Nearest Neighbor Dirichlet Mixtures

Shounak Chattopadhyay   SHOUNAK.CHATTOPADHYAY@DUKE.EDU
Department of Statistical Science
Duke University
Durham, NC 27708-0251, USA

Antik Chakraborty   ANTIK015@PURDUE.EDU
Department of Statistics
Purdue University
West Lafayette, IN 47907, USA

David B. Dunson   DUNSON@DUKE.EDU
Department of Statistical Science
Duke University
Durham, NC 27708-0251, USA

Editor: Erik Sudderth

Abstract

There is a rich literature on Bayesian methods for density estimation, which characterize the unknown density as a mixture of kernels. Such methods have advantages in terms of providing uncertainty quantification in estimation, while being adaptive to a rich variety of densities. However, relative to frequentist locally adaptive kernel methods, Bayesian approaches can be slow and unstable to implement in relying on Markov chain Monte Carlo algorithms. To maintain most of the strengths of Bayesian approaches without the computational disadvantages, we propose a class of nearest neighbor-Dirichlet mixtures. The approach starts by grouping the data into neighborhoods based on standard algorithms. Within each neighborhood, the density is characterized via a Bayesian parametric model, such as a Gaussian with unknown parameters. Assigning a Dirichlet prior to the weights on these local kernels, we obtain a pseudo-posterior for the weights and kernel parameters. A simple and embarrassingly parallel Monte Carlo algorithm is proposed to sample from the resulting pseudo-posterior for the unknown density. Desirable asymptotic properties are shown, and the methods are evaluated in simulation studies and applied to a motivating data set in the context of classification.

Keywords: Bayesian, Density estimation, Distributed computing, Embarrassingly parallel, Kernel density estimation, Mixture model, Quasi-posterior, Scalable

1. Introduction

Bayesian nonparametric methods provide a useful alternative to black box machine learning algorithms, having potential advantages in terms of characterizing uncertainty in inferences and predictions. However, computation can be slow and unwieldy to implement. Hence, it is important to develop simpler and faster Bayesian nonparametric approaches, and *hybrid* methods that borrow the best of both worlds. For example, if one could use the Bayesian machinery for uncertainty quantification and reduction of mean square errors

through shrinkage, while incorporating algorithmic aspects of machine learning approaches, one may be able to engineer a highly effective hybrid. The focus of this article is on proposing such an approach for density estimation, motivated by the successes and limitations of nearest neighbor algorithms and Bayesian mixture models.

Nearest neighbor algorithms are popular due to a combination of simplicity and performance. Given a set of n observations $\mathcal{X}^{(n)} = (X_1, \ldots, X_n)$ in $\mathbb{R}^p$, the density at x is estimated as $\hat{f}_{\mathrm{knn}}(x) = k/(nV_pR_k^p)$, where k is the number of neighbors of x in $\mathcal{X}^{(n)}$, $R_k = R_k(x)$ is the distance of x from its kth nearest neighbor in $\mathcal{X}^{(n)}$, and V_p is the volume of the p-dimensional unit ball (Loftsgaarden and Quesenberry, 1965; Mack and Rosenblatt, 1979). Refer to Biau and Devroye (2015) for an overview of related estimators and corresponding theory.

Nearest neighbor density estimators are a type of locally adaptive kernel density estimators. The literature on such methods identifies two broad classes: **balloon estimators** and **sample smoothing estimators**; see Scott (2015); Terrell and Scott (1992) for an overview. **Balloon estimators** characterize the density at a query point x using a bandwidth function $h(x)$; classical examples include the naive k-nearest neighbor density estimator (Loftsgaarden and Quesenberry, 1965) and its modification in Mack and Rosenblatt (1979). More elaborate balloon estimators face challenges in terms of choice of $h(x)$ and obtaining density estimators that do not integrate to 1. **Sample smoothing estimators** use n different bandwidths $h(X_i)$, one for each sample point X_i, to estimate the density at a query point x globally. By construction, sample smoothing estimators are *bona fide* density functions integrating to 1. To fit either the balloon or the sample smoothing estimator, one may compute an initial pilot density estimator employing a constant bandwidth and then use this pilot to estimate the bandwidth function (Breiman et al., 1977; Abramson, 1982). Another example of a locally adaptive density estimator is the local likelihood density estimator (Loader, 1996, 2006; Hjort and Jones, 1996), which fits a polynomial model in the neighborhood of a query point x to estimate the density at x, estimating the parameters of the local polynomial by maximizing a penalized local log-likelihood function. The above methods produce a point estimate of the density without uncertainty quantification (UQ).

Alternatively, there is a Bayesian literature on locally adaptive kernel methods, which express the unknown density as:

$$f(x) = \sum_{h=1}^m \pi_h \mathcal{K}(x; \theta_h), \quad \theta_h \sim P_0, \quad (\pi_h)_{h=1}^m \sim Q_0, \tag{1}$$

which is a mixture of m components, with the hth having probability weight π_h and kernel parameters θ_h; by allowing the location and bandwidth to vary across components, local adaptivity is obtained. A Bayesian specification is completed with prior P_0 for the kernel parameters and Q_0 for the weights. In practice, it is common to rely on an over-fitted mixture model (Rousseau and Mengersen, 2011), which chooses m as a pre-specified finite upper bound on the number of components, and lets

$$\pi = (\pi_1, \ldots, \pi_m)^{\mathrm{T}} \sim \mathrm{Dirichlet}(\alpha, \ldots, \alpha). \tag{2}$$

Augmenting component indices $c_i \in \{1, \ldots, m\}$ for $i = 1, \ldots, n$, a simple Gibbs sampler can be used for posterior computation, alternating between sampling (i) c_i from a multinomial

2

conditional posterior, for $i = 1, \ldots, n$; (ii) $\theta_h \mid - \sim P_0(\theta_h) \prod_{i:c_i=h} \mathcal{K}(X_i; \theta_h)$; and (iii) $\pi \mid - \sim \text{Dirichlet}(\alpha + n_1, \ldots, \alpha + n_m)$, with $n_h = \sum_{i=1}^{n} \mathbb{I}(c_i = h)$ for $h = 1, \ldots, m$.

Relative to frequentist locally adaptive methods, Bayesian approaches are appealing in automatically providing a characterization of uncertainty in estimation, while having excellent practical performance for a broad variety of density shapes and dimensions. However, implementation typically relies on Markov chain Monte Carlo (MCMC), with the Gibbs sampler sketched above providing an example of a common algorithm used in practice. Unfortunately, current MCMC algorithms for posterior sampling in mixture models tend to face issues with *slow mixing*, meaning the sampler can take a very large number of iterations to adequately explore different posterior modes and obtain sufficiently accurate posterior summaries.

MCMC inefficiency has motivated a literature on faster approaches, including sequential approximations (Wang and Dunson, 2011; Zhang et al., 2014) and variational Bayes (Blei and Jordan, 2006). These methods are order dependent, tend to converge to local modes, and/or lack theory support. Newton and Zhang (1999); Newton (2002) instead rely on predictive recursion. Such estimators are fast to compute and have theory support, but are also order dependent and do not provide a characterization of uncertainty. Alternatively, one can use a Polya tree as a conjugate prior (Lavine, 1992, 1994), and there is a rich literature on related multiscale and recursive partitioning approaches, such as the optional Polya tree (Wong and Ma, 2010). However, Polya trees have disadvantages in terms of sensitivity to a base partition and a tendency to favor spiky/erratic densities. These disadvantages are inherited by most of the computationally fast modifications.

This article develops an alternative to current locally adaptive density estimators, obtaining the practical advantages of Bayesian approaches in terms of uncertainty quantification and a tendency to have relatively good performance for a wide variety of true densities, but without the computational disadvantage due to the use of MCMC. This is accomplished with a *Nearest Neighbor-Dirichlet Mixture* (NN-DM) model. The basic idea is to rely on fast nearest neighbor search algorithms to group the data into local neighborhoods, and then use these neighborhoods in defining a Bayesian mixture model-based approach. Section 2 outlines the NN-DM approach and describes implementation details for Gaussian kernels. Section 3 provides some theory support for NN-DM. Section 4 contains simulation experiments comparing NN-DM with a rich variety of competitors in univariate and multivariate examples, including an assessment of UQ performance. Section 5 contains a real data application, and Section 6 a discussion.

2. Methodology

2.1 Nearest Neighbor Dirichlet Mixture Framework

Let $d(x_1, x_2)$ denote a distance metric between data points $x_1, x_2 \in \mathcal{X}$. For $\mathcal{X} = \mathbb{R}^p$, the Euclidean distance is typically chosen. For each $i \in \{1, 2, \ldots, n\}$, let $X_{i[j]}$ denote the jth nearest neighbor to X_i in the data $\mathcal{X}^{(n)} = (X_1, \ldots, X_n)$ such that $d(X_i, X_{i[1]}) \leq \ldots \leq d(X_i, X_{i[n]})$, with ties broken by increasing order of indices. By convention, we define $X_{i[1]} = X_i$. The indices on the k nearest neighbors to X_i are denoted as $\mathcal{N}_i = \{j : d(X_i, X_j) \leq d(X_i, X_{i[k]})\}$. Denote the set of data points in the ith neighborhood by

$\mathcal{S}_i = \{X_j : j \in \mathcal{N}_i\}$. In implementing the proposed method, we typically let the number of neighbors k vary as a function of n. When necessary, we use the notation k_n to express this dependence. However, we routinely drop the n subscript for notational simplicity.

Fixing $x \in \mathcal{X}$, we model the density of the data within the ith neighborhood using

$$f_i(x) = \mathcal{K}(x; \theta_i), \quad \theta_i \sim P_0, \tag{3}$$

where θ_i are parameters specific to neighborhood i that are given a global prior distribution P_0. To combine the $f_i(x)$s into a single global $f(x)$, similarly to equations (1)-(2), we let

$$f(x) = \sum_{i=1}^{n} \pi_i f_i(x), \quad \pi = (\pi_i)_{i=1}^{n} \sim \text{Dirichlet}(\alpha, \dots, \alpha), \quad \theta_i \sim P_0. \tag{4}$$

The key difference relative to standard Bayesian mixture model (1) is that in (4) we include one component for each data sample and assume that only the data in the k-nearest neighborhood of sample i will inform about θ_i. In contrast, (1) lacks any sample dependence, and we infer allocation of samples to mixture components in a posterior inference phase.

Given the restriction that only data in the ith neighborhood $\mathcal{S}_i$ inform about θ_i, the pseudo-posterior density $\widetilde{\Pi}_1(\theta_i; \mathcal{S}_i, P_0)$ of θ_i with data $\mathcal{S}_i$ and prior P_0 is

$$\widetilde{\Pi}_1(\theta_i; \mathcal{S}_i, P_0) \propto P_0(\theta_i) \prod_{j \in \mathcal{N}_i} \mathcal{K}(X_j; \theta_i), \tag{5}$$

where the right-hand side of (5) is motivated from Bayes' theorem. This pseudo-posterior is in a simple analytic form if P_0 is conjugate to $\mathcal{K}(x; \theta)$. The prior P_0 can involve unknown parameters and borrows information across neighborhoods; this reduces the large variance problem common to nearest neighbor estimators.

Since the neighborhoods are overlapping, proposing a pseudo-posterior update for π under (4) is not straightforward. However, one can define the number of effective members in the ith neighborhood $\mathcal{S}_i$ similar in spirit to the number of points in the hth cluster in mixture models of the form (1). By convention, we define the point X_i that generated its neighborhood $\mathcal{S}_i$ to be an effective member of that neighborhood. For any other data point X_j to be a effective member of the neighborhood generated by X_i for $j \neq i$, we require $X_j \in \mathcal{S}_i$ but $X_j \notin \mathcal{S}_u$ for all $u = 1, \dots, n$ such that $u \notin \{i, j\}$. That is, X_j lies in the neighborhood generated by X_i but does not lie in the neighborhood of any other X_u for $u \notin \{i, j\}$. In Section 3.2, we show that the number of effective member points defined as above approaches 1 as $n \to \infty$. This motivates the following Dirichlet pseudo-posterior density $\widetilde{\Pi}_2(\pi; \mathcal{X}^{(n)})$ for the neighborhood weights π:

$$\widetilde{\Pi}_2(\pi; \mathcal{X}^{(n)}) = \text{Dirichlet}(\pi \mid \alpha + 1, \dots, \alpha + 1), \tag{6}$$

where $\text{Dirichlet}(p \mid q_1, \dots, q_d)$ denotes the density of the Dirichlet distribution evaluated at p with parameters $(q_1, \dots, q_d)$. We provide a justification for the pseudo-posterior update (6) in Section 3.2. This distribution is inspired from the conditional posterior on the kernel weights in the Dirichlet mixture of equations (1)-(2), but we use n components and fix the effective number of samples allocated to each component at one.

Based on equations (3)-(6), our nearest neighbor-Dirichlet mixture produces a pseudo-posterior distribution for the unknown density $f(x)$ through simple distributions for the parameters characterizing the density within each neighborhood and for the weights. To generate independent Monte Carlo samples from the pseudo-posterior for f, one can simply draw independent samples of $(\theta_i)_{i=1}^{n}$ and π from (5) and (6) respectively, and plug these samples into the expression for $f(x)$ in (4). The resulting mechanism can be described as

$$
\begin{aligned}
\theta_i &\stackrel{ind}{\sim} \widetilde{\Pi}_1(\cdot\,; \mathcal{S}_i, P_0) \quad \text{for } i = 1, \ldots, n, \\
\pi &\sim \widetilde{\Pi}_2(\cdot\,; \mathcal{X}^{(n)}) \\
f(x) &= \sum_{i=1}^{n} \pi_i \mathcal{K}(x; \theta_i).
\end{aligned}
\tag{7}
$$

In (7), we denote the induced pseudo-posterior distribution on f by $f \sim \widetilde{\Pi}$. Although this is not exactly a coherent fully Bayesian posterior distribution, we claim that it can be used as a practical alternative to such a posterior in practice. This claim is backed up by theoretical arguments, simulation studies, and a real data application in the sequel.

2.2 Illustration with Gaussian Kernels

Suppose we have independent and identically distributed (iid) observations $\mathcal{X}^{(n)}$ from the density f, where $X_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$ and f is an unknown density function with respect to the Lebesgue measure on $\mathbb{R}^p$ for $p \geq 1$. Let $\mathbb{R}_+^{p \times p}$ denote the set of all real-valued $p \times p$ positive definite matrices. Fix $x \in \mathbb{R}^p$. We proceed by setting $\mathcal{K}(x; \theta)$ to be the multivariate Gaussian density $\phi_p(x; \eta, \Sigma)$, given by

$$
\phi_p(x; \eta, \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left\{ -(x - \eta)^{\mathrm{T}} \Sigma^{-1} (x - \eta)/2 \right\},
$$

where $\theta = (\eta, \Sigma)$, $\eta \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}_+^{p \times p}$. We first compute the neighborhoods $\mathcal{N}_i$ corresponding to X_i as in Section 2.1 and place a normal-inverse Wishart (NIW) prior on $\theta_i = (\eta_i, \Sigma_i)$, given by $(\eta_i, \Sigma_i) \sim \mathrm{NIW}_p(\mu_0, \nu_0, \gamma_0, \Psi_0)$ independently for $i = 1, \ldots, n$. That is, we let

$$
\eta_i \mid \Sigma_i \sim \mathrm{N}\left(\mu_0, \frac{\Sigma_i}{\nu_0} \right), \quad \Sigma_i \sim \mathrm{IW}_p(\gamma_0, \Psi_0),
$$

with $\mu_0 \in \mathbb{R}^p$, $\nu_0 > 0$, $\gamma_0 > p - 1$ and $\Psi_0 \in \mathbb{R}_+^{p \times p}$; for details about parametrization see Section I of the Appendix.

Monte Carlo samples from the pseudo-posterior of $f(x)$ can be obtained using Algorithm 1. The corresponding steps for the univariate case are provided in Section H of the Appendix. Although the pseudo-posterior distribution of $f(x)$ lacks an analytic form, we can obtain a simple form for its pseudo-posterior mean by integrating over the pseudo-posterior distribution of $(\theta_i)_{i=1}^{n}$ and π. Recall the definitions of μ_i and Ψ_i from Step 2 of Algorithm 1 and define $\Lambda_i = \{\nu_n(\gamma_n - p + 1)\}^{-1}(\nu_n + 1)\,\Psi_i$. Then the pseudo-posterior mean of $f(x)$ is given by

$$
\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} t_{\gamma_n - p + 1}(x; \mu_i, \Lambda_i),
\tag{8}
$$

- **Step 1:** For $i = 1, \ldots, n$, compute the neighborhood $\mathcal{N}_i$ for data point $X_i \in \mathbb{R}^p$ according to distance $d(\cdot, \cdot)$ with $(k-1)$ nearest neighbors in $\mathcal{X}^{-i} = \mathcal{X}^{(n)} \setminus \{X_i\}$.

- **Step 2:** Update the parameters for neighborhood $\mathcal{N}_i$ to $(\mu_i, \nu_n, \gamma_n, \Psi_i)$, where $\nu_n = \nu_0 + k$, $\gamma_n = \gamma_0 + k$,

$$\mu_i = \frac{1}{\nu_n} \left(\nu_0 \mu_0 + k \bar{X}_i \right), \quad \bar{X}_i = \frac{1}{k} \sum_{j \in \mathcal{N}_i} X_j, \text{ and}$$

$$\Psi_i = \Psi_0 + \sum_{j \in \mathcal{N}_i} (X_j - \bar{X}_i)(X_j - \bar{X}_i)^{\mathrm{T}} + \frac{k \nu_0}{\nu_n} (\bar{X}_i - \mu_0)(\bar{X}_i - \mu_0)^{\mathrm{T}}.$$

- **Step 3:** To compute the t-th Monte Carlo sample $f^{(t)}(x)$ of $f(x)$, sample $\pi^{(t)} \sim \mathrm{Dirichlet}(\alpha + 1, \ldots, \alpha + 1)$ and $(\eta_i^{(t)}, \Sigma_i^{(t)}) \sim \mathrm{NIW}_p(\mu_i, \nu_n, \gamma_n, \Psi_i)$ independently for $i = 1, \ldots, n$, and set

$$f^{(t)}(x) = \sum_{i=1}^{n} \pi_i^{(t)} \phi_p \left(x; \eta_i^{(t)}, \Sigma_i^{(t)} \right).$$

Algorithm 1: Nearest neighbor-Dirichlet mixture algorithm to obtain Monte Carlo samples from the pseudo-posterior of $f(x)$ with Gaussian kernel and normal-inverse Wishart prior.

where $t_\gamma(x; \mu, \Lambda)$ for $x \in \mathbb{R}^p$ is the p-dimensional Student's t-density with degrees of freedom $\gamma > 0$, location $\mu \in \mathbb{R}^p$ and scale matrix $\Lambda \in \mathbb{R}_+^{p \times p}$. We proceed with using Gaussian kernels and NIW conjugate priors when implementing the NN-DM for the remainder of the paper.

2.3 Hyperparameter Choice

The hyperparameters in the prior for the neighborhood-specific parameters need to be chosen carefully — we found results to be sensitive to γ_0 and Ψ_0. If non-informative values are chosen for these key hyperparameters, we tend to inherit typical problems of nearest neighbor estimators including lack of smoothness and high variance. Suppose $\Sigma \sim \mathrm{IW}_p(\gamma_0, \Psi_0)$ and for $i, j = 1, \ldots, p$, let Σ_{ij} and $\Psi_{0,ij}$ denote the i, jth entry of Σ and Ψ_0, respectively. Then $\Sigma_{jj} \sim \mathrm{IG}(\gamma_*/2, \Psi_{0,jj}/2)$ where $\gamma_* = \gamma_0 - p + 1$. For $p = 1$, the $\mathrm{IW}_p(\gamma_0, \Psi_0)$ density simplifies to an $\mathrm{IG}(\gamma_0/2, \gamma_0 \delta_0^2/2)$ density with $\delta_0^2 = \Psi_0/\gamma_0$. Thus borrowing from the univariate case, we set $\Psi_{0,jj} = \gamma_* \delta_0^2$ and $\Psi_{0,ij} = 0$ for all $i \neq j$, which implies that $\Psi_0 = (\gamma_* \delta_0^2) \, \mathbb{I}_p$ and we use leave-one-out cross-validation to select the optimum δ_0^2. With p dimensional data, we recommend fixing $\gamma_0 = p$ which implies a multivariate Cauchy prior predictive density. We choose the leave-one-out log-likelihood as the criterion function for cross-validation, which is closely related to minimizing the Kullback-Leibler divergence between the true and estimated density (Hall, 1987; Bowman, 1984). The explicit expression for the pseudo-posterior mean in (8) makes cross-validation computationally efficient. The description of a fast implementation is provided in Section G of the Appendix.

The proposed method has substantially faster runtime if one uses a default choice of hyperparameters. In particular, we found the default values $\mu_0 = 0_p, \nu_0 = 0.001, \gamma_0 = p$, and $\Psi_0 = \mathbb{I}_p$ to work well across a number of simulation cases, especially when the true density is smooth. Although using cross-validation to estimate Ψ_0 can lead to improved performance when the underlying density is spiky, cross-validation provides little to no gains for smooth true densities. Furthermore, with low sample size and increasing number of dimensions, we found this improvement to diminish rapidly. In order to obtain desirable uncertainty quantification in simulations and applications, we found small values of α to work well. As a default value, we recommend using $\alpha = 0.001$ for small samples and moderate dimensions.

The other key tuning parameter for NN-DM is the number of nearest neighbors $k = k_n$. The pseudo-posterior mean in (8) reduces to a single $t_{\gamma_n - p + 1}$ kernel if $k_n = n$. In contrast, $k_n = 1$ provides a sample smoothing kernel density estimate with a specific bandwidth function (Terrell and Scott, 1992). Therefore, the choice of k can impact the smoothness of the density estimate. To assess the sensitivity of the NN-DM estimate to the choice of k, we investigate how the out-of-sample log-likelihood of a test set changes with respect to k in Section 4.7. These simulations suggest that the proposed method is quite robust to the exact choice of k. In practice with finite samples and small dimensions, we recommend a default choice of $k_n = \lfloor n^{1/3} \rfloor + 1$ and $k_n = 10$ for univariate and multivariate cases, respectively. These values led to good performance across a wide variety of simulation cases as described in Section 4.

3. Theory

3.1 Asymptotic Properties

There is a rich literature on asymptotic properties of the posterior measure for an unknown density under Bayesian models, providing a frequentist justification for Bayesian density estimation; refer, for example to Ghosal et al. (1999), Ghosal and van der Vaart (2007). Unfortunately, the tools developed in this literature rely critically on the mathematical properties of fully Bayes posteriors, providing theoretical guarantees for a computationally intractable exact posterior distribution under a Bayesian model. Our focus is instead on providing frequentist asymptotic guarantees for our computationally efficient NN-DM approach, with this task made much more complex by the dependence across neighborhoods induced by the use of a nearest neighbor procedure.

We first focus on proving pointwise consistency of the pseudo-posterior of $f(x)$ induced by (7) for each $x \in [0, 1]^p$, using Gaussian kernels as in Section 2.2. We separately study the mean and variance of the NN-DM pseudo-posterior distribution, first showing that the pseudo-posterior mean in (8) is pointwise consistent and then that the pseudo-posterior variance vanishes asymptotically. The key idea behind our proof is to show that the pseudo-posterior mean is asymptotically close to a kernel density estimator with suitably chosen bandwidth for fixed p and $k_n \to \infty$ at a desired rate. The proof then follows from standard arguments leading to consistency of kernel density estimators. The NN-DM pseudo-posterior mean mimics a kernel density estimator only in the asymptotic regime; in finite sample simulation studies (refer to Section 4), NN-DM has much better performance. The detailed proofs of all results in this section are in the Appendix.

Consider independent and identically distributed data $\mathcal{X}^{(n)}$ from a fixed unknown density f_0 with respect to the Lebesgue measure on $\mathbb{R}^p$ equipped with the Euclidean metric, inducing the measure P_{f_0} on $\mathcal{B}(\mathbb{R}^p)$. We use $\widetilde{E}\{f(x)\}$, $\widetilde{\mathrm{var}}\{f(x)\}$, and $\widetilde{\mathrm{pr}}\{f(x) \in B\}$ to denote the mean of $f(x)$, variance of $f(x)$, and probability of the event $\{f(x) \in B\}$ for $B \in \mathcal{B}(\mathbb{R}^p)$, respectively, under the pseudo-posterior distribution of $f(x)$ implied by (7). We make the following regularity assumptions on f_0:

Assumption 1 (Compact support) *f_0 is supported on $[0,1]^p$.*

Assumption 2 (Bounded gradient) *f_0 is continuous on $[0,1]^p$ with $||\nabla f_0(x)||_2 \leq L$ for all $x \in [0,1]^p$ and some finite $L > 0$.*

Assumption 3 (Bounded sup-norm) *$||\log(f_0)||_\infty < \infty$.*

Our asymptotic analysis relies on analyzing the behaviour of the pseudo-posterior updates within each nearest neighborhood. We leverage on key results from Biau and Devroye (2015); Evans et al. (2002) which are based on the assumption that the true density has compact support as in Assumption 1. Assumption 2 ensures that the kernel density estimator has finite expectation. Versions of this assumption are common in the kernel density literature; for example, refer to Tsybakov (2009). Assumptions 1 and 3 imply the existence of $0 < a_1, a_2 < \infty$ such that $0 < a_1 < f_0(x) < a_2 < \infty$ for all $x \in [0,1]^p$, which is referred to as a positive density condition by Evans (2008); Evans et al. (2002). This is used to establish consistency of the proposed method and justify the choice of the pseudo-posterior distribution of the weights. These assumptions are standard in the literature studying frequentist asymptotic properties of nearest neighbor and Bayesian density estimators.

For $i = 1, \ldots, n$, recall the definitions of μ_i and Λ_i from (8):

$$\mu_i = \frac{\nu_0}{\nu_n}\mu_0 + \frac{k_n}{\nu_n}\bar{X}_i, \quad \Lambda_i = \frac{\nu_n + 1}{\nu_n(\gamma_n - p + 1)}\Psi_i,$$

where $\nu_n = \nu_0 + k_n, \gamma_n = \gamma_0 + k_n$, and $\bar{X}_i, \Psi_i$ are as in Algorithm 1. Define the bandwidth matrix

$$H_n = h_n^2 \mathbb{I}_p, \quad \text{where } h_n^2 = \frac{(\nu_n + 1)(\gamma_0 - p + 1)}{\nu_n(\gamma_n - p + 1)}\delta_0^2. \tag{9}$$

We have suppressed the dependence of μ_i and Λ_i on n for notational convenience. It is immediate that $h_n^2 \to 0$ if $k_n \to \infty$ as $n \to \infty$. Fix $x \in [0,1]^p$. To prove consistency of the pseudo-posterior mean, we first show that $\hat{f}_n(x)$ and $f_K(x) = (1/n)\sum_{i=1}^n t_{\gamma_n - p + 1}(x; X_i, H_n)$ are asymptotically close, that is we show that $E_{P_{f_0}}(|\hat{f}_n(x) - f_K(x)|) \to 0$ as $n \to \infty$. To obtain this result, we approximate μ_i by X_i and Λ_i by H_n using successive applications of the mean value theorem. Finally, we exploit the convergence of $f_K(x)$ to the true value $f_0(x)$ to obtain the consistency of $\hat{f}_n(x)$. The proof of convergence of $f_K(x)$ to $f_0(x)$ is provided in Section F of the Appendix. The precise statement regarding the consistency of the pseudo-posterior mean is given in the following theorem. Let $a \wedge b$ denote the minimum of a and b.

Theorem 4 *Fix $x \in [0,1]^p$. Let $k_n = o(n^{i_0})$ with $i_0 = \{2/(p^2 + p + 2)\} \wedge \{4/(p+2)^2\}$ such that $k_n \to \infty$ as $n \to \infty$, and $\nu_0 = o\{n^{-2/p}k_n^{(2/p)+1}\}$. Then, $\hat{f}_n(x) \to f_0(x)$ in P_{f_0}-probability as $n \to \infty$.*

8

We now look at the pseudo-posterior variance of $f(x)$. We let

$$R_n = \frac{\Gamma\{(\gamma_n - p + 2)/2\}}{\Gamma\{(\gamma_n - p + 1)/2\}} \left[\frac{\nu_n + 2}{4\pi\nu_n(\gamma_n - p + 2)}\right]^{p/2} \text{ and } D_n = \frac{(\gamma_n - p + 1)(\nu_n + 2)}{2(\gamma_n - p + 2)(\nu_n + 1)}. \quad (10)$$

For $i = 1, \ldots, n$, let $B_i = D_n \Lambda_i$ and define

$$\widehat{f}_{var}(x) = \frac{1}{n} \sum_{i=1}^{n} t_{\gamma_n - p + 2}(x; \mu_i, B_i). \quad (11)$$

As $n \to \infty$, we have $D_n \to 1/2$. Analogous steps to the ones used in the proof of Theorem 4 can be used to imply that $\widehat{f}_{var}(x) \to f_0(x)$ in P_{f_0}-probability. Also, as $n \to \infty$, $k_n^{(p-1)/2} R_n = \mathcal{O}(1)$ using Stirling's approximation. We now provide an upper bound on the pseudo-posterior variance of $f(x)$ which shows convergence of the pseudo-posterior variance to 0.

Theorem 5 *Let H_n be the bandwidth matrix defined in* (9). *Let R_n, D_n be as in* (10) *and $\widehat{f}_{var}$ be as in* (11). *Under Assumptions 1-3 with x, k_n, and ν_0 as in Theorem 4, we have*

$$\widetilde{\text{var}}\{f(x)\} \leq \frac{R_n D_n^{-p/2} \widehat{f}_{var}(x)}{|H_n|^{1/2}} \left\{\frac{1}{n(\alpha + 1) + 1} + \frac{1}{n}\right\}. \quad (12)$$

This implies $\widetilde{\text{var}}\{f(x)\} \to 0$ in P_{f_0}-probability as $n \to \infty$.

Refer to Sections B and C in the Appendix for proofs of Theorems 4 and 5, respectively. Pointwise pseudo-posterior consistency follows from Theorems 4 and 5, as shown below.

Theorem 6 *Let f_0 satisfy Assumptions 1-3 with x, k_n and ν_0 as in Theorem 4. Fix $\epsilon > 0$ and define the ϵ-ball around $f_0(x)$ by $U_\epsilon = \{y_* : |y_* - f_0(x)| \leq \epsilon\}$. Let $\widetilde{\text{pr}}\{f(x) \in U_\epsilon^c\}$ denote the probability of the set U_ϵ^c under the pseudo-posterior distribution of $f(x)$ as induced by* (7). *Then $\widetilde{\text{pr}}\{f(x) \in U_\epsilon^c\} \to 0$ in P_{f_0}-probability as $n \to \infty$.*

Proof Fix $\epsilon > 0$ and consider the ϵ-ball $U_\epsilon = \{y_* : |y_* - f_0(x)| \leq \epsilon\}$. Then by Chebychev's inequality, we have $\widetilde{pr}\{f(x) \in U_\epsilon^c\} \leq [\{\widehat{f}_n(x) - f_0(x)\}^2 + \widetilde{\text{var}}\{f(x)\}]/\epsilon^2 \to 0$ in P_{f_0}-probability as $n \to \infty$, using Theorems 4 and 5. ∎

We next focus on the limiting distribution of $f(x)$ for the univariate case. From Section H of the Appendix, the pseudo-posterior distribution of (η_i, σ_i^2) for $i = 1, \ldots, n$ is given by $\text{NIG}(\mu_i, \nu_n, \gamma_n/2, \gamma_n \delta_i^2/2)$, where μ_i, ν_n, γ_n are as before and

$$\gamma_n \delta_i^2 = \gamma_0 \delta_0^2 + \sum_{j \in \mathcal{N}_i} (X_j - \bar{X}_i)^2 + \frac{k_n \nu_0}{\nu_n}(\bar{X}_i - \mu_0)^2.$$

We establish in Theorem 7 that the limiting distribution of $f(x)$ is a Gaussian distribution with appropriate centering and scaling. This allows interpretation of $100(1 - \beta)\%$ pseudo-credible intervals as $100(1 - \beta)\%$ frequentist confidence intervals on average for large n.

Theorem 7 *Fix $x \in [0,1]$. Suppose f_0 satisfies Assumptions 1-3 and also satisfies $|f_0^{(4)}(x)| \leq C_0$ for all $x \in [0,1]$ for some finite $C_0 > 0$. Let k_n satisfy $k_n = o(n^{2/7})$ such that $n^{-2/9}k_n \to \infty$, h_n be as in (9) satisfying $h_n \to 0$, and $\alpha = \alpha_n \to \infty$, as $n \to \infty$. For $t \in \mathbb{R}$, define*

$$G_n(t) = \widetilde{\mathrm{pr}}\left[(nh_n)^{1/2}\left\{f(x) - \left(f_0(x) + \frac{h_n^2 f_0^{(2)}(x)}{2}\right)\right\} \leq t\right].$$

Then, we have

$$\lim_{n\to\infty} E_{P_{f_0}}\{G_n(t)\} = \Phi\left(t\,;0, \frac{f_0(x)}{2\pi^{1/2}}\right),$$

where $\Phi(t\,;0,\sigma^2)$ denotes the cumulative distribution function of the $N(0,\sigma^2)$ density.

For a proof of Theorem 7, we refer the reader to Section D of the Appendix.

3.2 Pseudo-Posterior Distribution of Weights

We investigate the rationale behind the pseudo-posterior update (6) of the weight π, which has a symmetric prior distribution $\pi \sim \mathrm{Dirichlet}(\alpha, \ldots, \alpha)$ as motivated in Section 1. As discussed in Section 1, the conditional update for the weights π in a finite Bayesian mixture model with m components given the cluster allocation indices $\{c_1, \ldots, c_n\}$ is obtained by $\mathrm{Dirichlet}(\alpha + n_1, \ldots, \alpha + n_m)$, where α is the prior concentration parameter and $n_h = \sum_{i=1}^{n} \mathbb{I}(c_i = h)$ is the number of data points allocated to the hth cluster. This is not true in our case as the k_n-nearest neighborhoods have considerable overlap between them. Instead, we consider the number of effective member data points in each of these neighborhoods.

Define the k_n-nearest neighborhood of X_i to be the set $\mathcal{S}_i = \{X_j : d(X_i, X_j) \leq d(X_i, X_{i[k_n]})\}$ where $X_{i[k_n]}$ is the k_n-th nearest neighbor of X_i in the data $\mathcal{X}^{(n)}$, following the notation in Section 2.1. We assume $d(\cdot, \cdot)$ is the Euclidean metric from here on, and let $R_i = d(X_i, X_{i[k_n]}) = ||X_i - X_{i[k_n]}||_2$ denote the distance of X_i from its k_n-th nearest neighbor in $\mathcal{X}^{(n)}$.

Let N_i denote the number of effective members in $\mathcal{S}_i$ as defined in Section 2.1. Then, we can express N_i as

$$N_i = 1 + \sum_{j \neq i} \mathbb{I}\left[X_j \in \mathcal{S}_i\,, \bigcap_{u \notin \{i,j\}} \{X_j \notin \mathcal{S}_u\}\right], \tag{13}$$

where $\mathbb{I}(A)$ is the indicator function of the set A. Under $X_1, \ldots, X_n \overset{iid}{\sim} f_0$, we have

$$E_{P_{f_0}}(N_1) = 1 + (n-1)P_{f_0}\left[X_2 \in \mathcal{S}_1\,, \bigcap_{u=3}^{n}\{X_2 \notin \mathcal{S}_u\}\right], \tag{14}$$

by symmetry. Furthermore, N_i are identically distributed for $i = 1, \ldots, n$. We now state a result which provides a motivation for our choice of the pseudo-posterior update of π. For two sequences of real numbers (a_n) and (b_n), we write $a_n \sim b_n$ if $|a_n/b_n| \to c_0$ as $n \to \infty$ for some constant $c_0 > 0$.

Theorem 8 *Suppose $X_1, \ldots, X_n \overset{iid}{\sim} f_0$ with f_0 satisfying Assumptions 1-3. Furthermore, suppose that $k_n \sim n^{i_0 - \epsilon}$ for some $\epsilon \in (0, i_0)$, where i_0 is as defined in Theorem 4. Then,*

$$\lim_{n \to \infty} n \, P_{f_0} \left[X_2 \in \mathcal{S}_1, \bigcap_{u=3}^{n} \{X_2 \notin \mathcal{S}_u\} \right] = 0. \tag{15}$$

Proof of Theorem 8 is in Section E of the Appendix. The above theorem suggests we asymptotically have only one effective member per neighborhood $\mathcal{S}_i$, namely the point X_i that itself generated this neighborhood. This result motivates our choice of the pseudo-posterior update of the weight vector π. We illustrate uncertainty quantification of the proposed method in finite samples in Section 4.4 with this choice of pseudo-posterior update of the weight vector π.

4. Simulation Experiments

4.1 Preliminaries

In this section, we compare the performance of the proposed density estimator with several other standard density estimators through several numerical experiments. We evaluate estimation performance based on the expected $\mathcal{L}_1$ distance (Devroye and Gyorfi, 1985). For the pair $(f_0, \hat{f})$, where f_0 is the true data generating density and $\hat{f}$ is an estimator, the expected $\mathcal{L}_1$ distance is defined as $\mathcal{L}_1(f_0, \hat{f}) = E_{P_{f_0}} \{\int |f_0(x) - \hat{f}(x)| \, dx\}$. We compute an estimate $\hat{\mathcal{L}}_1(f_0, \hat{f})$ of $\mathcal{L}_1(f_0, \hat{f})$ in two steps. First, we sample n training points $X_1, \ldots, X_n \sim f_0$ and obtain $\hat{f}$ based on this sample, and then further sample n_t independent test points $X_{n+1}, \ldots, X_{n+n_t} \sim f_0$ and compute

$$\hat{L} = \frac{1}{n_t} \sum_{i=1}^{n_t} \left| \frac{\hat{f}(X_{n+i})}{f_0(X_{n+i})} - 1 \right|.$$

In the second step, to approximate the expectation with respect to P_{f_0}, the first step is repeated R times. Letting $\hat{L}_r$ denote the estimate for the rth replicate, we compute the final estimate as $\hat{\mathcal{L}}_1(f_0, \hat{f}) = (1/R) \sum_{r=1}^{R} \hat{L}_r$. Then, it follows that $\hat{\mathcal{L}}_1(f_0, \hat{f}) \to \mathcal{L}_1(f_0, \hat{f})$ as $n_t, R \to \infty$, by the law of large numbers. In our experiments, we set $n_t = 500$ and $R = 20$. We let 0_p and $\mathbb{1}_p$ denote the vector with all entries equal to 0 and the vector with all entries equal to 1 in $\mathbb{R}^p$, respectively, for $p \geq 1$.

All simulations were carried out using the R programming language (R Core Team, 2018). For Dirichlet process mixture models, we collect $2,000$ Markov chain Monte Carlo (MCMC) samples after discarding a burn-in of $3,000$ samples using the `dirichletprocess` package (J. Ross and Markwick, 2019). The default implementation of the Dirichlet process mixture model in p dimensions in the `dirichletprocess` package uses multivariate Gaussian kernels and has the base measure as $\mathrm{NIW}_p(0_p, p, p, \mathbb{I}_p)$ with the Dirichlet concentration parameter having the Gamma$(2, 4)$ prior (West, 1992). For the nearest neighbor-Dirichlet mixture, $1,000$ Monte Carlo samples are taken. For the kernel density estimator, we select the bandwidth by the default plug-in method `hpi` for univariate cases and `Hpi` for multivariate cases (Sheather and Jones, 1991; Wand and Jones, 1994) using the package `ks` (Duong, 2020). We additionally consider the k-nearest neighbor estimator studied in Mack

and Rosenblatt (1979), setting the number of neighbors $k = n^{1/2}$, and the variational Bayes (VB) approximation to Dirichlet process mixture models (Blei and Jordan, 2006). We also compare with the optional Polya tree (OPT) (Wong and Ma, 2010) using the package `PTT`. For univariate cases, we consider the recursive predictive density estimator (RD) from Hahn et al. (2018), Polya tree mixtures (PTM) using the package `DPpackage` (Jara et al., 2011), and the sample smoothing kernel density estimator (A-KDE) using the package `quantreg`. Lastly, we also compare with the local likelihood density estimator (LLDE) using the package `locfit` for both univariate and multivariate cases. Dirichlet process mixture model hyperparameter values are kept the same in both the MCMC and variational Bayes implementations, with the number of components of the variational family set to 10 for all cases. We denote the nearest neighbor-Dirichlet mixture, Dirichlet process mixture (DPM) implemented with MCMC, kernel density estimator, variational Bayes approximation to the DPM, and k-nearest neighbor density estimator by NN-DM, DP-MC, KDE, DP-VB, and KNN, respectively, in tables and figures.

4.2 Univariate Cases

We set $n = 200, 500$ with $k_n = \lfloor n^{1/3} \rfloor + 1$. We consider 10 choices of f_0 from the R package `benchden` (Mildenberger and Weinert, 2012); the specific choices are Cauchy (CA), claw (CW), double exponential (DE), Gaussian (GS), inverse exponential (IE), lognormal (LN), logistic (LO), skewed bimodal (SB), symmetric Pareto (SP), and sawtooth (ST) with default choices of the corresponding parameters. The prior hyperparameter choices for the proposed method are $\mu_0 = 0, \nu_0 = 0.001, \gamma_0 = 1$; δ_0^2 is chosen via the cross-validation method of Section 2.3. Detailed numerical results are deferred to Table 5 in the Appendix. Instead, in Figure 1, we provide a visual summary of the performance of each method under consideration by forming a box plot of the estimated $\mathcal{L}_1$ errors of the methods across all the data generating densities. Methods with lower median as indicated by the solid line of the box plot, and smaller overall spread are preferable as they provide higher accuracy and also maintain such accuracy across a collection of true density cases. Results of KNN are omitted in Figure 1 due to much higher values compared to other methods. For the KDE and RD estimator, the plot and the table exclude the results for the heavy-tailed densities CA, IE, and SP due to very high $\mathcal{L}_1$ errors.

Overall, a major advantage of the proposed method is its versatility among the considered methods. The Bayesian nonparametric methods DP-MC, DP-VB, PTM, OPT, and RD are often close to NN-DM in terms of their performance when the true densities are smooth and do not display locally spiky behavior. However, the NN-DM performs better than other methods in densities where such local behavior is present and performs very close to the best estimator for either the smooth heavy-tailed or thin-tailed densities. The KDE and RD perform well when data are generated from a smooth underlying density. However, there are some cases where the error for KDE and RD is very high. For instance, when $n = 500$ and f_0 is the standard Cauchy (CA) density, the estimated $\mathcal{L}_1$ error for the KDE is 38501.85 and the algorithm for the RD estimate did not converge. Both the KDE and RD also perform poorly in very spiky multi-modal densities such as the ST. Compared to the LLDE and the A-KDE, the NN-DM displays similar performance in heavy-tailed and smooth densities when $n = 200$, with the NN-DM performing better for the spiky densities.
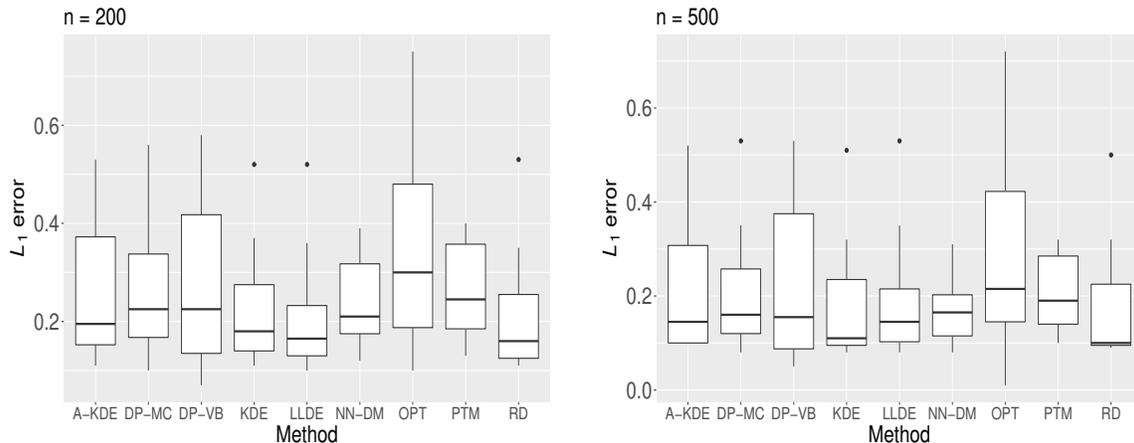
Figure 1: Box plots of $\hat{\mathcal{L}}_1(f_0, \hat{f})$ for the 10 different choices of the true density f_0 and different estimators $\hat{f}$ for univariate data. The box plots for KDE and RD exclude the heavy-tailed cases CA, IE, and SP.

However, when $n = 500$, the NN-DM shows significant improvements over the LLDE and the A-KDE for spiky densities such as the CW and the ST.

In Figure 2, we show the performance of the NN-DM estimator $\hat{f}_n$ (with hyperparameters chosen as described earlier) relative to the posterior mean under a DP-MC with default or hand-tuned hyperparameters, when 500 data points are generated from the sawtooth (ST) density. The Dirichlet process mixture with default hyperparameters is unable to detect the multiple spikes, merging adjacent modes to form larger clusters, perhaps due to inadequate mixing of the Markov chain Monte Carlo sampler or to the Gaussian kernels used in the mixture. As a result, we had to hand-tune the hyperparameters for the Dirichlet process mixture to obtain comparable performance with the NN-DM (without hand-tuning). We obtained the best results when changing the hyperparameters of the base measure of the DP-MC to NIG$(0, 0.01, 1, 1)$ while keeping the prior on α the same as before. This illustrates the deficiency of the DP-MC in estimating densities with spiky local behavior unless we hand-tune the hyperparameters, which requires knowledge of the true density. We also compare the performance of the two methods with a smoother test density in Figure 3, where the data are generated from a skewed bimodal (SB) distribution. Both the estimates are comparable, but the nearest neighbor-Dirichlet mixture provides better uncertainty quantification. Similar results are obtained for $n = 1000$, and hence are omitted.

4.3 Multivariate Cases

For the multivariate cases, we consider $n = 200$ and 1000. The number of neighbors is set to $k = 10$ and the dimension p is chosen from $\{2, 3, 4, 6\}$. Recall the definition of $\phi_p(x; \mu, \Sigma)$ from Section 2.2 and let $\Phi(x)$ be the cumulative distribution function of the standard Gaussian density. Let $S_0 = \rho \, \mathbb{1}_p \mathbb{1}_p^{\mathrm{T}} + (1 - \rho) \, \mathbb{I}_p$ with $\rho = 0.8$. Let $x = (x_1, \ldots, x_p)^{\mathrm{T}}$. We consider the following cases.

(1) *Mixture of Gaussians (MG)*: $f_0(x) = 0.4 \, \phi_p(x; m_1, S_0) + 0.6 \, \phi_p(x; m_2, S_0)$, where $m_1 =$
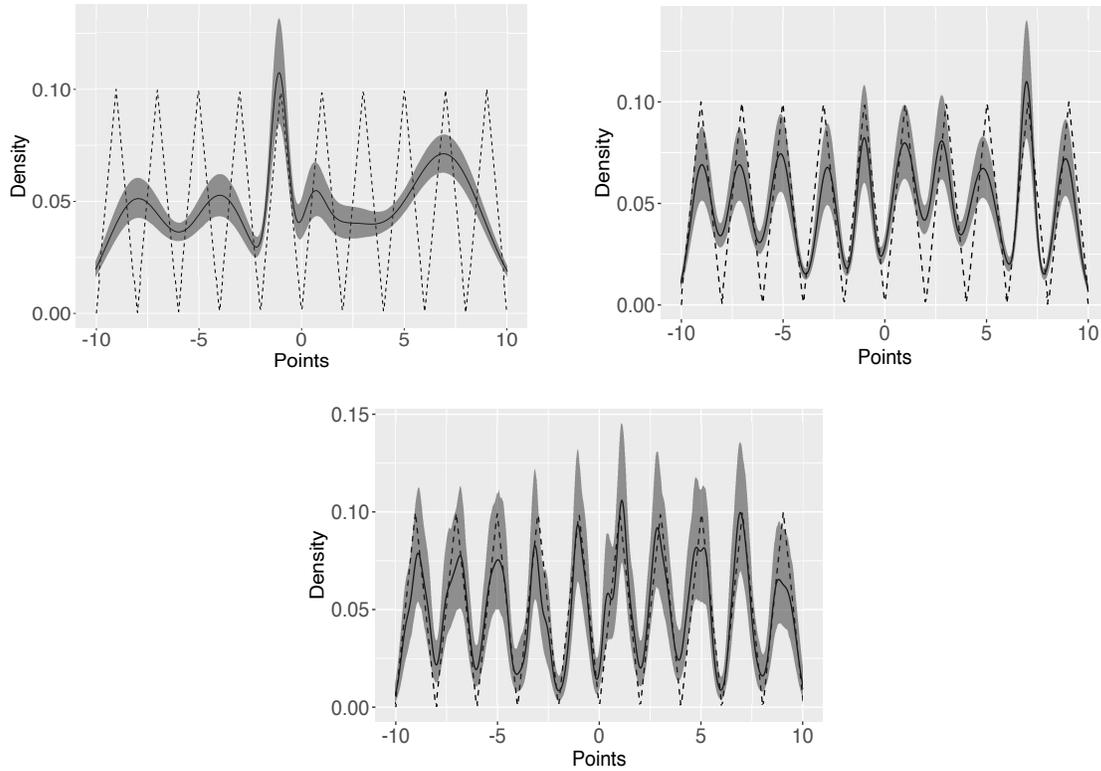
Figure 2: Plot comparing density estimates for the NN-DM and DP-MC for $n = 500$ samples generated from the sawtooth (ST) density. Shaded regions correspond to 95% (pseudo) posterior credible intervals. The true density is displayed using dotted lines. The top panel shows the performance of DP-MC with default hyperparameters on the left and with hand-tuned hyperparameters on the right. The bottom panel shows the performance of the NN-DM.
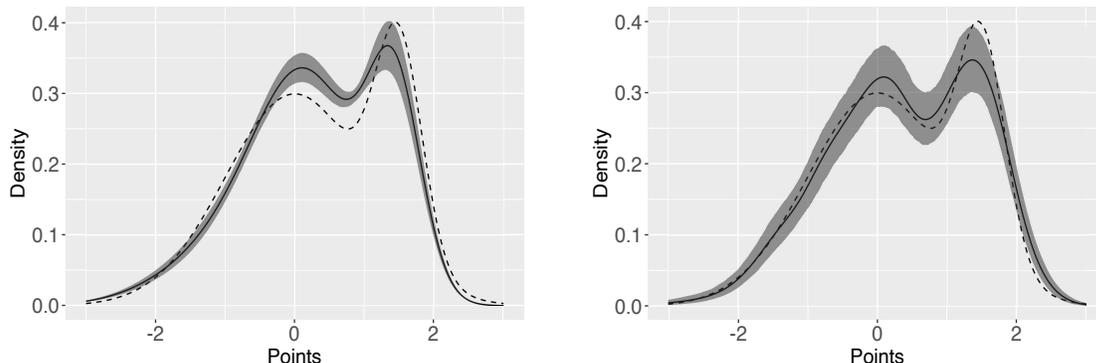
Figure 3: Similar to Figure 2, with data of sample size $n = 500$ generated from the skewed bimodal (SB) density. Left panel shows the DP-MC fit and the right panel shows the NN-DM fit.

$-2 \times \mathbb{1}_p, m_2 = 2 \times \mathbb{1}_p$.

(2) *Skew normal (SN)*: $f_0(x) = 2\phi_p(x; m_0, S_0)\Phi\{s_0^{\mathrm{T}} W^{-1}(x - m_0)\}$ (Azzalini, 2005), where W is the diagonal matrix with diagonal entries $W_{ii}^2 = S_{0,ii}$ for $i = 1, \ldots, p$. We choose $m_0 = 0_p$ and the skewness parameter vector $s_0 = 0.5 \times \mathbb{1}_p$.

(3) *Multivariate t-distribution (T)*: $f_0(x) = t_{d_0}(x; m_*, S_0)$ is the density of the p-dimensional multivariate Student's t-distribution. We set $d_0 = 10$ and $m_* = \mathbb{1}_p$.

(4) *Mixture of multivariate skew t-distributions (MST)*: $f_0(x) = 0.25\, t_{d_0}(x; m_1, S_0, s_0) + 0.75\, t_{d_0}(x; m_2, S_0, s_0)$. Here, $t_d(\cdot; \mu, S, s)$ is the skew t-density (Azzalini, 2005) with parameters d, μ, S, s, with d_0, s_0 defined as before and m_1, m_2 the same as in the first case.

(5) *Multivariate Cauchy (MVC)*: $f_0(x) \propto \{1 + (x - \mu_*)^{\mathrm{T}} S_0^{-1}(x - \mu_*)\}$ where $\mu_* = 0_p$.

(6) *Multivariate Gamma (MVG)*: $f_0(x) \propto c_{\Phi}(F_1(x_1), \ldots, F_p(x_p) \mid S_0) \prod_{j=1}^{p} f_j(x_j; \gamma_{j1}, \gamma_{j2})$ where f_j and F_j denote the density and distribution function of the univariate gamma distribution with shape parameter γ_{j1} and rate parameter γ_{j2}, respectively, for $j = 1, \ldots, p$ and $c_{\Phi}(\cdot \mid \Gamma)$ is as described in Song (2000). This is a Gaussian copula based construction of the multivariate gamma distribution. We set $\gamma_{j1} = \gamma_{j2} = 1$ for $j = 1, \ldots, p$.

The hyperparameters for the nearest neighbor-Dirichlet mixture are chosen as $\mu_0 = 0_p, \nu_0 = 0.001, \gamma_0 = p$, and $\Psi_0 = \{(\gamma_0 - p + 1)\delta_0^2\}\mathbb{I}_p = \delta_0^2\,\mathbb{I}_p$, where the optimal δ_0^2 is chosen via cross-validation as described in Section 2.3. Default hyperparameters as described in Section 4.1 are chosen for the MCMC and VB implementations of the DPM.

Similar to the univariate case, we defer the numerical results to Table 6 in the Appendix and in Figure 4 display a visual summary consisting of box plot of estimated $\mathcal{L}_1$ errors over the densities considered. The proposed method is very robust against a wide selection of true distributions, with its $\mathcal{L}_1$ error scaling nicely with the dimension. The KDE shows a noticeably sharp decline in performance - when the dimension is changed from 2 to 6, the average increase in $\mathcal{L}_1$ error is by factors of about 5 and 7 for sample sizes 200 and 1000, respectively. This is possibly due to lack of adaptive density estimation in higher dimensions using a single bandwidth matrix, since data in $\mathbb{R}^p$ become increasingly sparse with increasing p. As in the univariate case, we had to exclude the MVC density for the KDE due to the algorithm not converging. The performances of NN-DM, DP-MC, and
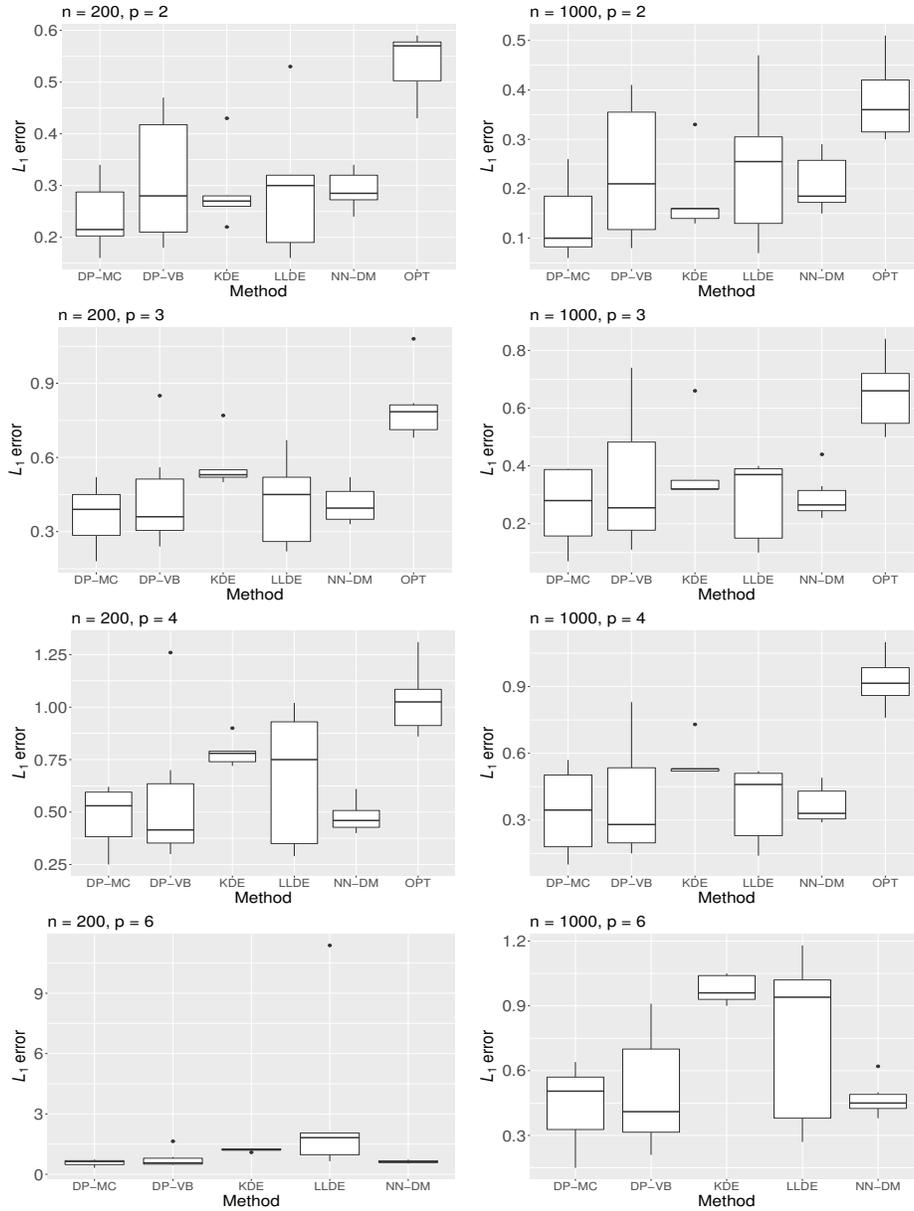
Figure 4: Box plots of $\hat{\mathcal{L}}_1(f_0, \hat{f})$ for the 6 different choices of the true density f_0 and different estimators $\hat{f}$ for multivariate data. The box plots for KDE and LLDE exclude the MVC density. The box plots for $p = 6$ exclude results from OPT.

DP-VB are quite competitive across densities, with NN-DM faring better than the DP-VB when estimating densities such as the MVC and the MVG. Furthermore, the NN-DM is hit the least significantly by the curse of dimensionality out of the three. This is particularly prominent for the DP-MC when the true density is either MG or MST with $n = 200$ and $p = 6$, and for the DP-VB when the true density is MVC. It is also important to keep in mind that the NN-DM provides similar results compared to the DP-MC while being at least an order of magnitude faster, as illustrated in Section 4.6. The performance of the OPT is hit quite significantly as the number of dimensions increases, along with the algorithm not converging for $p = 6$. The LLDE provides competitive results with the NN-DM in lower dimensions. However, in higher dimensions, the LLDE often does not converge, indicating lack of stability of the algorithm. We reported the average of the replicates for which the algorithm did converge. The results suggest that the performance of the LLDE is also affected quite drastically with increasing dimensions. When compared across all data generating cases considering the variation in densities, dimensions and sample sizes, the proposed method is seen to be more versatile than its competitors.

4.4 Accuracy of Uncertainty Quantification

In this section, we assess frequentist coverage of 95% pseudo-posterior credible intervals for the NN-DM and compare with coverage based on the 95% posterior credible intervals obtained from DP-MC and DP-VB. Ghosal and van der Vaart (2017) recommend investigating the frequentist coverage of Bayesian credible intervals. We do not include frequentist coverage for Polya tree mixtures (PTMs) and the optional Polya tree (OPT) due to the lack of available code. We consider the cases $p \in \{1, 2\}$ in our experiments with sample size $n = 500$. For each choice of density f_0, we fix $n_t = 200$ test points $\mathcal{X}_t = \{X_{t1}, \ldots, X_{tn_t}\}$ generated from the density f_0. With these fixed test points, we generate $n = 500$ data points in our sample for $R_{cov} = 200$ times and check the coverage of posterior/pseudo-posterior credible intervals obtained from the three methods. We implement the DP-MC with base measure $\mathrm{NIW}_p(0_p, 0.01, p, \mathbb{I}_p)$ and a Gamma$(2, 4)$ prior on the concentration parameter as in West (1992). These choices of hyperparameters were seen to give better frequentist coverage results than using the default values used in Sections 4.2 and 4.3. Same choices of hyperparameters are maintained for DP-VB. For the NN-DM, we take $k = 8$ in the univariate case and $k = 5$ in the bivariate case, $\alpha = 0.001$, and other hyperparameters chosen as before. We report the average coverage probability and average length of the (pseudo) credible intervals across all the points in the test data $\mathcal{X}_t$ in Tables 1 and 2 for the univariate and bivariate cases, respectively.

For univariate densities, both the DP-MC and DP-VB display severe under-coverage. In most of the cases, the DP-VB and NN-DM have similar width of (pseudo) credible intervals but the DP-VB displays dramatically lower coverage than the NN-DM. The under-coverage displayed by the DP-MC may be due to MCMC mixing issues. The NN-DM shows near nominal coverage in the smooth Gaussian (GS) and lognormal (LN) densities, while also attaining near nominal coverage in the skewed bimodal (SB), claw (CW), and sawtooth (ST) densities which are multi-modal. The shortcomings of DP-MC and DP-VB are especially noticeable when dealing with spiky densities such as the claw or sawtooth. For bivariate cases considered in Table 2 we see a similar trend; the NN-DM method provides uniformly

Method	CA	CW	DE	GS	IE
NN-DM	0.75 (0.05)	0.89 (0.21)	0.75 (0.06)	0.92 (0.08)	0.81 (0.11)
DP-MC	0.48 (0.02)	0.06 (0.01)	0.35 (0.02)	0.37 (0.01)	0.39 (0.04)
DP-VB	0.33 (0.05)	0.18 (0.07)	0.28 (0.07)	0.79 (0.05)	0.14 (0.04)

Method	LN	LO	SB	SP	ST
NN-DM	0.92 (0.17)	0.81 (0.03)	0.88 (0.10)	0.72 (0.01)	0.91 (0.05)
DP-MC	0.31 (0.05)	0.55 (0.03)	0.46 (0.03)	0.46 (0.01)	0.64 (0.03)
DP-VB	0.19 (0.15)	0.10 (0.03)	0.40 (0.10)	0.20 (0.01)	0.07 (0.01)

Table 1: Comparison of the frequentist coverage of 95% (pseudo) posterior credible intervals of the nearest neighbor-Dirichlet mixture and the MCMC and variational implementations of the Dirichlet process mixture for univariate data. Average length of the intervals are also provided for each case within parentheses. Number of replications and sample size are $R_{cov} = 200$ and $n_{cov} = 500$, respectively.

Method	MG	MST	MVC	MVG	SN	T
NN-DM	0.92 (0.04)	0.88 (0.03)	0.69 (0.03)	0.80 (0.31)	0.92 (0.06)	0.88 (0.03)
DP-MC	0.53 (0.01)	0.56 (0.01)	0.47 (0.01)	0.41 (0.16)	0.39 (0.02)	0.55 (0.01)
DP-VB	0.56 (0.03)	0.58 (0.03)	0.18 (0.02)	0.55 (0.26)	0.49 (0.05)	0.57 (0.02)

Table 2: Comparison of the frequentist coverage of 95% (pseudo) posterior credible intervals of the nearest neighbor-Dirichlet mixture and the MCMC and variational implementations of the Dirichlet process mixture for bivariate data. Average length of the intervals are also provided for each case within parentheses. Number of replications and sample size are $R_{cov} = 200$ and $n_{cov} = 500$, respectively.

better uncertainty quantification across all the densities considered. It is clear that in terms of frequentist uncertainty quantification, the NN-DM displays vastly superior coverage to the DP-MC and the DP-VB without inflating the interval width.

4.5 Comparison for High Dimensional Data

In addition to the above experiments, we performed a simulation experiment for high-dimensional data. Specifically, we set $n = 1000$, $p = 50$, and consider the same set of true densities in Section 4.3. We compared results from the proposed NN-DM method and the DP-VB. Due to severe computational time, we did not consider the DP-MC in this scenario. We also tried optional Polya trees (Wong and Ma, 2010) using the PTT package; however, the current implementation of the method breaks down in this high-dimensional setup. Due to numerical instability in estimating the $\mathcal{L}_1$ error in higher dimensions, we evaluate the methods in terms of their out-of-sample log-likelihood (OOSLL) instead (Gneiting and Raftery, 2007), on a test set of 500 data points. We report the average OOSLL over 30

Method	MG	SN	T	MST	MVC	MVG
NN-DM	-1.75	-1.74	-1.84	-1.84	-1.31	-1.36
DP-VB	-1.75	-1.74	-1.86	-1.84	-1.34	-1.36

Table 3: Out-of-sample log-likelihood ($\times 10^4$) of NN-DM and DP-VB on a test set of 500 points for 6 different multivariate densities considered in Section 4.3, for $n = 1000$ and $p = 50$. Greater out-of-sample log-likelihood is better.

replications in Table 3. The results indicate that both methods perform very similarly in terms of out-of-sample fit to the data, with the NN-DM outperforming the DP-VB when the true density is MVC. We also observed that for this experiment, the NN-DM methods with default choice of hyperparameters and with cross-validated choice of Ψ_0 have almost identical performance. For the NN-DM, we set $k = 12$ after carrying out a sensitivity analysis on k by considering $k = 5, 7, 10, 15$, and 20. The best results for the NN-DM were obtained for $k \in \{7, 10, 12\}$ with negligible difference in out-of-sample log-likelihoods between these three choices, with $k = 12$ performing the best.

4.6 Runtime Comparison

With n data points in p dimensions, the initial nearest neighbor allocation into n neighborhoods can be carried out in $\mathcal{O}(n \log n)$ steps (Vaidya, 1986; Ma and Li, 2019). Once the neighborhoods are determined with k_n points in each neighborhood, obtaining the neighborhood specific empirical means and covariance matrices has $\mathcal{O}(nk_np + nk_np^2) = \mathcal{O}(nk_np^2)$ complexity. Obtaining the pseudo-posterior mean (8) then requires inversion of n such $p \times p$ matrices to evaluate the multivariate t-density, with a runtime of $\mathcal{O}(np^3)$. Therefore, the total runtime to obtain the pseudo-posterior mean is of the order $\mathcal{O}(nk_np^2 + np^3)$. When we are interested in uncertainty quantification, we require Monte Carlo samples of the NN-DM, which are independently drawn from its pseudo-posterior. This involves sampling the Dirichlet weights, the neighborhood specific unknown mean and covariance matrix parameters of the Gaussian kernel, and evaluating a Gaussian density for each neighborhood, as outlined in Algorithm 1. To obtain M Monte Carlo samples, the combined complexity of this step is thus $\mathcal{O}(Mn + Mnp^3) = \mathcal{O}(Mnp^3)$. Overall the runtime complexity to obtain NN-DM samples is therefore $\mathcal{O}(Mnp^3 + nk_np^2 + np^3)$. For high dimensional scenarios, this runtime can be greatly improved by using a low rank matrix factorization of both the neighborhood specific empirical covariance matrices and the sampled covariance matrix parameters to make matrix inversion more efficient (Golub and van Loan, 1996). We now provide a detailed simulation study of runtimes of the proposed method, with all the simulations carried out on an M1 MacBook Pro with 16 GB of RAM.

We first focus on some runtime experiments comparing NN-DM and DP-MC. In our experiments, we focus on $p = 1$ and $p = 4$. The runtime for NN-DM consists of the time to estimate δ_0^2 by cross-validation as in Section 2.3 and then drawing samples from its pseudo-posterior. For both dimensions, the sample size is varied from $n = 200$ to $n = 1500$ in increments of 100. Data are generated from the standard Gaussian density (GS) for $p = 1$

and from a mixture of skew t-distributions with the parameters as described for the case MST in Section 4.3 for $p = 4$. For $p = 1$, we evaluate the two methods at 500 test points, while for $p = 4$ we evaluate the methods at 200 test points. The hyperparameters are kept the same as in Sections 4.2 and 4.3. We took 1000 Monte Carlo samples for the NN-DM and 2500 MCMC samples for the DP-MC with a burn-in of 1500 samples. We provide a figure summarising the results in Figure 5. In the top panel of Figure 5, we plot the average of the logarithm (base 10) of the run times of each approach for 10 independent replications. Corresponding $\mathcal{L}_1$ errors of the two methods is included in the bottom panel of Figure 5.

In Figure 5, the NN-DM is at least an order of magnitude faster than DP-MC. The time saved becomes more pronounced in the multivariate case, where for sample size 1500 the NN-DM is ~ 50 times faster. The gain in computing time does not come at the cost of accuracy as can be seen from the right panel; the proposed method maintains the same order of $\mathcal{L}_1$ error as the DP-MC in the univariate case and often outperforms the DP-MC in the multivariate case. We did not implement the Monte Carlo sampler for the proposed algorithm in parallel, but such a modification would substantially improve runtime. Bypassing cross-validation and choosing default hyperparameters instead as outlined in Section 2.3, NN-DM took 3.3 seconds and 16.4 seconds when $p = 1$ and $p = 4$, respectively, with sample size $n = 1500$. In the same scenario, DP-MC took 99.4 seconds and 1618.1 seconds for $p = 1$ and $p = 4$, respectively. Thus the NN-DM with default hyperparameters is about 30 times faster when $p = 1$ and almost 100 times faster when $p = 4$.

We also compare the runtime of the proposed method with three recent implementations of the DPM, namely the packages `bnpy` (Hughes and Sudderth, 2014), `DPMMSubClusters` (Dinari et al., 2019), and `vdpgm` (Kurihara et al., 2006) available for download at `https://kenichikurihara.com/variational-dirichlet-process-gaussian-mixture-model/`. These three packages implement variational approximations of the DPM posterior with different modifications. We also include the DP-MC and OPT for comparison. All the runtime results are comparable only up to machine and coding language differences. Amongst the competitor package implementations, the `NNDM` and `dirichletprocess` packages are the only ones providing (pseudo) posterior samples of the density estimate at a test point. We consider the average runtime of $R = 10$ replicates to fit a training data set of iid $N(0, 1)$ entries with $n = 1500$ and $p = 4$. For the NN-DM, DP-MC, and OPT, we consider 1000 (pseudo) posterior samples. Table 4 provides the runtimes for the different packages considered. Overall, the fastest implementation is observed for the `PTT` package. The next fastest implementations are the `NNDM` without cross-validation (CV), `DPMMSubClusters`, and `bnpy`. The runtime for `NNDM` with CV closely follows the previous implementations, with both `NNDM` with and without CV providing (pseudo) posterior samples. The major improvement in runtime for NN-DM is mainly due to the fact that neighborhood allocations are fixed here which is not the case for DP-MC.

4.7 Sensitivity to the Choice of k

In this subsection, we investigate the role of $k_n = k$ in finite samples for the proposed method. We consider $n = 200$ samples from the SP density in the univariate case and the MG density in the bivariate case. In each case, we fix a test set of $n_t = 500$ points, and evaluate the out-of-sample log-likelihood (OOSLL) of the test points for 20 different integer
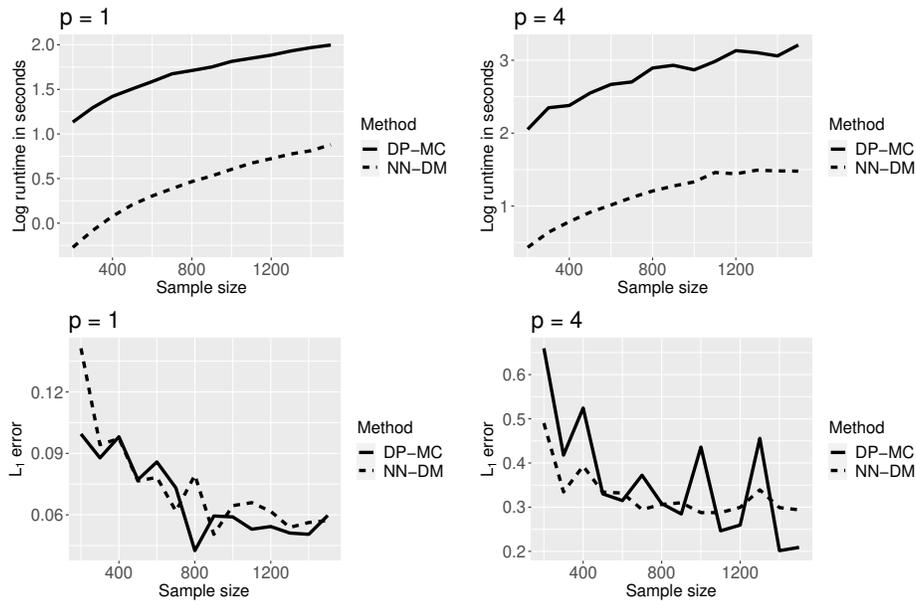
Figure 5: Runtime comparison of DP-MC and NN-DM in univariate case and for 4-dimensional data. Top panel shows runtimes in $\log_{10}$ scale whereas bottom panel shows corresponding $\mathcal{L}_1$ error. Sample size n is varied from 200 to 1500 in increments of 100.

Package (Language)	Average Runtime (s)	Provides Samples?
bnpy (Python)	5.79	No
DPMMSubClusters (Julia)	4.33	No
vdpgm (MATLAB)	58.38	No
NNDM (RCpp and R, with CV)	18.96	Yes
NNDM (Rcpp and R, without CV)	3.52	Yes
dirichletprocess (R)	1068.48	Yes
PTT (Rcpp and R)	0.59	No

Table 4: Table comparing the average runtimes of different packages for $n = 1500$, $p = 4$. NN-DM runtimes are provided both with and without cross-validation (CV), using the package NNDM developed by the authors.

values of k ranging from 2 to 50. Finally, we report results averaged from 10 independent replicates of this setup. We note that for each considered value of k, the parameter δ_0^2 was estimated using leave-one-out cross-validation. Figure 6 shows how the OOSLL averaged over replicates changes as a function of k for each density considered. The original OOSLL values of the test data points were scaled by the number of test points $n_t = 500$ for better representability.
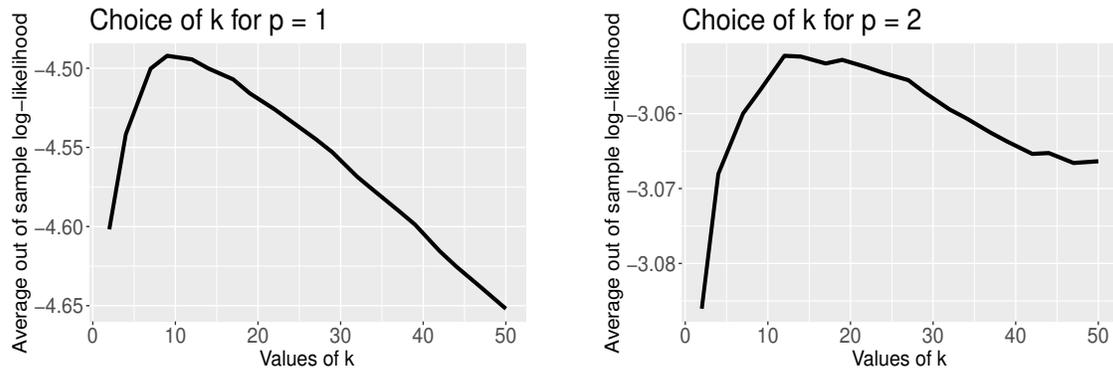
Figure 6: Average out-of-sample log-likelihood of 500 test points for the NN-DM as a function of k for one-dimensional and two-dimensional data. Number of samples and number of replications are $n = 200$ and $R = 10$, respectively.

For the univariate SP density, the optimal value of k which maximizes the average OOSLL is $\hat{k} = 9$. This is close to the choice of $k = 6$ as taken in Section 4.2. For the bivariate MG density, we observe that the choice of k maximizing the OOSLL is $\hat{k} = 12$, which is also close to the choice of $k = 10$ as taken in Section 4.3. For both the univariate and the bivariate case, the out-of-sample log-likelihood of the test set shows little variation with changing k. This indicates that the estimates obtained from the proposed method are quite robust to the particular choice of k.

5. Application

We apply the proposed density estimator to binary classification. Consider data $\mathcal{D} = \{(X_i, Y_i) : i = 1, \ldots, n\}$, where $X_i \in \mathbb{R}^p$ are p-dimensional feature vectors and $Y_i \in \{0, 1\}$ are binary class labels. To predict the probability that $y_0 = 1$ for a test point x_0, we use Bayes rule:

$$\mathrm{pr}(y_0 = 1 \mid x_0) = \frac{\tilde{f}_1(x_0) \, \mathrm{pr}(y_0 = 1)}{\tilde{f}_0(x_0) \, \mathrm{pr}(y_0 = 0) + \tilde{f}_1(x_0) \, \mathrm{pr}(y_0 = 1)}, \tag{16}$$

where $\tilde{f}_j(x_0)$ is the feature density at x_0 in class j and $\mathrm{pr}(y_0 = j)$ is the marginal probability of class j, for $j = 0, 1$. Based on n_t test data, we let $\widehat{\mathrm{pr}}(y_0 = 1) = (1/n_t) \sum_{i=1}^{n_t} Y_i$, with $\widehat{\mathrm{pr}}(y_0 = 0) = 1 - \widehat{\mathrm{pr}}(y_0 = 1)$. We use either the NN-DM pseudo-posterior mean $\hat{f}_n(\cdot)$, the DP-MC posterior mean $\hat{f}_{\mathrm{DP}}(\cdot)$, or the DP-VB posterior mean $\hat{f}_{\mathrm{VB}}(\cdot)$ for estimating the within class densities, and compare their classification performances in terms of sensitivity, specificity, and probabilistic calibration. We omit the KDE as to the best of our knowledge, no routine R implementation is available for data having more than 6 dimensions.

The high time resolution universe survey data (Keith et al., 2010) contain information on sampled pulsar stars. Pulsar stars are a type of neutron stars and their radio emissions are detectable from the Earth. These stars have gained considerable interest from the scientific community due to their several applications (Lorimer and Kramer, 2012). The data are publicly available from the University of California at Irvine machine learning repository.
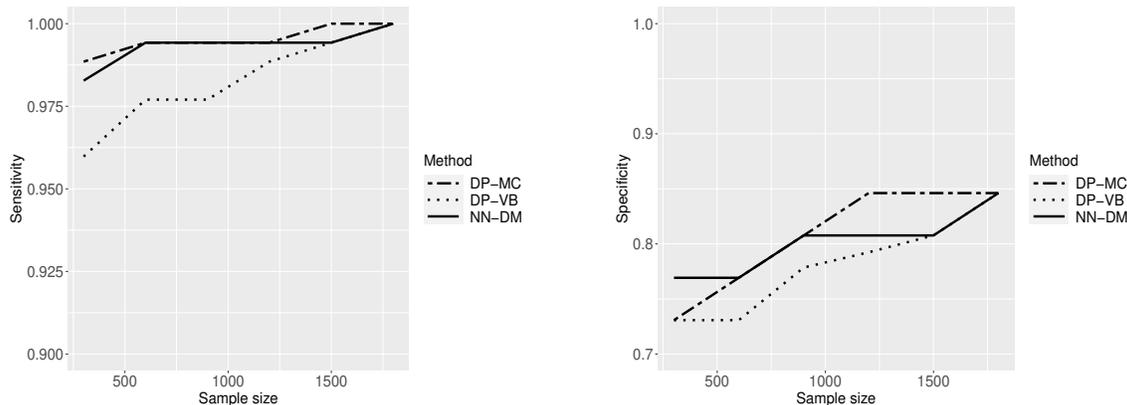
Figure 7: Sensitivity and specificity of the NN-DM, DP-MC, and DP-VB for the high time resolution universe survey data.

Stars are classified into pulsar and non-pulsar groups according to 8 attributes (Lyon, 2016). There are a total of 17898 instances of stars, among which 1639 are classified as pulsar stars.

We create a test data set of 200 stars, among which 23 are pulsar stars. The training size is then varied from 300 to 1800 in increments of 300, each time adding 300 training points by randomly sampling from the entire data leaving out the initial test set. In Figure 7, we plot the sensitivity and specificity of the three methods in consideration. All the methods exhibit similar sensitivity across various training sizes; the DP-MC has marginally better specificity for training sizes 1200 and 1500, while the NN-DM has better specificity for training sizes 300 and 600. Both the NN-DM and the DP-MC exhibit higher specificity and sensitivity than the DP-VB across all training sample sizes considered.

We also compare the methods using the Brier score, a proper scoring rule (Gneiting and Raftery, 2007) for probabilistic classification. Suppose for n_t test points and the ith Monte Carlo sample, p_i denotes the sampled $n_t \times 1$ probability vector for a generic method. We compute the normalized Brier score for the ith sample as $(1/n_t) \, ||p_i - Y_t||_2^2$, where Y_t is the vector of class labels in the test set. Then with T samples of p_i, $i = 1, \ldots, T$, we compute the mean Brier score for the three methods considered. The mean Brier score for each training size is shown in the right panel of Figure 8, which naturally shows a declining trend with increasing training size. There is little to choose between the three classifiers in terms of mean Brier score; the proposed method fairs equally well in terms of calibration of estimated test set probabilities with the MCMC implementation of the Dirichlet process. In the left panel of Figure 8, the receiver operating characteristic curve of the methods is shown for 1800 training samples. The area under the curve (AUC) for the NN-DM, the DP-MC and the DP-VB are 0.96, 0.95 and 0.96, respectively. For 1800 training samples, the computation time for the proposed method is about 13 minutes while for the DP-MC it is approximately 5 hours.

Hence, the proposed method is much faster, even without exploiting parallel computation. We also fitted the proposed method using the training set of all 17698 points; DP-MC was too slow in this case. The sensitivity and specificity of the proposed method
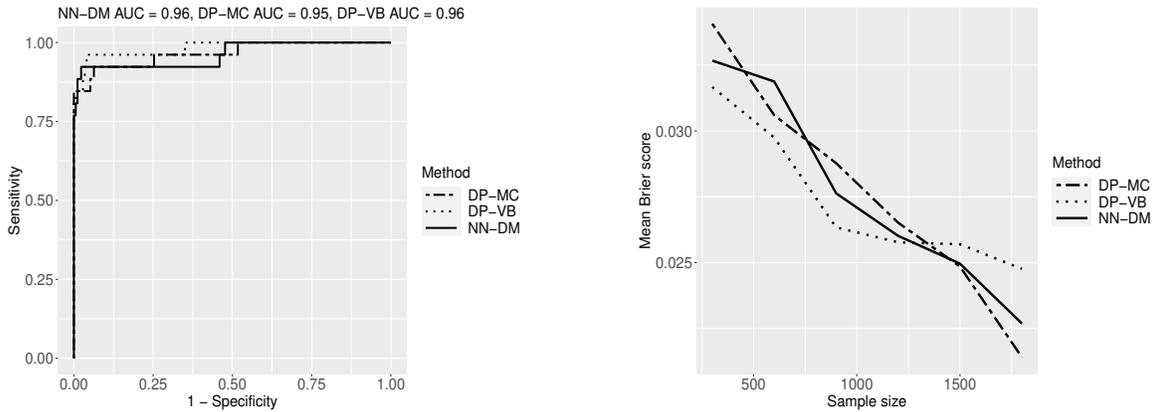
23

Figure 8: Left plot shows the receiver operating characteristic curve of the NN-DM, DP-MC, and DP-VB with 1800 training samples. Area under the curve is abbreviated as AUC. Right plot shows normalized Brier scores for the methods with varying training sample size.
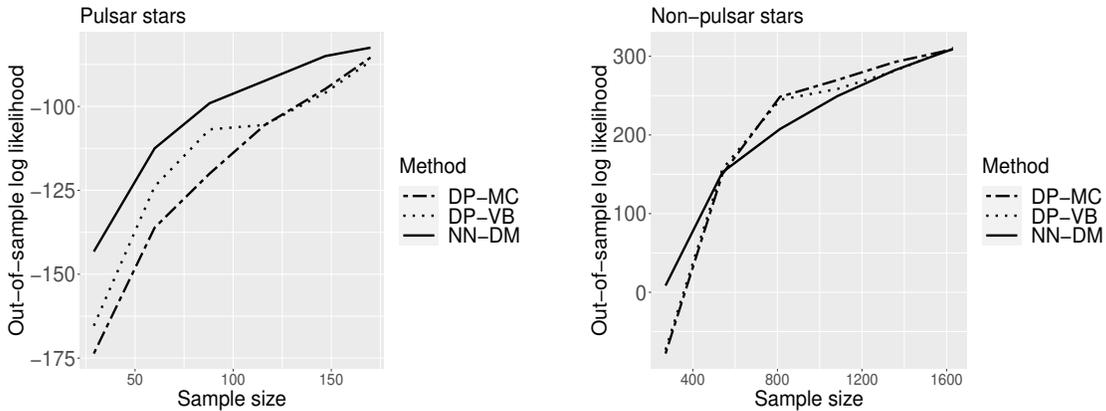

Figure 9: Left and right plots show the out-of-sample log-likelihoods of NN-DM, DP-MC, and DP-VB for the two different star types.

increased to 0.99 and 0.91, respectively. We additionally evaluated the methods in terms of the out-of-sample log-likelihood. The results are displayed in Figure 9. While the methods perform comparably in terms of their classification performance, NN-DM achieves a better fit overall, especially for the significantly less prevalent pulsar star type.

6. Discussion

The proposed nearest neighbor-Dirichlet mixture provides a useful alternative to Bayesian density estimation based on Dirichlet mixtures with much faster computational speed and stability in avoiding MCMC. MCMC can have very poor performance in mixture models and other multimodal cases, due to difficulty in mixing, and hence can lead to posterior

inferences that are unreliable. There is a recent literature attempting to scale up MCMC-based analyses in model-based clustering contexts including for Dirichlet process mixtures; refer, for example to Song et al. (2020) and Ni et al. (2020). However, these approaches are complex to implement and are primarily focused on the problem of clustering, while we are instead focused on flexible modeling of unknown densities.

The main conceptual disadvantage of the proposed approach is the lack of a coherent Bayesian posterior updating rule. However, we have shown that nonetheless the resulting pseudo-posterior can have appealing behavior in terms of frequentist asymptotic properties, finite sample performance, and accuracy in uncertainty quantification. In addition, it is important to keep in mind that Bayesian kernel mixtures have key disadvantages that are difficult to remove within a fully coherent Bayesian modeling framework. These include a strong sensitivity to the choice of kernel and prior on the weights on these kernels; refer, for example to Miller and Dunson (2019).

There are several important next steps. The first is to develop fast and robust algorithms for using the nearest neighbor-Dirichlet mixture not just for density estimation but also as a component of more complex hierarchical models. For example, one may want to model the residual density in regression nonparametrically or treat a random effects distribution as unknown. In such settings, one can potentially update other parameters within a Bayesian model using Markov chain Monte Carlo, while using algorithms related to those proposed in this article to update the nonparametric part conditionally on these other parameters.

Acknowledgments

Appendix

Appendix A. Prerequisites

We first introduce some notation with accompanying technical details which will be used hereafter. We denote the Frobenius norm and determinant of $A \in \mathbb{R}^{p \times p}$ by $||A||_F = \{\operatorname{tr}(A^{\mathrm{T}}A)\}^{1/2}$ and $|A|$, respectively. For $v \in \mathbb{R}^p$, one has $||vv^{\mathrm{T}}||_F = ||v||_2^2$ where $||a||_2 = (a^{\mathrm{T}}a)^{1/2}$ is the Euclidean norm of a. For two symmetric matrices $A, B \in \mathbb{R}^{p \times p}$, we say that $A \geq B$ if $A - B$ is positive semi-definite, that is $x^{\mathrm{T}}(A - B)x \geq 0$ for all $x \in \mathbb{R}^p, x \neq 0_p$ where $0_p = (0, \ldots, 0)^{\mathrm{T}}$. For a real symmetric matrix A_*, let the eigenvalues of A_* be $e_1(A_*), \ldots, e_p(A_*)$, arranged such that $e_1(A_*) \geq \ldots \geq e_p(A_*)$. If $A \geq B$, then it follows by the min-max theorem (Teschl, 2009) that for each $j = 1, \ldots, p$, we have $e_j(A) \geq e_j(B)$. In particular, we have $|A| \geq |B|$ and $||A||_F \geq ||B||_F$.

Now consider a true data generating density $X_1, \ldots, X_n \overset{iid}{\sim} f_0$ satisfying Assumptions 1-3 as in Section 3.1. Let $\mathcal{X}^{(n)} = (X_1, \ldots, X_n)$ and suppose f_0 induces the measure P_{f_0} on the Borel σ-field on $\mathbb{R}^p$, denoted by $\mathcal{B}(\mathbb{R}^p)$. We form the k-nearest neighborhood of X_i using the Euclidean norm for $i = 1, \ldots, n$. We also let k depend on n and express this dependence as k_n when required. However, we routinely drop this dependence for notational simplicity. For X_i, let Q_i be its kth nearest neighbor in $\mathcal{X}^{(n)}$ (for $k = 1$, $Q_i = X_i$) and let R_i be the distance between X_i and Q_i, given by $R_i = ||X_i - Q_i||_2$. Define the ball

$$B_i = \{y \in [0, 1]^p : 0 < ||y - X_i||_2 < R_i\}$$

and the probability

$$G(X_i, R_i) = \int_{B_i} f_0(u) \, du$$

of the ball B_i under P_{f_0}. Let $Y_1^{(i)} = X_i$ and $Y_2^{(i)}, \ldots, Y_{k-1}^{(i)}$ denote the rest of the interior points in B_i. Let the mean $\bar{X}_i$ and covariance matrix S_i of the ith neighborhood be

$$\bar{X}_i = \frac{1}{k_n} \left\{ \sum_{j=1}^{k-1} Y_j^{(i)} + Q_i \right\},$$

$$S_i = \frac{1}{k_n} \left\{ \sum_{j=1}^{k-1} (Y_j^{(i)} - \bar{X}_i)(Y_j^{(i)} - \bar{X}_i)^{\mathrm{T}} + (Q_i - \bar{X}_i)(Q_i - \bar{X}_i)^{\mathrm{T}} \right\}.$$

We observe that $(Y_2^{(i)}, \ldots, Y_{k-1}^{(i)}, Q_i)$ is identically distributed for $i = 1, \ldots, n$ since $X_1, \ldots, X_n$ are independent and identically distributed. Thus we only consider the case $i = 1$ from here on. For sake of brevity, denote $Y_u^{(1)}$ by Y_u for $u = 2, \ldots, k-1$ and Q_1 by Q.

Conditional on $X_1 = x_1 \in [0, 1]^p$ and $R_1 = r_1 > 0$, following Mack and Rosenblatt (1979), the conditional joint density of $Y_2, \ldots, Y_{k-1}$ and Q is

$$p(y_2, \ldots, y_{k-1}, q \mid x_1, r_1) = \left\{ \prod_{j=2}^{k-1} \frac{f_0(y_j)}{G(x_1, r_1)} \mathbb{I}(y_j \in B_1) \right\} \frac{f_0(q)}{G'(x_1, r_1)} \mathbb{I}(||q - x_1|| = r_1),$$

where $G'(x_1, r_1) = \partial G(x_1, r_1)/\partial r_1$ and $\mathbb{I}(A)$ denotes the indicator function of the event $A \in \mathcal{B}(\mathbb{R}^p)$. Thus conditional on X_1 and R_1, the random variables $Y_2, \ldots, Y_{k-1}$ are independent and identically distributed, and independent of Q.

Let the function $\rho(x_1, r_1) = r_1^{\kappa_1}$ where κ_1 is a non-negative integer. This function can be identified with $\phi(\cdot)$ in equation (11) of Mack and Rosenblatt (1979). In the results that follow, we will require the expectation of $\rho(x_1, r_1)$ under P_{f_0} for different choices of κ_1. To that end, we shall repeatedly make use of the equation (12) from Mack and Rosenblatt (1979) adapted to our setting:

$$E_{P_{f_0}}\{R_1^{\kappa_1} \mid X_1 = x_1\} =$$
$$\frac{(n-1)!}{(k-2)!(n-k)!}\int_0^1 \left\{\left(\frac{t}{C_p f_0(x_1)}\right)^{\kappa_1/p} + o(t^{\kappa_1/p})\right\} t^{k-2}(1-t)^{n-k}dt. \qquad (17)$$

Finally, we let $\widetilde{E}$ and $\widetilde{\mathrm{var}}$ denote the expectation and variance, respectively, of the NN-DM estimator $f(x)$ under the pseudo-posterior density $\widetilde{\Pi}$, described in (7). Conditioning notation under $\widetilde{\Pi}$ is as usual; for example, the conditional expectation

$$\widetilde{E}\{f(x) \mid \pi_1, \ldots, \pi_n\} = \sum_{i=1}^n \pi_i \widetilde{E}\{\phi_p(x; \eta_i, \Sigma_i)\},$$

where the expectation $\widetilde{E}\{\phi_p(x; \eta_i, \Sigma_i)\}$ is with respect to the pseudo-posterior density of (η_i, Σ_i) as described in Section 2.2.

Appendix B. Proof of Theorem 4

Suppose $X_1, \ldots, X_n$ are independent and identically distributed random variables generated from the density f_0 supported on $[0, 1]^p$ satisfying Assumptions 1-3. For $i = 1, \ldots, n$, recall the definitions of μ_i and Λ_i from (8):

$$\mu_i = \frac{\nu_0}{\nu_n}\mu_0 + \frac{k}{\nu_n}\bar{X}_i, \quad \Lambda_i = \frac{\nu_n + 1}{\nu_n(\gamma_n - p + 1)}\Psi_i.$$

We want to show that $\hat{f}_n(x) = (1/n)\sum_{i=1}^n t_{\gamma_n - p + 1}(x; \mu_i, \Lambda_i) \to f_0(x)$ in P_{f_0}-probability as $n \to \infty$ for any $x \in [0, 1]^p$, where $\hat{f}_n(x)$ is as described in (8). We first prove two propositions involving successive mean value theorem type approximations to $\hat{f}_n(x)$, which will imply the final result. We now state the two propositions, with accompanying proofs, before stating the final theorem.

Proposition 9 *Fix $x \in [0, 1]^p$. Let $f_A(x) = (1/n)\sum_{i=1}^n t_{\gamma_n - p + 1}(x; X_i, \Lambda_i)$. Also, let $k = o(n^{i_1})$ with $i_1 = 2/(p^2 + p + 2)$ and $\nu_0 = o(n^{-1/p}k^{(1/p)+1})$. Then, we have $E_{P_{f_0}}(|\hat{f}_n(x) - f_A(x)|) \to 0$ as $n \to \infty$.*

Proof Since the $(\Lambda_i)_{i=1}^n$ are identically distributed and $(\mu_i)_{i=1}^n$ are identically distributed, we have $E_{P_{f_0}}(|\hat{f}_n(x) - f_A(x)|) \leq E_{P_{f_0}}\{|t_{\gamma_n - p + 1}(x; \mu_1, \Lambda_1) - t_{\gamma_n - p + 1}(x; X_1, \Lambda_1)|\}$. The multivariate mean value theorem now implies that

$$E_{P_{f_0}}(|\hat{f}_n(x) - f_A(x)|) \leq E_{P_{f_0}}\left\{|\Lambda_1|^{-1/2}||\nabla t_{\gamma_n - p + 1}(\xi; 0_p, I_p)||_2 ||\Lambda_1^{-1/2}(X_1 - \mu_1)||_2\right\}, \qquad (18)$$

where $\nabla t_{\gamma_n-p+1}(\xi;0_p,I_p) = [\partial t_{\gamma_n-p+1}(x;0_p,\mathbb{I}_p)/\partial x]_\xi$ for some ξ in the convex hull of $\Lambda_1^{-1/2}(x-X_1)$ and $\Lambda_1^{-1/2}(x-\mu_1)$.

Using standard results and the min-max theorem, we have

$$||\Lambda_1^{-1/2}(X_1-\mu_1)||_2 \leq ||\Lambda_1^{-1/2}||_F\,||X_1-\mu_1||_2.$$

If we let $H_n = H = \{\nu_n(\gamma_n-p+1)\}^{-1}(\nu_n+1)\Psi_0 = h^2 I_p$ where $h^2 = h_n^2 = \{\nu_n(\gamma_n-p+1)\}^{-1}\{(\nu_n+1)(\gamma_0-p+1)\}\delta_0^2$ following the choice of Ψ_0 from Section 2.3, then it is clear that $\Lambda_1 \geq H$. Therefore, we have $||\Lambda_1^{-1/2}(X_1-\mu_1)||_2 \leq ||H^{-1/2}||_F\,||X_1-\mu_1||_2$. Straightforward calculations show that $||H^{-1/2}||_F = h^{-1}p^{1/2}$ and $||X_1-\mu_1||_2 \leq R_1 + \{\nu_n^{-1}(1+||\mu_0||_2)\nu_0\}$ where $R_1 = ||X_1-X_{1[k]}||_2$. Using Theorem 2.4 from Biau and Devroye (2015) for $p \geq 2$ and (17) for $p=1$, one gets

$$E_{P_{f_0}}(R_1^2) \leq d_p^2\left(\frac{k}{n}\right)^{2/p}, \tag{19}$$

for an appropriate constant $d_p > 0$. Thus, we have $E_{P_{f_0}}(R_1) \leq \{E_{P_{f_0}}(R_1^2)\}^{1/2} \leq d_p(k/n)^{1/p}$ for sufficiently large n. This implies that

$$E(||X_1-\mu_1||_2) \leq d_p\left(\frac{k}{n}\right)^{1/p} + o\left(\frac{k}{n}\right)^{1/p}. \tag{20}$$

We also have $|\Lambda_1|^{-1/2} \leq |H|^{-1/2} = h^{-p}$. Finally, simple calculations yield that

$$||\nabla t_{\gamma_n-p+1}(\xi;0_p,I_p)||_2 \leq L_{1,n,p}$$

where $L_{1,n,p} > 0$ satisfies $L_{1,n,p} \to (2\pi)^{-p/2}e^{-1/2}$ as $n \to \infty$. Plugging all these back in (18), we obtain a finite constant $L_{2,n,p} > 0$ such that

$$E_{P_{f_0}}(|\hat{f}_n(x)-f_A(x)|) \leq L_{2,n,p}(n^{-i_1}k)^{(p^2+p+2)/(2p)} + o\{(n^{-i_1}k)^{(p^2+p+2)/(2p)}\}, \tag{21}$$

which goes to 0 as $n \to \infty$, completing the proof. ∎

We now provide the second mean value theorem type approximation which approximates the random bandwidth matrix Λ_i in $f_A(x)$ by $H = H_n$ for each $i = 1,\ldots,n$.

Proposition 10 *Fix* $x \in [0,1]^p$. *Let* $f_K(x) = (1/n)\sum_{i=1}^n t_{\gamma_n-p+1}(x;X_i,H)$. *Also, let* $k = o(n^{i_2})$ *with* $i_2 = 4/(p+2)^2$ *and* $\nu_0 = o\{n^{-2/p}k^{(2/p)+1}\}$. *Then, we have* $E_{P_{f_0}}(|f_A(x)-f_K(x)|) \to 0$ *as* $n \to \infty$.

Proof Using the identically distributed properties of $(\Lambda_i)_{i=1}^n$ and $(X_i)_{i=1}^n$, we obtain $E_{P_{f_0}}(|f_A(x)-f_K(x)|) \leq E_{P_{f_0}}(|t_{\gamma_n-p+1}(x;X_1,\Lambda_1)-t_{\gamma_n-p+1}(x;X_1,H)|)$. Using the multivariate mean value theorem, we obtain that

$$E_{P_{f_0}}(|t_{\gamma_n-p+1}(x;X_1,\Lambda_1)-t_{\gamma_n-p+1}(x;X_1,H)|) \leq E_{P_{f_0}}(||M_1||_F\,||\Lambda_1-H||_F), \tag{22}$$

where $M_1 = [\partial\{t_{\gamma_n-p+1}(x;X_1,\Sigma)\}/\partial\Sigma]_{\Sigma_0}$ for some Σ_0, with Σ_0 in the convex hull of Λ_1 and H. Since $\Lambda_1 \geq H$, we immediately have $\Sigma_0 \geq H$ as well. Using the definitions of Λ_1 and H, we have

$$||\Lambda_1-H||_F \leq \frac{(\nu_n+1)}{\nu_n(\gamma_n-p+1)}\left\{\left|\left|\sum_{j\in\mathcal{N}_1}(X_j-\bar{X}_1)(X_j-\bar{X}_1)^{\mathrm{T}}\right|\right|_F + \frac{k\nu_0}{\nu_n}||\bar{X}_1\bar{X}_1^{\mathrm{T}}||_F\right\}.$$

28

Since $|| \sum_{j \in \mathcal{N}_1} (X_j - \bar{X}_1)(X_j - \bar{X}_1)^{\mathrm{T}}||_F \le \sum_{j \in \mathcal{N}_1} ||(X_j - \bar{X}_1)(X_j - \bar{X}_1)^{\mathrm{T}}||_F = \sum_{j \in \mathcal{N}_1} ||X_j - \bar{X}_1||_2^2 \le \sum_{j \in \mathcal{N}_1} R_1^2 = k R_1^2$, we get for sufficiently large n the following:

$$E_{P_{f_0}}(\, ||\Lambda_1 - H||_F \,) \le E_{P_{f_0}}(R_1^2) + o\left(\frac{k}{n}\right)^{2/p}, \tag{23}$$

$$\le d_p^2 \left(\frac{k}{n}\right)^{2/p} + o\left(\frac{k}{n}\right)^{2/p}, \tag{24}$$

using (19) and $\nu_0 = o\left\{n^{-2/p} k^{(2/p)+1}\right\}$. Taking partial derivatives of $\log\{t_{\gamma_n - p+1}(x; X_1, \Sigma)\}$ with respect to Σ evaluated at Σ_0 and taking Frobenius norm of both sides, we obtain

$$||t_{\gamma_n - p+1}^{-1}(x; X_1, \Sigma_0) \, M_1||_F \, \le h^{-2}(\gamma_n + 1)$$

for sufficiently large n. We now observe that

$$t_{\gamma_n - p+1}(x; X_1, \Sigma_0) \le c_{p, \gamma_n - p+1} |\Sigma_0|^{-1/2} \, \le c_{p, \gamma_n - p+1} |H|^{-1/2} \, = h^{-p} c_{p, \gamma_n - p+1},$$

where $c_{p,\beta} = (\pi \beta)^{-p/2} \{\Gamma(\beta/2)\}^{-1} \Gamma\{(\beta + p)/2\}$ for $p \ge 1, \beta > 0$. Note that $c_{p,\beta} \to (2\pi)^{-p/2}$ as $\beta \to \infty$ for any $p \ge 1$. This immediately implies that $||M_1||_F \, \le h^{-(p+2)} c_{p, \gamma_n - p+1}(\gamma_n + 1)$ for sufficiently large n. Plugging all these back in equation (22), we obtain for sufficiently large n, a finite $L_{3,n,p} > 0$ such that

$$E_{P_{f_0}}(\, |f_A(x) - f_K(x)| \,) \le L_{3,n,p}(n^{-i_2} k)^{(p+2)^2/(2p)} + o\{(n^{-i_2} k)^{(p+2)^2/(2p)}\}, \tag{25}$$

which goes to 0 as $n \to \infty$, proving the proposition. ∎

We now prove Theorem 4.

Proof [Theorem 4] $E_{P_{f_0}}(\, |\hat{f}_n(x) - f_K(x)| \,) \le E_{P_{f_0}}(\, |\hat{f}_n(x) - f_A(x)| \,) + E_{P_{f_0}}(\, |f_A(x) - f_K(x)| \,)$ by the triangle inequality. Using Propositions 9 and 10, we obtain that $E_{P_{f_0}}(\, |\hat{f}_n(x) - f_K(x)| \,) \to 0$ as $n \to \infty$. From Section F of the Appendix, we obtain $f_K(x) \to f_0(x)$ in P_{f_0}-probability. This immediately implies that given the conditions on k, ν_0, and for any $x \in [0,1]^p$, we have $\hat{f}_n(x) \to f_0(x)$ in P_{f_0}-probability. ∎

Appendix C. Proof of Theorem 5

Proof Fix $x \in [0,1]^p$. For $i = 1, \ldots, n$, let $z_i = \phi_p(x \, ; \eta_i, \Sigma_i)$ and suppose $z^{(n)} = (z_1, \ldots, z_n)^{\mathrm{T}}$. Then, we have $f(x) = \sum_{i=1}^n \pi_i z_i = z^{(n)\mathrm{T}} \pi^{(n)}$ where $\pi^{(n)} = (\pi_1, \ldots, \pi_n)^{\mathrm{T}}$. We begin with the identity

$$\widetilde{\mathrm{var}}\{f(x)\} = \widetilde{\mathrm{var}}[\widetilde{E}\{f(x) \mid z^{(n)}\}] + \widetilde{E}[\widetilde{\mathrm{var}}\{f(x) \mid z^{(n)}\}]. \tag{26}$$

We start with the first term on the right hand side of (26). Observe that $z_1, \ldots, z_n$ are independent under $\widetilde{\Pi}$ and $\widetilde{E}(\pi_i) = 1/n$ for $i = 1, \ldots, n$. Thus, we have

$$
\widetilde{\mathrm{var}}[\widetilde{E}\{f(x) \mid z^{(n)}\}] = \widetilde{\mathrm{var}}\left(\frac{1}{n}\sum_{i=1}^{n} z_i\right)
$$

$$
= \frac{1}{n^2}\sum_{i=1}^{n}\widetilde{\mathrm{var}}(z_i)
$$

$$
\leq \frac{1}{n^2}\sum_{i=1}^{n}\widetilde{E}(z_i^2)
$$

$$
= \frac{1}{n^2}\sum_{i=1}^{n} R_n |B_i|^{-1/2} t_{\gamma_n - p + 2}(x; \mu_i, B_i),
$$

since for $i = 1, \ldots, n$, we have

$$
\widetilde{E}(z_i^2) = R_n |B_i|^{-1/2} t_{\gamma_n - p + 2}(x; \mu_i, B_i), \tag{27}
$$

where

$$
R_n = \frac{\Gamma\{(\gamma_n - p + 2)/2\}}{\Gamma\{(\gamma_n - p + 1)/2\}}\left[\frac{\nu_n + 2}{4\pi\nu_n(\gamma_n - p + 2)}\right]^{p/2}, \quad B_i = D_n \Lambda_i,
$$

and $D_n = \{2(\gamma_n - p + 2)(\nu_n + 1)\}^{-1}(\gamma_n - p + 1)(\nu_n + 2)$. To obtain (27), we integrate over the pseudo-posterior distribution of $(\eta_i, \Sigma_i)_{i=1}^{n}$, namely $\mathrm{NIW}(\mu_i, \nu_n, \gamma_n, \Psi_i)$. For $i = 1, \ldots, n$, since $|\Lambda_i| \geq |H_n|$, we have $|B_i| \geq D_n^p |H_n|$. Letting $\widehat{f}_{var}(x) = (1/n)\sum_{i=1}^{n} t_{\gamma_n - p + 2}(x; \mu_i, B_i)$, we have

$$
\widetilde{\mathrm{var}}[\widetilde{E}\{f(x) \mid z^{(n)}\}] \leq \frac{R_n D_n^{-p/2}\widehat{f}_{var}(x)}{n|H_n|^{1/2}}. \tag{28}
$$

We now analyze the second term on the right hand side of (26). Recall that $\pi^{(n)}$ is independent of $z^{(n)}$ under $\widetilde{\Pi}$. Let Σ_π denote the pseudo-posterior covariance matrix of $\pi^{(n)}$. Standard results yield $\Sigma_\pi = V_n\{(1 - C_n)\mathbb{I}_n + C_n \mathbb{1}_n \mathbb{1}_n^{\mathrm{T}}\}$, where $V_n = (n-1)/[n^2\{n(\alpha+1)+1\}]$, and $C_n = -1/(n-1)$. Then, we have

$$
\widetilde{E}[\widetilde{\mathrm{var}}\{f(x) \mid z^{(n)}\}] = \widetilde{E}[z^{(n)\mathrm{T}}\Sigma_\pi z^{(n)}]. \tag{29}
$$

Using the expression for Σ_π along with (29), we obtain,

$$
\widetilde{E}[\widetilde{\mathrm{var}}\{f(x) \mid z^{(n)}\}] = \frac{1}{n(\alpha+1)+1}\widetilde{E}\left\{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2\right\}, \tag{30}
$$

where $\bar{z} = (1/n)\sum_{i=1}^{n} z_i$. We now have

$$
\widetilde{E}[\widetilde{\mathrm{var}}\{f(x) \mid z^{(n)}\}] = \frac{1}{n\{n(\alpha+1)+1\}}\left\{\sum_{i=1}^{n}\widetilde{E}(z_i^2) - n\widetilde{E}(\bar{z}^2)\right\}
$$

$$
\leq \frac{1}{n\{n(\alpha+1)+1\}}\sum_{i=1}^{n}\widetilde{E}(z_i^2)
$$

$$
= \frac{1}{n\{n(\alpha+1)+1\}}\sum_{i=1}^{n} R_n |B_i|^{-1/2} t_{\gamma_n - p + 2}(x; \mu_i, B_i),
$$

30

using (27). Using $|B_i| \geq D_n^p |H|$ for $i = 1, \ldots, n$ as before, we have

$$\widetilde{E}[\widetilde{\text{var}}\{f(x) \mid z^{(n)}\}] \leq \frac{R_n D_n^{-p/2} \widehat{f}_{var}(x)}{\{n(\alpha + 1) + 1\}|H_n|^{1/2}}. \tag{31}$$

Combining (28) and (31) and putting the results back in (26), we have the desired result. If we let $n \to \infty$, we immediately obtain that $\widetilde{\text{var}}\{f(x)\} \to 0$ in P_{f_0}-probability. ∎

Appendix D. Proof of Theorem 7

Proof We have iid data $\mathcal{X}^{(n)} = (X_1, \ldots, X_n)$ such that $X_1, \ldots, X_n \overset{iid}{\sim} f_0$, with f_0 satisfying Assumptions 1-3 for $p = 1$. Given the NN-DM estimator $f(x) = \sum_{i=1}^n \pi_i \phi(x; \eta_i, \sigma_i^2)$, we define the simplified NN-DM density estimator to be

$$g(x) = \frac{1}{n} \sum_{i=1}^n \phi(x; \eta_i, \sigma_i^2),$$

The simplified estimator $g(x)$ can be interpreted as a version of $f(x)$ with the Dirichlet weights being replaced by their pseudo-posterior mean. That is, $g(x) = \widetilde{E}\{f(x) \mid (\eta_1, \sigma_1^2), \ldots, (\eta_n, \sigma_n^2)\}$. The pseudo-posterior distribution of $g(x)$ is induced through the pseudo-posterior distributions of $\{(\eta_i, \sigma_i^2)\}_{i=1}^n$. The pseudo-posterior mean is of the form

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\lambda_i} t_{\gamma_n}\left(\frac{x - \mu_i}{\lambda_i}\right),$$

where $\lambda_i = \{(\nu_n + 1)/\nu_n\}^{1/2} \delta_i$. Let $h_n = (\nu_n \gamma_n)^{-1/2} (\nu_n + 1)^{1/2} (\gamma_0 \delta_0^2)^{1/2}$. Then

$$(nh_n)^{1/2} E_{P_{f_0}} |\hat{f}_n(x) - f_K(x)| \to 0, \tag{32}$$

for $k_n = o(n^{2/7})$ and $k_n \to \infty$ as $n \to \infty$ from Section B of the Appendix, where

$$f_K(x) = \frac{1}{nh_n} \sum_{i=1}^n t_{\gamma_n}\left(\frac{x - X_i}{h_n}\right).$$

We want to investigate the asymptotic distribution of $f(x)$ as $n \to \infty$. For that, we first investigate the asymptotic distribution of the simplified NN-DM estimator $g(x)$, and then show that $f(x)$ and $g(x)$ are asymptotically close in P_{f_0}-probability.

To derive the asymptotic distribution of $g(x)$, we begin with the asymptotic distribution of $f_K(x)$, which can be expressed as $f_K(x) = n^{-1} \sum_{i=1}^n u_{in}$, where $u_{in} = h_n^{-1} t_{\gamma_n}\{(x - X_i)/h_n\}$. Using Lyapunov's central limit theorem and denoting convergence in distribution under f_0 by d_0, we have

$$\frac{f_K(x) - E_{P_{f_0}}\{f_K(x)\}}{[\text{var}_{P_{f_0}}\{f_K(x)\}]^{1/2}} \overset{d_0}{\to} N(0, 1)$$

if

$$\frac{(\sum_{i=1}^n \rho_{in})^{1/r}}{(\sum_{i=1}^n \tau_{in}^2)^{1/2}} \to 0, \text{ as } n \to \infty, \tag{33}$$

for some $r > 2$, where $\rho_{in} = E|u_{in} - E(u_{in})|^r$ and $\tau_{in}^2 = E\{u_{in} - E(u_{in})\}^2$ for $i = 1, \ldots, n$. By standard calculations, we have

$$\tau_{in}^2 = \frac{f_0(x)}{h_n} \int t_{\gamma_n}^2(u)du + o\left(\frac{1}{h_n}\right).$$

For $r = 3$,

$$\rho_{in} \leq \frac{8f_0(x)}{h_n^2} \int t_{\gamma_n}^3(u)du + o\left(\frac{1}{h_n^2}\right).$$

It is straightforward to see that $\int t_{\gamma_n}^r(u)du / \int t_{\gamma_n}(u)du = \mathcal{O}(1)$ for any $r \geq 1$. So, Lyapunov's condition is satisfied as the ratio in this case satisfies $\mathcal{O}\{(nh_n)^{-1/6}\}$ and $nh_n \to \infty$. Additionally, $|\tau_{in}^2 - \{f_0(x)/h_n\} \int \phi^2(u)\, du| \to 0$. So by a combination of Lyapunov's central limit theorem and Slutsky's theorem, we have

$$(nh_n)^{1/2}\left[f_K(x) - E_{P_{f_0}}\{f_K(x)\}\right] \overset{d_0}{\to} \mathrm{N}\left(0, \frac{f_0(x)}{2\pi^{1/2}}\right), \tag{34}$$

since $\int \phi^2(u)\, du = (2\pi^{1/2})^{-1}$. From the calculations in Section F of the Appendix, we can expand the Taylor series to two more terms to obtain

$$E_{P_{f_0}}\left\{f_K(x) - f_0(x) - \frac{h_n^2 f_0''(x)}{2}\right\} = \mathcal{O}(h_n^4),$$

since $|f_0^{(4)}(x)| \leq C_0$ for all $x \in [0, 1]$. Thus,

$$(nh_n)^{1/2}\left[f_K(x) - \left\{f_0(x) + \frac{h_n^2 f_0^{(2)}(x)}{2}\right\}\right] \overset{d_0}{\to} \mathrm{N}\left(0, \frac{f_0(x)}{2\pi^{1/2}}\right), \tag{35}$$

provided $n^{-2/9}k_n \to \infty$ as $n \to \infty$, implying $(nh_n)^{1/2}h_n^4 \to 0$.

We now argue that $(nh_n)^{1/2}|g(x) - \hat{f}_n(x)| \to 0$ in P_{f_0}-probability. For this, we first look at

$$E_{P_{f_0}}\left[nh_n\{g(x) - \hat{f}_n(x)\}^2\right] = nh_n\, E_{P_{f_0}}\left[\widetilde{E}\left\{(g(x) - \hat{f}_n(x))^2\right\}\right]$$
$$= nh_n\, E_{P_{f_0}}\left[\widetilde{\mathrm{var}}\{g(x)\}\right],$$

since $\widetilde{E}\{g(x)\} = \hat{f}_n(x)$. The pseudo-posterior variance of $g(x)$ is given by

$$\widetilde{\mathrm{var}}\{g(x)\} = \frac{1}{n^2}\sum_{i=1}^{n} \widetilde{\mathrm{var}}\{Z_i(x)\},$$

where $Z_i(x) = \phi(x; \eta_i, \sigma_i^2)$ for $i = 1, \ldots, n$. It is straightforward to show that

$$\widetilde{\mathrm{var}}\{Z_i(x)\} \sim |\Delta_n|\frac{1}{\widetilde{\lambda}_i^2}\, t_{2\gamma_n + 1}\left(\frac{x - \mu_i}{\widetilde{\lambda}_i}\right), \tag{36}$$

32

as $n \to \infty$, where $\widetilde{\lambda}_i^2 = \lambda_i^2/2$ for $i = 1, \ldots, n$ and

$$\Delta_n = \frac{u_{\gamma n}^2}{u_{2\gamma_n+1}} - \frac{1}{(2\pi)^{1/2}},$$

with $u_d = \Gamma\{(d+1)/2\}/\{(d\pi)^{1/2}\Gamma(d/2)\}$ being the normalizing constant of the Student's t-density with degrees of freedom $d > 0$. Using Stirling's approximation, $\Delta_n \to 0$ as $n \to \infty$. This immediately implies

$$nh_n \widetilde{\mathrm{var}}\{g(x)\} \leq |\Delta_n| v_g(x),$$

where

$$v_g(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\widetilde{\lambda}_i} t_{2\gamma_n+1}\left(\frac{x - \mu_i}{\widetilde{\lambda}_i}\right).$$

Using the techniques of Section B of the Appendix, it can be shown that $E_{P_{f_0}}\{v_g(x)\} \to f_0(x)$ as $n \to \infty$. Therefore, we have

$$\begin{aligned} E_{P_{f_0}}\left[nh_n\{g(x) - \hat{f}_n(x)\}^2\right] &= nh_n E_{P_{f_0}}\left[\widetilde{E}\left\{(g(x) - \hat{f}_n(x))^2\right\}\right] \\ &= nh_n E_{P_{f_0}}\left[\widetilde{\mathrm{var}}\{g(x)\}\right] \\ &\leq E_{P_{f_0}}\{|\Delta_n|v_g(x)\} \\ &\to 0, \end{aligned}$$

as $n \to \infty$. A simple application of Chebychev's inequality implies $(nh_n)^{1/2}|g(x) - \hat{f}_n(x)| \to 0$ in P_{f_0}-probability as $n \to \infty$. Combining this with (32) and (35) and using Slutsky's theorem, we obtain the desired result for $g(x)$.

We now demonstrate that $f(x)$ and $g(x)$ are asymptotically close to derive the same result for $f(x)$. We start out with

$$\begin{aligned} \mathrm{var}_{P_{f_0}}\left[(nh_n)^{1/2}\{f(x) - g(x)\}\right] &= nh_n E_{P_{f_0}}\left[\widetilde{\mathrm{var}}\{f(x) - g(x)\}\right] \\ &= nh_n E_{P_{f_0}}\left[\widetilde{\mathrm{var}}\left\{\sum_{i=1}^{n}\left(\pi_i - \frac{1}{n}\right)Z_i(x)\right\}\right] \end{aligned} \tag{37}$$

We now focus on the term inside $E_{P_{f_0}}$ in (37) and further decompose it as

$$\begin{aligned} \widetilde{\mathrm{var}}\left\{\sum_{i=1}^{n}\left(\pi_i - \frac{1}{n}\right)Z_i(x)\right\} &= \widetilde{E}\left[\widetilde{\mathrm{var}}\left\{\sum_{i=1}^{n}\left(\pi_i - \frac{1}{n}\right)Z_i(x) \mid \pi^{(n)}\right\}\right] \\ &+ \widetilde{\mathrm{var}}\left[\widetilde{E}\left\{\sum_{i=1}^{n}\left(\pi_i - \frac{1}{n}\right)Z_i(x) \mid \pi^{(n)}\right\}\right], \end{aligned} \tag{38}$$

where $\pi^{(n)} = (\pi_1, \ldots, \pi_n)^{\mathrm{T}}$. In (38), let Ξ_{1n} and Ξ_{2n} be as follows:

$$\Xi_{1n} = \widetilde{E}\left[\widetilde{\mathrm{var}}\left\{\sum_{i=1}^{n}\left(\pi_i - \frac{1}{n}\right)Z_i(x) \mid \pi^{(n)}\right\}\right],$$

33

$$\Xi_{2n} = \widetilde{\text{var}} \left[\widetilde{\text{E}} \left\{ \sum_{i=1}^{n} \left(\pi_i - \frac{1}{n} \right) Z_i(x) \mid \pi^{(n)} \right\} \right].$$

Thus, we can write (37) as

$$\text{var}_{P_{f_0}} \left[(nh_n)^{1/2} \{ f(x) - g(x) \} \right] = nh_n E_{P_{f_0}}(\Xi_{1n}) + nh_n E_{P_{f_0}}(\Xi_{2n}). \tag{39}$$

It is straightforward to see that $\Xi_{1n} = \sum_{i=1}^{n} \widetilde{\text{var}}(\pi_i) \widetilde{\text{var}}\{Z_i(x)\}$. As $n \to \infty$, we use the fact that $\pi_i \sim \text{Beta}(\alpha_n + 1, (n-1)(\alpha_n + 1))$ under $\widetilde{\Pi}$ and (36), to get

$$nh_n E_{P_{f_0}}(\Xi_{1n}) \sim \frac{n \Delta_n}{n(\alpha_n + 1) + 1} \widetilde{\Xi}_{1n}, \tag{40}$$

for some $\widetilde{\Xi}_{1n}$ satisfying $\widetilde{\Xi}_{1n} \to f_0(x)$ and $\Delta_n \to 0$ as $n \to \infty$. Therefore, $nh_n E_{P_{f_0}}(\Xi_{1n}) \to 0$ as $n \to \infty$. For the second part, let $d_i(x) = (1/\lambda_i) t_{\gamma_n}\{(x - \mu_i)/\lambda_i\}$ so that the pseudo-posterior mean $\hat{f}_n(x) = (1/n) \sum_{i=1}^{n} d_i(x)$. We first observe that

$$\begin{aligned}
\Xi_{2n} &= \widetilde{\text{var}} \left[\sum_{i=1}^{n} \left(\pi_i - \frac{1}{n} \right) \frac{1}{\lambda_i} t_{\gamma_n} \left(\frac{x - \mu_i}{\lambda_i} \right) \right] \\
&= \widetilde{\text{var}} \left[\sum_{i=1}^{n} \pi_i d_i(x) \right] \\
&= \frac{1}{n(\alpha_n + 1) + 1} \left[\frac{1}{n} \sum_{i=1}^{n} \{ d_i(x) - \hat{f}_n(x) \}^2 \right] \\
&\leq \frac{1}{n(\alpha_n + 1) + 1} \left[\frac{1}{n} \sum_{i=1}^{n} d_i^2(x) \right].
\end{aligned}$$

It now follows from some algebra that

$$\frac{1}{n} \sum_{i=1}^{n} d_i^2(x) \leq \frac{d_0(n)\gamma_n}{2\gamma_n + 1} \left(\frac{2\gamma_n + 1}{\gamma_n} \right)^{1/2} \frac{1}{h_n} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\tilde{\lambda}_i} t_{2\gamma_n + 1} \left(\frac{x - \mu_i}{\tilde{\lambda}_i} \right) \right],$$

where $d_0(n) \to (2\pi)^{-1/2}$ as $n \to \infty$. Therefore, we have

$$nh_n E_{P_{f_0}}(\Xi_{2n}) \leq \mathcal{O} \left\{ \frac{1}{2\pi^{1/2}} \frac{n}{n(\alpha_n + 1) + 1} \widetilde{\Xi}_{2n} \right\}, \tag{41}$$

for some $\widetilde{\Xi}_{2n}$ satisfying $\widetilde{\Xi}_{2n} \to f_0(x)$ as $n \to \infty$. By the conditions of the theorem, we have $nh_n E_{P_{f_0}}(\Xi_{2n}) \to 0$ as $n \to \infty$. This, along with (40) substituted in (39) provides $(nh_n)^{1/2} |f(x) - g(x)| \to 0$ in P_{f_0}-probability. This implies the desired result for $f(x)$ using Slutsky's theorem. As a result, we can interpret pseudo-credible intervals to be frequentist confidence intervals, on average, asymptotically. ∎

Appendix E. Proof of Theorem 8

E.1 A Property of the k_n-Nearest Neighbor Distance

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} f_0$ with f_0 a density on $\mathbb{R}^p$ satisfying Assumptions 1-3. We denote the induced probability measure P_{f_0} by P_0 in this section for the sake of convenience. We define the smoothed k-nearest neighborhood of X_i as $\mathcal{B}_i = \{y \in \mathbb{R}^p : ||X_i - y||_2 \leq R_i\}$, where $R_i = ||X_i - X_{i[k_n]}||_2$ is the Euclidean distance between X_i and the k_n-nearest neighbor of X_i for $i = 1, \ldots, n$. By symmetry, $R_1, \ldots, R_n$ are identically distributed. Suppose $r_n = (k_n/n)^{1/p}$ and define the quasi-neighborhood $\tilde{\mathcal{B}}_i(r) = \{y \in \mathbb{R}^p : ||X_i - y||_2 \leq r\}$, where the random variables R_i have been replaced by $r \geq 0$. Let

$$\omega_x(r) = \int\limits_{\{y : ||y-x||_2 \leq r\}} f_0(y) \, dy.$$

The positive density condition on f_0 obtained from Assumptions 1 and 3 (Evans et al., 2002; Evans, 2008) ensures the existence of $A > 1$ and $\rho > 0$ such that for all $0 \leq r \leq \rho$ and for all $x \in [0, 1]^p$,

$$\frac{r^p}{A} \leq \omega_x(r) \leq Ar^p. \tag{42}$$

We first state a Lemma proving some important properties of R_1. We next use this Lemma to prove Theorem 8. Recall that two non-negative sequences (a_n) and (b_n) are said to be asymptotically equivalent if $|a_n/b_n| \to c_0$ for some $c_0 > 0$, denoted by $a_n \sim b_n$.

Lemma 11 *Define $i_0 = \{2/(p^2+p+2)\} \wedge \{4/(p+2)^2\}$ as in Theorem 4. Assume $k_n \sim n^{i_0-\epsilon}$ for some $\epsilon \in (0, i_0)$. Suppose $\delta > 0$ satisfies*

$$\delta < (1 - i_0 + \epsilon)^{-1} - 1.$$

Define

$$r_n = \left(\frac{k_n}{n}\right)^{1/p}, \quad c_n = \frac{1}{(Ae)^{1/p}} r_n^{1+\delta}, \quad \text{and} \quad n_0 = \left\lfloor \left\{ \frac{1 - \frac{i_0-\epsilon}{2}}{\delta(1 - i_0 + \epsilon)} + 1 \right\}^{\frac{1}{i_0-\epsilon}} \right\rfloor + 1.$$

Then, the following results hold:

(i) $p_n = P_0(R_1 \leq c_n) = \mathcal{O}\left[k_n^{-1/2} \left(\frac{k_n}{n}\right)^{(k_n-1)\delta} \right]$ as $n \to \infty$.

(ii) $\sum\limits_{n=n_0+1}^{\infty} p_n < \infty$. That is, $\{p_n\}_{n=1}^{\infty}$ is summable.

(iii) $P\left[\limsup\limits_{n\to\infty} \{R_1 \leq c_n\} \right] = 0$.

(iv) $nc_n^p \to \infty$ as $n \to \infty$.

Proof

(i) Note that $c_n \leq \rho$ for sufficiently large n. From Lemma 4.1 of Evans et al. (2002) we have

$$P_0(R_1 \leq c_n \mid X_i = x) = \int\limits_0^{\omega_x(c_n)} (k_n - 1)\binom{n-1}{k_n - 1} y^{k_n - 2}(1-y)^{n-k_n} dy$$

$$\leq \binom{n-1}{k_n - 1} \omega_x(c_n)^{k_n - 1}$$

$$\sim k_n^{-1/2} \left(\frac{k_n}{n}\right)^{(k_n - 1)\delta},$$

for any $x \in [0,1]^p$, using (42). This immediately implies

$$P_0(R_1 \leq c_n) = \int_{[0,1]^p} P_0(R_i \leq c_n \mid X_i = x) f_0(x)\, dx \leq \mathcal{O}\left[k_n^{-1/2}\left(\frac{k_n}{n}\right)^{(k_n-1)\delta}\right].$$

(ii) For $n > n_0$, we have $p_n = \mathcal{O}\{n^{-(1+\Theta_n)}\}$ for a sequence $\Theta_n \to \infty$, $\Theta_n > 0$. This ensures that $\sum_{n=n_0+1}^{\infty} p_n < \infty$.

(iii) Since $\sum_{n=1}^{\infty} p_n < \infty$, a direct application of the first Borel-Cantelli lemma proves the statement.

(iv) We have, using the condition on δ,

$$nc_n^p = \frac{n}{Ae}\left(\frac{k_n}{n}\right)^{1+\delta}$$

$$= \frac{1}{Ae}\frac{k_n^{1+\delta}}{n^\delta}$$

$$\to \infty, \quad \text{as } n \to \infty.$$

$\blacksquare$

We now use the above Lemma to prove Theorem 8. The key idea is to leverage the fact that $R_i > c_n$ for all $i = 1, \ldots, n$ with probability 1 for all but finite n.

E.2 Number of Effective Member Points in Each Neighborhood

We now prove Theorem 8.

Proof Using (iii) from Lemma 11, for $i = 1, \ldots, n$, we have an integer $\widetilde{N}_i$ such that for all $n \geq \widetilde{N}_i$, $P_0(R_i > c_n) = 1$. However, since $R_1, \ldots, R_n$ are identically distributed, $\widetilde{N}_1 = \ldots = \widetilde{N}_n = \widetilde{N}$, say. Thus, for all $i = 1, \ldots, n$, we have $P_0(R_i > c_n) = 1$ for all $n \geq \widetilde{N}$. This immediately implies that $P_0\left[\bigcap_{i=1}^n \{R_i > c_n\}\right] = 1 - P_0\left[\bigcup_{i=1}^n \{R_i \leq c_n\}\right] \geq 1 - \sum_{i=1}^n P_0[R_i \leq c_n] = 1$ for all $n \geq \widetilde{N}$, which shows $P_0\left[\bigcap_{i=1}^n \{R_i > c_n\}\right] = 1$. Therefore,

we have

$$nQ_n = nP_0 \left[X_2 \in \mathcal{B}_1, \bigcap_{i=3}^{n} \{X_2 \notin \mathcal{B}_i\}, \bigcap_{i=1}^{n} \{R_i > c_n\} \right]$$

$$\leq nP_0 \left[\bigcap_{i=3}^{n} \{X_2 \notin \tilde{\mathcal{B}}_i(c_n)\} \right]$$

$$= nP_0 \left[\bigcap_{i=3}^{n} \{\|X_2 - X_i\| > c_n\} \right]$$

$$= n \int \theta_n^{n-2}(x) \, f_0(x) \, dx,$$

where $\theta_n(x) = 1 - \omega_x(c_n)$, since $X_1, \ldots, X_n \overset{iid}{\sim} f_0$. Using (42), we have $\theta_n(x) \leq 1 - (c_n^p/A)$ for all $x \in [0,1]^p$. Given the conditions on k, it follows that as $n \to \infty$,

$$\log n + n \log \left(1 - \frac{c_n^p}{A} \right) \sim \log n - \frac{n^\xi}{A^2 e} \to -\infty,$$

for all $x \in [0,1]^p$, where $\xi = 1 - (1 + \epsilon - i_0)(1 + \delta) > 0$. Therefore, we have

$$nQ_n = n \int \theta_n^{n-2}(x) f_0(x) \, dx$$

$$\leq n \left[1 - \frac{c_n^p}{A} \right]^{n-2} \int f_0(x) \, dx$$

$$= \mathcal{O} \left[\exp \left\{ \log n + n \log \left(1 - \frac{c_n^p}{A} \right) \right\} \right]$$

$$\to 0,$$

as $n \to \infty$. This proves the result. ∎

Appendix F. Proof of Consistency of Kernel Estimator

Define the standard multivariate t-density with $d > 0$ degrees of freedom to be $g_d(x) = t_d(x; 0_p, \mathbb{I}_p)$. Since $H = H_n = h_n^2 \mathbb{I}_p$ as defined in Section 3.1 is diagonal, it immediately follows that $t_{\gamma_n - p + 1}(x; \mu, H) = h_n^{-p} g_{\gamma_n - p + 1}\{h_n^{-1}(x - \mu)\}$. The following lemma proves the consistency of any such generic kernel density estimator with t kernel depending on n, say

$$f_K(x) = \frac{1}{nw^p} \sum_{i=1}^{n} g_{\gamma_n - p + 1} \left(\frac{x - X_i}{w} \right),$$

where the bandwidth $w = w_n$ satisfies $w_n \to 0$ and $nw_n^p \to \infty$ as $n \to \infty$, with independent and identically distributed data $X_1, \ldots, X_n \sim f_0$ satisfying Assumptions 1-3.

Lemma 12 *Suppose w_n is a sequence satisfying $w_n \to 0$ and $nw_n^p \to \infty$ as $n \to \infty$. Let $f_K(x) = (nw_n^p)^{-1} \sum_{i=1}^{n} g_{\gamma_n - p + 1}\{w_n^{-1}(x - X_i)\}$. Then $f_K(x) \to f_0(x)$ in P_{f_0}-probability for each $x \in [0,1]^p$.*

Proof It is enough to show that $E_{P_{f_0}}\{f_K(x)\} \to f_0(x)$ and $\text{var}_{P_{f_0}}\{f_K(x)\} \to 0$ as $n \to \infty$. Let us start first with $E_{P_{f_0}}\{f_K(x)\}$. We have

$$
\begin{aligned}
E_{P_{f_0}}\{f_K(x)\} &= E_{P_{f_0}} \left\{ \frac{1}{w_n^p} g_{\gamma_n-p+1} \left(\frac{x-X_1}{w_n} \right) \right\} \\
&= \int_{[0,1]^p} \frac{1}{w_n^p} g_{\gamma_n-p+1} \left(\frac{y-x}{w_n} \right) f_0(y)\, dy \\
&= \int_{\left[-\frac{x}{w_n}, \frac{1-x}{w_n}\right]^p} g_{\gamma_n-p+1}(u)\, f_0(x+w_n u)\, du, \\
&= \int_{\left[-\frac{x}{w_n}, \frac{1-x}{w_n}\right]^p} g_{\gamma_n-p+1}(u)\, \{f_0(x) + w_n u^{\mathrm{T}} \nabla f_0(\xi)\}\, du \\
&= f_0(x) \int_{\left[-\frac{x}{w_n}, \frac{1-x}{w_n}\right]^p} g_{\gamma_n-p+1}(u)\, du + w_n \int_{\left[-\frac{x}{w_n}, \frac{1-x}{w_n}\right]^p} g_{\gamma_n-p+1}(u) u^{\mathrm{T}} \nabla f_0(\xi)\, du, \\
&= f_0(x)\{1 - o_n(1)\} + w_n \mathcal{O}_n(1),
\end{aligned}
$$

using the mean value theorem and Polya's theorem (Pólya, 1920) along with Assumption 2 to bound $\nabla f_0(\cdot)$. As $n \to \infty$, this implies that $E_{P_{f_0}}\{f_K(x)\} \to f_0(x)$ since $w_n \to 0$ as $n \to \infty$.

The variance may be dealt with in a similar manner. Following the same steps as before we get

$$
\begin{aligned}
\text{var}_{P_{f_0}}\{f_K(x)\} &= \frac{1}{n} \text{var}_{P_{f_0}} \left\{ \frac{1}{w_n^p} g_{\gamma_n-p+1} \left(\frac{x-X_1}{w_n} \right) \right\} \leq \frac{1}{n} E_{P_{f_0}} \left\{ \frac{1}{w_n^{2p}} g_{\gamma_n-p+1}^2 \left(\frac{x-X_1}{w_n} \right) \right\} \\
&\leq \frac{1}{n w_n^{2p}} \int_{[0,1]^p} g_{\gamma_n-p+1}^2 \left(\frac{y-x}{w_n} \right) f_0(y)\, dy \\
&\leq \frac{1}{n w_n^p} \int_{\left[-\frac{x}{w_n}, \frac{1-x}{w_n}\right]^p} g_{\gamma_n-p+1}^2(u) \{f_0(x) + w_n u^{\mathrm{T}} \nabla f_0(\xi)\}\, du, \\
&\leq \frac{f_0(x) \mathcal{O}_n(1)}{n w_n^p},
\end{aligned}
$$

which shows that the variance goes to 0 as $n \to \infty$, since $n w_n^p \to \infty$ as $n \to \infty$. ∎

For the nearest neighbor-Dirichlet mixture, recall $f_K(x) = (1/n) \sum_{i=1}^n t_{\gamma_n-p+1}(x; X_i, H_n)$ from Section 3.1 of the main document, where $H_n = h_n^2 \mathbb{I}_p$ and $h_n^2 = \{\nu_n(\gamma_n-p+1)\}^{-1}\{(\nu_n + 1)(\gamma_0-p+1)\}\delta_0^2$. Here, the bandwidth h_n satisfies $h_n \to 0$ and $nh_n^p \to \infty$ as $n \to \infty$. Lemma 12 then shows that $f_K(x)$ converges to $f_0(x)$ in P_{f_0}-probability as $n \to \infty$.

Appendix G. Cross-validation

G.1 Algorithm for Leave-one-out Cross-validation

Consider independent and identically distributed data $X_1, \ldots, X_n \in \mathbb{R}^p \sim f$ with f having the nearest neighbor-Dirichlet mixture formulation. The prior of the neighborhood parameters (η_i, Σ_i) following Sections 2.2 and 2.3 is $(\eta_i, \Sigma_i) \sim \text{NIW}_p(\mu_0, \nu_0, \gamma_0, \Psi_0)$ where $\Psi_0 = (\gamma_* \delta_0^2)\, \mathbb{I}_p$ with $\gamma_* = \gamma_0 - p + 1$. We use the pseudo-posterior mean in (8) to compute

leave-one-out log-likelihoods $\mathbf{L}(\delta_0^2)$ for different choices of the hyperparameter δ_0^2, choosing $\delta_{0,\mathrm{CV}}^2 = \arg\sup_{\delta_0^2} \mathbf{L}(\delta_0^2)$ to maximize this criteria. The details of the computation of $\mathbf{L}(\delta_0^2)$ for a fixed δ_0^2 are provided in Algorithm 2.

- **Step 1:** Consider data $\mathcal{X}^{(n)} = (X_1, \ldots, X_n)$ where $X_i \in \mathbb{R}^p$, $p \geq 1$.

  Fix the number of neighbors k and other hyperparameters μ_0, ν_0, γ_0.

- **Step 2:** For $i \in \{1, \ldots, n\}$, consider the data set leaving out the ith data point, given by $\mathcal{X}^{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$.

  Compute the pseudo-posterior mean density estimate at X_i, namely $\hat{f}_{-i}(X_i)$, using $\mathcal{X}^{-i}$ and (8); let $\mathbf{L}_i(\delta_0^2) = \hat{f}_{-i}(X_i)$.

- **Step 3:** Compute the leave-one-out log-likelihood given by

$$\mathbf{L}(\delta_0^2) = \frac{1}{n}\sum_{i=1}^{n}\log\{\mathbf{L}_i(\delta_0^2)\}.$$

- **Step 4:** For $\delta_0^2 > 0$, obtain $\delta_{0,\mathrm{CV}}^2 = \arg\sup_{\delta_0^2} \mathbf{L}(\delta_0^2)$.

Algorithm 2: Leave-one-out cross-validation for choosing the hyperparameter δ_0^2 in nearest neighbor-Dirichlet mixture method.

G.2 Fast Implementation of Cross-validation

In Algorithm 2, the nearest neighborhood specification for each $\mathcal{X}^{-i}$ is different for $i = 1, \ldots, n$. However, we bypass this computation by initially forming a neighborhood of size $(k+1)$ for each data point using the entire data and storing the respective neighborhood means and covariance matrices. Suppose for X_i, the indices of the $(k+1)$-nearest neighbors are given by $\widetilde{\mathcal{N}}_i = \{j \in \{1, \ldots, n\} : \|X_i - X_j\|_2 \leq \|X_i - X_{i[k+1]}\|_2\}$, arranged in increasing order according to their distance from X_i with $X_{i[1]} = X_i$. Define the neighborhood mean $m_i = \{1/(k+1)\}\sum_{j \in \widetilde{\mathcal{N}}_i} X_j$ and the neighborhood covariance matrix $S_i = (k+1)^{-1}\{\sum_{j \in \widetilde{\mathcal{N}}_i}(X_j - m_i)(X_j - m_i)^{\mathrm{T}}\}$. Then, to form a k-nearest neighborhood for the new data $\mathcal{X}^{-i}$, a single pass over the initial neighborhoods $\widetilde{\mathcal{N}}_i$ is sufficient to update the new neighborhood means and covariance matrices. Below, we describe the update for the neighborhood means $m_j^{(-i)}$ and covariance matrices $S_j^{(-i)}$ for $j = 1, \ldots, n$ and $j \neq i$, considering the data $\mathcal{X}^{-i}$. For $j = 1, \ldots, n$ and $j \neq i$, we have,

$$m_j^{(-i)} = \begin{cases} (1/k)\{(k+1)m_j - X_{j[k+1]}\} & \text{if } i \notin \widetilde{\mathcal{N}}_j, \\ (1/k)\{(k+1)m_j - X_i\} & \text{if } i \in \widetilde{\mathcal{N}}_j. \end{cases}$$

$$S_j^{(-i)} = \begin{cases} S_j - \{(k+1)/k\}(m_j - X_{j[k+1]})(m_j - X_{j[k+1]})^{\mathrm{T}} & \text{if } i \notin \widetilde{\mathcal{N}}_j, \\ S_j - \{(k+1)/k\}(m_j - X_i)(m_j - X_i)^{\mathrm{T}} & \text{if } i \in \widetilde{\mathcal{N}}_j. \end{cases} \quad (43)$$

Appendix H. Algorithm with Gaussian Kernels for Univariate Data

For $p = 1$, we have a univariate Gaussian density $\phi(x; \eta_i, \sigma_i^2)$ in neighborhood i and normal-inverse gamma priors $(\eta_i, \sigma_i^2) \sim \text{NIG}(\mu_0, \nu_0, \gamma_0/2, \gamma_0\delta_0^2/2)$ independently for $i = 1, \ldots, n$, with $\mu_0 \in \mathbb{R}$ and $\nu_0, \gamma_0, \delta_0^2 > 0$. That is,

$$\eta_i \mid \sigma_i^2 \sim \text{N}\left(\mu_0, \frac{\sigma_i^2}{\nu_0}\right), \quad \sigma_i^2 \sim \text{IG}\left(\frac{\gamma_0}{2}, \frac{\gamma_0\delta_0^2}{2}\right).$$

Monte Carlo samples from the pseudo-posterior of the unknown density f at any point x can be generated following the steps of Algorithm 3.

- **Step 1:** Compute the k-nearest neighborhood $\mathcal{N}_i$ for X_i with $X_{i[1]} = X_i$, using the distance $d(\cdot, \cdot)$.

- **Step 2:** Update the parameters for neighborhood $\mathcal{N}_i$ to $(\mu_i, \nu_n, \gamma_n/2, \gamma_n\delta_i^2/2)$, where $\nu_n = \nu_0 + k$, $\gamma_n = \gamma_0 + k$,

$$\mu_i = \frac{\nu_0\mu_0 + k\bar{X}_i}{\nu_n}, \quad \bar{X}_i = \frac{1}{k}\sum_{j \in \mathcal{N}_i} X_j,$$

  and $\gamma_n\delta_i^2 = \gamma_0\delta_0^2 + \sum_{j \in \mathcal{N}_i}\left(X_j - \bar{X}_i\right)^2 + k\nu_0\nu_n^{-1}\left(\mu_0 - \bar{X}_i\right)^2$.

- **Step 3:** To compute the t-th Monte Carlo sample of $f(x)$, sample $\pi^{(t)} \sim \text{Dirichlet}(\alpha + 1, \ldots, \alpha + 1)$ and $(\eta_i^{(t)}, \sigma_i^{(t)2}) \sim \text{NIG}\left(\mu_i, \nu_n, \gamma_n/2, \gamma_n\delta_i^2/2\right)$ independently for $i = 1, \ldots, n$, and set

$$f^{(t)}(x) = \sum_{i=1}^n \pi_i^{(t)}\phi(x; \eta_i^{(t)}, \sigma_i^{(t)2}).$$

Algorithm 3: Nearest neighbor-Dirichlet mixture algorithm to obtain Monte Carlo samples from the pseudo-posterior of $f(x)$ with Gaussian kernel and normal-inverse gamma prior.

Appendix I. Inverse Wishart Parametrization

The parametrization of the inverse Wishart density defined on the set of all $p \times p$ matrices with real entries used in this article is given as follows. Suppose $\gamma > p - 1$ and Ψ is a $p \times p$ positive definite matrix. If $\Sigma \sim \text{IW}_p(\gamma, \Psi)$, then Σ has the following density function:

$$g(\Sigma) = \begin{cases} \dfrac{|\Psi|^{\gamma/2}}{2^{\gamma p/2}\Gamma_p\left(\frac{\gamma}{2}\right)}|\Sigma|^{-(\gamma+p+1)/2}\,\text{etr}\left(-\frac{1}{2}\Psi\Sigma^{-1}\right) & \text{if } \Sigma \text{ is positive definite,} \\[20pt] 0 & \text{otherwise,} \end{cases}$$

where $\Gamma_p(\cdot)$ is the multivariate gamma function given by

$$\Gamma_p(a) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\left(a + \frac{1-j}{2}\right),$$

for $a \geq (p-1)/2$ and the function $\mathrm{etr}(A) = \exp\{\mathrm{tr}(A)\}$ for a square matrix A. When $p = 1$, the $\mathrm{IW}_p(\gamma, \Psi)$ density is the same as the $\mathrm{IG}(\gamma/2, \gamma\delta^2/2)$ density, where $\delta^2 = \Psi/\gamma$. The $\mathrm{IW}_p(\gamma, \Psi)$ distribution has mean $\Psi/(\nu - p - 1)$ for $\nu > p + 1$ and mode $\Psi/(\nu + p + 1)$.

Appendix J. $\mathcal{L}_1$ Error Tables in Sections 4.2 and 4.3

Sample size	Estimator	CA	CW	DE	GS	IE	LN	LO	SB	SP	ST
	NN-DM	0.20	0.31	0.19	0.12	0.36	0.20	0.13	0.16	0.30	0.31
200	NN-DM (default)	0.21	0.37	0.17	0.12	0.34	0.20	0.14	0.17	0.31	0.32
	DP-MC	0.17	0.37	0.14	0.10	0.36	0.22	0.13	0.23	0.27	0.55
	KDE	-	0.37	0.16	0.12	-	0.18	0.11	0.18	-	0.52
	KNN	5.99	0.58	0.59	0.28	3.46	0.54	0.48	0.39	6.02	0.46
	DP-VB	0.20	0.35	0.15	0.08	0.53	0.25	0.11	0.11	0.44	0.57
	RD	-	0.35	0.13	0.12	-	0.16	0.11	0.16	-	0.53
	PTM	0.29	0.27	0.18	0.13	0.38	0.22	0.13	0.20	0.40	0.39
	LLDE	0.19	0.36	0.14	0.10	-	0.15	0.10	0.18	-	0.55
	OPT	0.32	0.36	0.28	0.17	0.55	0.21	0.02	0.18	0.75	0.52
	A-KDE	0.22	0.35	0.15	0.14	0.46	0.16	0.11	0.17	0.38	0.53
	NN-DM	0.16	0.17	0.13	0.08	0.30	0.16	0.10	0.10	0.24	0.20
500	NN-DM (default)	0.16	0.36	0.12	0.09	0.30	0.17	0.10	0.12	0.25	0.22
	DP-MC	0.11	0.35	0.10	0.08	0.27	0.18	0.09	0.13	0.22	0.53
	KDE	-	0.32	0.11	0.08	-	0.15	0.08	0.11	-	0.51
	KNN	3.62	0.47	0.48	0.27	3.39	0.40	0.30	0.31	5.64	0.35
	DP-VB	0.14	0.33	0.11	0.05	0.48	0.19	0.08	0.08	0.45	0.55
	RD	-	0.32	0.10	0.09	-	0.13	0.09	0.10	-	0.50
	PTM	0.24	0.19	0.14	0.10	0.32	0.19	0.11	0.14	0.32	0.30
	LLDE	0.17	0.35	0.11	0.08	-	0.15	0.08	0.14	-	0.53
	OPT	0.27	0.31	0.16	0.12	0.51	0.16	0.01	0.14	0.72	0.46
	A-KDE	0.16	0.32	0.10	0.10	0.40	0.13	0.10	0.10	0.27	0.52

Table 5: Comparison of the methods in terms of $\mathcal{L}_1$ error in the univariate case. Number of test points and replications considered are $n_t = 500$ and $R = 20$, respectively.

Sample size	Density / Dimension	MG 2	MG 3	MG 4	MG 6	MST 2	MST 3	MST 4	MST 6	MVC 2	MVC 3	MVC 4	MVC 6	MVG 2	MVG 3	MVG 4	MVG 6	SN 2	SN 3	SN 4	SN 6	T 2	T 3	T 4	T 6
200	NN-DM	0.29	0.39	0.46	0.63	0.27	0.37	0.43	0.59	0.33	0.48	0.52	0.62	0.33	0.50	0.61	0.74	0.23	0.32	0.40	0.56	0.26	0.33	0.38	0.52
	NN-DM (default)	0.29	0.40	0.47	0.65	0.28	0.38	0.45	0.61	0.37	0.49	0.61	0.80	0.38	0.50	0.62	0.75	0.24	0.34	0.42	0.58	0.26	0.33	0.40	0.53
	DP-MC	0.20	0.36	0.62	0.67	0.22	0.42	0.55	0.68	0.31	0.46	0.51	0.60	0.34	0.52	0.61	0.73	0.16	0.18	0.25	0.32	0.21	0.26	0.34	0.44
	KDE	0.28	0.52	0.79	1.29	0.27	0.55	0.72	1.21	-	-	-	-	0.43	0.77	0.90	1.09	0.22	0.50	0.78	1.26	0.26	0.53	0.74	1.21
	KNN	1.93	3.82	4.80	18.83	7.28	8.22	9.25	11.05	4.92	7.31	15.24	20.5	3.30	3.65	4.75	5.54	2.21	5.16	8.37	10.04	2.97	6.37	10.22	17.54
	DP-VB	0.29	0.38	0.41	0.50	0.24	0.36	0.44	0.59	0.48	0.85	1.28	1.69	0.45	0.58	0.71	0.86	0.17	0.23	0.31	0.52	0.19	0.29	0.34	0.46
	LLDE	0.32	0.52	0.93	1.82	0.28	0.67	1.02	11.38	0.53	-	-	-	0.32	0.45	0.75	0.97	0.16	0.22	0.29	0.65	0.17	0.26	0.35	2.05
	OPT	0.58	0.79	1.04	-	0.57	1.08	1.31	-	0.59	0.82	1.01	-	0.43	0.69	0.86	-	0.48	0.68	0.88	-	0.57	0.78	1.10	-
1000	NN-DM	0.18	0.26	0.34	0.46	0.19	0.27	0.32	0.44	0.28	0.33	0.46	0.50	0.29	0.44	0.49	0.62	0.15	0.22	0.29	0.42	0.17	0.24	0.30	0.38
	NN-DM (default)	0.22	0.30	0.37	0.48	0.21	0.29	0.33	0.44	0.31	0.41	0.52	0.66	0.36	0.48	0.55	0.68	0.22	0.29	0.36	0.45	0.22	0.28	0.33	0.39
	DP-MC	0.08	0.39	0.57	0.58	0.11	0.18	0.21	0.47	0.26	0.39	0.48	0.54	0.21	0.38	0.51	0.64	0.06	0.07	0.10	0.15	0.09	0.15	0.17	0.28
	KDE	0.16	0.32	0.52	0.96	0.16	0.35	0.53	1.04	-	-	-	-	0.33	0.66	0.73	0.93	0.13	0.32	0.53	1.05	0.14	0.32	0.52	0.90
	KNN	0.92	2.62	4.01	15.28	5.96	6.48	7.04	9.63	4.68	6.29	13.7	17.04	2.01	2.59	3.88	4.09	1.89	4.39	6.84	8.15	2.37	5.30	9.66	13.28
	DP-VB	0.25	0.29	0.33	0.36	0.15	0.24	0.25	0.45	0.42	0.74	0.82	0.91	0.38	0.54	0.61	0.77	0.10	0.12	0.15	0.20	0.10	0.15	0.18	0.31
	LLDE	0.31	0.37	0.52	1.02	0.22	0.40	0.46	1.18	0.47	-	-	-	0.29	0.39	0.51	0.94	0.07	0.10	0.14	0.27	0.10	0.15	0.23	0.38
	OPT	0.39	0.72	1.00	-	0.43	0.84	1.10	-	0.51	0.72	0.89	-	0.30	0.60	0.85	-	0.31	0.50	0.76	-	0.33	0.53	0.94	-

Table 6: Comparison of the methods in terms of $\mathcal{L}_1$ error in the multivariate case. Number of test points and replications considered are $n_t = 500$ and $R = 20$, respectively.

42

References

Ian S Abramson. On bandwidth variation in kernel estimates-a square root law. *The Annals of Statistics*, 10(4):1217–1223, 1982.

Adelchi Azzalini. The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, 32(2):159–188, 2005.

Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer, 2015.

David M Blei and Michael I Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 2006.

Adrian W Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360, 1984.

Leo Breiman, William Meisel, and Edward Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19(2):135–144, 1977.

Luc Devroye and Laszlo Gyorfi. *Nonparametric Density Estimation: the L_1 view*. Wiley Series in Probability and Statistics, 1985.

Or Dinari, Angel Yu, Oren Freifeld, and John Fisher. Distributed MCMC inference in Dirichlet process mixture models using Julia. In *2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, pages 518–525, 2019. doi: 10.1109/CCGRID.2019.00066.

Tarn Duong. *ks: Kernel Smoothing*, 2020. URL `https://CRAN.R-project.org/package=ks`. R package version 1.11.7.

Dafydd Evans. A law of large numbers for nearest neighbour statistics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 464(2100):3175–3192, 2008.

Dafydd Evans, Antonia J Jones, and Wolfgang M Schmidt. Asymptotic moments of near–neighbour distance distributions. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 458(2028):2839–2849, 2002.

Subhashis Ghosal and Aad van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697–723, 2007.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*, volume 44. Cambridge University Press, 2017.

Subhashis Ghosal, Jayanta K Ghosh, and RV Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158, 1999.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Gene H Golub and Charles F van Loan. *Matrix Computations*. John Hopkins University Press, 3rd edition, 1996.

P Richard Hahn, Ryan Martin, and Stephen G Walker. On recursive Bayesian predictive distributions. *Journal of the American Statistical Association*, 113(523):1085–1093, 2018.

Peter Hall. On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, 15 (4):1491–1519, 1987.

Nils Lid Hjort and M Chris Jones. Locally parametric nonparametric density estimation. *The Annals of Statistics*, pages 1619–1647, 1996.

Michael C Hughes and Erik B Sudderth. Bnpy: Reliable and scalable variational inference for Bayesian nonparametric models. In *Proceedings of the NIPS Probabilistic Programimming Workshop, Montreal, QC, Canada*, pages 8–13, 2014.

Gordon J. Ross and Dean Markwick. *dirichletprocess: Build Dirichlet Process Objects for Bayesian Modelling*, 2019. URL `https://CRAN.R-project.org/package=dirichletprocess`. R package version 0.3.1.

Alejandro Jara, Timothy Hanson, Fernando Quintana, Peter Müller, and Gary Rosner. DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*, 40(5):1–30, 2011. URL `http://www.jstatsoft.org/v40/i05/`.

MJ Keith, A Jameson, W Van Straten, M Bailes, S Johnston, M Kramer, A Possenti, SD Bates, NDR Bhat, M Burgay, et al. The High Time Resolution Universe Pulsar Survey I, System configuration and initial discoveries. *Monthly Notices of the Royal Astronomical Society*, 409(2):619–627, 2010.

Kenichi Kurihara, Max Welling, and Nikos Vlassis. Accelerated variational Dirichlet process mixtures. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL `https://proceedings.neurips.cc/paper/2006/file/2bd235c31c97855b7ef2dc8b414779af-Paper.pdf`.

Michael Lavine. Some aspects of Polya tree distributions for statistical modelling. *Annals of Statistics*, 20(3):1222–1235, 1992.

Michael Lavine. More aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, 22(3):1161–1176, 1994.

Clive Loader. *Local regression and likelihood*. Springer Science & Business Media, 2006.

Clive R Loader. Local likelihood density estimation. *The Annals of Statistics*, 24(4):1602–1618, 1996.

Don O Loftsgaarden and Charles P Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.

Duncan Ross Lorimer and Michael Kramer. *Handbook of Pulsar Astronomy*, 2012.

Robert James Lyon. *Why are pulsars hard to find?* PhD thesis, The University of Manchester (United Kingdom), 2016.

Hengzhao Ma and Jianzhong Li. A true O($n \log n$) algorithm for the all-k-nearest-neighbors problem. In *International Conference on Combinatorial Optimization and Applications*, pages 362–374. Springer, 2019.

YP Mack and Murray Rosenblatt. Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9(1):1–15, 1979.

Thoralf Mildenberger and Henrike Weinert. The benchden package: Benchmark densities for nonparametric density estimation. *Journal of Statistical Software*, 46(14):1–14, 2012. URL http://www.jstatsoft.org/v46/i14/.

Jeffrey W Miller and David B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2019.

Michael A Newton. On a nonparametric recursive estimator of the mixing distribution. *Sankhyā: The Indian Journal of Statistics, Series A*, 64(2):306–322, 2002.

Michael A Newton and Yunlei Zhang. A recursive algorithm for nonparametric analysis with missing data. *Biometrika*, 86(1):15–26, 1999.

Yang Ni, Yuan Ji, and Peter Müller. Consensus Monte Carlo for random subsets using shared anchors. *Journal of Computational and Graphical Statistics*, 29(4):1–12, 2020.

Georg Pólya. Über den zentralen grenzwertsatz der wahrscheinlichkeitsrechnung und das momentenproblem. *Mathematische Zeitschrift*, 8(3-4):171–181, 1920.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

Judith Rousseau and Kerrie Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710, 2011.

David W Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, 2015.

Simon J Sheather and Michael C Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):683–690, 1991.

Hanyu Song, Yingjian Wang, and David B. Dunson. Distributed Bayesian clustering using finite mixture of mixtures. *arXiv preprint*, page arXiv:2003.13936, 2020.

Peter Xue-Kun Song. Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics*, 27(2):305–320, 2000.

George R Terrell and David W Scott. Variable kernel density estimation. *The Annals of Statistics*, pages 1236–1265, 1992.

Gerald Teschl. Mathematical methods in quantum mechanics. *Graduate Studies in Mathematics*, 99:106, 2009.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, NY, 1st edition, 2009.

Pravin M. Vaidya. An optimal algorithm for the all-nearest-neighbors problem. In *27th Annual Symposium on Foundations of Computer Science*, pages 117–122, 1986.

Matt P Wand and M Chris Jones. Multivariate plug-in bandwidth selection. *Computational Statistics*, 9(2):97–116, 1994.

Lianming Wang and David B. Dunson. Fast Bayesian inference in Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 20(1):196–216, 2011.

Mike West. *Hyperparameter estimation in Dirichlet process mixture models*. Duke University ISDS Discussion Paper# 92-A03, 1992.

Wing H Wong and Li Ma. Optional Polya tree and Bayesian inference. *The Annals of Statistics*, 38(3):1433–1459, 2010.

Xiaole Zhang, David J Nott, Christopher Yau, and Ajay Jasra. A sequential algorithm for fast fitting of Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 23(4):1143–1162, 2014.