Quasi-Bayesian Nonparametric Density Estimation via Autoregressive Predictive Updates

Sahra Ghalebikesabi[1] **Chris Holmes**[2] **Edwin Fong**[*2] **Brieuc Lehmann**[*3]

[1]University of Oxford
[2]Novo Nordisk
[3]University College London

Abstract

Bayesian methods are a popular choice for statistical inference in small-data regimes due to the regularization effect induced by the prior. In the context of density estimation, the standard nonparametric Bayesian approach is to target the posterior predictive of the Dirichlet process mixture model. In general, direct estimation of the posterior predictive is intractable and so methods typically resort to approximating the posterior distribution as an intermediate step. The recent development of quasi-Bayesian predictive copula updates, however, has made it possible to perform tractable predictive density estimation without the need for posterior approximation. Although these estimators are computationally appealing, they struggle on non-smooth data distributions. This is due to the comparatively restrictive form of the likelihood models from which the proposed copula updates were derived. To address this shortcoming, we consider a Bayesian nonparametric model with an autoregressive likelihood decomposition and a Gaussian process prior. While the predictive update of such a model is typically intractable, we derive a quasi-Bayesian update that achieves state-of-the-art results in small-data regimes.

1 INTRODUCTION

Modelling the joint distribution of multivariate random variables with density estimators is a central topic in modern unsupervised machine learning research [Durkan et al., 2019, Papamakarios et al., 2017]. As well as providing insight into the statistical properties of the data, density estimates are used in a number of downstream applications, including image restoration [Zoran and Weiss, 2011], density-based clustering [Scaldelai et al., 2022], and simulation-based inference [Lueckmann et al., 2021]. In small-data regimes, Bayesian methods are a popular choice for a wide range of machine learning tasks, including density estimation, thanks to their attractive generalization capacities. For density estimation, the typical Bayesian approach is to target the *Bayesian predictive density*, $p_n(x) = \int f(x|\theta)\pi_n(\theta)d\theta$, where π_n denotes the posterior density of the model parameters θ after observing $x_1, \ldots, x_n$, and f denotes the likelihood function.

De Finetti's representation theorem [De Finetti, 1937, Hewitt and Savage, 1955] states that an exchangeable joint density fully characterises a Bayesian model, which then implies a sequence of predictive densities. Further, Fong et al. [2021] recently showed that a sequence of predictive densities can be sufficient for full Bayesian posterior inference. This provides theoretical motivation for an iterative approach to Bayesian predictive density estimation by updating the predictive $p_{i-1}(x)$ to $p_i(x)$ given observation x_i for $i = 1, \ldots, n$. The idea of recursive Bayesian updates goes back to at least Hill [1968], but was only recently made more widely applicable through the relaxation of the assumption of exchangeability in favour of conditionally identically distributed [Berti et al., 2004] sequences.

Here, we focus on a particular class of one-step-ahead predictive updates $p_{i-1}(x) \to p_i(x)$ based on bivariate copulas, which were first introduced by Hahn et al. [2018] for univariate data, and extended by Fong et al. [2021] to the multivariate setting and to regression analyses. This class of updates is inspired by Bayesian models and thus retains many desirable Bayesian properties, such as coherence and regularization. However, we emphasize that the copula updates do not correspond exactly, nor approximately, to a traditional Bayesian likelihood-prior model, and we thus refer to them

[*]equal contribution

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

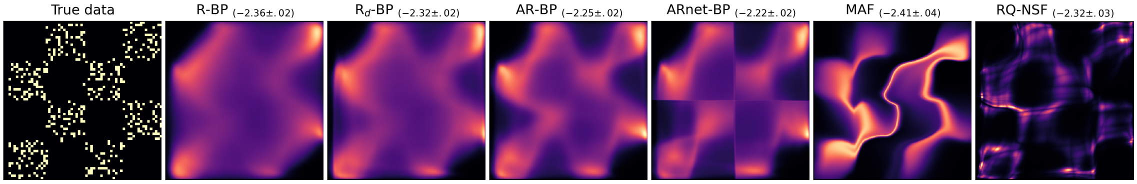

Figure 1: Density estimates of 600 observations from a chessboard distribution, reported with mean and standard deviation of test log likelihoods. For larger training sizes, see Supplement C.2. Our methods, AR-BP and ARnet-BP, outperform R-BP and AR neural networks.

as *quasi-Bayesian* [Fortini and Petrone, 2020]. The most related Bayesian density estimator proposed to date, henceforth referred to as the **R**ecursive **B**ayesian **P**redictive (R$_d$-BP), lacks flexibility to model highly complex data distributions (see Figure 1). This is because the existing copula updates rely on a Gaussian copula with a single scalar bandwidth parameter, corresponding to a Bayesian model with a likelihood that factorizes over dimensions. In contrast, popular neural network based approaches, such as masked autoregressive flows (MAFs) [Papamakarios et al., 2017], and rational-quadratic neural spline flows (RQ-NSFs) [Durkan et al., 2019] can struggle in small-data regimes (see Figure 1).

Contributions This motivates our main contribution, namely the formulation of a more flexible autoregressive (AR) copula update based on which we propose a new Dirichlet Process Mixture Model (DPMM) inspired density estimator. In particular:

- By considering a DPMM with an AR likelihood and a Gaussian process (GP) prior, we formulate a tractable copula update with a novel *data-dependent bandwidth* based on the Euclidean metric in data space. Our method, **A**utoregressive **R**ecursive **B**ayesian **P**redictives (AR-BP), outperforms traditional density estimators on tabular data with up to 63 features, and 10,000 samples.
- We observe in practice that the Euclidean metric used in AR-BP can be inadequate for highly non-smooth data distributions. For such cases, we propose using an AR neural network [Bengio and Bengio, 1999, Frey et al., 1998, Germain et al., 2015, Larochelle and Murray, 2011] that maps the observations into a latent space before bandwidth estimation. This introduces additional non-linearity through the dependence of the bandwidth on the data, leading to a density estimator, ARnet-BP, that is more accurate on non-smooth densities.

2 BACKGROUND

We briefly recap predictive density estimation via bivariate copula updates, before describing a particular such update inspired by DPMMs.

2.1 UNIVARIATE PREDICTIVE DENSITY UPDATES

To compute predictive densities quickly, Hahn et al. [2018] propose an iterative approach. For $x \in \mathbb{R}$, any sequence of Bayesian posterior predictive densities $p_i(x)$ with likelihood f and posterior π_i, conditional on $x_{1:i}$, can be expressed as

$$p_i(x) = \int f(x|\theta)\pi_i(\theta)d\theta = p_{i-1}(x)h_i(x, x_i), \qquad (1)$$

for some bivariate function $h_i(x, x_i)$ [Hahn et al., 2018]. Rearranging for h_i, we have

$$h_i(x, x_i) = \frac{p_i(x)}{p_{i-1}(x)} \overset{(a)}{=} \frac{p_{i-1}(x|x_i)}{p_{i-1}(x)} \overset{(b)}{=} \frac{p_{i-1}(x, x_i)}{p_{i-1}(x)p_{i-1}(x_i)} \quad (2)$$

where (a) holds by definition, and (b) $p_{i-1}(x, x_i) = p_{i-1}(x|x_i)p_{i-1}(x_i) = p_i(x)p_{i-1}(x_i)$ holds by Bayes' law. Hahn et al. [2018] show that $h_i(x, x_i)$ is the transformation of a bivariate copula density. A *bivariate copula* is a bivariate cumulative distribution function (CDF) $C : [0, 1]^2 \to [0, 1]$ with uniform marginal distributions that is used to characterise the dependence between two random variables independent of their marginals:

Theorem 1 (Sklar's theorem [Sklar, 1959])**.** *For any bivariate density $f(y_1, y_2)$ with continuous marginal CDFs, $F_1(y_1)$ and $F_2(y_2)$, and marginal densities $f_1(y_1)$ and $f_2(y_2)$, there exists a unique bivariate copula C with density c such that*

$$f(y_1, y_2) = c\left\{F_1(y_1), F_2(y_2)\right\} f_1(y_1)f_2(y_2).$$

Applying the copula factorization from Sklar's theorem to (2) yields that there exists some bivariate copula density c_i such that $p_{i-1}(x, x_i) = c_i\{P_{i-1}(x), P_{i-1}(x_i)\}p_{i-1}(x)p_{i-1}(x_i)$, and thus $h_i(x, x_i) = c_i\{P_{i-1}(x), P_{i-1}(x_i)\}$, where P_{i-1} is the CDF corresponding to the predictive density p_{i-1}. Given prior π and likelihood f, Equation 2 suggests that the update function can be written as

$$h_i(x, x_i) = \frac{\int f(x|\theta)f(x_i|\theta)\pi_{i-1}(\theta)d\theta}{\int f(x|\theta)\pi_{i-1}(\theta)d\theta \int f(x_i|\theta)\pi_{i-1}(\theta)d\theta}.$$

For each Bayesian model, there is thus a unique sequence of symmetric copula densities $c_i(u,v) = c_i(v,u)$. This sequence has the property that $c_n(\cdot,\cdot) \to 1$ converges to a constant function as $n \to \infty$, ensuring that the predictive density converges asymptotically with sample size n.

In general, the above equation is intractable due to the posterior so it is not possible to compute the iterative update in (1) for fully Bayesian models. Alternatively, we will consider sequences of h_i that match the Bayesian model for $i = 1$, but not for $i > 1$. As mentioned above, this copula update no longer corresponds to a Bayesian model, nor are the resulting predictive density estimates approximations to a Bayesian model. Nevertheless, if the copula updates are *conditionally identically distributed*, they still exhibit desirable Bayesian characteristics such as coherence and regularization, and are hence referred to as *quasi-Bayes*. Please refer to Berti et al. [2004] for details.

2.2 MULTIVARIATE PREDICTIVE DENSITY UPDATES

The above arguments cannot directly be extended to multivariate $x \in \mathbb{R}^d$ since h_i cannot necessarily be written as $c_i\{P_{i-1}(x), P_{i-1}(x_i)\}$ for $d > 1$. However, (2) still holds, and recursive predictive updates with bivariate copulas as building blocks can be derived explicitly given a pre-defined likelihood model and a prior, which we now exhibit.

Hahn et al. [2018] and Fong et al. [2021] propose to use DPMMs as a general-use nonparametric model. The DPMM [Escobar, 1988, Escobar and West, 1995] can be written as

$$f(x|G) = \int_{\Theta} K(x|\theta)\, dG(\theta), \text{ with } G \sim \mathrm{DP}(c, G_0) \quad (3)$$

where $\theta \in \Theta = \mathbb{R}^d$ are parameter vectors, the prior assigned to G is a Dirichlet process (DP) prior with base measure G_0 and concentration parameter $c > 0$ [Ferguson, 1973], and $K(x|\theta)$ is a user-specified kernel (not to be confused with the covariance function of a GP). In particular, Fong et al. [2021] consider the base measure $G_0 = \mathcal{N}(0, \tau^{-1} I_d)$ for some precision parameter $\tau \in \mathbb{R}_{>0}$, and the factorized kernel $K(x|\theta) = \mathcal{N}(x|\theta, I_d)$ where I_d is the d-dimensional identity matrix. The likelihood is then

$$f(x|G) = \int \prod_{j=1}^{d} \mathcal{N}\left(x^j \mid \theta^j, 1\right) dG(\theta), \quad (4)$$

where the dimensions of x are conditionally independent given θ. Following Hahn et al. [2018], we denote the dimension j of a vector y with y^j. We note that the strong assumption of a factorised kernel form drastically impacts the performance of the regular DPMM and also influences the form and modelling capacity of the corresponding copula update.

This model inspires the following recursive predictive density update $p_i(x) = h_i(x, x_i) p_{i-1}(x)$ for which the first $d' \in \{1, \ldots, d\}$ marginals take on the form

$$\begin{aligned} \frac{p_i(x^{1:d'})}{p_{i-1}\left(x^{1:d'}\right)} &= 1 - \alpha_i + \alpha_i \prod_{j=1}^{d'} c\left(u^j_{i-1}(x^j), v^j_{i-1}; \rho_0\right), \quad (5) \\ u^j_{i-1}(x^j) &:= P_{i-1}\left(x^j \mid x^{1:j-1}\right), \\ v^j_{i-1} &:= P_{i-1}\left(x^j_i \mid x^{1:j-1}_i\right), \end{aligned}$$

where $c(u, v; \rho_0)$ is the bivariate Gaussian copula density with correlation $\rho_0 = 1/(1+\tau)$, p_0 can be any chosen prior density, and $\alpha_i = \left(2 - \frac{1}{i}\right) \frac{1}{i+1}$ (see Supplement A and Fong et al. [2021]). Note that the above update requires a specific ordering of the feature dimensions, and the Gaussian copula follows from the Gaussian distribution in the kernel and G_0 for the DPMM. Unlike the DPMM, there are now no underlying parameters (beyond ρ_0) in the copula update as we have integrated out θ, so we do not carry out clustering directly. While ρ_0 is a scalar here, Fong et al. [2021] also consider the setting with a distinct bandwidth parameter for each dimension. We refer to these recursive Bayesian predictives as R$_d$-BP, or simply R-BP if the dimensions share a single bandwidth.

3 AR-BP: AUTOREGRESSIVE BAYESIAN PREDICTIVES

For smooth data distributions, the recursive update defined in (5) generates density estimates that are highly competitive against other popular density estimation procedures such as kernel density estimation (KDE) and DPMM [Fong et al., 2021]. Moreover, the iterative updates provide a fast estimation alternative to fitting the full DPMM through Markov chain Monte Carlo (MCMC). When considering more structured data, however, performance suffers due to the choices of the factorized kernel $K(\cdot|\theta) = \mathcal{N}(\cdot|\theta, I_d)$ and simple base measure $G_0 = \mathcal{N}(0, \tau^{-1} I_d)$ in the DPMM. These choices induce a priori independence between the data dimensions, and are thus insufficiently flexible to capture more complex dependencies.

3.1 BAYESIAN MODEL FORMULATION

We therefore propose employing more general kernels and base measures in the DPMM and show that these inspire a more general tractable recursive predictive update. In particular, we allow the kernel to take on an autoregressive structure

$$K(x|\theta) = \prod_{j=1}^{d} \mathcal{N}\left(x^j \mid \theta^j\left(x^{1:j-1}\right), 1\right), \tag{6}$$

where $\theta^j : \mathbb{R}^{j-1} \to \mathbb{R}$ is now an unknown mean *function*, and not scalar, for dimension x^j, which we allow to depend on the previous $j-1$ dimensions of x. Thus, specifying our DPMM requires a base measure supported on the function space in which $(\theta^1, \ldots, \theta^d)$ is valued. We specify this base measure as a product of independent GP priors on the functional parameters

$$\theta^j \sim \mathrm{GP}(0, \tau^{-1}k^j) \text{ for } j = 1, ..., d \tag{7}$$

where $k^j : \mathbb{R}^{j-1} \times \mathbb{R}^{j-1} \to \mathbb{R}$ and k^j can be any given covariance function that takes as input a pair of $x^{1:j-1}$ values. In practice, we use the same functional form of k for each j, so we will drop the superscript j. For later convenience, we have also written the scaling term τ^{-1} explicitly. We highlight that for $j = 1$, $\theta^1 \sim \mathcal{N}(0, \tau^{-1})$. Under this choice, the mean of the normal kernels in the DPMM for each dimension j is thus a flexible function of the first $j-1$ dimensions $x^{1:j-1}$, on which we elicit independent GP priors. The conjugacy of the GP with the Gaussian DPMM kernel in (6) is crucial for deriving a tractable density update.

Remark. The proposed DPMM kernel in (6) is in fact more flexible than a general multivariate kernel, $K(x \mid \theta) = \mathcal{N}(x \mid \theta, \Sigma)$. This is because the multivariate kernel also implies an AR form like (6) but where the parameters θ^j are restricted to be linear in $x^{1:j-1}$; see Wade et al. [2014] for details.

3.2 ITERATIVE PREDICTIVE DENSITY UPDATES

Computing the Bayesian posterior predictive density induced by the DPMM with kernel given by (6) and base measure given by (7) through posterior estimation is *intractable* and requires MCMC. However, as before, we can utilize the model to derive tractable iterative copula updates. In Supplement A.1, we derive the corresponding recursive predictive density update $p_i(x) = h_i(x, x_i)p_{i-1}(x)$ for the first d' marginals and show that it takes on the form

$$\frac{p_i(x^{1:d'})}{p_{i-1}(x^{1:d'})} = 1 - \alpha_i + \alpha_i \cdot \prod_{j=1}^{d'} c\left(u_{i-1}^j(x^j), v_{i-1}^j; \rho^j(x^{1:j-1}, x_i^{1:j-1})\right), \tag{8}$$

with $u_{i-1}^j(x^j), v_{i-1}^j$ defined as in (5), $\alpha_i = \left(2 - \frac{1}{i}\right)\frac{1}{i+1}$, and the bandwidth given by

$$\rho^j(x^{1:j-1}, x_i^{1:j-1}) = \rho_0 k\left(x^{1:j-1}, x_i^{1:j-1}\right), \tag{9}$$

for $\rho_0 = 1/(1+\tau)$, and $\rho_i^1 = \rho_0$. Where appropriate, we henceforth drop the argument x for brevity. The conditional CDFs u_{i-1}^j can also be computed through an iterative closed form expression similarly to (8) (Supplement B.3). Please see Figure 2 for a simplified overview of the density estimation pipeline.

```
(a) Train: Estimate $v_i^j = P_{i-1}(x_i^{1:j})$ for each $i$
Initialise $u_0^j(x_i) \leftarrow \Phi(x_i^j)$
For each preceding observation $x_k$ with $k < i$:
    For each feature $j$:
        Compute data-dependent bandwidth $\rho^j(x_i^{1:j}, x_k^{1:j})$ (9)
        Update conditional CDF $u_i^j(x_k) := P_i^j(x_k)$ based on the similarity between $u_{i-1}^j(x_k)$ and $v_{i-1}^j$ (18)
Set $v_i^j \leftarrow u_i^j(x_i)$ for all $j$
(b) Test: Estimate predictive at test point $p_n(z)$
Initialise $u_0^j(z) \leftarrow \Phi(z^j)$
For each train observation $x_i$:
    For each feature $j$:
        Compute data-dependent bandwidth $\rho^j(x_i^{1:j}, z^{1:j})$ (9)
        Update conditional CDF $u_i^j(z) := P_i^j(z)$ based on the similarity between $u_{i-1}^j(z)$ and $v_{i-1}^j$ (18)
    Update predictive density $p_{i-1}(z) \to p_i(z)$ (8)
```

Figure 2: Simplified summary of AR-BP. We repeat the training update for each train datum x_i to estimate $v_i^j = P_{i-1}(x_i^{1:j})$. These are needed at test time to update from $p_{i-1}(z) \to p_i(z)$. All steps are averaged over different feature and sample permutations. The main step that induces autoregression in the observations is highlighted pink. Please see Supplement B.3 for detailed algorithms.

Note that the estimation is identical to the update given in (5) induced by the factorized DPMM kernel, except for the main difference that the bandwidth ρ is *no longer a constant*, but is now *data-dependent*. More precisely, the bandwidth for dimension j is a transformation of the GP covariance function k on the first $j-1$ dimensions. The additional flexibility afforded by the inclusion of k enables us to capture more complex dependency structures, as we do not enforce a-priori independence between the dimensions of the parameter θ. Similarly to the extension of R-BP to R$_d$-BP, we can also define AR$_d$-BP by introducing dimension dependence in ρ_0. Finally, we highlight that extending R-BP to mixed data is possible as given in Appendix E.1.3 of Fong et al. [2021], which also extends naturally to AR-BP.

Remark. The data-dependent bandwidth also appears when starting from other Bayesian nonparametric models, such as dependent DPs and GPs (see Supplement

A.2.2 for the derivation).

Our approach can be viewed as a Bayesian version of an online KDE procedure. To see this, note that a KDE trained on $i-1$ observations – yielding the density estimate $q_{i-1}(x)$ – can be updated after observing the i^{th} observation x_i via $q_i(x) = (1-\alpha_i)q_{i-1}(x) + \alpha_i d\,(x, x_i)$, where $\alpha_i = 1/i$ and $d(\cdot,\cdot)$ denotes the kernel of the KDE. Rather than adding a weighted kernel term directly, AR-BP instead adds an adaptive kernel that depends on a notion of distance between x and x_i based on the predictive CDFs conditional on $x_{1:i-1}$.

To better understand the importance of the data-dependent bandwidth, we compare the conditional predictive mean of R-BP and AR-BP in the bivariate setting $X \times Y$. Under the simplifying assumption of Gaussian predictive densities, we show in Supplement A.3 that the conditional mean of $Y \mid X$ is given by

$$\mu_i(x) = \mu_{i-1}(x) + \alpha_i(x, x_i)\rho(x, x_i)(y_i - \mu_{i-1}(x_i)),$$

$$\alpha_i(x, x_i) = \frac{\alpha_i c(P_{i-1}(x), P_{i-1}(x_i); \rho)}{1 - \alpha_i + \alpha_i c(P_{i-1}(x), P_{i-1}(x_i); \rho)}.$$

Note that $\rho(x, x_i) = \rho_0$ for R-BP. Intuitively, the updated mean is the previous mean plus a residual term at y_i scaled by some notion of distance between x and x_i. For R-BP, this distance between x and x_i depends only on their predictive CDF values through $\alpha_i(x, x_i)$. This can result in undesirable behaviour as shown in the upper plot in Figure 3(a), where the peak of $\alpha_i(x, x_i)$, as a function of x, is not centred at x_i. Counterintuitively, there is thus an $x > x_i$ where $\mu_i(x)$ is updated more than at the actual observed $x = x_i$. This follows from the lack of focus on *conditional* density estimates for R-BP, which is alleviated by AR-BP. In the AR case, $\rho(x, x_i)$ takes into account the Euclidean distance between x and x_i in the data space. We see in the lower plot in Figure 3(a) that the peak is closer to x_i. Figure 3(b) further demonstrates this difference on another toy example - we see that R-BP struggles to fit a linear conditional mean function for $n = 4$, focussing density in data sparse regions, while AR-BP succeeds to assign significant density only to points on the data manifold.

Training the update parameters In order to compute the predictive density $p_n(x^*)$, we require the vector of conditional CDFs $[v_1^j, \dots, v_{n-1}^j]$ where $v_i^j = P_i(x_{i+1}^j \mid x_{i+1}^{1:j-1})$. Given a bandwidth parameterization, obtaining this vector thus amounts to model-fitting, and each v_i^j requires $i-1$ iterations (Supplement B.3), for $i \in \{1, \dots, n\}$. We note that the order of samples and dimensions influences the prediction performance in AR density estimators [Vinyals et al., 2015]. In practice, averaging over different permutations of these improves performance (Supplement B.3). Full implementation details can be found in Supplement B.

Computational complexity The above procedure results in a computational complexity of $\mathcal{O}(Mdn^2)$ at the training stage where M is the number of permutations. At test time, we have already obtained the necessary conditional prequential CDFs v_n^j in computing the prequential log-likelihood above. As a result, we have a computational complexity $\mathcal{O}(Mdn)$ for each test observation. Note that the introduction of a data-dependent bandwidth does not increase the computational complexity at train or test time relative to R-BP and only adds a negligible factor to the computational time for the calculation of the bandwidth.

3.3 BANDWIDTH PARAMETERISATION

The choice of covariance function in (7) provides substantial modelling flexibility in our AR-BP framework. Moreover, the additional parameters associated with the covariance function allow us to tune the implied covariance structure according to the observed data. This formulation enables us to draw upon the rich literature on the choice of covariance functions for Gaussian processes [Williams and Rasmussen, 2006]. For simplicity we only consider the most popular such choice here, but study the more flexible rational-quadratic covariance in Supplement C.2. The radial basis function (RBF) covariance function is defined as $k_\ell(x^{1:j-1}, x'^{1:j-1}) = \exp[-\sum_{\kappa=1}^{j-1}\{(x^\kappa - x'^\kappa)/\ell^\kappa\}^2]$, where $\ell \in \mathbb{R}_{>0}^{d-1}$ is the length scale.

Neural parameterisation As we saw in the motivating example of the density estimation of a chessboard distribution in Figure 1, the RBF kernel can restrict the capacity of the predictive density update to capture intricate nonlinearities if the training data size is not sufficient. While the parameterization of the bandwidth in (9) was initially derived via the first predictive update for a DPMM, all we require is that the bandwidth function $\rho^j : \mathbb{R}^{j-1} \times \mathbb{R}^{j-1} \to \mathbb{R}$ lies in (0,1). We would also like $\rho^j(x^{1:j-1}, x'^{1:j-1})$ to take larger values when $x^{1:j-1}$ and $x'^{1:j-1}$ are ‘close’ in some sense. Motivated by this observation, we now consider more expressive bandwidth functions that can lead to increased predictive performance. In particular, we formulate an AR neural network $f_w : \mathbb{R}^d \to \mathbb{R}^{d \times d'}$ for $d' \in \mathbb{N}$ with the property that the j^{th} row of the output depends only on the first $j-1$ dimensions of the input. Let $Z = f_w(x)$ and denoting z^j to be the j^{th} row of the matrix Z, the covariance function is then computed as $\rho^j(x^{1:j-1}, x'^{1:j-1}) = \rho_0 \exp(-\sum_{\kappa=1}^{j-1} ||z^\kappa - z'^\kappa||_2^2)$.

Numerous AR neural network models have been extensively used for density estimation [Dinh et al., 2014, Huang et al., 2018, Kingma et al., 2016]. In our experiments, we use a relatively simple model with parameter

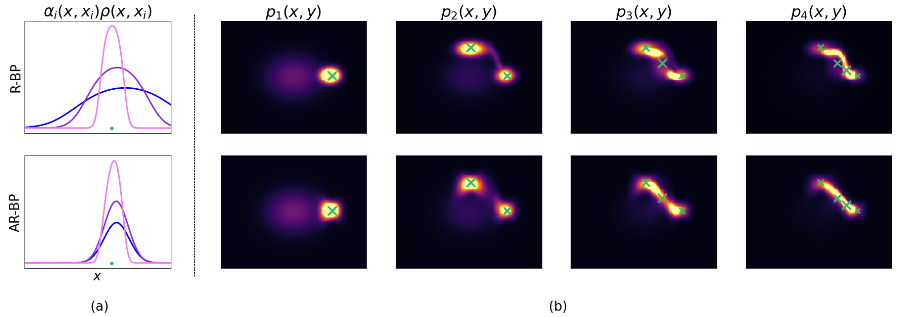

Figure 3: (a) Plots of $\alpha_i(x, x_i)\rho(x, x_i)$ for R-BP and AR-BP for $\rho_0 \in \{0.5, 0.7, 0.95\}$ (—,—,—) with new observation x_i (●). Note that $\rho(x, x_i) = \rho_0$ for R-BP, and $\ell = 1$ for AR-BP. (b) Density plots for R-BP and AR-BP trained on 4 sequential data points (×). Both figures show that the update of R-BP, unlike AR-BP, is not centred around the new datum.

sharing inspired by NADE, an AR neural network designed for density estimation [Larochelle and Murray, 2011]. More advanced properties like the permutation invariance of MADE [Papamakarios et al., 2017] create an additional overhead that cannot be used in the copula formulation as the predictive update is not permutation-invariant. We refer to Bayesian predictive densities estimated using AR neural networks as *ARnet Bayesian predictives* (ARnet-BP).

Tuning the bandwidth function Recall that the bandwidths $\rho_i(\cdot, \cdot)$ are parameterised by ρ_0 and the parameters of the chosen covariance functions or neural embedders. For AR-BP, these are the length scales ℓ of the RBF covariance function, while for ARnet-BP, these are the parameters w of the AR neural network. We fit these tunable parameters in a data-driven approach by maximising the prequential [Dawid, 1997] log-likelihood $\sum_{i=1}^{n} \log p_{i-1}(x_i)$ which is analogous to the Bayesian marginal likelihood – the tractable predictive density allows us to compute this exactly, and this approach is analogous to empirical Bayes. Specifically, we use gradient descent optimisation with Adam, sampling a different random permutation of the training data at each optimisation step (Supplement B.3).

4 RELATED WORK

Our work falls into the broad area of multivariate density estimation [Scott, 2015]. While AR networks have been previously used directly for the task of density estimation [Bengio and Bengio, 1999, Germain et al., 2015, Larochelle and Murray, 2011], we use them to elicit a data-dependent bandwidth in the predictive update to mitigate the smoothing effect observed in AR-BP. Neural network based approaches, however, often underperform in small-data regimes. Deep learning approaches that do target few-shot density estimation require complex meta-learning and pre-training pipelines [Gu et al., 2020, Reed et al., 2017].

Our work directly extends the contributions of Hahn et al. [2018] and Fong et al. [2021] through an alternative specification of the nonparametric Bayesian model in the recursive predictive update scheme. R-BP has recently been used for nonparametric solvency risk prediction [Hong and Martin, 2019], and survival analysis [Fong and Lehmann, 2022]. Berti et al. [2021a,b, 2004] also focus on univariate predictive updates in the Bayesian nonparametric paradigm, specifically exploring the use of the conditionally identically distributed condition as a relaxation of the standard exchangeability assumption. Other studies have investigated quasi-Bayesian updates in the special case of the mixing distribution in nonparametric mixture models [Dixit and Martin, 2022, Fortini and Petrone, 2020, Martin, 2018, Tokdar et al., 2009], though these typically focus on univariate or low-dimensional spaces. See also Martin [2021] for a survey.

Finally, copulas are a well-studied tool for modelling the correlations in multivariate data (see e.g. Kauermann et al. [2013], Ling et al. [2020], Nelsen [2007]). Copula density estimation aims to construct density estimates whose univariate marginals are uniform [Gijbels and Mielniczuk, 1990], and often focus on modelling strong tail dependencies [Wiese et al., 2019]. In contrast, we employ bivariate copulas for generic multivariate density estimation as a tool to model the correlations between subsequent subjective predictive densities, rather than across the data dimensions directly.

5 EXPERIMENTS

We demonstrate the benefits of AR-BP, AR$_d$-BP and ARnet-BP for density estimation and prediction tasks in an experimental study with five baseline approaches

Table 1: Average NLL with standard error over five runs on data sets analysed by Fong et al. [2021].

	WINE	BREAST	PARKIN	IONO	BOSTON
n/d	89/12	97/14	97/16	175/30	506/13
KDE	$13.69_{\pm 0.00}$	$10.45_{\pm 0.24}$	$12.83_{\pm 0.27}$	$32.06_{\pm 0.00}$	$8.34_{\pm 0.00}$
DPMM (Diag)	$17.46_{\pm 0.6}$	$16.26_{\pm 0.71}$	$22.28_{\pm 0.66}$	$35.30_{\pm 1.28}$	$7.64_{\pm 0.09}$
DPMM (Full)	$32.88_{\pm 0.82}$	$26.67_{\pm 1.32}$	$39.95_{\pm 1.56}$	$86.18_{\pm 10.22}$	$9.45_{\pm 0.43}$
MAF	$39.60_{\pm 1.41}$	$10.13_{\pm 0.40}$	$11.76_{\pm 0.45}$	$140.09_{\pm 4.03}$	$56.01_{\pm 27.74}$
RQ-NSF	$38.34_{\pm 0.63}$	$26.41_{\pm 0.57}$	$31.26_{\pm 0.31}$	$54.49_{\pm 0.65}$	$-2.20_{\pm 0.11}$
R-BP	$13.57_{\pm 0.04}$	$7.45_{\pm 0.02}$	$9.15_{\pm 0.04}$	$21.15_{\pm 0.04}$	$4.56_{\pm 0.04}$
R$_d$-BP	$13.32_{\pm 0.01}$	$6.12_{\pm 0.05}$	$7.52_{\pm 0.05}$	$19.82_{\pm 0.08}$	$-13.50_{\pm 0.59}$
AR-BP	$13.45_{\pm 0.05}$	$6.18_{\pm 0.05}$	$8.29_{\pm 0.11}$	$17.16_{\pm 0.25}$	$-0.45_{\pm 0.77}$
AR$_d$-BP	$\mathbf{13.22}_{\pm \mathbf{0.04}}$	$\mathbf{6.11}_{\pm \mathbf{0.04}}$	$\mathbf{7.21}_{\pm \mathbf{0.12}}$	$16.48_{\pm 0.26}$	$\mathbf{-14.75}_{\pm \mathbf{0.89}}$
ARnet-BP	$14.41_{\pm 0.11}$	$6.87_{\pm 0.23}$	$8.29_{\pm 0.17}$	$\mathbf{15.32}_{\pm \mathbf{0.35}}$	$-5.71_{\pm 0.62}$

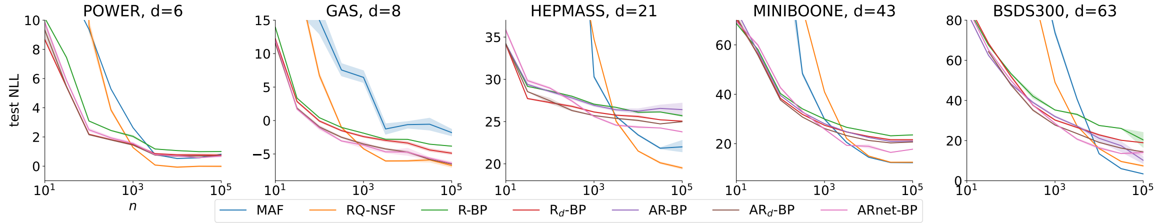

Figure 4: Average NLL and standard errors over 10 runs for training sets of different size. Our models outperform neural methods for data sets up to 10,000 samples.

and 13 different data sets. The code and data is available at `https://github.com/sghalebikesabi/autoregressive-bayesian-predictives`. See Supplement C for additional experimental details and results, including a sensitivity study, an ablation study, further illustrative examples, a preliminary investigation into image examples, and an empirical study of the computational complexity of the proposed methods.

5.1 DENSITY ESTIMATION

We compared our models against KDEs [Parzen, 1962], DPMMs [Rasmussen, 1999], MAFs [Papamakarios et al., 2017] and RQ-NSFs [Durkan et al., 2019]. The hyperparameters of the baselines were tuned with cross-validation. Unless otherwise specified, we use respectively 10 permutations over samples and features to average the quasi-Bayesian estimates. We did not see substantial improvements with more permutations. We use the same few hyperparameters (initialisation of $\rho_0, l_1, \ldots, l_d$, number of permutations, neural network architecture, and learning rate) on all data sets as our method is robust to their choice. See Supplement C.1 for further information.

Data sets analysed by Fong et al. [2021] See Table 1 for the negative log-likelihood (NLL) estimated on five UCI data sets [Asuncion and Newman, 2007] of small size with up to 506 samples, as investigated by Fong et al. [2021]. Our proposed methods display highly competitive performance: AR$_d$-BP achieved the best test NLL on four of the data sets, while ARnet-BP prevailed on ionosphere.

Data sets analysed by Papamakarios et al. [2017] A number of UCI data sets have become the standard evaluation benchmark for deep AR models [Durkan et al., 2019, Huang et al., 2018, Papamakarios et al., 2017]. These include low-dimensional data sets with up to 63 features, but at least 29,000 with up to 10^6 samples. In many circumstances, data sets of such a data size are not available. To investigate performance as a function of sample size, we trained the models on subsets of the full data set. We do not report results for the KDEs and the DPMM estimators here as these estimators performed significantly worse than the other approaches. Similarly, we do not report deep learning results for sample sizes smaller than 10^2. See Supplement C.2 for complete results.

In the small-data regime, we observe that the R-BP methods significantly outperform the neural density estimators (Figure 4). As the sample size increases, the gap in performance decreases until eventually the neural density estimators outcompete the R-BP methods. The performance between the R-BP methods and our proposed AR extensions is largely similar, though we note that the AR-BP methods were generally more effective on the GAS dataset.

5.2 SUPERVISED LEARNING

R-BP methods, including AR-BP, can be used for prediction tasks such as regression and classification [Fong

Table 2: Average NLL over five runs reported with standard error for supervised tasks

		Regression			Classification	
	BOSTON	CONCR	DIAB	IONO	PARKIN	MNIST01
n/d	506/13	1,030/8	442/10	351/33	195/22	12,031/784
Linear	$0.87_{\pm 0.03}$	$0.99_{\pm 0.01}$	$1.07_{\pm 0.01}$	$0.33_{\pm 0.01}$	$0.38_{\pm 0.01}$	$0.003_{\pm 0.000}$
GP	$0.42_{\pm 0.08}$	$0.36_{\pm 0.02}$	$1.06_{\pm 0.02}$	$0.30_{\pm 0.02}$	$0.42_{\pm 0.02}$	$0.035_{\pm 0.000}$
MLP	$1.42_{\pm 1.01}$	$2.01_{\pm 0.98}$	$3.32_{\pm 4.05}$	$0.26_{\pm 0.05}$	$0.31_{\pm 0.02}$	$\mathbf{0.003_{\pm 0.000}}$
R-BP	$0.76_{\pm 0.09}$	$0.87_{\pm 0.03}$	$1.05_{\pm 0.03}$	$0.26_{\pm 0.01}$	$0.37_{\pm 0.01}$	$0.015_{\pm 0.001}$
R$_d$-BP	$0.40_{\pm 0.03}$	$0.42_{\pm 0.00}$	$1.00_{\pm 0.02}$	$0.34_{\pm 0.02}$	$0.27_{\pm 0.03}$	$0.018_{\pm 0.001}$
AR-BP	$0.52_{\pm 0.13}$	$0.42_{\pm 0.01}$	$1.06_{\pm 0.02}$	$0.21_{\pm 0.02}$	$0.29_{\pm 0.02}$	$0.015_{\pm 0.001}$
AR$_d$-BP	$\mathbf{0.37_{\pm 0.10}}$	$0.39_{\pm 0.01}$	$\mathbf{0.99_{\pm 0.02}}$	$\mathbf{0.20_{\pm 0.02}}$	$0.28_{\pm 0.03}$	$0.017_{\pm 0.001}$
ARnet-BP	$0.45_{\pm 0.11}$	$\mathbf{-0.03_{\pm 0.00}}$	$1.41_{\pm 0.07}$	$0.24_{\pm 0.04}$	$\mathbf{0.26_{\pm 0.04}}$	$0.014_{\pm 0.001}$

et al., 2021]. In short, this is achieved by estimating the conditional predictive density $p_n(y|x)$ of the labels y directly by assuming a dependent Dirichlet process likelihood. See Supplement B.2 for details. Again, we follow the experimental set-up of Fong et al. [2021], and additionally report results on the MNIST data set, restricted to digits of class 0 and 1. We report the conditional test NLL $-\frac{1}{n'}\sum_i \log p_n(y_i^*|x_i^*)$ for a test set $\{(x_1^*, y_1^*), \ldots, (x_{n'}^*, y_{n'}^*)\}$. We compared our models against a GP, a linear Bayesian model (Linear), and a one-hidden-layer multilayer perceptron (MLP) on several classification and regression tasks. To get a distribution over the predicted outcome in the regression case, we trained an ensemble over 10 MLPs. Our proposed methods were again highly competitive (Table 2). AR$_d$-BP performed best on two regression tasks and one classification task. ARnet-BP was substantially better than the remaining methods on CONCR and also performed best on the PARKIN. On the other hand, the MLP model was best on MNIST.

6 DISCUSSION

Although Bayesian methods generally perform well in the small sample setting, the conventional Bayesian approach to density estimation, i.e. DPMM estimation via the posterior predictive, is computationally intensive. Here, we set out to propose a computationally efficient density estimator as an alternative to DPMM density estimation. We recommend its use for tabular data sets of up to 63 features, and 10,000 observations. Such data set sizes are ubiquitous in healthcare, finance, hyperparameter tuning, and survey data applications.

We expand upon the tractable recursive copula updates of Fong et al. [2021], Hahn et al. [2018] by incorporating regression methods, such as kernels and neural networks. This introduces a data-dependent bandwidth, thus increasing the flexibility of this class of models, with little computational overhead compared to R-BP. More generally, it would be of interest to integrate other machine learning methods with recursive copula updates. Furthermore, other Bayesian nonparametric models may inspire other recursive copula updates–see Appendix A.2 for an example based on GPs.

An appealing feature of AR-BP is that it requires no manual hyperparameter tuning. Further, on small data sets, AR-BP shows state-of-the-art generalization and is faster than competing deep learning models. It significantly increases the modelling capacity of the baseline R-BP via a data-dependent bandwidth. Additionally, ARnet-BP provides a useful illustration of how powerful neural network models can be incorporated into R-BP methods to improve density estimation. Future work can investigate alternative architectures for structured data. Our work adds to the rich body of density estimators and thus we do not anticipate any additional negative societal impact arising from our proposal.

This strong performance of AR-BP (and other copula methods) in the small data regime is likely due to its Bayesian-like regularization towards an initial density p_0, as shown in the weighted sum in (8). Its weaker performance in the large data regime may be due to the importance of the sequence α_i which governs how regularization decays, but further theoretical work is needed to understand AR-BP's asymptotic behaviour. A limitation of R-BP methods, including AR-BP, is the quadratic time dependence on the number of training observations. Subsampling techniques thus offer a particularly promising avenue to reduce the overall computational cost and warrant further investigation. Although the recursive updates depend on the sample and covariate ordering, it is possible to alleviate this dependence though by estimating the R-BP over multiple permutations in parallel, as we have done in the above experiments. Nevertheless, the algorithm is relatively fast: with a single GPU, we were able to train models with 100,000 observations in less than an hour.

The use of a GP prior greatly increases the flexibility of our framework. Moreover, it opens the door to future research to incorporate ideas from the vast GP literature to further boost performance in high-dimensional settings. Our use of the RBF kernel was illustrative; other kernels are discussed in Appendix C.2 where we find that the RBF kernel performs best. For example, we anticipate that the use of recent advances in convolutional kernels [Van der Wilk et al., 2017] would be particularly suited for computer vision tasks.

Acknowledgements

SG is a student of the EPSRC CDT in Modern Statistics and Statistical Machine Learning (EP/S023151/1) and receives funding from the Oxford Radcliffe Scholarship and Novartis. CH is supported by The Alan Turing Institute, Health Data Research UK, the Medical Research Council UK, the EPSRC through the Bayes4Health programme Grant EP/R018561/1, and AI for Science and Government UK Research and Innovation (UKRI). EF is supported by Novo Nordisk. BL is supported by the UK Engineering and Physical Sciences Research Council through the Bayes4Health programme (grant number EP/R018561/1) and gratefully acknowledges funding from Jesus College, Oxford.

REFERENCES

Arthur Asuncion and David Newman. UCI machine learning repository, 2007.

Yoshua Bengio and Samy Bengio. Modeling high-dimensional discrete data with multi-layer neural networks. *Advances in Neural Information Processing Systems*, 12, 1999.

Patrizia Berti, Luca Pratelli, and Pietro Rigo. Limit theorems for a class of identically distributed random variables. *The Annals of Probability*, 32(3):2029–2052, 2004.

Patrizia Berti, Emanuela Dreassi, Fabrizio Leisen, Pietro Rigo, and Luca Pratelli. Bayesian predictive inference without a prior. *arXiv preprint arXiv:2104.11643*, 2021a.

Patrizia Berti, Emanuela Dreassi, Luca Pratelli, and Pietro Rigo. A class of models for Bayesian predictive inference. *Bernoulli*, 27(1):702–726, 2021b.

A Philip Dawid. Prequential analysis. *Encyclopedia of Statistical Sciences*, 1:464–470, 1997.

Bruno De Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, 7:1–68, 1937.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Vaidehi Dixit and Ryan Martin. A prticle filter algorithm for nonparametric estimation of multivariate mixing distributions. *arXiv preprint arXiv:2204.01646*, 2022.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

Michael David Escobar. *Estimating the means of several normal populations by nonparametric estimation of the distribution of the means.* PhD thesis, Yale University, 1988.

Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

Edwin Fong and Brieuc Lehmann. A predictive approach to bayesian nonparametric survival analysis. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 6990–7013. PMLR, 28–30 Mar 2022. URL `https://proceedings.mlr.press/v151/fong22a.html`.

Edwin Fong, Chris Holmes, and Stephen G Walker. Martingale posterior distributions. *To appear at the Journal of the Royal Statistical Society: Series B (with discussion)*, 2021.

Sandra Fortini and Sonia Petrone. Quasi-bayes properties of a procedure for sequential learning in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1087–1114, 2020.

Brendan J Frey, J Frey Brendan, and Brendan J Frey. *Graphical models for machine learning and digital communication.* MIT press, 1998.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pages 881–889. PMLR, 2015.

Iène Gijbels and Jan Mielniczuk. Estimating the density of a copula function. *Communications in Statistics - Theory and Methods*, 19(2):445–464, January 1990. ISSN 0361-0926. doi: 10.1080/03610929008830212. URL `https://doi.org/10.1080/03610929008830212`.

Ke Gu, Yonghui Zhang, and Junfei Qiao. Ensemble meta-learning for few-shot soot density recognition. *IEEE Transactions on Industrial Informatics*, 17(3): 2261–2270, 2020.

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.

Edwin Hewitt and Leonard J Savage. Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, 80(2):470–501, 1955.

Bruce M Hill. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63(322): 677–691, 1968.

Liang Hong and Ryan Martin. Real-time Bayesian non-parametric prediction of solvency risk. *Annals of Actuarial Science*, 13(1):67–79, 2019.

Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, pages 2078–2087. PMLR, 2018.

Göran Kauermann, Christian Schellhase, and David Ruppert. Flexible copula density estimation with penalized hierarchical b-splines. *Scandinavian Journal of Statistics*, 40(4):685–705, 2013.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.

Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 29–37. JMLR Workshop and Conference Proceedings, 2011.

Chun Kai Ling, Fei Fang, and J Zico Kolter. Deep archimedean copulas. *Advances in Neural Information Processing Systems*, 33:1535–1545, 2020.

Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, pages 343–351. PMLR, 2021.

Ryan Martin. On nonparametric estimation of a mixing density via the predictive recursion algorithm. *arXiv preprint arXiv:1812.02149*, 2018.

Ryan Martin. A survey of nonparametric mixing density estimation via the predictive recursion algorithm. *Sankhya B*, 83(1):97–121, 2021.

R.B. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer New York, 2007. ISBN 978-0-387-28678-5.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

Carl Rasmussen. The infinite gaussian mixture model. *Advances in neural information processing systems*, 12, 1999.

Scott Reed, Yutian Chen, Thomas Paine, Aäron van den Oord, SM Eslami, Danilo Rezende, Oriol Vinyals, and Nando de Freitas. Few-shot autoregressive density estimation: Towards learning to learn distributions. *arXiv preprint arXiv:1710.10304*, 2017.

D Scaldelai, LC Matioli, SR Santos, and M Kleina. Multiclusterkde: a new algorithm for clustering based on multivariate kernel density estimation. *Journal of Applied Statistics*, 49(1):98–121, 2022.

David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

M Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

Surya T Tokdar, Ryan Martin, and Jayanta K Ghosh. Consistency of a recursive estimate of mixing distributions. *The Annals of Statistics*, pages 2502–2522, 2009.

Mark Van der Wilk, Carl Edward Rasmussen, and James Hensman. Convolutional gaussian processes. *Advances in Neural Information Processing Systems*, 30, 2017.

Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*, 2015.

Sara Wade, Stephen G Walker, and Sonia Petrone. A predictive study of dirichlet process mixture models for curve fitting. *Scandinavian Journal of Statistics*, 41(3):580–605, 2014.

Magnus Wiese, Robert Knobloch, and Ralf Korn. Copula & marginal flows: Disentangling the marginal from its joint. *arXiv preprint arXiv:1907.03361*, 2019.

Christopher K Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. Number 3 in 2. MIT press Cambridge, MA, 2006.

Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *2011 International Conference on Computer Vision*, pages 479–486. IEEE, 2011.