Statistica Sinica Preprint No: SS-2022-0410	
Title	Optimal Averaging Estimation for Density Functions
Manuscript ID	SS-2022-0410
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0410
Complete List of Authors	Peng Lin, Jun Liao, Zudi Lu, Kang You and Guohua Zou
Corresponding Authors	Guohua Zou
E-mails	ghzou@amss.ac.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

Statistica Sinica

Optimal averaging estimation for density functions

Peng Lin[a,*], Jun Liao[b,*], Zudi Lu[c,d], Kang You[e] and Guohua Zou[e,¶]

[a] *Shandong University of Technology*

[b] *Renmin University of China*

[c] *University of Southampton*

[d] *City University of Hong Kong*

[e] *Capital Normal University*

Abstract: Extraction of information from data is critical in the age of data science. Probability density function theoretically provides comprehensive information on the data. But, practically, different probability density models, either parametric or non-parametric, can often characterize partial features on the data, e.g., owing to model bias or less efficiency in estimation. In this paper we suggest a framework to optimally combine different density models to catch the comprehensive data features by a new information criterion (IC) based unsupervised learning approach. Our optimal information extraction is in the sense that the resultant density averaging or selected density minimises the Kullback–Leibler (KL) information loss function. Differently from the usual supervised learning IC for model selection or averaging, we first need to derive an estimator of the KL loss function in our setting, which takes the Akaike and Takeuchi information criteria as two special cases. A feasible density model averaging (DMA) procedure is accordingly suggested, with the DMA estimation achieving the lowest possible KL loss asymptotically. Further, the consistency of the weights of the DMA estimator

*Co-first authors.

¶Corresponding author: Guohua Zou. Email: *ghzou@amss.ac.cn*.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

tending to the optimal averaging weights minimizing the KL distance is obtained, and the convergence rate of our empirical weights is also derived. Simulation studies show that the DMA performs overall better and more robustly than the commonly used parametric or nonparametric density models, including kernel, finite mixture, logarithmic scoring rule and selection methods for density estimation in the literature. The real data analysis further demonstrates the performance of the proposed method.

Key words and phrases: Asymptotic optimality, Density estimation, Density averaging, Weight choice.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

1. Introduction

Extracting knowledge from data is key to data science (c.f., Baimuratov et al., 2019). As is well known, density function of a data-generating random variable theoretically provides comprehensive information on the data, estimation of which is hence a fundamental problem in Econometrics, Statistics and data science. Usually there are two types of approaches that are widely applied to estimate a density function. One is parametric approach depending on a given parametric form of the density function family, which may suffer from model bias, and the other is nonparametric approach like kernel, nearest neighbor and other density estimation methods depending on some local parameters such as bandwidths for kernel estimation, possibly suffering from less efficiency in estimation. Hence, practically, different probability density models, either parametric or non-parametric, can often characterize partial features on the data. From the learning perspective, probability density estimation is a kind of unsupervised learning, and how to learn to optimally combine different density models to catch the comprehensive data features is an important but difficult task.

As is well addressed in the literature (e.g., Baimuratov et al., 2019), data processing techniques require a set of tools for evaluating knowledge extracted from data. In unsupervised learning it is impossible to use referential or predictive estimation, and the only reliable way to evaluate results of unsupervised

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

learning is information estimation. It is well known that information estimation unfortunately suffers from under-fitting and over-fitting; see Baimuratov et al. (2019) and the related references therein.

In this paper we are hence concerned with how to optimally combine different density models to extract more comprehensive data features. Our optimal information extraction is in the sense that the resultant density averaging or selected density minimises the Kullback–Leibler (KL) information loss function. In particular, with the parametric approach, we generally have different families of density functions such as normal, log-normal, and gamma, and we usually need to select one from them. Many methods have been proposed for selecting an appropriate density function from different families of candidate density functions. For instance, Claeskens and Hjort (2008) used AIC (Akaike, 1973) and BIC (Schwarz, 1978) to select a density model for real data, and Saumard and Navarro (2021) provided a novel model selection method in the context of density estimation, which is a correction of the AIC criterion by developing a new penalty term. Similarly, in nonparametric density estimation, often a local smoothing parameter also needs to be selected. Although such selection strategies may be sensible when the true density model is included in the candidate density model set, it is often not wise when the candidate models compete with each other. This is because different probability density models often have different features

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

that may partially capture the pattern on the data. In fact, a real data set can often be approximately described by different density functions. Therefore, to avoid the risk of "putting all eggs in one basket", a more rational procedure is to combine all the possible candidate density models by averaging, i.e., density model averaging (DMA). Concerning some specific feature, e.g., clustering, of the data, Baimuratov et al. (2019) have also proposed a new method for evaluating unsupervised learning results, which is based on the Bayesian criterion for optimal decision and an objective prior probability distribution of partitions. Under a reasonable knowledge on the objective prior of partitions, their method was shown to perform well with their problem and prevention of under-fitting and over-fitting. However, we may lack such prior knowledge for our density combination for more comprehensive data features in practice. Different from the references above, our main objective is to suggest a framework to optimally averaging different density models to more fully capture the data features in information extraction. In fact, from the aspect of estimation or forecast, model selection can be seen as a special case of the model averaging with the selected model weight being one while zeros for other models.

In the setting of regression modelling under supervised learning, the idea of model averaging (MA) has been popular. Generally, estimation and prediction risks can be substantially reduced by model averaging instead of model selection

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

(Hansen, 2014). In particular, frequentist MA methods play an important role; see Hjort and Claeskens (2003) for early relevant study. The optimal frequentist MA was proposed in the seminal work of Hansen (2007). There, he derived the Mallows model averaging (MMA) criterion, which is an unbiased estimator of the mean squared error up to a constant, and showed that the MMA estimator is asymptotically optimal in terms of minimizing the squared error loss. Wan et al. (2010) further demonstrated such an asymptotic optimality for MMA in a more general setting. For heteroscedastic error cases, see Hansen and Racine (2012) for the Jackknife MA and Liu and Okui (2013) for robust MA. For time series cases, the reader is referred to Hansen (2008) for the application of MMA in prediction, Cheng and Hansen (2015) for the cross-validation MA prediction with factor-augmented regression, and Liao et al. (2019) for the leave-subject-out cross-validation MA for vector autoregressive models. Zhang et al. (2016) further developed a weight choice criterion for generalized linear models. For high-dimensional regression cases, one is referred to Ando and Li (2014). Recently, Li et al. (2015) and Chen et al. (2018) have moreover developed semiparametric model averaging procedures for nonlinear dynamic high-dimensional time series modelling and forecasting.

Unlike the setting of regression modelling averaging under supervised learning in the literature, we are suggesting an unsupervised learning framework to opti-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

mally averaging different density models to more fully capture the data features. In the unsupervised learning setting, we need to directly study the estimation of the density without the information on correct answers (e.g., the class label) (Maggioni and Murphy, 2019), so the learning for this case is challenging (Hastie et al., 2009; Safarinejadian et al., 2010). Differently from the regression setting, the averaging for which is mostly squared error loss based, we suggest the DMA procedure by proposing a new IC from the KL information loss function. Starting from the KL distance between the true density and its averaging estimator, we will propose and derive an appropriate, feasible estimator of the KL loss, which takes the Akaike and Takeuchi information criteria as two special cases (c.f. Remark 1 in Subsection 3.1). Like other information criteria, this DMA criterion consists of two terms with the first being negative logarithm of the averaging likelihood and the second penalizing the complexities of the candidate density functions. Thus, the proposed criterion shares the useful property with the frequently used information criteria (e.g., AIC).

It is worth pointing out that the derivation of our proposed DMA criterion is not trivial and involves quite complicated technical details because the candidate density functions have no specific expressions and thus the methods used for deriving common regression model averaging criteria (like MMA) no longer apply. Based on the techniques of Taylor expansion and some large sample theories, we

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

first obtain the approximate bias between the KL loss and its empirical estimator. Then such a bias serves as the penalty term of the DMA and accordingly our DMA criterion yields. Earlier, Hall and Mitchell (2007) also considered the density averaging forecast, who selected weights by the logarithmic scoring rule, i.e., minimizing the empirical estimator of the KL loss directly (see Geweke and Amisano (2011) as well). Such a method ignores the bias mentioned above and may lead to the poor estimation and forecast for finite sample sizes. In addition, the KL loss was applied by Zhang et al. (2015) and Zhang et al. (2016) for regression model averaging. It can be found that the former requires the normal distribution assumption for the derivation of the weight choice criterion and the latter constructs the weight choice criterion in an intuitive way, which also implies the difficulty of developing weight choice criterion for general models like those considered in the current paper.

We will theoretically verify that the DMA procedure is asymptotically optimal in the sense that the DMA estimator achieves the lowest possible KL loss asymptotically. Further, we will derive the rate of convergence of the estimated DMA weights to the optimal ones that minimize the KL loss for parametric DMA. Not only so, based on the optimal DMA criterion and the frequently used kernel density estimation, a data driven semiparametric density averaging method is also developed, which combines parametric and nonparametric density estima-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

tors adaptively. Our simulation studies will show that the DMA performs overall better and more robustly than the commonly used parametric or nonparametric density models, including kernel, finite mixture, logarithmic scoring rule and selection methods for density estimation in the literature. The analysis of real data will further demonstrate the performance of the proposed method.

The remainder of this paper is organized as follows. Section 2 proposes the framework of the DMA criterion. Section 3 develops the estimation of the KL loss based DMA criterion with the asymptotic optimality of the resultant density averaging estimator established and the rate of the DMA based weights tending to the optimal weights minimizing the KL distance derived. Extension of the DMA method to semiparametric density averaging estimation is also considered. Section 4 demonstrates the performance of DMA by both simulation trials and the real data example. Section 5 concludes. The proofs of the theorems are contained in the supplementary document.

2. The framework of optimal density averaging criterion

Let $X_1, ..., X_n$ be independent and identically distributed (i.i.d.) data from the distribution generating random variable X. Denote by g the true density of the random variable X, and suppose that there are candidate model density functions that, extracting the information from the data, are $f_1(x, \theta_1)$,..., and $f_M(x, \theta_M)$,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

where $\theta_m = (\theta_{m1}, ..., \theta_{mp_m})'$ is the parameter vector for the mth model density and p_m is the dimension of θ_m with $m = 1, ..., M$. Define the information loss of the KL distance between g and $f_m(x, \theta_m)$, $m = 1, ..., M$, as

$$KL(g, f_m(x, \theta_m)) = \int g(x) \log \frac{g(x)}{f_m(x, \theta_m)} dx,$$

based on which we seek the optimal information extraction from the data.

For the mth candidate model $f_m(x, \theta_m)$, let $\widehat{\theta}_m$ be the maximum likelihood estimator of θ_m based on the data $\{X_i\}$, and define the theoretically optimal parameter vector

$$\theta_{m0} = \text{argmin}_{\theta_m}\{KL(g, f_m(x, \theta_m))\}. \tag{2.1}$$

In order to more comprehensively catch the features from the different candidate densities for the data, we propose an optimal density averaging approach by combining the candidate model densities to approximate the true density g. Let $w \in \mathcal{W} = \{w = (w_1, \cdots, w_M)' \in [0, 1]^M : \sum_{m=1}^M w_m = 1\}$. The averaging estimator for g is

$$f(x, \hat{\theta}, w) = \sum_{m=1}^{M} w_m f_m(x, \hat{\theta}_m),$$

where $\hat{\theta} = (\hat{\theta}_1', ..., \hat{\theta}_M')'$. Let $R(\hat{\theta}, w) = \int g(x) \log f(x, \hat{\theta}, w) dx$. Then the KL distance between g and $f(x, \hat{\theta}, w)$ is given by

$$KL(g, f(x, \hat{\theta}, w)) = \int g(x) \log g(x) dx - R(\hat{\theta}, w), \tag{2.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

in which the first term is unrelated to candidate models. Thus, the optimal weights which minimize (2.2) will minimize $-R(\hat{\theta}, w)$. Correspondingly, the theoretical density averaging for g is

$$f(x, \theta_0, w) = \sum_{m=1}^{M} w_m f_m(x, \theta_{m0}), \tag{2.3}$$

where $\theta_0 = (\theta'_{10}, ..., \theta'_{M0})'$, with the optimal vector of weights $w^0 = (w_{10}, ..., w_{M0})'$, which minimizes $KL(g, f(x, \theta_0, w))$, minimizing $-R(\theta_0, w)$, w.r.t. $w \in \mathcal{W}$.

Before ending this section, we comment that the DMA procedure based density function looks like the form of the well-known finite mixture models (FMM) for density estimation (McLachlan and Peel, 2000; Chen and Khalili, 2008), but they are essentially different. Note that the FMM method estimates the true density $f(x)$ that is supposed to be a mixture of a finite number of density functions, i.e., $f(x) = \sum_{j=1}^{S} \lambda_j f_j(x, \theta_j)$, where S is some positive integer, λ_j is the mixing probability for the jth component $f_j(x, \theta_j)$, and θ_j is the associated parameters in $f_j(x, \theta_j)$. Usually, the mixture components are assumed to be from the same class, such as Gaussian (i.e., all of $f_j(x, \theta_j)$, $1 \leq j \leq S$, are Gaussian). The parameters $\{\lambda_j, \theta_j\}$ are estimated by maximizing the log-likelihood function $\sum_{i=1}^{n} \log\left\{\sum_{j=1}^{S} \lambda_j f_j(X_i, \theta_j)\right\}$. To achieve such a maximum, the EM algorithm is often needed. Differently from the FMM, the DMA combines different density functions by the asymptotically optimal weights in the sense of minimizing the KL distance, hence it will extract useful information as much as possible with fea-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

tures of different density functions from the data. Moreover, the implementation of the DMA is very convenient without need of the EM algorithm. In addition, the order S of the FMM is selected by some criterion (e.g., BIC) in practice and thus some drawbacks of model selection may also remain for the FMM. To some extent, an FMM can be viewed as a special version of the DMA based density estimation method; see Remark 1.

With the optimal criterion framework above, unfortunately the optimal weights in (2.3) are infeasible to estimate from minimizing (2.2) because of the unknown $R(\hat{\theta}, w)$ depending on the (unknown) true density g. We will hence need to consider how to estimate them first.

3. Estimation of optimal density model averaging

In the following, we will suggest an estimator of $R(\hat{\theta}, w)$, and the feasible weights will be obtained based on this estimator. Let

$$Q_n = E(R(\hat{\theta}, w)) = E \int g(x) \log f(x, \hat{\theta}, w) dx.$$

Then the empirical estimator of Q_n is defined as

$$\widehat{Q}_n(\hat{\theta}, w) = n^{-1} \sum_{i=1}^{n} \log f(X_i, \hat{\theta}, w).$$

We first consider the parametric case in Subsection 3.1 and extension to nonparametric case in Subsection 3.2.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

3.1 Parametric density model averaging estimation

3.1.1 Estimation

To study $E(\widehat{Q}_n(\hat{\theta}, w) - Q_n)$, we will first consider the parametric density averaging by giving the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$, where $\theta_0 = (\theta'_{10}, ..., \theta'_{M0})'$. To facilitate the derivation of the weight choice criterion, we assume that the dimension of parameters $\sum_{m=1}^{M} p_m$ is fixed, but it is not necessary for the theoretical analysis. Let

$$u_m(x, \theta_m) = \frac{\partial \log f_m(x, \theta_m)}{\partial \theta_m},$$

$$I_m(x, \theta_m) = \frac{\partial^2 \log f_m(x, \theta_m)}{\partial \theta_m \theta'_m},$$

$J_m(\theta_{m0}) = -EI_m(X, \theta_{m0})$, and $\widehat{U}_m(\theta_{m0}) = n^{-1} \sum_{i=1}^{n} u_m(X_i, \theta_{m0})$ with

$$E\left(u_m(X_i, \theta_{m0})\right) = 0 \tag{3.1}$$

by the definition of θ_{m0} in (2.1). Further, denote $J^{-1}(\theta_0) = diag\left(J_1^{-1}(\theta_{10}), ..., J_M^{-1}(\theta_{M0})\right)$, $\widehat{U}(\theta_0) = \left(\widehat{U}_1(\theta_{10})', ..., \widehat{U}_M(\theta_{M0})'\right)'$, and $K(\theta_0) = p\lim_{n\to\infty} \mathrm{Var}(\sqrt{n}\widehat{U}(\theta_0))$. Then, we obtain from Lemma 1 in Section S2 of the supplementary material that

$$\hat{\theta} = \theta_0 + J^{-1}(\theta_0)\widehat{U}(\theta_0) + o_p(n^{-1/2}), \tag{3.2}$$

and

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N_k(0, J^{-1}(\theta_0)K(\theta_0)J^{-1}(\theta_0)), \tag{3.3}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

the k-dimensional ($k = \sum_{m=1}^{M} p_m$) normal distribution with mean 0 and covariance matrix $J^{-1}(\theta_0)\ K(\theta_0)J^{-1}(\theta_0)$.

Now denote

$$u(x, \theta, w) = \frac{\partial \log f(x, \theta, w)}{\partial \theta},$$

and

$$I(x, \theta, w) = \frac{\partial^2 \log f(x, \theta, w)}{\partial \theta \theta'},$$

and let $Q(\theta_0, w) = \int g(x) \log f(x, \theta_0, w) dx$, $Z(X_i, \theta_0, w) = \log f(X_i, \theta_0, w) - Q(\theta_0, w)$, $\widehat{Z}(\theta_0, w) = n^{-1} \sum_{i=1}^{n} Z(X_i, \theta_0, w)$, and $\widehat{U}(\theta_0, w) = n^{-1} \sum_{i=1}^{n} u(X_i, \theta_0, w)$.

By Taylor expansion and (3.3), we have

$$\begin{aligned} \widehat{Q}_n(\hat{\theta}, w) &\approx n^{-1} \sum_{i=1}^{n} \Big\{ \log f(X_i, \theta_0, w) + u(X_i, \theta_0, w)'(\hat{\theta} - \theta_0) \\ &\quad + \frac{1}{2}(\hat{\theta} - \theta_0)' I(X_i, \theta_0, w)(\hat{\theta} - \theta_0) \Big\} \\ &= Q(\theta_0, w) + \widehat{Z}(\theta_0, w) + \widehat{U}(\theta_0, w)'(\hat{\theta} - \theta_0) \\ &\quad - \frac{1}{2}(\hat{\theta} - \theta_0)' \widehat{J}(\theta_0, w)(\hat{\theta} - \theta_0), \end{aligned} \tag{3.4}$$

where $\widehat{J}(\theta_0, w) = -n^{-1} \sum_{i=1}^{n} I(X_i, \theta_0, w)$.

In addition, it follows from Taylor expansion and (3.3) again that

$$\begin{aligned} R(\hat{\theta}, w) &\approx \int g(x) \Big\{ \log f(x, \theta_0, w) + u(x, \theta_0, w)'(\hat{\theta} - \theta_0) \\ &\quad + \frac{1}{2}(\hat{\theta} - \theta_0)' I(x, \theta_0, w)(\hat{\theta} - \theta_0) \Big\} dx \end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

$$= Q(\theta_0, w) - \frac{1}{2}(\hat{\theta} - \theta_0)' J(\theta_0, w)(\hat{\theta} - \theta_0) + \int g(x) u(x, \theta_0, w)'(\hat{\theta} - \theta_0) dx, \tag{3.5}$$

where $J(\theta_0, w) = -EI(X_i, \theta_0, w)$.

Supposing that $E \|I(X_i, \theta_0, w)\| < \infty$, we have $\widehat{J}(\theta_0, w) \xrightarrow{p} J(\theta_0, w)$ because $\{X_i, i = 1, ..., n\}$ are i.i.d. Then, by (3.4) and (3.5), we have

$$\widehat{Q}_n(\hat{\theta}, w) - R(\hat{\theta}, w) \approx \widehat{Z}(\theta_0, w) + \widehat{U}(\theta_0, w)'(\hat{\theta} - \theta_0) - \int g(x) u(x, \theta_0, w)'(\hat{\theta} - \theta_0) dx. \tag{3.6}$$

Therefore,

$$\begin{aligned} & E(\widehat{Q}_n(\hat{\theta}, w) - Q_n) \\ \approx\ & E\left\{ \widehat{U}(\theta_0, w)'(\hat{\theta} - \theta_0) - \int g(x) u(x, \theta_0, w)'(\hat{\theta} - \theta_0) dx \right\} \\ =\ & E\left[\left\{ \widehat{U}(\theta_0, w) - \int g(x) u(x, \theta_0, w) dx \right\}' \left(\hat{\theta} - \theta_0 \right) \right] \\ \approx\ & E\left[\left\{ \widehat{U}(\theta_0, w) - \int g(x) u(x, \theta_0, w) dx \right\}' \left\{ J^{-1}(\theta_0) \widehat{U}(\theta_0) \right\} \right] \\ =\ & trE\left[\left\{ J^{-1}(\theta_0) \widehat{U}(\theta_0) \right\} \left\{ \widehat{U}(\theta_0, w) - \int g(x) u(x, \theta_0, w) dx \right\}' \right], \end{aligned} \tag{3.7}$$

where the first and second approximations are obtained from (3.6) and $E\left(\widehat{Z}(\theta_0, w)\right) = 0$, and (3.2), respectively.

Let $Z = (Z_1', Z_2')' = \left[\left\{ J^{-1}(\theta_0) \widehat{U}(\theta_0) \right\}', \left\{ \widehat{U}(\theta_0, w) - \int g(x) u(x, \theta_0, w) dx \right\}' \right]'$. Supposing that $u_m(X, \theta_{m0})$ and $u(X, \theta_0, w)$ have finite covariance matrices, by

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

the Central Limit Theorem (CLT) for i.i.d. variables, $\sqrt{n}Z$ has an asymptotic normal distribution $N(0, \Sigma)$, where

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}$$

with $\Sigma_{11} = \text{Var}\left(\sqrt{n}Z_1\right)$, $\Sigma_{22} = \text{Var}\left(\sqrt{n}Z_2\right)$, and

$$\Sigma_{12} = \text{Cov}\left(\sqrt{n}Z_1, \sqrt{n}Z_2\right). \tag{3.8}$$

Accordingly, (3.7) can be simply expressed as

$$E\left(\widehat{Q}_n(\hat{\theta}, w) - Q_n\right) \approx tr\left(\Sigma_{12}\right)/n. \tag{3.9}$$

In Section S1 of the supplementary material, we show that $tr\left(\Sigma_{12}\right)$ can be approximated by

$$\sum_{m=1}^{M} E\left\{\frac{w_m f_m(X, \theta_{m0})}{f(X, \theta_0, w)}\right\} tr\left\{J_m^{-1}(\theta_{m0}) K_m(\theta_{m0})\right\}, \tag{3.10}$$

where $K_m(\theta_{m0}) = E\left\{u_m(X, \theta_{m0}) u_m(X, \theta_{m0})'\right\}$. As a result, combining (3.9) and (3.10), we yield that

$$nE\left(\widehat{Q}_n(\hat{\theta}, w) - Q_n\right) \approx \sum_{m=1}^{M} n^{-1} \sum_{i=1}^{n} \left\{\frac{w_m f_m(X_i, \widehat{\theta}_m)}{f(X_i, \widehat{\theta}, w)}\right\} tr\left\{\widehat{J}_m^{-1}(\widehat{\theta}_m) \widehat{K}_m(\widehat{\theta}_m)\right\},$$

where

$$\widehat{J}_m(\hat{\theta}_m) = -n^{-1} \sum_{i=1}^{n} I_m(X_i, \widehat{\theta}_m)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

and

$$\widehat{K}_m(\hat{\theta}_m) = n^{-1}\sum_{i=1}^{n} u_m(X_i, \widehat{\theta}_m)u_m(X_i, \widehat{\theta}_m)'.$$

Thus, we propose the following density model averaging (DMA) criterion

$$\begin{aligned} C^{\mathrm{DMA}}(w) &= -n\widehat{Q}_n(\hat{\theta}, w) + \sum_{m=1}^{M} n^{-1}\sum_{i=1}^{n}\left\{\frac{w_m f_m(X_i, \widehat{\theta}_m)}{f(X_i, \widehat{\theta}, w)}\right\} tr\left\{\widehat{J}_m^{-1}(\widehat{\theta}_m)\widehat{K}_m(\widehat{\theta}_m)\right\} \\ &= -\sum_{i=1}^{n}\log f(X_i, \widehat{\theta}, w) + \sum_{m=1}^{M} w_m d_m, \end{aligned} \tag{3.11}$$

in which,

$$d_m = n^{-1}\sum_{i=1}^{n}\left\{\frac{f_m(X_i, \widehat{\theta}_m)}{f(X_i, \widehat{\theta}, w)}\right\} tr\left\{\widehat{J}_m^{-1}(\widehat{\theta}_m)\widehat{K}_m(\widehat{\theta}_m)\right\}. \tag{3.12}$$

The resultant weight estimator is defined as $\hat{w} = \mathrm{argmin}_{w\in\mathcal{W}} C^{\mathrm{DMA}}(w)$.

Remark 1. If the weight vector is set as that with the m-th element being one and others being zeros, then the DMA criterion becomes the following form

$$-\sum_{i=1}^{n}\log f_m(X_i, \widehat{\theta}_m) + tr\left\{\widehat{J}_m^{-1}(\widehat{\theta}_m)\widehat{K}_m(\widehat{\theta}_m)\right\},$$

which is the Takeuchi information criterion (TIC) (Takeuchi, 1976). Furthermore, if $f_m(x, \theta_{m0})$ is the true density function, then the DMA criterion is the AIC by (S1.1) in the supplementary material. In addition, when the candidate density set consists of the mixture models (discussed in Section 1) with different number of mixing components, the DMA based density estimator becomes an

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

FMM averaging estimator and reduces to the usual FMM for the special weight vector mentioned above. So FMM can be included in our current framework (see Section 4.2 for the pertinent example). An exhaustive analysis (theoretical and practical) on DMA for this case is left for future research.

Remark 2. The implementation of our method is convenient and the computational burden is not heavy. In addition, when M is very large, some methods can be utilized to screen candidate density functions before performing our procedure. For instance, we can use the 'top m' method of Yuan and Yang (2005) to obtain the candidate density functions, which have 'top m' AIC or BIC scores. The resultant number of candidate density functions is at most $2m$.

Remark 3. Note that d_m in (3.12) can be viewed as the penalty for the mth density model, and the term in d_m, $n^{-1}\sum_{i=1}^{n}\left\{f_m(X_i,\widehat{\theta}_m)/f(X_i,\widehat{\theta},w)\right\}$, is an adjusted factor which varies for different models. Intuitively, this adjusted factor is the averaged proportion of the mth candidate density in the density averaging estimator, which means that if the proportion is large, the penalty d_m for this density model will be also large. Thus, the DMA will overcome the overfitting automatically.

In the above analyses, the specific forms of the candidate density functions are not assumed. Two illustrating examples on the proposed method DMA can be found in Section S7 of the supplementary material.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

3.1.2 Asymptotic properties

Denote $\mathrm{KL}(w) = \mathrm{KL}(g, f(x, \hat{\theta}, w))$. In the following, we establish the asymptotic optimality for DMA.

Let $\lambda_{\min}(B)$ and $\lambda_{\max}(B)$ be the minimum and maximum singular value of a general real matrix B. Define $\|B\| = tr^{1/2}(B'B)$ for a real matrix B. Let $\mathrm{KL}^*(w) = \mathrm{KL}(g, f(x, \theta_0, w))$, $\xi_n = \inf_{w \in \mathcal{W}} \mathrm{KL}^*(w)$, and Θ_m be the compact parameter space of θ_m. $\aleph$ denotes some neighbourhood of θ_0. In the following, c and C denote two generic constants.

We need the following regularity conditions for Theorem 1.

Condition (C.1). The random variable X has a bounded compact support $\mathcal{C}$. Furthermore, $c \leq \inf_{x \in \mathcal{C}} \inf_{\theta_m \in \Theta_m} f_m(x, \theta_m) \leq \sup_{x \in \mathcal{C}} \sup_{\theta_m \in \Theta_m} f_m(x, \theta_m) \leq C$ uniformly in m $(1 \leq m \leq M)$.

Condition (C.2). $\sum_{i=1}^{n} \left\| \frac{\partial \log f_m(X_i, \theta_m)}{\partial \theta_m} |_{\theta_m = \theta_{m0}} \right\| = O_p(n)$ for $m = 1, ..., M$.

Condition (C.3). $E\left[\sup_{w \in \mathcal{W}} \sup_{\theta \in \aleph} \lambda_{\max}\left\{\frac{\partial^2 \log f(X_i, \theta, w)}{\partial \theta \partial \theta'}\right\}\right] < C$.

Condition (C.4). $E\left\{\sup_{w \in \mathcal{W}} \sup_{\theta \in \aleph} \left\| \frac{\partial \log f(X, \theta, w)}{\partial \theta} \right\|\right\} < C$.

Condition (C.5). $\lambda_{\max}\left\{\widehat{K}_m(\widehat{\theta}_m)\right\} = O_p(1)$, and $\lambda_{\max}\left\{\widehat{J}_m^{-1}(\widehat{\theta}_m)\right\} = O_p(1)$, uniformly in m.

Condition (C.6). $1/(n^{\frac{1-\delta}{2}} \xi_n) = O(1)$ for some $0 < \delta < 1$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

The first part of Condition (C.1) is on the boundedness of random variable X which holds in most real applications. Since $\mathcal{C}$ is bounded, the second part of Condition (C.1) is reasonable. The similar condition can be found in Chen et al. (2018). Condition (C.2) is an assumption on the boundedness of $1/n\sum_{i=1}^{n}\|\partial \log f_m(X_i,\theta_m)/\partial\theta_m|_{\theta_m=\theta_{m0}}\|$, which can be rigorously established under certain regularity conditions. Condition (C.3) is similar to Assumption 1.6 of Hansen (2016) if $f(x,\theta,w)$ is the true density for X_i. Conditions (C.4) is a standard moment bound for the asymptotic theory. Condition (C.5) implies that maximum singular values of $\widehat{K}_m(\widehat{\theta}_m)$ and $\widehat{J}_m^{-1}(\widehat{\theta}_m)$ are bounded. As an example, supposing that the m^*th model ($1 \le m^* \le M$), $f_{m^*}(x,\theta_{m^*0})$ with $\theta_{m^*0}=(\theta_{m^*01},\theta_{m^*02})'$, is the true density of normal distribution $N(\theta_{m^*01},\theta_{m^*02})$ generating data $\{X_i\}$, we have $\lambda_{\max}\{K_{m^*}(\theta_{m^*0})\} = \max\{1/\theta_{m^*02}, 1/(2\theta_{m^*02}^2)\}$ and $\lambda_{\max}\left\{J_{m^*}^{-1}(\theta_{m^*0})\right\} = \max\{\theta_{m^*02}, 2\theta_{m^*02}^2\}$, and hence Condition (C.5) is reasonable since $\widehat{K}_{m^*}(\widehat{\theta}_{m^*})$ and $\widehat{J}_{m^*}(\widehat{\theta}_{m^*})$ can approach $K_{m^*}(\theta_{m^*0})$ and $J_{m^*}(\theta_{m^*0})$ under some conditions, respectively. Condition (C.6) illustrates that $n\xi_n$ increases at a rate faster than $n^{1/2}$, which is similar to Condition C.3 of Zhang et al. (2016).

Theorem 1. *Under Conditions (C.1)-(C.6) and Conditions 1-2 in the supple-*

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

mentary material, for fixed $p_{\max} = \max\{p_1, ..., p_M\}$ *and* M*, we have*

$$\frac{KL(\widehat{w})}{\inf_{w \in \mathcal{W}} KL(w)} \xrightarrow{p} 1,$$

as $n \to \infty$*.*

Theorem 1 shows that the DMA based density averaging estimator is asymptotically optimal in the sense of achieving the lowest KL loss.

Next, we consider the case with diverging $p_{\max}$ and M, and establish the following theorem. We list the additional conditions for Theorem 2.

Condition (C.7). $\max_{1 \le m \le M} \sum_{i=1}^{n} \left\| \frac{\partial \log f_m(X_i, \theta_m)}{\partial \theta_m} |_{\theta_m = \theta_{m0}} \right\| = O_p(n p_{\max}^{1/2})$.

Condition (C.8). $E\left\{ \sup_{w \in \mathcal{W}} \sup_{\theta \in \aleph} \left\| k^{-1/2} \frac{\partial \log f(X, \theta, w)}{\partial \theta} \right\| \right\} = O(1)$, where $k = \sum_{m=1}^{M} p_m$.

Condition (C.9). $\max_{1 \le m \le M} \|\hat{\theta}_m - \theta_{m0}\| = O_p(p_{\max}^{1/2} n^{-1/2})$.

Condition (C.10). $p_{\max} M / (n^{\frac{1-\delta}{2}} \xi_n) = O(1)$ for some $0 < \delta < 1$.

Conditions (C.7) and (C.8) are the corresponding versions of Conditions (C.2) and (C.4), respectively, for the cases with diverging M and $p_{\max}$. Condition (C.9) is a high level condition which can be proved under some weak conditions; see White (1982) and Zhang et al. (2016) for the relevant verifications. Condition (C.10) is stronger than Condition (C.6) and is the same as it when $p_{\max} M$ is fixed.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

Theorem 2. *Under Conditions (C.1), (C.3), (C.5), and (C.7)-(C.10), if* $M\log(n^{\frac{1-\delta}{2}}\log n)/n^{\delta}\to 0$, *then for diverging* $p_{\max}$ *and* M,

$$\frac{KL(\widehat{w})}{\inf_{w\in\mathcal{W}} KL(w)}\xrightarrow{p}1,$$

as $n\to\infty$.

Theorem 2 shows that, under regularity conditions, the asymptotic optimality established in Theorem 1 still holds for diverging $p_{\max}$ and M.

In the following, we study the consistency and convergence rate of the DMA based weights. Recall that $\mathrm{KL}^*(w)=\mathrm{KL}(g,f(x,\theta_0,w))$ and the optimal weight vector $w^0=\mathrm{argmin}_{w\in\mathcal{W}}\mathrm{KL}^*(w)$. We suppose that w^0 is an interior point of $\mathcal{W}$ for the following theorem, which gives the rate of the DMA based weights tending to the infeasible optimal weight vector w^0. We consider the case with diverging $p_{\max}$ and M. Obviously, the theory is also valid for fixed $p_{\max}$ and M.

We list the conditions for Theorem 3. Let $F(X_i,\theta)=(f_1(X_i,\theta_1),...,f_M(X_i,\theta_M))'$ and $F(X_i,\widehat{\theta})=(f_1(X_i,\widehat{\theta}_1),...,f_M(X_i,\widehat{\theta}_M))'$.

Condition (C.11). $E\left\{\sup_{1\le m\le M} f_m^4(X_i,\theta_{m0})\right\}=O(1)$.

Condition (C.12). $E\left\{\sup_{w\in\mathcal{W}}\sup_{\theta\in\aleph}\left\|k^{-1/2}\frac{\partial\log f(X_i,\theta,w)}{\partial\theta}\right\|^2\right\}=O(1)$.

Condition (C.13). $E\left\{\sup_{w\in\mathcal{W}}\sup_{\theta\in\aleph} f^{-4}(X_i,\theta,w)\right\}=O(1)$.

Condition (C.14). $\left(E\sup_{\theta\in\aleph}\left[\lambda_{\max}\left\{\frac{\partial F(X_i,\theta)}{\partial\theta'}\right\}\right]^4\right)^{1/4}=O(M^{1/2}k^{1/2})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

Condition (C.15). $kp_{\max}Mn^{-1} = O(1)$.

Condition (C.16). For some positive constant κ_1, $\lambda_{\min}\left(\sum_{i=1}^{n} F(X_i,\widehat{\theta})F(X_i,\widehat{\theta})'/n\right) > \kappa_1 > 0$ in probability tending to 1.

Conditions (C.11), (C.12) and (C.13) are standard moment bounds for the asymptotic theory, where Condition (C.12) is similar to and sightly stronger than Condition (C.8). Furthermore, it is evident that under Condition (C.1), we have $E\left\{\sup_{1\le m\le M} f_m^4(X_i,\theta_{m0})\right\} = O(1)$ and $E\left\{\sup_{w\in\mathcal{W}}\sup_{\theta\in\aleph} f^{-4}(X_i,\theta,w)\right\} = O(1)$, i.e., Conditions (C.11) and (C.13) are naturally satisfied. Therefore, Conditions (C.11) and (C.13) can be replaced by Condition (C.1) but the formers are relatively weaker. Condition (C.14) is mild since $\frac{\partial F(X_i,\theta)}{\partial\theta'}$ is the $M\times k$ matrix. Conditions (C.15) restricts the relationship among $\{p_{\max}, M, n\}$; for fixed $p_{\max}$ and M, Condition (C.15) is naturally satisfied. Condition (C.16) requires that the minimum singular value of $\sum_{i=1}^{n} F(X_i,\widehat{\theta})F(X_i,\widehat{\theta})'/n$ is bounded asymptotically (Fan and Peng, 2004; Bickel and Levina, 2008).

Theorem 3. *If Conditions (C.5), (C.9) and (C.11)-(C.16) hold, then there exists a local minimizer $\hat{w}$ of $C^{DMA}(w)$ such that*

$$\|\hat{w} - w^0\| = O_p\left(k^{1/2}p_{\max}^{1/2}Mn^{-1/2+\alpha}\right),$$

where $k = \sum_{m=1}^{M} p_m$ and $\alpha < 1/2$ is a positive constant.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

Theorem 3 shows that $\hat{w}$ approaches the optimal weight vector w^0 at the rate no slower than $k^{1/2}p_{\max}^{1/2}Mn^{-1/2+\alpha}$, where $\alpha > 0$ can be sufficiently small. For fixed $p_{\max}$ and M, we have $\|\hat{w} - w^0\| = O_p\left(n^{-1/2+\alpha}\right)$.

3.2 Extension to semiparametric density averaging estimation

As introduced in Section 1, the kernel density estimation (KDE) is one of the most important density estimation approaches. As a nonparametric method, KDE may capture some useful information that could not be utilized by parametric DMA in some cases, which is also reflected in our simulation study. So in this section, we extend the proposed DMA method above to semiparametric density averaging estimation by combining parametric DMA and KDE. Specifically, let the kernel density estimator of the true density function $g(x)$ be

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),$$

where $K(\cdot)$ is a kernel function and h represents the bandwidth, which can be determined by some bandwidth selection methods, such as the rules of thumb, unbiased cross validation, biased cross validation, and smoothed bootstrap (see Jones et al. (1996) and the references therein). Then, the semiparametric density averaging estimator is defined as $\tilde{f}(x,\rho) = (1-\rho)f(x,\hat{\theta},\hat{w}) + \rho f_h(x)$, where ρ is a tuning parameter satisfying $0 \le \rho \le 1$, which controls the balance between DMA and KDE.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

Further, let the KL distance between g and $\tilde{f}(x,\rho)$ be

$$\begin{aligned}\mathrm{KL}(g,\tilde{f}(x,\rho)) &= \int g(x)\log\frac{g(x)}{\tilde{f}(x,\rho)}dx \\ &= \int g(x)\log g(x)dx - \int g(x)\log\tilde{f}(x,\rho)dx,\end{aligned}$$

where $\int g(x)\log g(x)dx$ is unrelated to ρ. Clearly, the ideal ρ is obtained by minimizing $-\int g(x)\log\tilde{f}(x,\rho)dx$, but unfortunately, it is not feasible in practice. To address the issue of selecting ρ, we suggest the following cross validation approach. Let $\tilde{f}^{[-i]}(X_i,\rho) = (1-\rho)f(X_i,\hat{\theta}^{[-i]},\hat{w}^{[-i]}) + \rho f_h^{[-i]}(X_i)$, where $\hat{\theta}^{[-i]}$, $\hat{w}^{[-i]}$ and $f_h^{[-i]}(X_i)$ are the same as $\hat{\theta}$, $\hat{w}$ and $f_h(X_i)$, respectively, except that the formers are obtained with X_i removed. Then, the feasible estimator $\tilde{\rho}$ of ρ is obtained by

$$\tilde{\rho} = \underset{0\le\rho\le1}{\operatorname{argmin}}\left\{-\sum_{i=1}^{n}\log\tilde{f}^{[-i]}(X_i,\rho)\right\}.$$

The resultant semiparametric density averaging estimator is given by

$$\tilde{f}(x,\tilde{\rho}) = (1-\tilde{\rho})f(x,\hat{\theta},\hat{w}) + \tilde{\rho}f_h(x).$$

Next, we explore the convergence of $\tilde{\rho}$. Define the optimal weight coefficient for the semiparametric density averaging estimator as

$$\rho_0 = \underset{0\le\rho\le1}{\operatorname{argmin}}\left\{\mathrm{KL}\left(g,\tilde{f}^0(x,\rho)\right)\right\},$$

where $\tilde{f}^0(x,\rho) = (1-\rho)f(x,\theta_0,w^0) + \rho Ef_h(x)$. We suppose that the minimum of $\lim_{n\to\infty}\mathrm{KL}\left(g,\tilde{f}^0(x,\rho)\right)$ on $[0,1]$ is achieved uniquely.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

We list the conditions for Theorem 4.

Condition (C.17). $n^{-1}\sum_{i=1}^{n}\left\{\log \tilde{f}^{[-i]}(X_i,\rho)-\log \tilde{f}^{0}(X_i,\rho)\right\}\xrightarrow{a.s.}0$ uniformly in ρ.

Condition (C.18). There exists a function $H(x)$ such that $EH(X)<\infty$, and $\left|\log \tilde{f}^{0}(x,\rho)\right|<H(x)$ for all x and ρ.

Condition (C.19). There exists a continuous function $T(\rho)$ such that

$$\sup_{0\le\rho\le 1}\left|E\left[\log \tilde{f}^{0}(X,\rho)\right]-T(\rho)\right|\to 0.$$

Condition (C.17) requires that the density estimator DMA (KDE) based on the data with the ith observation removed approaches the limit of the entire data based DMA (KDE). Although such a condition is a high-level assumption, it is reasonable for the large sample cases. Similar conditions can be found in Hansen and Racine (2012). Condition (C.18) is the regularity assumption on the moment bound; see, for example, Ferguson (1996) and Hansen (2016). Condition (C.19) is mild because $E\left[\log \tilde{f}^{0}(X,\rho)\right]$ approaches its limit when n is sufficiently large.

Theorem 4. *If Conditions (C.17) - (C.19) are satisfied, then we have* $\tilde{\rho}-\rho_0\xrightarrow{p}0$.

Theorem 4 shows that the weight estimator that combines the DMA and KDE approaches the optimal weight in the sense of the KL distance for large sample sizes.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

4. Numerical data examples

We demonstrate our proposed method by both simulation and real data examples.

4.1 Simulations

In the context of density estimation, AIC and BIC have been widely used, e.g., for the selection of finite mixture model; see Leroux (1992), Ishwaran et al. (2001), among others. Specifically, AIC and BIC choose the density functions minimizing

$$\mathrm{AIC}^{(m)} = -2\sum_{i=1}^{n} \log f_m(X_i, \widehat{\theta}_m) + 2p_m,$$

and

$$\mathrm{BIC}^{(m)} = -2\sum_{i=1}^{n} \log f_m(X_i, \widehat{\theta}_m) + p_m \log n,$$

respectively, and SAIC assigns weights

$$w_m = exp(-\mathrm{AIC}^{(m)}/2)/\sum_{m=1}^{M} exp(-\mathrm{AIC}^{(m)}/2)$$

to the mth density function. SBIC is the same as SAIC except with $\mathrm{AIC}^{(m)}$ replaced by $\mathrm{BIC}^{(m)}$. Also, we consider the corrected AIC criterion developed by Saumard and Navarro (2021) and such a criterion is labeled as IAIC. Specifically, IAIC consists of two terms, where the first term is the negative logarithm likelihood for the mth candidate model and the second term (i.e., the penalty term) for the mth model is $(1+Ca_m)p_m$, where p_m is the dimension of parametric vector in

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

the mth model, $a_m = \max\{\sqrt{p_m \log(n+1)/n}, \sqrt{\log(n+1)/p_m}, \log(n+1)/p_m\}$, $1 \le m \le M$, and C is a constant. As discussed by Saumard and Navarro (2021), considering the prediction accuracy and computation efficiency, we take $C = 1$.

As we know, the kernel density estimation (KDE) is commonly used in practice and the bandwidth choice is necessary for KDE. Here, we perform KDE using Gaussian kernel and the bandwidths are selected by the four well-known methods: the rules of thumb, unbiased cross validation, biased cross validation and the method of Sheather and Jones (1991) (Jones et al., 1996). To obtain such four bandwidth selectors, the four functions (bw.nrd, bw.ucv, bw.bcv and bw.SJ) with all the default settings in R package "stats" are used and the associated KDEs are denoted as K.nrd, K.ucv, K.bcv and K.SJ, respectively. In addition, to implement the FMM method (i.e., the finite mixture model) mentioned before, we apply the "densityMclust" function with all the default setting in the R language, which performs the density estimation by finite mixture of Gaussian components. Also, we consider the adaptive mixing strategy for the density aggregation (Yang, 2000;Yang, 2004), which is labeled as ADA. To be specific, divide the sample into two parts with the first part D_1 and the second part D_2, and then the ADA weight for the jth density function is given by $w_j = \frac{\pi_j \prod_{i \in D_2} \hat{f}_j(X_i; D_1)}{\sum_{m=1}^{M} \pi_m \prod_{i \in D_2} \hat{f}_m(X_i; D_1)}$, where $\hat{f}_j(X_i; D_1)$ is the estimator of true density $f(X_i)$ $(i \in D_2)$ based on the data in D_1 and $\pi_j \ge 0$ satisfying $\sum_{j=1}^{M} \pi_j = 1$. Data splittings are done 100 times and

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

then the averaged weight estimator is computed for the ADA method. We set $\pi_j = 1/M$, $j = 1, ..., M$, and the sample size of the first part to be $0.5n$ for ADA. In this section, we compare the finite sample performance of these density estimation methods.

To evaluate these methods, we calculate the KL loss as follows:

$$n_0^{-1} \sum_{i=1}^{n_0} \left\{ \log g(X_i^*) - \log f(X_i^*, \hat{\theta}, w) \right\},$$

where $n_0 = 500$ and $\{X_i^*\}_1^{n_0}$ are independent of $\{X_i\}_1^n$ ($n = 200, 300, 500$) but from the same distribution as $\{X_i\}_1^n$.

The candidate density functions include those of the normal, log-normal, exponential, gamma, Cauchy distributions, and the mixture models with two and three mixing components (the variances in the mixing components are set to be equal). Assume that the true distribution function of X is given by the following cases:

1. Mixture of Log-normal distribution (LN (μ, σ^2) with mean μ and standard deviation σ of the logarithm) and normal distribution (N (μ, σ^2) with mean μ and standard deviation σ) $\lambda_1 \times$LN $(0.5, 0.5^2) + \lambda_2 \times$N $(4, 0.5^2)$ for $\lambda_1 = 0, 0.3, 0.5, 0.7$ and $\lambda_2 = 1 - \lambda_1$;

2. Mixture of LN distribution and gamma distribution (Gamma (a,b) with parameters shape a and scale $1/b$) $\lambda_1 \times$LN $(0.5, 0.5^2) + \lambda_2 \times$Gamma $(2, 1)$;

3. Mixture of LN and beta distributions $\lambda_1 \times$LN $(0.5, 0.5^2) + \lambda_2 \times$Beta $(2, 2)$,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

where the density of Beta (a, b) is $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$ with $a > 0$, $b > 0$ and $0 < x < 1$;

4. Mixture of beta distribution and exponential distribution (E (a) with mean $1/a$) $\lambda_1 \times$Beta $(2, 2) + \lambda_2 \times$ E(1).

Based on 500 replications, the simulation results are shown in Figures 1-4 in Section S10 of the supplementary material. Because either K.nrd or K.bcv performs the best among the four KDE methods, i.e., K.nrd, K.ucv, K.bcv, K.SJ, for the most cases considered here, we only present the results for K.nrd and K.bcv.

The findings are summarized as follows:

(1) DMA, and AIC/BIC/SAIC/SBIC/IAIC/ADA. From Figures 1-4, we see that, DMA dominates AIC, BIC, SAIC and SBIC for most cases with $\lambda_1 \neq 0$, and the superiority is evident. Moreover, Figures 1-4 show that DMA performs better than IAIC for most cases with $\lambda_1 \neq 0$ and the improvement over IAIC is often remarkable. For instance, in Case 1, when $\lambda_1 = 0.3$, DMA is superior to IAIC with a little margin; when $\lambda_1 = 0.5$ and 0.7, the superiority of DMA is obvious. Also, as expected, the risks of IAIC and DMA reduce with the sample size n increasing.

From Figures 1-4, it is seen that, for $\lambda_1 \neq 0$, ADA and DMA are comparable for Cases 2 and 3, but DMA often performs better in Cases 1 and 4 especially

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

for the relatively large sample sizes (e.g., $n = 500$). Also, these two methods are superior to the model selection methods for most cases. This implies that model averaging will often reduce the prediction risk relative to model selection. Moreover, using the optimal weights may further reduce such risks.

Further, when $\lambda_1 = 0$, all the model selection and averaging methods (i.e., AIC, BIC, SAIC, SBIC, IAIC, ADA and DMA) have similar performance. This is reasonable. In fact, in Cases 1, 2, and 4, $\lambda_1 = 0$ means that the candidate model set includes the true density model, which often has the best performance in the current simulation study. Accordingly, both model selection and averaging methods tend to choose the true density model and hence have the close results. In Case 3, although the true density model is not included in the candidate set, the true distribution of the data has a simple form and the true density can be approximated well by a single candidate density, which leads to the similar simulation results for all the model selection and averaging methods.

(2) DMA and FMM. For Case 1 with $\lambda_1 = 0$ shown in Figure 1 and Case 4 with $\lambda_1 = 0.7$ shown in Figure 4, DMA is comparable or slightly inferior to FMM. However, for the other cases, DMA usually outperforms FMM by a large margin.

(3) DMA and KDE (K.nrd/K.bcv). For most cases, DMA performs better than K.nrd and K.bcv remarkably; see Figures 1-4. In addition, K.bcv has better performance than DMA in Case 1 with $\lambda_1 = 0.7$ for $n = 500$ but the variation of

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

DMA is often smaller (see Figure 1); for Case 3 with $\lambda_1 = 0$, the DMA and KDE methods have similar performance and the former produces the slightly lower risks.

In addition, we have also conducted a simulation comparison of DMA and the logarithmic scoring rule (LS) of Hall and Mitchell (2007) discussed in the Introduction section, where the LS based weight vector is obtained by minimizing the first part of (3.11) over $\mathcal{W}$, i.e., the LS density estimator is $f(x, \hat{\theta}, \tilde{w})$ with $\tilde{w} = \arg\min_{w \in \mathcal{W}} \{-\sum_{i=1}^{n} \log f(X_i, \widehat{\theta}, w)\}$. Let $n = 50, 100$ and 200. The other settings are the same as those described before. We only present the simulation results for Case 1 (reported in Figure 5 in Section S10 of the supplementary material), and the simulation results for the other cases are similar to those for Case 1. From Figure 5, it is observed that DMA is clearly superior to LS for most cases with small and moderate sample sizes ($n = 50, 100$) and they are comparable for the relatively large sample size ($n = 200$).

Overall, the DMA has a desirable performance in our simulation study, which suggests that DMA is worth recommending for density estimation.

4.2 Empirical applications

In this section, we investigate the life length data in Roman Egypt which have been analysed based on model selection by Claeskens and Hjort (2008). This data

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

set contains the age at death for 141 Egyptian mummies in the Roman period dating from around year 100 B.C. For this life length data, Claeskens and Hjort (2008) tried using AIC to select an appropriate density model. Although the selected model is the best in terms of AIC score, it seems not perfect for this data since some fluctuations are not captured by the selected model as Claeskens and Hjort (2008) commented. Here, we utilize the DMA method to obtain a more appropriate density estimation.

Specifically, as in Example 2.6 of Claeskens and Hjort (2008), the candidate set includes the density functions of the exponential, gamma, and log-normal distributions used in Section 4.1, and the Gompertz Models 1 and 2 whose density functions are $f(t) = e^{-\int_0^t h(s)ds}h(t)$ with $h(t) = ae^{bt}$ and $k + ae^{bt}(k + ae^b > 0)$, respectively, where t is the life length and $\{a, b, k\}$ denote the parameters. Claeskens and Hjort (2008) concluded that the Gompertz Model 1 is the best in terms of AIC and perhaps an acceptable approximation, but Claeskens and Hjort (2008) also realized that some fluctuations for this data, such as the extra mortality at age around 25, were not captured by such a model. The histogram for the full data and the Gompertz Model 1 based density estimation curve indicate this issue (see the left graph of Figure 6 in Section S10 of the supplementary material). To make a more accurate estimation, we include the mixture models with two and three mixing components (the variances in the mixing components

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

are set to be unequal) in the candidate set, because both seem to be sensible for describing the structure of this data; see the right graph of Figure 6. So, there are seven candidate models totally. Now, we use the model selection and averaging methods (AIC, BIC, SAIC, SBIC, LS, IAIC, ADA and DMA), KDE methods (K.nrd, K.ucv, K.bcv, and K.SJ) and FMM to estimate the density for this data, where the implementations of these methods are the same as in Section 4.1 (except that considering the relatively small sample size, we set the size of the first part sample for ADA to be $0.9n$ so that it can be performed). To assess the estimation accuracy, the data are randomly split into two parts with the first being the training sample of size $n = 100$ and the rest for testing. The log-likelihood of the test sample (denoted by EL) is computed as $\sum_{i=1}^{141-n} \log \hat{f}(X_i^*)$, where $\{X_i^*\}$ is the test sample and $\hat{f}(X_i^*)$ represents a density estimator. We repeat this process 2000 times for the eleven methods and then obtain the distributions of the EL values.

It is seen from Figure 7 in Section S10 of the supplementary material that DMA, IAIC, LS, AIC, SAIC and FMM have clearly much higher EL values (say, in median) than the other methods. Although the variations of BIC, SBIC, K.nrd, K.bcv and K.SJ are slightly smaller than that of DMA, the DMA method is superior to them remarkably in terms of median of EL values. Further, in order to compare the performance of DMA with those of IAIC, LS, AIC, SAIC and

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

FMM, we examine the kernel density estimates (by R code 'density') of the EL values for DMA, IAIC, LS, AIC, SAIC and FMM, which are plotted in Figure 8 in Section S10 of the supplementary material. It is clear from Figure 8 that the density of the EL values of DMA looks very different from those of IAIC, LS, AIC, SAIC and FMM (which appear to be quite similar). The former has much higher peaks and lighter tails, implying that our proposed DMA method performs more stable. So DMA performs the best for this data globally. Finally, for the full data, the DMA based density estimation curve is shown on the left of Figure 6, which seems to be a more reasonable approximation to the true density than the Gompertz Model 1 used by Claeskens and Hjort (2008).

5. Conclusion

In this paper, we have developed a model averaging procedure of density estimation, DMA, for optimal extraction of information from data under unsupervised learning. The asymptotic optimality of the DMA estimator has been established. The convergence rate of the DMA based weights tending to the optimal weights minimizing the KL distance has also been derived. We have also proposed a semiparametric density averaging method, which combines parametric and nonparametric density estimators in a data driven fashion. Sufficient simulation trials show that the DMA has desirable finite sample performance that is robust for

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

different simulating density models. The real data example also supports the DMA method.

Many issues deserve to be further investigated. For instance, we propose the DMA and derive its asymptotic properties for the independent and identically distributed case. How to extend the DMA method to the dependent data case warrants the future researches.

In addition, different nonparametric density estimations (e.g., the kernel method, local likelihood method and wavelet estimation) may have different advantages, so how to combine these nonparametric estimation methods will also be an interesting topic for study.

Supplementary Material

The Supplementary Material contains the derivation of $tr\left(\Sigma_{12}\right)$, Lemma 1, the proofs of all theorems, some illustrating examples, the explanations on the technical conditions, the discussion on the different density aggregation methods and the numerical results (Figures 1–8).

Acknowledgments

We thank the editor, associate editor and two referees for their helpful comments and suggestions. This work was partially supported by the National Natural Science Foundation of China (Grant nos. 12001534, 12426308, 12031016

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

and 71971131). Lu's work was partially supported by the European Research Agency's Marie Curie Career Integration Grant (Grant no. PCIG14-GA-2013-631692). Zou's work was also partially supported by the Beijing Outstanding Young Scientist Program (Grant no. JWZQ20240101027).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Caski, F., editors, *Second International Symposium on Information Theory*. Akademiai Kiado, Budapest.

Ando, T. and Li, K. C. (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association*, 109:254–265.

Baimuratov, I., Shichkina, Y., Stankova, E., Zhukova, N., and Than, N. (2019). A Bayesian information criterion for unsupervised learning based on an objective prior. In *Computational Science and Its Applications – ICCSA 2019*. Springer International Publishing.

Bickel, P. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics*, 36:199–227.

Chen, J. and Khalili, A. (2008). Order selection in finite mixture models with a nonsmooth penalty. *Journal of the American Statistical Association*, 103:1674–1683.

Chen, J., Li, D., Linton, O., and Lu, Z. (2018). Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. *Journal of the American Statistical Association*, 113:919–932.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

Cheng, X. and Hansen, B. E. (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics*, 186:280–293.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32:928–961.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman and Hall, London.

Geweke, J. and Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, 164:130–141.

Hall, S. G. and Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, 23:1–13.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, 75:1175–1189.

Hansen, B. E. (2008). Least squares forecast averaging. *Journal of Econometrics*, 146:342–350.

Hansen, B. E. (2014). Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics*, 5:495–530.

Hansen, B. E. (2016). Efficient shrinkage in parametric models. *Journal of Econometrics*, 190:115–132.

Hansen, B. E. and Racine, J. (2012). Jackknife model averaging. *Journal of Econometrics*, 167:38–46.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American*

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

Statistical Association, 98:879–899.

Ishwaran, H., James, L., and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96:1316–1332.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91:401–407.

Leroux, B. (1992). Consistent estimation of a mixing distribution. *Annals of Statistics*, 20:1350–1360.

Li, D., Linton, O., and Lu, Z. (2015). A flexible semiparametric forecasting model for time series. *Journal of Econometrics*, 187:345–357.

Liao, J., Zong, X., Zhang, X., and Zou, G. (2019). Model averaging based on leave-subject-out cross-validation for vector autoregressions. *Journal of Econometrics*, 209:35–60.

Liu, Q. and Okui, R. (2013). Heteroskedasticity-robust c_p model averaging. *The Econometrics Journal*, 16:463–472.

Maggioni, M. and Murphy, J. M. (2019). Learning by unsupervised nonlinear diffusion. *Journal of Machine Learning Research*, 20:1–56.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

Safarinejadian, B., Menhaj, M. B., and Karrari, M. (2010). Distributed unsupervised gaussian mixture learning for density estimation in sensor networks. *IEEE Transactions on Instrumentation and Measurement*, 59:2250–2260.

Saumard, A. and Navarro, F. (2021). Finite sample improvement of Akaike's information criterion.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0410

IEEE Transactions on Information Theory, 67:6328–6343.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, 53:683–690.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku [Mathematical Sciences] (in Japanese)*, 153:12–18.

Wan, A. T. K., Zhang, X., and Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics*, 156:277–283.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25.

Yang, Y. (2000). Mixing strategies for density estimation. *The Annals of Statistics*, 28:75–87.

Yang, Y. (2004). Combining forecasting procedures: Some theoretical results. *Econometric Theory*, 20:176–222.

Yuan, Z. and Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association*, 100:1202–1214.

Zhang, X., Yu, D., Zou, G., and Liang, H. (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association*, 111:1775–1790.

Zhang, X., Zou, G., and Carroll, R. (2015). Model averaging based on Kullback–Leibler distance. *Statistica Sinica*, 25:1583–1598.