jun 2026.

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2026.2689052

Statistical properties of mean relative entropy and its applications

Lei Huang [a], Danting Wang[a] and Lanpeng Ji [b]

[a]Department of Statistics, School of Mathematics, Southwest Jiaotong University, Chengdu, People's Republic of China; [b]Schoolof Mathematics, University of Leeds, Leeds, UK

ABSTRACT

A fundamental concept in information theory is information entropy, which is used to quantify the degree of uncertainty associated with random variables. Building upon this, relative entropy serves as a measure of the discrepancy between two probability distributions and has been widely studied in statistics. The function of relative entropy in the field of parameter estimation has been extensively investigated. This paper extends existing research by deriving minimum mean relative entropy estimators for the parameters of the Gamma, Laplace, and Rayleigh distributions. Furthermore, we introduce the residual mean relative entropy, a novel measure based on mean relative entropy, and apply it to model comparison. To estimate this measure, we employ kernel density estimation and bootstrap methods. Simulation experiments and empirical analysis demonstrate that the proposed residual mean relative entropy provides an effective new criterion for model comparison.

ARTICLE HISTORY
Received 20 August 2025
Accepted 10 June 2026

KEYWORDS
Parameter estimation; mean relative entropy; kernel density estimation; bootstrap; model comparison

1. Introduction

Information entropy is a fundamental concept in information theory and is used to measure the uncertainty of random variables (Shannon, 1948). Information entropy, however, is not limited to the field of information theory. It is also frequently applied in a variety of other domains, such as data analysis, artificial intelligence, and statistics (Cover & Thomas, 1991). In addition, a variety of entropy-related concepts have been derived from information entropy and applied to address a range of statistical problems. Hu et al. (2010) generalized information entropy to clear ordered classification and fuzzy ordered classification problems, and proposed information measures for ordering mutual information and fuzzy ordering mutual information. Namdari and Li (2019) studied a variety of entropy measures, including information entropy, and showed that entropy measures have good predictive capacity for stochastic processes with uncertainty.

In particular, Kullback and Leibler (1951) introduced relative entropy as a measure to compare two probability distributions, which is another significant and frequently applied extension of information entropy in statistics. Arizono and Ohta (1989) introduced an

CONTACT Lei Huang stahl@swjtu.edu.cn Department of Statistics, School of Mathematics, Southwest Jiaotong University, Chengdu 611756, People's Republic of China

extended test method for normal fitting based on relative entropy. Compared with the sample entropy-based test method proposed by Vasicek (1976), this method can be applied not only to complex hypothesis problems, but also to simple hypothesis problems. Mühlenbein and Höns (2005) showed that the principle of minimum relative entropy plays an important role in distribution algorithm estimation. In addition, the role of relative entropy in parameter estimation has been extensively studied. Parsian and Nematollahi (1996) investigated estimators of scale parameters based on the principle of relative entropy minimization and extended them to Bayesian estimation. Gupta and Srivastava (2010) studied the Bayesian estimation method of parameters based on relative entropy, and derived the parameter estimators under a variety of distributions, such as the uniform and Gaussian distributions. Mean relative entropy (MRE), introduced by Zhang (2021), serves as a measure of estimation error. Although mean square error (MSE) is commonly used in parameter estimation to assess the accuracy of estimated parameters, Zhang (2021) pointed out that MRE is invariant under the one-to-one transformation of both parameters and data and performs better than MSE in estimator selection.

However, Zhang (2021) derived minimum MRE estimators only for a limited set of distributions, leaving the results incomplete. Building on this work, this paper extends the theoretical development by obtaining minimum MRE estimators for three commonly used distributions, namely the Gamma, Laplace, and Rayleigh distributions, based on a given sample.

In addition, although Zhang (2021) discussed the theoretical superiority of the minimum MRE estimator, the practical application of MRE was not further examined. Motivated by the fact that residuals are widely used in model diagnostics (Liu et al., 2016), such as in the comparison of residual MSE, this paper introduces an approach that employs the MRE of residuals (rMRE) for model comparison. Unlike MSE, rMRE compares models by measuring the discrepancy between the probability density distributions of the residuals and the true error terms. To obtain these distributions, the density of the residuals is estimated using kernel density estimation (Parzen, 1962; Rosenblatt, 1956), while the density of the error terms is approximated via bootstrap resampling (Efron, 1992), due to the limited number of residual observations and the unknown form of the true error distribution. The use of bootstrap in non-parametric statistics and various other fields has been extensively studied in the literature (DiCiccio & Efron, 1996; Hall, 1988; Hall & Wilson, 1991; Politis & Romano, 1994). To further assess the viability of the proposed approach, several linear regression model comparisons are conducted in this paper.

The following are the contributions of this paper in comparison with previous research: (1) The minimum MRE estimators of three commonly used distributions are derived. (2) A new criterion for model comparison, termed rMRE, is proposed. (3) An estimation framework for rMRE, combining kernel density estimation and bootstrap resampling, is developed and shown to produce consistent and robust results across multiple kernel choices.

The structure of this paper is organized as follows. Section 2 introduces the concept of MRE and its properties. Section 3 presents the theoretical derivation of the minimum MRE estimators for the Gamma, Laplace, and Rayleigh distributions. In Section 4, a method for estimating rMRE is proposed. To verify and demonstrate our proposals, simulation studies and two empirical analyses are conducted in Sections 5 and 6. Finally, Section 7 concludes the paper.

2. Definition and properties of mean relative entropy

2.1. Mean relative entropy

Let the population probability density function of X be $f(x, \eta)$, and let $x_1, x_2, \ldots, x_n$ be a random sample of the population X. Let $\hat{\eta} = \hat{\eta}(\boldsymbol{X})$ be an arbitrary estimator of the parameter η. If X has a common support, for any η and $\hat{\eta}$, the MRE of $\hat{\eta}$ is defined as Zhang (2021)

$$\mathrm{MRE}(\hat{\eta}) = E\left[\log \frac{f(X, \eta)}{f(X, \hat{\eta})}\right]. \tag{1}$$

2.2. Properties of mean relative entropy

Lemma 2.1: *The mean relative entropy is always non-negative.*

Lemma 2.2: *Under any one-to-one parameter transformation, the MRE remains invariant.*

Lemma 2.3: *For any estimator* $\hat{\eta} = \hat{\eta}(\boldsymbol{X})$,

$$\mathrm{MRE}(\hat{\eta}) \geq \mathrm{MRE}(\hat{\eta}_T),$$

where $\hat{\eta}_T = E[\hat{\eta} \mid T(\boldsymbol{X})]$ *is the conditional expectation of* $\hat{\eta}$ *given* $T(\boldsymbol{X})$, *a sufficient statistic for* η.

Lemma 2.4: *If* $T(\boldsymbol{X})$ *is a complete and sufficient statistic for* η, *then for any estimator* $\hat{\eta} = \hat{\eta}(\boldsymbol{X})$, $\eta_T = E[\hat{\eta} \mid T(\boldsymbol{X})]$ *is the unique minimum MRE estimator of* η *with mean* $E(\hat{\eta})$.

Lemmas 2.1–2.4 (Zhang, 2021) demonstrate that the MRE possesses several desirable properties. Based on these results, this paper will extend the theoretical study of minimum MRE estimators to several commonly used distributions in Section 3.

3. Minimum MRE estimators for three common distributions

3.1. The minimum MRE estimator for the scale parameter of the Gamma distribution

Theorem 3.1: *If* $X \sim \mathrm{Gamma}(\alpha, \lambda)$, *with density*

$$f(x) = \frac{1}{\Gamma(\alpha)\lambda^{\alpha}} x^{\alpha-1} \mathrm{e}^{-x/\lambda}, \quad \alpha > 0,$$

then based on a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, *the minimum MRE estimator of* λ *is*

$$\hat{\lambda} = \frac{1}{n\alpha - 1} \sum_{i=1}^{n} x_i.$$

Proof: For any estimator $\hat{\lambda} = \hat{\lambda}(\mathbf{x})$,

$$\begin{aligned}\mathrm{MRE}(\hat{\lambda}) &= E\left[\log\frac{f(x,\lambda)}{f(x,\hat{\lambda})}\right]\\ &= E\left[\alpha\log\frac{\hat{\lambda}}{\lambda} + \left(\frac{1}{\hat{\lambda}} - \frac{1}{\lambda}\right)x\right]\\ &= \alpha\, E\left[\log\frac{\hat{\lambda}}{\lambda} + \frac{\lambda}{\hat{\lambda}} - 1\right].\end{aligned}$$

Note that $T = \sum_{i=1}^{n} x_i$ is a sufficient statistic for λ, and

$$\mathrm{MRE}(T) = \alpha\, E\left[\log\frac{T}{\lambda} + \frac{\lambda}{T} - 1\right].$$

Due to the additivity property of the Gamma distribution, $T \sim \mathrm{Gamma}(n\alpha, \lambda)$, so $1/T \sim \mathrm{IGamma}(n, \lambda)$, and

$$E\left[\frac{1}{T}\right] = \frac{1}{(n\alpha - 1)\lambda}.$$

Then

$$\mathrm{MRE}(T) = \alpha\, E\left[\log\frac{T}{\lambda} + \frac{1}{n\alpha - 1} - 1\right].$$

According to Lemma 2.3, the minimum estimator for λ must be a function of a sufficient statistic. Let c be a constant such that the minimum MRE of cT is achieved. Then

$$\mathrm{MRE}(cT) = \alpha\, E\left[\log\frac{cT}{\lambda} + \frac{\lambda}{cT} - 1\right] = \mathrm{MRE}(T) + \alpha\left[\frac{1-c}{c(n\alpha - 1)} + \log c\right],$$

which is minimized at $c = \frac{1}{n\alpha - 1}$.

Therefore, the minimum MRE estimator of λ is

$$\hat{\lambda} = \frac{1}{n\alpha - 1}\sum_{i=1}^{n} x_i.$$

■

3.2. The minimum MRE estimator for the parameter of the Laplace distribution

Theorem 3.2: *If* $X \sim \mathrm{Laplace}(\mu, \lambda)$, *with density*

$$f(x) = \frac{1}{2\lambda}\mathrm{e}^{-|x-\mu|/\lambda}, \quad \mu = 0,\ \lambda > 0,$$

then based on a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, *the minimum MRE estimator of* λ *is*

$$\hat{\lambda} = \frac{1}{n-1}\sum_{i=1}^{n} |x_i|.$$

Proof: We first find a sufficient statistic for λ.

Since $X \sim \text{Laplace}(0, \lambda)$, the joint density of the sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is

$$f(x_1, x_2, \ldots, x_n; \lambda) = \prod_{i=1}^{n} \frac{1}{2\lambda} \mathrm{e}^{-|x_i|/\lambda} = (2\lambda)^{-n} \mathrm{e}^{-T/\lambda}, \quad T = \sum_{i=1}^{n} |x_i|.$$

Define $g(T, \lambda) = (2\lambda)^{-n} \mathrm{e}^{-T/\lambda}$ and $h(x_1, x_2, \ldots, x_n) = 1$. Therefore, $T = \sum_{i=1}^{n} |x_i|$ is a sufficient statistic for λ.

For any estimator $\hat{\lambda} = \hat{\lambda}(\mathbf{x})$,

$$\begin{aligned} \text{MRE}(\hat{\lambda}) &= E\left[\log \frac{f(x, \lambda)}{f(x, \hat{\lambda})}\right] \\ &= E\left[\log \frac{\hat{\lambda}}{\lambda} + \left(\frac{1}{\hat{\lambda}} - \frac{1}{\lambda}\right)|x|\right] \\ &= E\left[\log \frac{\hat{\lambda}}{\lambda} + \frac{\lambda}{\hat{\lambda}} - 1\right]. \end{aligned}$$

Thus,

$$\text{MRE}(T) = E\left[\log \frac{T}{\lambda} + \frac{\lambda}{T} - 1\right].$$

Since $|X| \sim \text{Exp}(\lambda)$ and an exponential distribution can be regarded as a Gamma distribution with shape parameter 1, by the additivity property of the Gamma distribution, we have $T \sim \Gamma(n, \lambda)$, so $1/T \sim \text{IGamma}(n, \lambda)$, and

$$E\left[\frac{1}{T}\right] = \frac{1}{(n-1)\lambda}.$$

Substituting this result gives

$$\text{MRE}(T) = E\left[\log \frac{T}{\lambda} + \frac{1}{n-1} - 1\right].$$

Now, the objective is to find a constant c such that the minimum MRE of cT is achieved. Then

$$\text{MRE}(cT) = E\left[\log \frac{cT}{\lambda} + \frac{\lambda}{cT} - 1\right] = \text{MRE}(T) + \frac{1-c}{c(n-1)} + \log c,$$

which is minimized at $c = \frac{1}{n-1}$.

Therefore, the minimum MRE estimator of λ is

$$\hat{\lambda} = \frac{1}{n-1} \sum_{i=1}^{n} |x_i|.$$

■

3.3. The minimum MRE estimator for the parameter of the Rayleigh distribution

Theorem 3.3: *If* $X \sim \text{Rayleigh}(\sigma)$, *let* $\theta = \sigma^2$, *so that the density is*

$$f(x) = \frac{x}{\theta} e^{-x^2/(2\theta)}, \quad x > 0,\ \theta > 0.$$

Then based on a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, *the minimum MRE estimator of* θ *is*

$$\hat{\theta} = \frac{1}{2(n-1)} \sum_{i=1}^{n} x_i^2.$$

Proof: We first find a sufficient statistic for θ.

The joint density of the sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is

$$f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} \frac{x_i}{\theta} e^{-x_i^2/(2\theta)} = \frac{\prod_{i=1}^{n} x_i}{\theta^n} e^{-T/(2\theta)}, \quad T = \sum_{i=1}^{n} x_i^2.$$

Define $g(T, \theta) = \theta^{-n} e^{-T/(2\theta)}$ and $h(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} x_i$. Therefore, $T = \sum_{i=1}^{n} x_i^2$ is a sufficient statistic for θ.

For any estimator $\hat{\theta} = \hat{\theta}(\mathbf{x})$,

$$\begin{aligned} \text{MRE}(\hat{\theta}) &= E\left[\log \frac{f(x,\theta)}{f(x,\hat{\theta})}\right] \\ &= E\left[\log \frac{\hat{\theta}}{\theta} + \left(\frac{1}{2\hat{\theta}} - \frac{1}{2\theta}\right) x^2\right] \\ &= E\left[\log \frac{\hat{\theta}}{\theta} + \frac{\theta}{\hat{\theta}} - 1\right]. \end{aligned}$$

Thus,

$$\text{MRE}(T) = E\left[\log \frac{T}{\theta} + \frac{\theta}{T} - 1\right].$$

Since $T = \sum_{i=1}^{n} X_i^2 \sim \text{Gamma}(n, 2\theta)$, we have $1/T \sim \text{IGamma}(n, 2\theta)$ and

$$E\left[\frac{1}{T}\right] = \frac{1}{2(n-1)\theta}.$$

Substituting this gives

$$\text{MRE}(T) = E\left[\log \frac{T}{\theta} + \frac{1}{2(n-1)} - 1\right].$$

Now, we find a constant c such that the minimum MRE of cT is achieved. Then

$$\text{MRE}(cT) = E\left[\log \frac{cT}{\theta} + \frac{\theta}{cT} - 1\right] = \text{MRE}(T) + \frac{1-c}{2(n-1)c} + \log c,$$

which is minimized at $c = \frac{1}{2(n-1)}$.

Therefore, the minimum MRE estimator of θ is

$$\hat{\theta} = \frac{1}{2(n-1)} \sum_{i=1}^{n} x_i^2.$$

■

Zhang (2021) highlighted that, unlike MSE, MRE not only remains invariant under one-to-one transformations of the parameters but also demonstrates superior performance in estimator selection, making it a more reliable and appropriate measure of estimation error.

MRE essentially quantifies the difference between two probability density distributions. In parametric estimation, it provides a natural basis for deriving minimum-error estimators for distributional parameters. However, in non-parametric estimation, MRE also plays an important role. In practical applications, researchers often need to compare different models. For instance, in multiple linear regression modelling, commonly used model selection criteria include the adjusted coefficient of determination, Akaike information criterion (AIC), Bayesian information criterion (BIC), C_p statistic, among others (Akaike, 1974; Ghosh & Yang, 1988; Mallows, 2000). In addition, residuals are pivotal for model diagnostics (Ling, 1984), and a widely used approach is to assess models through the MSE of residuals. Beyond summary measures such as MSE, it is also helpful to examine how the residuals are distributed. Since the distribution of residuals should resemble that of the true error terms for a well-specified model, the discrepancy between these two distributions provides an additional perspective for model comparison. If the error distribution were known, such discrepancies could be evaluated directly using MRE. In practice, however, the error distribution is rarely specified (Chen et al., 2003), which motivates the need for an alternative approach. To overcome this limitation, we introduce a method for estimating rMRE when the statistical distribution of the error terms is unknown.

4. Kernel-bootstrap estimation of rMRE

4.1. The rMRE estimation method

Consider a general regression model with n observations. The relationship between the response variable y and the explanatory variables $x_1, x_2, \ldots, x_p$ is given by

$$y_i = f(x_{i1}, x_{i2}, \ldots, x_{ip}) + \varepsilon_i, \tag{2}$$

where ε_i is the error term with mean 0, and the error terms are assumed to be uncorrelated.

For a multiple linear regression model, we have

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \tag{3}$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are unknown regression coefficients and ε_i represents the error term.

After estimating the parameters by least squares, we obtain

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}, \tag{4}$$

and the residuals (or out-of-sample prediction errors) are

$$e_i = y_i - \hat{y}_i. \tag{5}$$

In many cases, the statistical distribution of the error terms may often be unknown, which poses a challenge in determining the probability density function of the error terms and

the residuals. To address this limitation, this paper utilizes the kernel density estimation (Parzen, 1962; Rosenblatt, 1956) to obtain the empirical probability density function of the residuals.

Let $x_1, x_2, \ldots, x_n$ be independent and identically distributed samples drawn from some univariate distribution with an unknown density f. Its kernel density estimator is

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{6}$$

where K is a non-negative kernel function and $h > 0$ is the bandwidth.

Then, bootstrap is used to approximate the true statistical distribution of the error terms (Efron, 1992).

Suppose the dataset X consists of N data points. Each time a new sample with a data size of N is generated through sampling with replacement, let X^j denote the new sample obtained in the j-th sampling. The probability density function of the dataset X^j can be recalculated using Equation (6). This sampling step is repeated B times (typically $B \geq 1000$). Finally, the average of $\{\hat{f}_{hj}(x), j = 1, \ldots, B\}$ is taken as an estimate of the true probability density of the data generated by X. Thus, the true probability density function of the error terms $f_h(\mathbf{e})$ is represented as follows:

$$f_h(\mathbf{e}) = \frac{1}{B} \sum_{j=1}^{B} \hat{f}_{hj}(\mathbf{e}), \tag{7}$$

where B represents the number of sampling iterations, and $\hat{f}_{hj}(.)$ represents the empirical probability density function of the residuals obtained from the j-th sampling.

Thus, rMRE($\mathbf{e}$) is represented as follows:

$$\text{rMRE}(\mathbf{e}) = E \log \frac{f_h(\mathbf{e})}{\hat{f}h(\mathbf{e})} = \frac{1}{N} \sum_{i=1}^{N} \log \frac{\frac{1}{B} \sum_{j=1}^{B} \hat{f}_{hj}(e_i)}{\hat{f}_h(e_i)}, \tag{8}$$

where N denotes the number of residual data points, B the number of sampling iterations, and e_i ($i = 1, 2, \ldots, N$) the residual.

The algorithm and steps for estimating rMRE are as follows:

Algorithm 1 rMRE estimation procedure

1: Estimate the empirical residual density $\hat{f}_h(\mathbf{e})$ using Formula (6).
2: Estimate the bootstrap-averaged density $f_h(\mathbf{e})$ using Formula (7).
3: Compute rMRE using Formula (8).

4.2. Computational considerations and scalability

The KDE–bootstrap procedure described in Section 4.1 is computationally intensive when the sample size N is large, because the density estimation must be recomputed for each of the B bootstrap samples. While this is feasible for moderate sample sizes, it becomes challenging for large-scale datasets.

Recent advances provide practical solutions to this problem. Zheng et al. (2013) proposed a block-based, parallelizable framework for KDE that allows the density estimator

$$\hat{f}_h(x) = \frac{1}{N}\sum_{i=1}^{N} K_h(x - x_i)$$

to be computed efficiently on partitioned data blocks, reducing computational time while maintaining high accuracy. In addition, representative subsampling methods have been developed to further reduce the data size for KDE computation. In particular, Zhang et al. (2023) introduced an optimal-transport-based approach to select a small but representative subsample whose empirical distribution approximates that of the full dataset. The authors provide theoretical guarantees on the convergence of the subsample-based KDE and propose data-adaptive bandwidth selection rules, demonstrating both accuracy and computational efficiency in numerical experiments.

These developments indicate that, even for large N, our rMRE estimation method can be efficiently implemented by combining distributed or parallel KDE computation with a carefully selected representative subsample. Therefore, the proposed rMRE approach remains computationally feasible and practically applicable in large-scale data settings.

Remark 4.1 (Consistency of rMRE): The proposed rMRE is a plug-in estimator constructed from kernel density estimates of the error and residual densities. Under standard regularity assumptions for kernel density estimation–smooth densities that are bounded and bounded away from zero, with bandwidths satisfying $h \to 0$ and $nh \to \infty$–the kernel estimators converge uniformly to their population densities. Under these same boundedness conditions, the KL-type functional that defines rMRE is continuous with respect to uniform convergence. These properties are well established in the classical literature (Billingsley, 1999; Parzen, 1962; Pollard, 1984; Rosenblatt, 1956; Silverman, 1978). Consequently, once the kernel density estimators converge to the true densities of the errors and residuals, the plug-in rMRE estimator also converges to its population counterpart. Thus, the consistency of rMRE follows naturally from the well-established consistency properties of kernel density estimators.

In order to further validate the feasibility of rMRE as a criterion for model comparison, we will take the multiple linear regression models as examples for analysis in the following simulation and empirical study, where the residual data are sampled 1000 times.

5. Simulation

In this simulation, we first specified the true multiple regression model and randomly generated two true explanatory variables, x_{is1} and x_{is2} $(i = 1, 2, \ldots, 100)$. We also generated an error term ε, and thus the true model was defined as follows:

$$y = 4 + 2x_{s1} + 3x_{s2} + \varepsilon. \tag{9}$$

Next, we randomly generated two spurious explanatory variables x_{is3}, x_{is4}, $(i = 1, 2, \ldots, 100)$. We then conducted regression analyses of y on the true variables and on all variables, respectively. The regression results are shown in Tables 1–2.

Table 2 indicates that the full model exhibits severe overfitting, thereby undermining the interpretability and reliability of the MSE in this context. However, based on the method

Table 1. Regression results of the true model in the simulation.

	Estimate	Std. error	*t* value	Pr(> |*t*|)
Intercept	3.9458	0.2117	18.637	$< 2\times10^{-16}$***
x_{s1}	1.9044	0.2309	8.246	8.04×10^{-13}***
x_{s2}	3.0897	0.2180	14.172	$< 2\times10^{-16}$***

$p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2. Regression results of the full model in the simulation.

	Estimate	Std. error	*t* value	Pr(> |*t*|)
Intercept	4.000	1.455×10^{-16}	2.749×10^{16}	$< 2\times10^{-16}$***
x_{s1}	2.000	1.586×10^{-16}	1.261×10^{16}	$< 2\times10^{-16}$***
x_{s2}	3.000	1.485×10^{-16}	2.021×10^{16}	$< 2\times10^{-16}$***
x_{s3}	1.695×10^{-16}	1.523×10^{-16}	1.113	0.269
x_{s4}	2.000	1.383×10^{-16}	1.446×10^{16}	$< 2\times10^{-16}$***

$p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3. Comparison of rMRE results between the full and true models in the simulation.

	rMRE	
Kernel function	Full model	True model
Gaussian	2.325×10^{-2}	$\mathbf{1.319\times10^{-3}}$
Epanechnikov	2.246×10^{-2}	$\mathbf{9.648\times10^{-4}}$
triangular	2.705×10^{-2}	$\mathbf{1.187\times10^{-3}}$
biweight	2.380×10^{-2}	$\mathbf{1.061\times10^{-3}}$
cosine	2.392×10^{-2}	$\mathbf{1.116\times10^{-3}}$
optcosine	2.319×10^{-2}	$\mathbf{8.632\times10^{-4}}$

proposed in this paper, the two models can still be compared through rMRE estimation. Thus, we conducted 200 repetitions of five-fold cross-validation for both the true model and the full model. Additionally, considering that the choice of kernel function may influence the estimation of rMRE, we explored and compared the average rMRE values across the 200 five-fold cross-validations under different kernel functions. The results are shown in Table 3.

The results presented in Table 3 indicate that the true model consistently achieved a smaller rMRE than the full model. Additionally, the rMRE estimates obtained using six different kernel functions were largely consistent, suggesting that the rMRE-based comparison exhibits robustness.

6. Empirical analysis

6.1. Example 1

In Example 1, we utilized the diabetes dataset from the `lars` package in R. The dataset contains information on 442 patients with diabetes, including variables such as age, sex, BMI (body mass index), mean blood pressure, six blood serum measurements (TC, LDL, HDL, TCH, LTG, GLU), and the progression of diabetes disease (y) (Efron et al., 2004). In this dataset, BMI, mean blood pressure, and the six blood serum measurements are standardized.

We first fitted a regression model using all available explanatory variables. The regression results are shown in Table 4. Table 5 and Figure 1 show the correlations among the variables, illustrating the strength of the correlations through both absolute values and colour intensity.

Table 4. Regression results of the full model in Example 1.

	Estimate	Std. error	*t* value	Pr(> |*t*|)
Intercept	152.13	2.58	59.06	$< 2\times10^{-16}$***
Age	−10.01	59.75	−0.17	0.87
Sex	−239.82	61.22	−3.91	0.0001***
BMI	519.84	66.53	7.81	4.30×10^{-14}***
Map	324.39	65.42	4.96	1.02×10^{-6}***
TC	−792.18	416.68	−1.90	0.06 .
LDL	476.746	339.035	1.41	0.16
HDL	101.05	212.53	0.48	0.63
TCH	177.06	161.48	1.10	0.27
LTG	751.28	171.90	4.37	1.56×10^{-5}***
GLU	67.63	65.98	1.03	0.31

$p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5. Correlation matrix of all variables in Example 1.

	Age	Sex	BMI	Map	TC	LDL	HDL	TCH	LTG	GLU
Age	1.00	0.17	0.19	0.34	0.26	0.22	−0.08	0.20	0.27	0.30
Sex	0.17	1.00	0.09	0.24	0.04	0.14	−0.38	0.33	0.15	0.21
BMI	0.19	0.09	1.00	0.40	0.25	0.26	−0.37	0.41	0.45	0.39
Map	0.34	0.24	0.40	1.00	0.24	0.19	−0.18	0.26	0.39	0.39
TC	0.26	0.04	0.25	0.24	1.00	0.90	0.05	0.54	0.52	0.33
LDL	0.22	0.14	0.26	0.19	0.90	1.00	−0.20	0.66	0.32	0.29
HDL	−0.08	−0.38	−0.37	−0.18	0.05	−0.20	1.00	−0.74	−0.40	−0.27
TCH	0.20	0.33	0.41	0.26	0.54	0.66	−0.74	1.00	0.62	0.42
LTG	0.27	0.15	0.45	0.39	0.52	0.32	−0.40	0.62	1.00	0.46
GLU	0.30	0.21	0.39	0.39	0.33	0.29	−0.27	0.42	0.46	1.00

Table 4 shows that half of the variables are not significant. Together with the correlation patterns in Table 5, this suggests that further variable selection is required to obtain a more robust and reliable model.

Therefore, we adopted the BIC principle to select explanatory variables based on the full dataset. The final selected model includes six variables: *Sex*, *BMI*, *Map*, *TC*, *LDL*, and *LTG*. The regression results of the selected model are reported in Table 6.

Next, we conducted 200 repetitions of five-fold cross-validation for both the full model and the selected model. Typically, we calculate the MSE of the two groups of residuals and compare their values. The model corresponding to the residuals with the lower MSE is generally considered to have better predictive performance. In this paper, we estimated the rMRE of the two groups of residuals separately. In addition, considering that the choice of kernel function may influence the estimation of rMRE, we explored and compared the average rMRE from the 200 cross-validation repetitions using different kernel functions.

The results presented in Table 7 demonstrate that both the rMRE and MSE of the selected model are lower than those of the full model. All variables retained in the selected model are more significant. In Example 1, the model selected by rMRE exhibits higher interpretability and predictive performance. This implies that rMRE can also serve as an effective criterion for model comparison. The consistency of the results from the estimation of rMRE using different kernel functions indicates the robustness of rMRE.

6.2. Example 2

In Example 2, we utilized a dataset that includes PM2.5 measurements along with associated

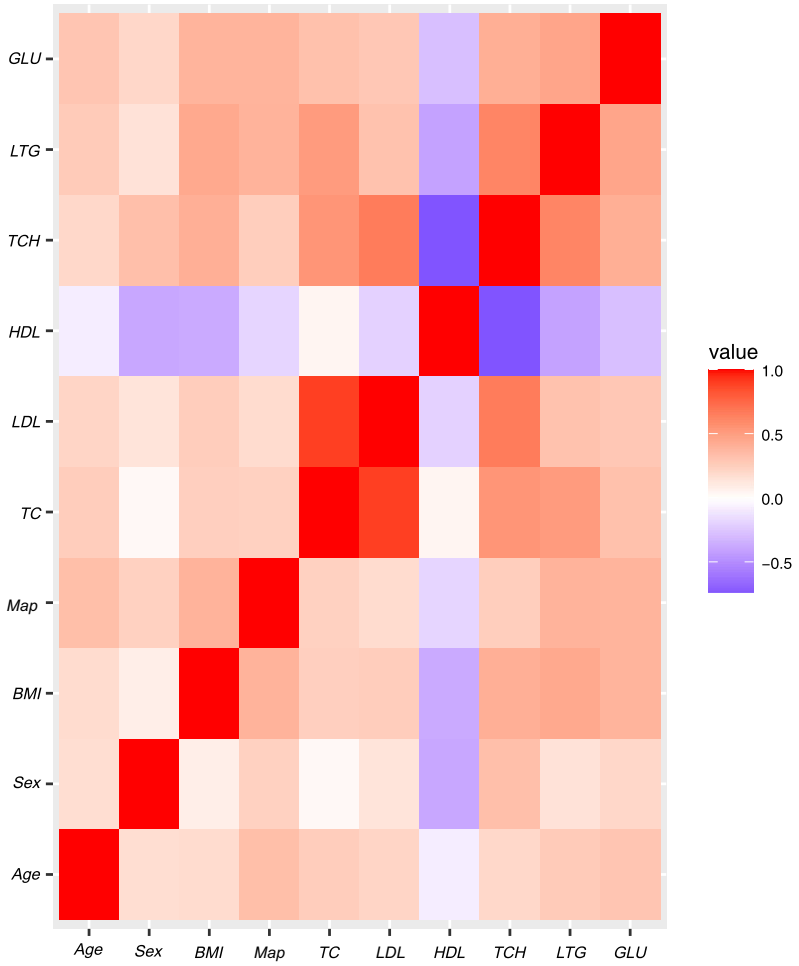

Figure 1. Heatmap of correlation coefficients among variables in Example 1.

Table 6. Regression results of the selected model in Example 1.

	Estimate	Std. error	*t* value	Pr(> |*t*|)
Intercept	152.133	2.57	59.16	$< 2\times10^{-16}$***
Sex	−226.51	59.86	−3.78	1.76×10^{-4}***
BMI	529.87	65.62	8.08	6.69×10^{-15}***
Map	327.22	62.69	5.22	2.79×10^{-7}***
TC	−757.94	160.44	−4.72	3.12×10^{-6}***
LDL	538.59	146.74	3.67	2.72×10^{-4}***
LTG	804.19	80.17	10.03	$< 2\times10^{-16}$***

$p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

meteorological data recorded at 2:00 p.m. in Beijing in 2011. The climate condition index data used in this paper were obtained from Dataset (https://archive.ics.uci.edu/datasets). After excluding 28 observations with missing PM2.5 values, a total of 337 valid data points were retained for analysis. In this dataset, the wind direction was represented as categorical variables, specifically NW, SE, NE, and CV (calm and variable). For analytical convenience, we

Table 7. Comparison of results between the full and selected models in Example 1.

	rMRE		MSE	
Kernel function	full model	selected model	full model	selected model
Gaussian	4.1×10^{-4}	$\mathbf{2.4\times10^{-4}}$	3013.85	**2970.14**
Epanechnikov	4.8×10^{-4}	$\mathbf{2.2\times10^{-4}}$	3013.85	**3011.32**
triangular	4.4×10^{-4}	$\mathbf{2.1\times10^{-4}}$	3011.65	**2968.44**
biweight	4.9×10^{-4}	$\mathbf{2.3\times10^{-4}}$	3011.04	**2968.68**
cosine	4.6×10^{-4}	$\mathbf{2.3\times10^{-4}}$	3011.20	**2969.09**
optcosine	4.9×10^{-4}	$\mathbf{2.6\times10^{-4}}$	3011.54	**2969.11**

Table 8. Descriptive statistics of PM2.5 index and meteorological data at 2:00 p.m. in Beijing, China (2011).

	Mean	SD	Min	1st quarter	Median	3rd quarter	Max
PM2.5(μg/m^3)	87.11	87.44	4.00	20.00	55.00	123.00	477.00
DEWP(° C)	1.13	14.67	−25.00	−12.00	0.00	15.00	28.00
TEMP(° C)	17.37	11.46	−4.00	7.00	19.00	28.00	35.00
PRES(hPa)	1016	10.98	994	1007	1017	1025	1040
NW	0.31	0.46	0.00	0.00	0.00	1.00	1.00
SE	0.43	0.50	0.00	0.00	0.00	1.00	1.00
NE	0.08	0.27	0.00	0.00	0.00	0.00	1.00
lws(m/s)	25.72	50.97	0.89	3.13	6.71	21.90	377.27

Table 9. Regression results of the full model in Example 2.

	Estimate	Std. error	t value	Pr(> $\lvert t\rvert$)
Intercept	1218.03	772.58	1.58	0.12
DEWP	3.89	0.51797	7.504	5.85×10^{-13}***
TEMP	−5.18	0.77	−6.76	6.49×10^{-11}***
PRES	−1.03	0.75	−1.37	0.17
NW	−32.16	12.80	−2.51	0.01*
SE	29.51	10.97	2.69	0.008**
NE	−38.91	16.49	−2.36	0.02*
lws	−0.12	0.09	−1.39	0.17

$p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

used dummy variables to represent the three explicit directions NW, SE, and NE, taking values 1 or 0. When all three dummy variables took the value 0, the wind direction was classified as CV. Table 8 shows the descriptive statistics of daily data regarding PM2.5 index (y), dew point (DEWP), temperature (TEMP), pressure (PRES), wind direction, and cumulative wind speed (lws) in Beijing during 2011.

We first conducted a regression analysis using the full model with all explanatory variables. The regression results are shown in Table 9. Table 10 and Figure 2 display the correlations among the variables, illustrating the strength of the correlations through both absolute values and colour intensity. From Table 9, we can see that several variables were not statistically significant. Similar to Example 1, this suggests that a further selection of explanatory variables is necessary to refine the model. Table 11 presents the results of the regression analysis after selecting variables using the BIC principle. Following the selection process, it was found that the intercept term and the explanatory variables, specifically DEWP, TEMP, and SE, remained highly significant in the model. Similarly, we conducted 200 repetitions of five-fold cross-validation for the full regression model and the selected model. The results of this analysis are summarized in Table 12 below.

Table 10. Correlation matrix of all variables in Example 2.

	DEWP	TEMP	PRES	NW	SE	NE	lws
DEWP	1.00	0.82	−0.79	−0.47	0.42	0.02	−0.33
TEMP	0.82	1.00	−0.87	−0.35	0.31	0.006	−0.24
PRES	−0.79	−0.87	1.00	0.30	−0.32	0.009	0.23
NW	−0.47	−0.35	0.30	1.00	−0.58	−0.19	0.48
SE	0.42	0.31	−0.32	−0.58	1.00	−0.25	−0.24
NE	0.02	0.006	0.009	−0.19	−0.25	1.00	−0.07
lws	−0.33	−0.24	0.23	0.48	−0.24	−0.07	1.00

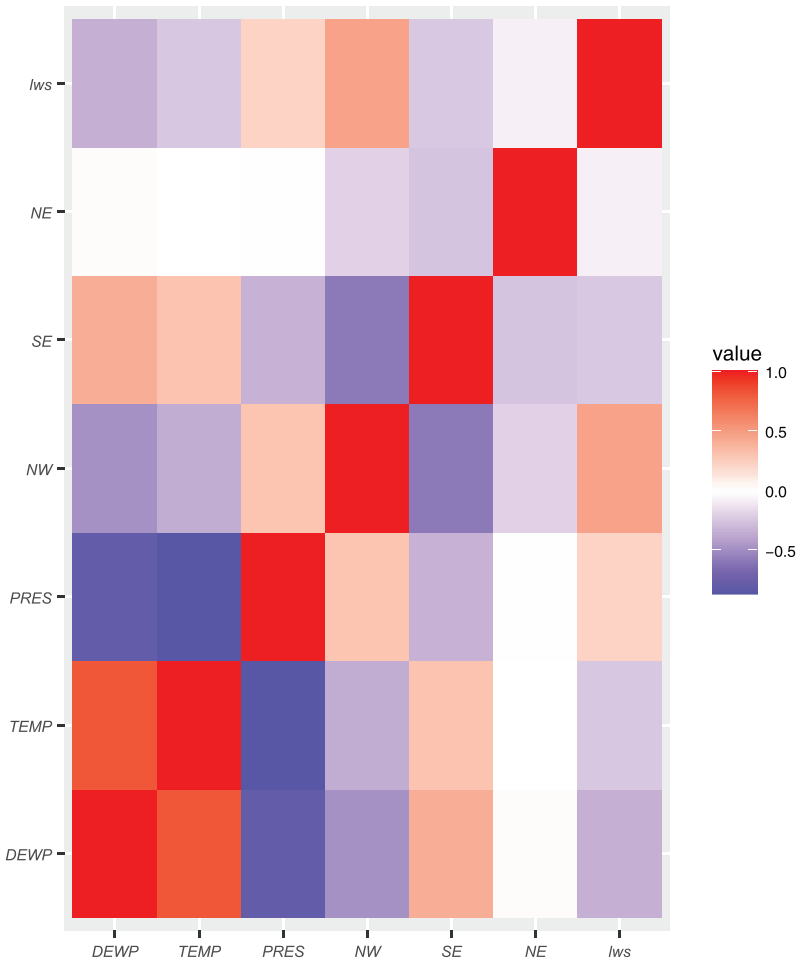

Figure 2. Heatmap of correlation coefficients among variables in Example 2.

Table 11. Regression results of the selected model in Example 2.

	Estimate	Std. error	*t* value	Pr(> |*t*|)
Intercept	139.43	11.48	12.14	$< 2\times10^{-16}$***
DEWP	4.40	0.49	9.00	$< 2\times10^{-16}$***
TEMP	−4.58	0.60	−7.64	2.37×10^{-13}***
SE	51.73	8.69	5.95	6.65×10^{-9}***

$p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 12. Comparison of results between the full and selected models in Example 2.

	rMRE		MSE	
Kernel function	full model	selected model	full model	selected model
Gaussian	1.524×10^{-2}	$\mathbf{1.431 \times 10^{-2}}$	**5083.23**	5209.81
Epanechnikov	1.583×10^{-2}	$\mathbf{1.501 \times 10^{-2}}$	**5075.40**	5212.81
triangular	1.459×10^{-2}	$\mathbf{1.436 \times 10^{-2}}$	**5082.05**	5208.23
biweight	1.553×10^{-2}	$\mathbf{1.473 \times 10^{-2}}$	**5077.74**	5211.21
cosine	1.510×10^{-2}	$\mathbf{1.472 \times 10^{-2}}$	**5082.28**	5209.93
optcosine	1.570×10^{-2}	$\mathbf{1.491 \times 10^{-2}}$	**5076.61**	5210.60

Table 12 shows that although the full model exhibits a slightly lower MSE than the selected model, its rMRE is consistently higher. This indicates that relying solely on MSE may lead to misleading conclusions. The presence of insignificant variables in the full model further highlights the necessity of variable selection. In Example 2, the model selected based on rMRE exhibits greater interpretability and competitive predictive performance. Combined with the findings of Example 1, this example further highlights the potential and practical significance of rMRE as a criterion for model comparison. In the future, rMRE can serve as an effective criterion for evaluating and comparing models.

7. Conclusion

This paper investigated the theoretical and practical applications of mean relative entropy (MRE) in statistical estimation. Based on the work of Zhang (2021), we derived minimum MRE estimators for the scale parameter of the Gamma distribution, the scale parameter of the Laplace distribution, and the parameter of the Rayleigh distribution. These results illustrate that MRE provides a principled approach for parameter estimation.

In addition, this paper proposed a non-parametric estimation method for model residuals, referred to as rMRE, which combines kernel density estimation and bootstrap. Simulation studies and empirical analyses demonstrated that rMRE effectively distinguishes between overfitted and true models, favouring models with greater interpretability. Compared with classical measure such as MSE, rMRE provides a more comprehensive measure of model fit by assessing the discrepancy between the distributions of residuals and the true error terms. Furthermore, we also briefly discussed the theoretical properties of rMRE and its stable performance in large-scale data settings, showing that it remains a practical and reliable criterion for model comparison.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the National Natural Science Foundation of China [grant number 12571312]; and the Central Government Fund for Guiding Local Scientific and Technological Development [grant number 2024ZYD0019]; and the Fundamental Research Funds for the Central Universities [grant numbers 2682026GH009; 2682025ZTPY012; 2682026CX005].

ORCID

Lei Huang http://orcid.org/0000-0002-7513-7461
Lanpeng Ji https://orcid.org/0000-0002-7790-7765

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705
Arizono, I., & Ohta, H. (1989). A test for normality based on Kullback-Leibler information. *The American Statistician*, *43*(1), 20–22.
Billingsley, P. (1999). *Convergence of probability measures* (2nd ed.). John Wiley & Sons.
Chen, G., Gott, J. R., & Ratra, B. (2003). Non-Gaussian error distribution of Hubble constant measurements. *Publications of the Astronomical Society of the Pacific*, *115*(813), 1269–1279. https://doi.org/10.1086/pasp.2003.115.issue-813
Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory* (2nd ed.). John Wiley & Sons.
DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*(3), 189–228. https://doi.org/10.1214/ss/1032280214
Efron, B. (1992). Bootstrap methods: Another look at the Jackknife. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics: Methodology and distribution* (pp. 569–593). Springer.
Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499. https://doi.org/10.1214/009053604000000067
Ghosh, M., & Yang, M. C. (1988). Simultaneous estimation of Poisson means under entropy loss. *The Annals of Statistics*, *16*(1), 278–291. https://doi.org/10.1214/aos/1176350705
Gupta, M., & Srivastava, S. (2010). Parametric Bayesian estimation of differential entropy and relative entropy. *Entropy*, *12*(4), 818–843. https://doi.org/10.3390/e12040818
Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, *16*(3), 927–953.
Hall, P., & Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, *47*(2), 757–762. https://doi.org/10.2307/2532163
Hu, Q., Guo, M., Yu, D., & Liu, J. (2010). Information entropy for ordinal classification. *Science China Information Sciences*, *53*(6), 1188–1200. https://doi.org/10.1007/s11432-010-3117-7
Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694
Ling, R. F. (1984). Residuals and influence in regression. *Technometrics*, *26*(4), 413–415. https://doi.org/10.1080/00401706.1984.10487996
Liu, S., Ma, T., & Liu, Y. (2016). Sensitivity analysis in linear models. *Special Matrices*, *4*(1), 225–232. https://doi.org/10.1515/spma-2016-0021
Mallows, C. L. (2000). Some comments on C_p. *Technometrics*, *42*(1), 87–94.
Mühlenbein, H., & Höns, R. (2005). The estimation of distributions and the minimum relative entropy principle. *Evolutionary Computation*, *13*(1), 1–27. https://doi.org/10.1162/1063656053583469
Namdari, A., & Li, Z. S. (2019). A review of entropy measures for uncertainty quantification of stochastic processes. *Advances in Mechanical Engineering*, *11*(6), 1687814019857350. https://doi.org/10.1177/1687814019857350
Parsian, A., & Nematollahi, N. (1996). Estimation of scale parameter under entropy loss function. *Journal of Statistical Planning and Inference*, *52*(1), 77–91. https://doi.org/10.1016/0378-3758(95)00026-7
Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, *33*(3), 1065–1076. https://doi.org/10.1214/aoms/1177704472
Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, *89*(428), 1303–1313. https://doi.org/10.1080/01621459.1994.10476870
Pollard, D. (1984). *Convergence of stochastic processes*. Springer.
Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, *27*(3), 832–837. https://doi.org/10.1214/aoms/1177728190

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/bltj.1948.27.issue-3

Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics*, *6*(1), 177–184. https://doi.org/10.1214/aos/1176344076

Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society: Series B (Methodological)*, *38*(1), 54–59. https://doi.org/10.1111/j.2517-6161.1976.tb01566.x

Zhang, J. (2021). The mean relative entropy: An invariant measure of estimation error. *The American Statistician*, *75*(2), 117–123. https://doi.org/10.1080/00031305.2018.1543139

Zhang, J., Meng, C., Yu, J., Zhang, M., Zhong, W., & Ma, P. (2023). An optimal transport approach for selecting a representative subsample with application in efficient kernel density estimation. *Journal of Computational and Graphical Statistics*, *32*(1), 329–339. https://doi.org/10.1080/10618600.2022.2084404

Zheng, Y., Jestes, J., Phillips, J. M., & Li, F. (2013). Quality and efficiency for kernel density estimates in large data. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (pp. 433–444). Association for Computing Machinery.