Learning Sparse Mixture Models

Fatima Antarou Ba

March 30, 2022

This work approximates high-dimensional density functions with an ANOVA-like sparse structure by the mixture of wrapped Gaussian and von Mises distributions. When the dimension $d$ is very large, it is complex and impossible to train the model parameters by the usually known learning algorithms due to the curse of dimensionality. Therefore, assuming that each component of the model depends on an a priori unknown much smaller number of variables than the space dimension $d$, we first define an algorithm that determines the mixture model’s set of active variables by the Kolmogorov-Smirnov and correlation test. Then restricting the learning procedure to the set of active variables, we iteratively determine the set of variable interactions of the marginal density function and simultaneously learn the parameters by the Kolmogorov and correlation coefficient statistic test and the proximal Expectation-Maximization algorithm. The learning procedure considerably reduces the algorithm’s complexity for the input dimension $d$ and increases the model’s accuracy for the given samples, as the numerical examples show.

Contents

1 Introduction 2
   1.1 Prior work .......................................................... 2
   1.2 Our contribution ...................................................... 3
   1.3 Outline of paper .................................................... 3

2 Preliminaries and notation 3

3 Sparse Mixture Models 3
   3.1 Sparse additive Model ............................................. 3
   3.2 Determination of Active Variables .............................. 8

4 Learning Sparse Mixture Models 13

1TU Berlin, Straße des 17. Juni 136, D-10587 Berlin, Germany, {fatimaba}@math.tu-berlin.de.
1 Introduction

One of the most recurrent problem of multivariate function approximation theory problems is the curse of dimensionality. An algorithm is said to face the curse of dimensionality if the algorithm depends exponentially on the dimension of the data. In order to circumvent or solve the problem, several authors have focused on the study of sparse functions with respect to their arguments. A widespread example appears in compressive sensing [19], where the target function can be spanned precisely by assuming that the input vector is sparse with respect to the $\|\cdot\|_0$-norm. Another theory assumes that the target function can be decomposed into a sum or product of much smaller dimensional functions [15, 2, 4]. In the specific case of ANOVA decomposition [14], it is assumed that only a minimal number of ANOVA terms whose dimension is minimal compared to $d$ are relevant. This implies that the function to be approximated can be factorized into a sum of functions that depend on only a limited number of variables [16, 9], i.e., only a certain number of variables interact with each other. Thus the notion of superposition dimension and truncation dimension [19] was introduced to penalize the number of ANOVA terms (equal to $2^d$) and, in addition, the dimension of each of them. Several fruitful pieces of research have been done in this sense, as in regression problems [17, 9] and density function approximations [6, 7, 3]. Therefore, we introduced finite sparse mixtures models which are inspired by the ANOVA decomposition of sparse functions. Indeed we assume that each mixture component may only depend on a smaller of variables interaction than the space dimension $d$.

1.1 Prior work

Given $\{x^n\}_n$ data of a multivariate random variable $X \in \mathbb{T}^d$ of potentially very large dimension, the objective of our work is to approximate the density function $f$ through a mixture of wrapped Gaussian or von Mises distribution models. One of the best known methods is Expectation-Maximization which maximizes the likelihood of the data. It should be noted that in the case where the dimension is high it is impossible to apply
the algorithm naively without prior knowledge of the sparsity of the density function \( f \). Thus, in a previous paper [6] we tried to take into account the sparsity assumption of the density function \( f \) of the mixture model. The algorithm proved to be very effective in approximating periodic B-splines, the first Friedman function and in image classification.

1.2 Our contribution

This current paper is an extension of our previous work "Sparse ANOVA Inspired Mixture Models" [6]. In particular we deal with improvement of learning algorithm by first determining the active variables of the density function. Then we restricted the study to the set of active variables. This approach is even more efficient if we assume that some variables do not play any role in the approximation of \( f \). This considerably reduces the computational time and space. Thus we will assign masses to the variables according to the amount of information they contain. Thus it is possible to obtain an accurate approximate the density function by its marginal which contains the variables with the most information.

1.3 Outline of paper

Section 2 introduced the notation. In section 3 we have introduced a sparse mixture model from the parametric family of multivariate wrapped Gaussian and von Mises distribution. Furthermore, we have derived the marginal and the conditional density function of the wrapped Gaussian Distribution, which will later help us approximate the target density function iteratively.

In section 4 we have implemented an algorithm that determines the set of active variables of a sparse mixture model by the Kolmogorov-Smirnov and correlation coefficient test. Therefore the model learning can be restricted to active variables set \( A \), which will considerably reduce the complexity of the model training if we assume that \( |A| \ll d \).

In section 4 we will define an Algorithm that will iteratively estimate the set of interacting variables and the parameters of the marginal density function as well. Later in section 5 we will test our model on sparse mixtures of wrapped Gaussian, B-splines function, and the California Housing prices data.

2 Preliminaries and notation

3 Sparse Mixture Models

3.1 Sparse additive Model

Under similar assumption as [6, section 2], we try to approximate the density function \( g \) of an unknown distribution given a finite number of weighted samples \( X = \{(x^n, w_n)\}_{n \in [N]} \) by a finite dimensional sparse mixture model, whose probability density function (pdf)
is given by

\[ f(x \mid \alpha, \theta) = \sum_{k=1}^{K} \alpha_{uk} p(x_{uk} \mid \theta_{uk}) \]

where \( u_k \in U \subset \mathcal{P}([d]), \alpha = (\alpha_{uk})_{k=1,\ldots,K} \in \Delta_K, \theta = (\theta_{uk})_{u_k \in U} \) and \( p(\cdot \mid \theta_{uk}) \) is a probability density function with \(|u_k|\)-dimensional parameter \( \theta_{uk} \). We will here consider samples \((x_n)_{n \in \mathbb{N}}\) which are equally weighted, i.e \( w_n = 1 \) for all \( n \in \mathbb{N} \). Similarly to \[6\] we will also assume that the index set \( u_k \) may not be pairwise different, i.e there may exist \( k, t \in [K] \), such that \( u_t = u_k \) but \( t \neq k \). Thus denotes by \( K_u \) the number of mixture components \( p(\cdot \mid \theta_{uk}) \) such that \( u = u_k, k = 1, \ldots, K \). Mixture models, whose density function has the form \(1\) are called sparse mixture model (sparse MM). The parametric family of sparse wrapped Gaussian distribution with both diagonal and full covariance matrix on one side and the family of sparse von Mises distribution on the other will be used to approximate the unknown target density function. Recall that the wrapped Gaussian distribution is obtained by wrapping the Gaussian distribution around the torus. Indeed if \( Y \) is a Gaussian distributed random variable (RV), the corresponding wrapped Gaussian RV \( X \) is given by \( X = Y \bmod T = Y - T[\frac{Y}{T}] \), where \( T > 0 \) denotes the period. Since we are interested in approximating 1-periodic functions on the unit torus, then the wrapped random variable becomes \( X = Y - [Y] \). There pdf are defined as

\[ pG(x_{uk} \mid \theta_{uk} = (\mu_{uk}, \Sigma_{uk})) = \sum_{l \in \mathbb{Z}^{|u_k|}} N(x_{uk} + l \mid \mu_{uk}, \Sigma_{uk}) = Nw(x_{uk} \mid \mu_{uk}, \Sigma_{uk}), \]

where \((N_w) N\) denotes the pdf of the \(|u_k|\)-dimensional (wrapped) Gaussian distribution with mean \( \mu_{uk} \in \mathbb{T}^{|u_k|} \) and symmetric positive definite (SPD) covariance matrix \( \Sigma_{uk} := \Sigma_{u_k u_k} \in \mathbb{R}^{|u_k| \times |u_k|} \). If the wrapped Gaussian distribution has a diagonal covariance matrix \( \Sigma = \text{diag}(\Sigma_1, \ldots, \Sigma_{u_k}) \) then its pdf is simplified to a product of univariate wrapped Gaussian density function as

\[ pG(x_{uk} \mid \theta_{uk} = (\mu_{uk}, \Sigma_{uk})) = \prod_{l \in \mathbb{Z}^{|u_k|}} \prod_{i \in u_k} N(x_i + l_i \mid \mu_i, \Sigma_i) = \prod_{i \in u_k} Nw(x_i \mid \mu_i, \Sigma_i), \]

where \((N_w) N\) is the pdf of the univariate (wrapped) Gaussian density function with parameters \( \mu_i \) and \( \sigma^2 = \Sigma_i \). Since it is practically impossible to numerically compute the probability function of the wrapped Gaussian distribution, and due to the assumption on its covariance matrix, which is positively definite, it can been shown that

\[ Nw(x_{uk} \mid \mu_{uk}, \Sigma_{uk}) \approx N^B_w(x_{uk} \mid \mu_{uk}, \Sigma_{uk}), \]

where

\[ N^B_w(x_{uk} \mid \mu_{uk}, \Sigma_{uk}) := \sum_{l \in (-B, B] \cap \mathbb{Z}^{|u_k|}} N(x_{uk} + l \mid \mu_{uk}, \Sigma_{uk}) \]

for a suitably chosen \( B \in \mathbb{N} \). For instance \[8, 10\] has derived some values of \( B \) depending on the standard deviation \( \sigma \) for \( T = 2\pi, d = 1 \), where \( N^B_w(x_{uk} \mid \mu_{uk}) \) approximates gut
the ground truth density function \( \mathcal{N}_w(x_{u_k} | \mu_{u_k}) \). It has been showed by \cite{10} for 
\[ B = \begin{cases} 1, & \text{if } \sigma \geq 2\pi, \\ 0, & \text{otherwise}, \end{cases} \]
and by \cite{8} for 
\[ B = \begin{cases} 1, & \text{if } \sigma < 2\pi/3, \\ 2, & \text{if } 2\pi/3 \leq \sigma < 4\pi/3, \end{cases} \]
the approximation is very accurate. As the space dimension increases, then \( B \) also increased. Thus we will consider the truncated function \( \mathcal{N}_w^B(x_{u_k} | \mu_{u_k}, \Sigma_{u_k}) \) instead of \( \mathcal{N}_w(x_{u_k} | \mu_{u_k}, \Sigma_{u_k}) \) in the rest of the paper.

**Remark 3.1.** The pdf of the wrapped Gaussian distribution from \cite{2} can be interpreted as the marginal density function with respect to \( (X_u, L_u) \in (\mathbb{T}^{u_k}, \mathbb{Z}^{u_k}) \) of the joint pdf
\[ f(x, l | \mu, \Sigma) = \mathcal{N}(x + l | \mu, \Sigma), \]
where \( \mu \) and \( \Sigma \) are the wrapped normal distribution parameters of \( X \), and the hidden variable \( L_u \) denotes the number of winding, i.e \( Y = X + L \sim \mathcal{N}(\cdot | \mu + L, \Sigma) \).

The von Mises distribution, which represents the restriction of the pdf of an isotropic normal distribution to the unit circle has the pdf
\[ p_M(x_{u_k} | \theta_{u_k} = (\mu_{u_k}, \kappa_{u_k})) = \prod_{i \in u_k} \frac{1}{I_0(\kappa_i)} \exp \left( \kappa_i \cos \left( 2\pi(x_i - \mu_i) \right) \right), \]
where \( \mu_{u_k} \in \mathbb{T}^{u_k} \) represents mean and \( \kappa_{u_k} \in \mathbb{R}_+^{u_k} \) and \( I_0 \) is the modified Bessel function of the first kind of order 0.

To ensure a good approximation accuracy by the parametric family of sparse mixture models of wrapped and von Mises distribution, we assume furthermore that the ground function \( g \) is smooth enough and has a compact support, since Gaussian Mixture Models (GMM) has proved to be good approximators for continuous density functions with compact support \cite{1}.

Under the same assumption as above we will derive a form of the marginal density function of \( f \), where \( f \) is defined as \cite{1}. Indeed we will introduce later in section \cite{4} an algorithm that iteratively approximate the marginal density function of \( f \).

**Definition 3.2.** Let \( X \in \mathbb{T}^d \) be a continuous random variable with probability density function \( f \). For every \( u \subset [d] \), the marginal probability density function with respect to \( X_u := (X_j)_{j \in u} \) is defined as
\[ f_{X_u}(x_u | \theta_u) = P_u(f(x | \theta) = \int_{\mathbb{T}^{d-|u|}} f(x_u, x_{u^c} | \theta)dx_{u^c}, \]
where \( u^c = [d] \setminus u \) and \( P_u : \mathbb{T}^d \to \mathbb{T}^{|u|} \) is the projection operator.
The linearity of the integral immediately implies that the marginal distribution of a mixture model is equal to the mixture of the marginal of each mixture component and thus the linearity of the projection operator. Since two different components may have the same marginal (i.e. the same parameters), then they are put together by summing their mixing weights. This implies, that number of components of the marginal is smaller or equal to the number of components of the ground mixture model. Furthermore if the multivariate random variable $X$ is componentwise independent or (wrapped) Gaussian distributed with parameter $\theta$ then the marginal distribution with respect to the subset of random variable $X_u, u \subset [d]$ is of the same family as the ground distribution with parameters $\theta_u = (\theta_j)_{j \in u}$. For the special case of wrapped Gaussian distribution with dependent random variables the assumption also holds. Before stating the theorem on the marginals of sparse mixture models of wrapped Gaussian or von Mises distributions, let us recall first the marginal and conditional distribution of a multivariate wrapped Gaussian distribution.

**Lemma 3.3.** Let $u \subset [d], n := |u|$ and $n_c = d - n$. Let furthermore $X = (X_u, X_{u^c}) \in (T^n, T^{n_c})$ be a multivariate continuous random variable of a wrapped Gaussian distribution, i.e

$$X \sim N_w(\mu, \Sigma),$$

with parameters

$$\mu := \begin{pmatrix} \mu_u \\ \mu_{u^c} \end{pmatrix}, \Sigma := \begin{pmatrix} \Sigma_{uu} & \Sigma_{uu^c} \\ \Sigma_{u^c u} & \Sigma_{u^c u^c} \end{pmatrix},$$

such that $\mu_u \in T^n, \mu_{u^c} \in T^{n_c}$ are the mean and $\Sigma_{uu} \in \mathbb{R}^{n \times n}, \Sigma_{u^c u^c} \in \mathbb{R}^{n_c \times n_c}, \Sigma_{uu^c} \in \mathbb{R}^{n \times n_c}$ are positive definite covariance matrices parameters. Then the marginal distribution of $X_u$ and $X_{u^c}$ are also a wrapped Gaussian distribution, such that

$$X_u \sim N_w(\mu_u, \Sigma_{uu}), X_{u^c} \sim N_w(\mu_{u^c}, \Sigma_{u^c u^c}).$$

The conditional distribution of $X_u \mid X_{u^c}$ is given by

$$X_{u^c} \mid (X_u, L_u) = (x_u, l_u) \sim N_w(\bar{\mu}, \bar{\Sigma}),$$

with

$$\bar{\mu} = \mu_{u^c} + \Sigma_{u^c u}(\Sigma_{uu})^{-1}(x_u + l_u - \mu_u),$$

$$\bar{\Sigma} = \Sigma_{u^c u^c} - \Sigma_{u^c u}(\Sigma_{uu})^{-1}\Sigma_{uu^c},$$

where $L_u$ denotes the $|u|$-dimensional winding number [8].

**Theorem 3.4.** Let $X \in T^d$ be a continuous random variable of a sparse mixture model of wrapped Gaussian distribution, with density function $f : T^d \rightarrow \mathbb{R}_+$. Let furthermore $X_u := (X_j)_{j \in u}, u \subset [d]$ be an $|u|$-dimensional random variable and $\xi_k := u \cap u_k$. Then
the marginal distribution with respect to $X_u$ is also a sparse mixture model with density function

$$f_{X_u} = \sum_{t=1}^{T_u} \alpha_t p(\cdot \mid \theta_{u_t}),$$

where $T_u \leq K$ and $u_t$ is element of

$$U_{X_u} = \{v \cap u \mid v \in U\}.$$

and $U$ is the collection of the indices of interacting variables of $f$. The mixing weights and density functions of the marginal distribution are respectively

$$\alpha_t = \sum_{k=1}^{K} \alpha_k \chi_{\{\theta_{u_t} = \theta_k\}}, \quad p(\cdot \mid \theta_{u_t}) = \begin{cases} 1, & \text{if } u_t = \emptyset, \\ P_{\xi_k}p(\cdot \mid \theta_{u_k}), & \text{otherwise}, \end{cases}$$

such that $\sum_{t} \alpha_t = 1$ and for $u_t \neq \emptyset$

$$\theta_{u_t} = (\mu_{u_t}, \Sigma_{u_t}) = (\mu_{\xi_k}, \Sigma_{\xi_k})$$

Proof. By definition of the marginal density function and by the linearity of the integral, the marginal density of the mixture model holds

$$f_{X_u} = \int_{T_d(-u)} f_{dx_{u'}} = \sum_{k=1}^{K} \alpha_k \int_{T_d(-u)} p(\cdot \mid \theta_{u_k}) dx_{u'},$$

$$= \sum_{k=1}^{K} \alpha_k P_{\xi_k} p(\cdot \mid \theta_{u_k}) = \sum_{k=1}^{K} \alpha_k P_{\xi_k} p(\cdot \mid \theta_{u_k}).$$

The definition of the marginal density function implies that

$$p(\cdot \mid \theta_{u_t}) = \begin{cases} 1, & \text{if } \xi_k = \emptyset, \\ P_{\xi_k}p(\cdot \mid \theta_{u_k}), & \text{otherwise}. \end{cases}$$

The theorem on conditional wrapped Gaussian distribution, implies that for each $k \in [K]$ the marginal distribution of each mixture component $k$ is a wrapped Gaussian distribution with parameter

$$\theta_{u_k \cap u} = (\mu_{\xi_k}, \Sigma_{\xi_k})$$

if $\xi_k \neq \emptyset$. Thus the marginal of the mixture model yields

$$f_{X_u} = \sum_{k=1}^{K} \alpha_k P_{\xi_k} p(\cdot \mid \theta_{\xi_k}),$$

where $p(\cdot \mid \theta_{\xi_k})$ denotes the probability function of the wrapped Gaussian with parameter $\theta_{\xi_k}$. If there exists $k \neq t, k, t \in [k]$, such that $\theta_{\xi_k} = \theta_{u_t \cap u}$ then combine both components by summing up their weights and reduce the number of mixture component to one. ☐

Theorem 3.4 shows, that the marginal of a sparse mixture model of a parametric family of wrapped Gaussian or von Mises distribution may contain the uniform distribution as mixing component.
3.2 Determination of Active Variables

Assuming that the above assumptions are fulfilled, we can considerably reduce the complexity of learning the parameters of the sparse mixture models by removing the independent uniform distributed random variables. Indeed a random variables

$$X = (X_A, X_{A^c}) \in \mathbb{T}^d$$

such that

$$A \subset [d], X_A \text{ and } X_{A^c}$$

are independent yields the Bayes theorem

$$h(x) = p_1(x_A | \theta_1)p_2(x_{A^c} | x_A, \theta_2)$$

where

$$p_1$$

denotes the marginal density function with respect to

$$X_A$$

and

$$p_2$$

the conditional density function of

$$X_A | X_{A^c}.$$ By assumption

$$p_2(x_{A^c} | x_A, \theta_2) = p_2(x_{A^c} | \theta_2)$$

since both random variables are independent. If we further assume that

$$p_1$$

is a sparse density function having the form (1) and

$$p_2$$

is the uniform density function then

$$h(x) = \sum_{k=1}^{K} \alpha_k p(x_{u_k} | \theta_k),$$

since the multivariate uniform density function on \( \mathbb{T} \) is equal to 1 everywhere. Therefore we can introduce the notion of active and inactive variables for sparse density functions.

**Definition 3.5.** Let \( X \in \mathbb{T}^d \) be a multivariate random variable with density function

$$f(x) = \sum_{k=1}^{K} \alpha_k p(x_{u_k} | \theta_k).$$

a mixture model of wrapped or von Mises mixture model. The set \( \mathcal{A} \) of active variables of \( f \) by

$$\mathcal{A} := \{ i \in [d] | \exists u \in U : i \in u \} .$$

and any random variable \( X_i \) such that \( i \in \mathcal{A} \) is called active.

Thus an active random variable \( X_i \) is either non uniformly distributed or dependant to some \( X_j \) such that \( j \in \mathcal{A} \). Otherwise the random variable is called inactive. Taking as example the density function of the sparse mixture model defined in (1), the active set of each mixture component is given by

$$\mathcal{A}_{u_k} := \{ i \in u_k \},$$

which yields

$$\mathcal{A} = \bigcup_{k=1,...,K} \mathcal{A}_{u_k}.$$ Based on \( \mathcal{X} \) we can determine iteratively \( \mathcal{A} \) by checking which features variables are non uniform distributed with the help of Kolmogorov-Smirnov test or which depends to the non-uniform random variables. Since the independence is generally not trivial, we will only test the random variables by correlations. We can explicitly determine the active set of density function \( f \) given a large enough number of weighted samples by Algorithm [1]

To better understand the concept let us consider two density functions of mixture of wrapped Gaussian distribution which will study in detail along the paper.
Algorithm 1 Determine the active set

1: Input: \(d\)-dimensional weighted samples \(X = \{(x^n, w_n)_{n \in [N]}\}\) with probability density function \(f\), \(\varepsilon_{KS}, \varepsilon_c > 0\).
2: Output: the active set \(A_f\).
3: Determine the index set \(S\) of uniform distributed variables by:
4: for \(i = 1, \cdots, d\) do
5: if \(D((x^n_i, w_n)_{n \in [N]}) \leq \varepsilon_{KS}\) then
6: \(S \leftarrow \{i\}\)
7: end if
8: end for
9: Set \(A_f := [d] \setminus S\)
10: for \(i \in S\) do
11: Compute correlation coefficient vector \(C(X_i, X_A)\) between \(X_i\) and \(X_A\)
12: if \(C(X_i, X_j) \geq \varepsilon_c\) for some \(j \in A_f\) then
13: \(A_f \leftarrow \{i\}\)
14: end if
15: end for

Example 3.6. Consider a 15-dimensional density functions

\[
f_j(x) := \sum_{k=1}^2 \alpha_k p_{u_k}(x_{u_k} | \mu_k, \Sigma_k), \quad p(x_{u_k} | \mu_k, \Sigma_k) := \sum_{l \in \mathbb{Z}^{||u_k||}} N(x_{u_k} + l | \mu_k, \Sigma_k),
\]

where \(j = 1, 2\). The first function \(f^1\) has the parameters

\[
U := \{u_1, u_2\} = \{\{0, 1, 8\}, \{0, 14\}\},
\]

\[
\alpha := (0.7, 0.3),
\]

\[
\mu := \frac{1}{2} ((1, 1, 1)^T , (.5, .5)^T),
\]

\[
\Sigma_k := \sigma^2 I_{|u_k|}, \sigma^2 := 0.001.
\]

The second function \(f^2\) has the parameters

\[
U := \{u_1, u_2, u_3\} = \{\{0, 1, 8\}, \{0, 14\}, \{5\}\},
\]

\[
\alpha := (0.7, 0.2, 0.1),
\]

\[
\mu := \frac{1}{2} ((1, 1, 1)^T , (0.5, 0.5)^T, 0.3),
\]

\[
\Sigma_k := \sigma^2 I_{|u_k|}, \sigma^2 := 0.001.
\]

Following the definition of an active variable, we can directly read the active set from the function definition which are respectively

\[
A_{f^1} = \{0, 1, 8, 14\}, A_{f^2} = \{0, 1, 5, 8, 14\}.
\]
Applying formally Algorithm 1, the plot of the Kolmogorov-Smirnov distance of the weighted samples along each dimension, shows that the variables, whose indices are elements of \(\{0, 1, 8, 14\}\) are non uniform distributed for the first density function \(f^1\) and the variables with index in \(\{0, 1, 8, 5, 14\}\) are non uniformly distributed by the second function \(f^2\). Since all variables are uncorrelated as the correlation it shows.

Assuming that the active set \(A_f\) of the sparse mixture model is already known and increasingly ordered, we can iterate over the index \(i \in A_f\) of the active variables to determine the marginal distribution of the subset \(X_i\) of the random variable \(X\) with probability density function \(f\). For sake of simplicity we will Denote by \(P_i\) the marginal probability density function with respect to \(X_i\), where \(\tilde{i} := [i] \cap A_f\). Let \(r(i) = |\tilde{i}|\) represents the position of \(i\) in \(A_f\) and by the same way the number of iterations. Theorem 3.4 implies that for each \(i \in A_f\) the index set of all interacting variables of the marginal mixture model with respect to \(X_i\) is equal to

\[
U^{r(i)} = \left\{ u^{r(i)} := u \cap \tilde{i} \mid u \in U \right\},
\]

with parameters set

\[
\theta^{r(i)} = \left\{ \theta_{u_k \cap \tilde{v}} := (\theta_j)_{j \in u_k \cap \tilde{v}}, k = 1, \ldots, K \mid u_k \in U \text{ and } v = u^{r(i)}_k \in U^{r(i)} \right\}.
\]
Note that for \( r^{(i)} = \{ \} \), the function \( p(\cdot | \theta^{(i)}_u) \) is the probability density function of the uniform distribution, and for \( r(i) = |A_f| \), the marginal density function \( P[f] := f_i \) is equal to \( f \). By definition of the marginal mixture model, it follows that the active set \( A_f \subseteq A_f \) and \( A_f^{(i)} \subseteq A_f^k \) for all \( k \in [K] \) and \( i \in A_f \). Thus we can further define for a fix \( i \in A_f \) the residual active set of the ground function \( f \) with respect to the marginal density function by

\[
A^{r(i)}_f := A_f \setminus A_f^i,
\]

and the residual active set of each mixture component density function \( p(\cdot | \theta_u^k), u_k \in U, k = 1, \ldots, K \) with respect to its marginal counterpart by

\[
A^{r(i),u_k}_f := A_f^u \setminus A^{u_k}_f,
\]

where \( \theta_w = \theta^r_u, w \in U^{r(i)} \). This notion of residual active set will be useful to considerably reduced the complexity of the algorithm presented in section 4, which approximate iteratively the marginal density function of the sparse mixture model. These new notions can be illustrated by two concrete examples. Indeed in the following we will consider two 15-dimensional density function of sparse wrapped Gaussian mixture models. We will compute their marginal distributions with respect to the subset of random variables \( X_u, \emptyset \neq u \subset [d] \).

In the following we will introduce two notions of effective dimension, when dealing with very high dimensional sparse functions.

**Definition 3.7.** Let \( f : \mathbb{T}^d \to \mathbb{R} \) be a function and \( \eta \in (0, 1] \). The superposition dimension \( d_s > 0 \) at level \( \eta \) is defined as

\[
d_s := \arg \min_{s \in [d-1]} \frac{\sum_{u \subseteq [d]} \sigma^2(f_u)}{\sigma^2(f)} \geq \eta,
\]

where the \(|u|\)-dimensional functions

\[
f_u = \sum_{v \subseteq u} (-1)^{|u|-|v|} P_v f
\]

are the ANOVA-term of the function \( f, P_v \) denotes the projection operator of definition 3.2 and \( \sigma^2(\cdot) \) the variance of the corresponding functions. The second notion of effective dimension is the truncation dimension, which is defined as

\[
d_t := \arg \min_{s \in [d-1]} \frac{\sum_{u \subseteq [s]} \sigma^2(f_u)}{\sigma^2(f)} \geq \eta.
\]

We will combine later in section ?? these two notions of effective dimension to introduce an assumption of sparsity criterion for the density functions, we want to approximate. First, the function \( f : \mathbb{T}^d \to \mathbb{R} \) from (3.6) can be rewritten as

\[
f = \sum_{w \in W} g_w,
\]
where $W = U \subset P([d])$, and $g_w$ are linear combination of lower-dimensional functions depending only on variables with index set in $w$. For those class of density functions, it has been shown in [3, proposition 2.1], that the ANOVA-decomposition of $f$ is equal to

$$f = \sum_{u \in W} f_u,$$

where $W$ denotes the set of all $w \in W$ and all their subsets. Then the superposition dimension defined in (3) is also

$$d_s = \arg \min_{s \in [d-1]} \frac{\sum_{\emptyset \neq u \subseteq W, |u| \leq s} \sigma^2(f_u)}{\sigma^2(f)},$$

and the truncation dimension

$$d_t = \arg \min_{s \in [d-1]} \frac{\sum_{\emptyset \neq u \subseteq W, |u| \leq s} \sigma^2(f_u)}{\sigma^2(f)},$$

where

$$W_s := \{ u \subseteq [d] \mid u \in W \text{ and } u \subseteq [s] \}.$$

Considering the ANOVA decomposition of the marginal density function $P\xi f$ of $f$ associated to an arbitrary but fixed $i \in [d-1], \xi := [i]$, it follows that

$$P\xi f = \sum_{u \subseteq [d]} (P\xi f)_u,$$

where

$$(P\xi f)_u = \sum_{v \subseteq u} (-1)^{|u| - |v|} P_v P\xi f. \quad (4)$$

We know by definition, that $P\xi f$ depends only on the variables $x_\xi$. Lemma [3, proposition 2.1] implies that $(P\xi f)_u = 0$ for all $u \subseteq [d]$, such that $u$ is not included in $\xi$. Thus

$$P\xi f = \sum_{u \subseteq \xi} (P\xi f)_u.$$

For $v \subseteq u \subseteq \xi$ it holds that $P_v P\xi f = P_v f$. Hence equation (4) implies that $(P\xi f)_u = f_u$ for all $u \subseteq \xi$. Thus with [16, Lemma 2.9] the truncation dimension defined in (3.7) yields

$$d_t = \arg \min_{s \in [d-1]} \frac{\sigma^2(P_s f)}{\sigma^2(f)}.$$

Using this, we can introduce an iterative algorithm, which can approximate the marginal density function $P[i] f$ for any $i \in A_f$ of the form (1), under the assumption that a large enough number of samples are provided. The function $f$ is an accurate approximation to
the ground function \( g \). If \( f \) is sparse in sense of equation 1, then there exists an element \( v \subset [d], |v| \ll d \) such that

\[
\| g - P_v f \| \leq \| g - f \| + \| f_{X_v} - I \| \| P_v f \| \to 0,
\]

and the maximal number of interacting variables are very small with respect to the space dimension.

4 Learning Sparse Mixture Models

In the rest of this paper, we assume that all variables \( x_1, \ldots, x_d \) in (1) are active. For learning the sparse MM, we propose an algorithm which iteratively approximates the marginals \( f_{X_{r-1}} f \) for \( r = 1, \ldots, d \). In the following, we give an idea of the algorithm by describing its first two steps. Let the samples \( \{ x^n = (x^n_k)_{k=1}^d : n = 1, \ldots, N \} \) be given.

**Step 1:** Find an approximation of the first marginal by

\[
f_1(x_1) = \alpha_{0}^1 + \sum_{k=1}^{K_1} \alpha_k^1 p(x_1 | \mu_k^1, \sigma_k^1)
\]

from the samples \( \{ x^n_1 : n = 1, \ldots, N \} \) as follows:

1.1 Determine \( K_1 \) by the BIC method described in Appendix 7.4.

1.2 Apply a univariate EM algorithm to compute \( (\alpha_k^1, \mu_k^1, \sigma_k^1), k = 1, \ldots, K_1 \) and \( \alpha_0^1 \) and to determine the probability \( \beta_{k,n}^1, n = 1, \ldots, N, k = 0, \ldots, K_1 \) that \( x^n_1 \) belongs to the \( k \)-th mixture component.

**Step 2:** Find an approximation of the first two marginals by the following steps:

2.1 For each \( k = 0, \ldots, K_1 \) determine if the weighted samples \( \{ \beta_{k,n}^1 x^n_2 : n = 1, \ldots, N \} \) are uniformly distributed and uncorrelated by the Kolmogorov-Smirnov test in Appendix 7.2 and correlation estimate in Appendix 7.1. Then we get

\[
\{0, \ldots, K_1\} = K_{nu} \cup K_u,
\]

where \( K_{nu} \) denote the indices of those mixture summands in [5], where the samples are not uniformly distributed and \( K_u \) the other ones.

2.2 For each \( k \in K_{nu} \) and samples \( \{ \beta_{n,k}^1 x_2 : n = 1, \ldots, N \} \) determine

\[
f_k^2(x_2) = \alpha_{0,0}^2 + \sum_{l=1}^{L_k} \alpha_{k,l}^2 p(x_2 | \mu_{k,l}^2, \sigma_{k,l}^2)
\]

by computing
- \( L_k \) by the BIC method described in Appendix 7.4
- \((\alpha_{k,l}^2, \mu_{k,l}^2, \sigma_{k,l}^2), l = 1, \ldots, L_k \) and \( \alpha_{k,0}^2 \) by a univariate EM algorithm. These parameters will be used as initial ones in the next EM step.

2.3 Case 1: If \( 0 \in K_u \), set \( p(x_1|\mu_0^1, \sigma_0^1) := 1 \) and compute the parameters \((\alpha_{k,l}^{1,2}, \mu_{k,l}^{1,2}, \Sigma_{k,l}^{1,2}) \) and determine the probability \( \beta_{kl,n}^{1,2}, n = 1, \ldots, N, l = 1, \ldots, L_k, k \in K_{nu} \), in the MM

\[
\begin{align*}
  f_{k}^{1,2}(x_1, x_2) &= \alpha_0^1 + \sum_{k \in K_u} \alpha_k^1 p(x_1|\mu_k^1, \sigma_k^1) \\
  &+ \sum_{k \in K_{nu}} \left( \sum_{l=1}^{L_k} \alpha_{k,l}^{1,2} p(x_1, x_2|\mu_{k,l}^{1,2}, \Sigma_{k,l}^{1,2}) + \alpha_{k,0}^{1,2} p(x_1|\mu_k^1, \sigma_k^1) \right)
\end{align*}
\]

with \( \alpha_{k,0} := \alpha_{k,0}^1 \) and initialization for \( l = 1, \ldots, L_k \) as

\[
\begin{align*}
  \alpha_{k,l}^{1,2} := \alpha_k^1 \alpha_{k,l}^1, & \quad \mu_{k,l}^{1,2} := \left( \frac{\mu_k^1}{\mu_{k,l}^1} \right), \quad \Sigma_{k,l}^{1,2} := \begin{pmatrix} \sigma_k^1 & 0 \\ 0 & \sigma_{k,l}^2 \end{pmatrix} \quad \text{(6)}
\end{align*}
\]

Case 2: If \( 0 \in K_{nu} \), compute the parameters \((\alpha_{k,l}^{1,2}, \mu_{k,l}^{1,2}, \Sigma_{k,l}^{1,2}) \) and determine the probability \( \beta_{kl,n}^{1,2}, n = 1, \ldots, N, l = 1, \ldots, L_k, k \in K_{nu} \) in the MM

\[
\begin{align*}
  f_{k}^{1,2}(x_1, x_2) &= \alpha_{0,0}^1 + \sum_{l=1}^{L_0} \alpha_{0,l}^{1,2} p(x_2|\mu_{0,l}^{1,2}, \sigma_{0,l}^{1,2}) + \sum_{k \in K_u} \alpha_k^1 p(x_1|\mu_k^1, \sigma_k^1) \\
  &+ \sum_{k \in K_{nu}\setminus\{0\}} \left( \sum_{l=1}^{L_k} \alpha_{k,l}^{1,2} p(x_1, x_2|\mu_{k,l}^{1,2}, \Sigma_{k,l}^{1,2}) + \alpha_{k,0}^{1,2} p(x_1|\mu_k^1, \sigma_k^1) \right)
\end{align*}
\]

We use the same initialization \( \text{(6)} \) for \( k \in K_{nu}\setminus\{0\} \) and

\[
\begin{align*}
  \alpha_{0,l}^{1,2} := \alpha_{0,0}^1 \alpha_{0,l}^1, & \quad \mu_{0,l}^{1,2} := \mu_{0,l}^1, \quad \sigma_{0,l}^{1,2} := \sigma_{0,l}^2.
\end{align*}
\]

If we use a MM with wrapped Gaussians with just diagonal covariance matrices, Step 2.3 is superfluous and the new parameters are those from the initialization.

**Remark 4.1.** If we consider the sparse mixture model of diagonal wrapped Gaussian or von Mises distribution then the estimation step 12 and 18 in algorithm \[\square\] will be resumed to fitting univariate marginal distribution. This will considerably increase the computation (time and storage) complexity.
Algorithm 2 Active set detection and parameters estimate

1. **Input:** \((x_1, \ldots, x_N) \in \mathbb{R}^d, (x_1, \ldots, x_N) \in \mathbb{R}^N, K_{\text{max}} \in \mathbb{N}, \varepsilon_{KS}, \varepsilon_c > 0\)
2. **Output:** \(U, \hat{\theta}\)
3. Set \(U = \{u = \{\}\}, \alpha = (1), \theta = (\{\})\), \(r(-1) = 0, n_u = 1\)
4. for \(i = 1, \ldots, d\) do
5. \(\text{Set } K = \sum_{u \in U(i-1)} n_u, U_{\text{new}}^{(i)} = \{\}, U_{\text{fix}}^{(i)} = \{\}, \theta_{\text{new}}^{(i)} = \{\}, \theta_{\text{fix}}^{(i)} = \{\}\)
6. for \(k = 0, \ldots, K\) do
7. Update the samples weights w.r.t the \(k\)-component of \(P_{[i-1]}\)
   \[\tilde{w}_n^{i,k} = w_n^{\beta_n^{i,k}}, \text{ where } \beta_n^{i,k} = \frac{\alpha_{k}p(x_n^{i,k} | \theta_k)}{\sum_{t=1}^{K} \alpha_{t}p(x_n^{i,k} | \theta_t)}\]
8. Determine the corresponding residual active set
   \[\overline{A}_{fix}^{ik} = \left\{ i \in \overline{A}_{fix}^{ik} \mid D_N \left((x_t^{n}, \tilde{w}_n^{i,k})_{n \in [N]}\right) \geq \varepsilon_{KS}\right\}\]
9. if \(i \in \overline{A}_{fix}^{ik}\) then
10. Add to \(U_{\text{new}}^{(i/2)} \leftarrow u_{(i/2)} = u_k \cup \{i\}\)
11. Determine the number of the mixture components with the index set \(u_k^{(i)}\)
   \[n_k = \arg \min_{k=1, \ldots, K_{\text{max}}} \text{BIC} \left((x_t^{n}, \tilde{w}_n^{i,k})_{n \in [N]}\right)\]
   \[\left(\alpha_{j}, \theta_{j}^{(i)}\right)_{j \in [n_k]} = \text{Prox-EM} \left((x_t^{n}, \tilde{w}_n^{i,k})_{n \in [N]}, (\theta_{t}^{(i)})_{t \in [n_k]} \cup \theta_{\gamma_2}\right)\]
12. Set \(\theta = \left\{\theta_{u_k}^{(i-1)}, \theta_{j}^{(i)}\right\}_{j \in [n_k]}\) and add \(\theta_{\text{new}}^{(i/2)} \leftarrow \theta\)
13. else
14. Add \(U_{\text{fix}}^{(i/2)} \leftarrow u_k\) and \(\theta_{\text{fix}}^{(i)} \leftarrow \theta_{u_k}^{(i-1)}\)
15. end if
16. end for
17. Compute \(\tilde{w}_n = w_n \cdot \beta_{n,G2}\) the posterior probabilities \(\beta_{n,G2}\) of the samples according to section 7.3, where \(G_2 = \left(U_{\text{new}}^{(i/2)}, \theta_{\text{new}}^{(i/2)}\right)\)
18. Update the parameters with the EM-Algorithm [3]
   \[U_{\text{new}}^{(i)}, \theta_{\text{new}}^{(i)} = \text{Prox-EM} \left((x_t^{n}, \tilde{w}_n^{i,k})_{n \in [N]}, \theta_{\text{new}}^{(i/2)}, \gamma_1\right)\]
19. Set \(U_{\text{fix}}^{(i)} = U_{\text{new}}^{(i)} \cup U_{\text{fix}}^{(i)}\) and \(\theta_{\text{fix}}^{(i)} = \theta_{\text{new}}^{(i)} \cup \theta_{\text{fix}}^{(i)}\)
20. end for
| N samples \cdot 10^2 | Method           | Truth | $\varepsilon_{KS}$ | $\varepsilon_c$ | $\gamma_1 \cdot 10^{-3}$ | $\gamma_2 \cdot 10^{-3}$ |
|-------------------|-------------------|-------|---------------------|-----------------|--------------------------|--------------------------|
| 75                | wrapped Gaussian  | a)    | 4.7                 | 3.0             |                          |                          |
|                   |                   | b)    |                     |                 |                          |                          |
|                   | von Mises         | a)    | 0.1                 | 5.0             |                          |                          |
|                   |                   | b)    |                     |                 |                          |                          |
| 100               | wrapped Gaussian  | a)    | 5.0                 | 3.0             |                          |                          |
|                   |                   | b)    |                     |                 |                          |                          |
|                   | von Mises         | a)    | 5.0                 | 3.0             |                          |                          |
|                   |                   | b)    |                     |                 |                          |                          |

Table 1: Hyper-parameters for the first two test functions

5 Experimental Results

In the following we will apply Algorithm \[2\] to determine the collection of variable interactions and the associated mixture components parameters for the test functions defined in ??, the product of B-splines function and the California Housing data. To evaluate the model the log-likelihood the training data and the test will be compared with each other. Furthermore the relative $L^p, p = 1, 2$ errors between the ground truth function $f$ and the approximated model $\hat{p}$ on unknown test data will be computed. Recall that the relative $L^p$-error is defined as

$$e_{L^p}(\hat{p}, f) = \frac{\|\hat{p} - f\|_{L^p}}{\|f\|_{L^p}},$$

where $\|\cdot\|_{L^p}$ will be determined via the Monte-Carlo integration, i.e

$$\|f\|_{MC} = \frac{1}{N_{MC}} \sum_{n=1}^{N_{MC}} |f(x^n)|^p$$

and

$$\|\hat{p} - f\|_{MC} = \frac{1}{N_{MC}} \sum_{n=1}^{N_{MC}} |\hat{p}(x^n) - f(x^n)|^p$$

where $(x^n)_{n \in N_{MC}}$ are uniformly distributed samples on $\mathbb{T}^d$. For large value of $N_{MC}$ the Monte-Carlo norm $\|\cdot\|_{MC}$ yields an accurate approximation of $\|\cdot\|_{L^p}$. Therefore $N_{MC} = 10^5$ will be used to compute the relative errors.

To train the function from example\[3,6\] $N_{tr} = 75 \cdot 10^2, 10^4$ training samples has been drawn by the rejection sampling method\[6,20\]. The parameters used by Algorithm \[2\] to learn the model are in Table \[1\]. The approximated model from Algorithm \[2\] given in Table \[3\] are obviously an accurate approximation of the ground truth mixture models. The negative log-likelihood and the relative error in Table \[2\] prove this.

5.1 B-spline functions

In this section we will approximate the function $f^3 : \mathbb{T}^9 \to \mathbb{R}$ defined as $f^3(x) = \tilde{f}^3(x)/\|\tilde{f}^3\|_{L_1}$, such that

$$\tilde{f}^3(x) = B_2(x_1)B_4(x_3)B_6(x_8) + B_2(x_2)B_4(x_5)B_6(x_6) + B_2(x_4)B_4(x_7)B_6(x_9),$$
In the following we want to apply the California Housing Prices to our model. The data contain information from the 1990 California census. The goal is to predict the median

| Truth | Method | \( \mathcal{L}_f(x^1, \ldots, x^N) \) | \( \mathcal{L}_\hat{f}(x^1, \ldots, x^N) \) | \( \varepsilon_{L_1}(\hat{p}, f) \) | \( \varepsilon_{L_2}(\hat{p}, f) \) |
|-------|-------|----------------------|----------------------|------------------|------------------|
| a)    | wrapped | 36546.7 ± 190.5 | 36553.7 ± 189.3 | 0.0343 ± 0.0069 | 0.0312 ± 0.0059 |
| a)    | comp. wrapped | 36546.7 ± 190.5 | 36551.9 ± 189.5 | 0.0329 ± 0.0076 | 0.0311 ± 0.0069 |
| a)    | von Mises | 36588.8 ± 201.4 | 36592.7 ± 199.8 | 0.0325 ± 0.0072 | 0.0312 ± 0.0123 |
| b)    | wrapped | 33591.0 ± 261.2 | 33609.9 ± 258.2 | 0.0371 ± 0.0061 | 0.0302 ± 0.0073 |
| b)    | comp. wrapped | 33591.0 ± 261.2 | 33598.7 ± 258.4 | 0.0356 ± 0.0079 | 0.0304 ± 0.0097 |
| b)    | von Mises | 33647.5 ± 274.5 | 33514.7 ± 498.4 | 0.0505 ± 0.0501 | 0.0285 ± 0.0094 |

Table 2: Approximation of \( f^i, i = 1, 2 \) in example 3.6 by sparse mixture models of wrapped Gaussian and von Mises distribution from section 3. Top: \( N = 7500 \), Bottom: \( N = 10000 \). Average value of the log likelihood function and relative \( L_q \) errors, \( q = 1, 2 \) for \( N_{te} = 10^4 \) training sets top: \( N_{tr} = 7500 \), bottom: \( N_{tr} = 10000 \).

where \( B_i, i \in \text{set 2, 4, 6} \) To train the model \( N_{te} = 10^4 \) samples with density function \( f^3 \) has been drawn by the rejection sampling method. The ordered indices set of active variables are

\[
\mathcal{A}_{f^3} = \{8, 6, 9, 3, 5, 7, 1, 2, 4\}.
\]

and the set of coupling indices

\[
\mathcal{U} = \{\{1, 3, 8\}, \{2, 5, 6\}, \{4, 7, 9\}\}.
\]

Note that by the von Mises distribution there may exits more that one component (some times 2 mixture components) with the same coupling variables. Similarly to the first two examples \( N_{te} = 10^3 \) uniform samples has been used to estimate the model. The negative likelihood of the samples, the relative \( L_1 \) and \( L_2 \) errors are given in table 4. We can also train the model by the usual EM-Algorithm 4 without any sparsity assumption on the density function. The computation time is

\[
\mathcal{U} = \{\{1, 3, 8\}, \{2, 5, 6\}, \{4, 7, 9\}\}.
\]

5.2 California Housing Prices

In the following we want to apply the California Housing Prices to our model. The data contain information from the 1990 California census. The goal is to predict the median...
Table 3: Parameters estimation of $f^i, i = 1, 2$ in example 3.6. Top: $N = 7500$, Bottom: $N = 10000$, with the training parameters $\varepsilon_c = 0.1, \gamma_{prox} = 10^{-3}$ for all density functions and $\varepsilon_{KS} = 3.4$ for the class of wrapped normal density function and $\varepsilon_{KS} = 2.7$ for the von Mises distribution function.
Table 4: Approximation of $f^k$ by sparse mixture models of wrapped Gaussian and von Mises distribution from section 3. Average value of the log likelihood function, relative $L_q$ errors, $q = 1, 2$ and Mean Square Error (MSE) for $N_{tr} = 10^5$ training sets. Top: negative Log-likelihood, relative errors. Bottom: MSE for all models and for naive EM-Algorithm.

| Method       | MSE       | Time(s)  |
|--------------|-----------|----------|
| wrapped      | 0.0257 ± 0.0023 | 97.3 ± 6.6 |
| comp. wrapped| 0.0247 ± 0.0019 | 102.4 ± 3.5 |
| von Mises    | 0.0915 ± 0.0067 | 152.6 ± 13.3 |
| Naiv EM-Algorithm | 10727 ± 560 |

The total number of samples $N_{CH}$ is 20640. The input variable will be denoted by $x$ and the target variable by $y$. A data preprocessing step has cleaned the data and using using min-max scaler to rescale the data to fit $\mathbb{T}$. We assume that the input features $x^n, n \in N_{CH}$ are the samples of a $8$-dimensional random variable $X \in \mathbb{R}^d$ and the target variable $y^n, n \in N_{CH}$ a sample from $Y \in \mathbb{T}$.

To be able to train the data with Algorithm 2 we assume that the target density function can be written as convex combination of linear model, i.e

$$y = \varphi(x) = \sum_{k=1}^{K} \alpha_k \left( M_k x_{u_k} + N_k \right),$$

where for all $k \in [K]$ $M_k \in \mathbb{R}^{1 \times |u_k|}$ and $N_k \in \mathbb{R}$. Recall the regression model can be rewritten as conditional expectation $\varphi(x) = E(y|x)$, where the conditional expectation is defined as

$$E(y|x) = \int_{\mathbb{T}} y p_{Y|X}(y | x, \theta) dy$$

and $p_{Y|X}(y|x, \theta)$ defined the conditional density function of $Y|X$. Furthermore, we assume that the joint density function is sparse with the form \[1\] If $p_{Y|X}$ denotes the conditional density function of a wrapped Gaussian distribution, we know by Theorem 3.3 that

$$E(y|x, l) = \sum_k \alpha_k(x) \left( \mu_{uk} + \Sigma_{uk}^{-1}(\Sigma_{uk} + l_{uk} - \mu_{uk})^{-1}(x_{uk} + l_{uk} - \mu_{uk}) \right)$$
Table 5: Mean Square Error of California Housing Prices by SPAMM and other regression Models

| Method                      | MSE  |
|-----------------------------|------|
| Wrapped                     | 0.1803 |
| Linear Regression           | 0.4478 |
| Lasso Regression            | 0.7193 |
| Ridge Regression            | 0.6056 |
| Decision Tree Regression    | 0.5912 |
| Random Forest Regression    | 0.4478 |

where the weights $\alpha_k$ are defined as

$$\alpha_k(x) = \frac{\tilde{\alpha}_k p_{X_{u_k}}(x_{u_k}|\theta_k)}{\sum_{k=1}^{K} \tilde{\alpha}_k p_{X_{u_k}}(x_{u_k}|\theta_k)}.$$  

Note that the density function

$$p_{X_A}(x|\theta) = \sum_{k=1}^{K} \tilde{\alpha}_k p_{X_{u_k}}(x_{u_k}|\theta_k)$$

is the marginal density function with respect to $X_A$ with parameters $(\alpha_k, \mu_{u_k}, \Sigma_{u_k u_k})_{k \in [K]}$ and

$$f(z) = p(z|\theta) = p(x_A, y|\theta) = \sum_{k=1}^{K} \alpha_k p(y, x_{u_k}|\theta_k)$$

the joint density function $Z_A = (X_A, Y)$. Similarly to the previous section $A$ denotes the set of active variable of $f$.

Obviously the set of active variable are the variables $X_i$ such that $X_i$ and $Y$ are correlated and $X_i$ is non uniformly distributed. Therefore the index set of active variables are

$$A_{CH} = \{1, 3, 6, 9\},$$

does not the collection of coupling variables are

$$U = \{\{1, 3, 9\}\}$$

and the MSE of the approximated model is given in Table 5. Comparing this model with some other regression model show a better approximation result that Linear Regression, Lasso Regression, Ridge Regression, Decision Tree Regression and Random Forest Regression.

6 Conclusion

This paper introduces an efficient algorithm that can accurately estimate the parameters of a sparse mixture model of wrapped Gaussian and von Mises distribution based on
the input samples. Assuming that each component of the multivariate density function depends only on a certain number of interacting variables, which is also unknown, we have iteratively determined the set $\mathcal{U}$ using statistical tests and the model parameters using Expectation-Maximization. Incorporating this sparsity assumption speeds up the learning procedure in the case where the dimension $d$ is very large and even provides a better approximation accuracy. Indeed this yields a better approximation accuracy and is more efficient than the usual expectation maximization for the mixtures of wrapped Gaussian, the B-spline function, and the California housing data, as shown in the numerical results. However, the approximation relies on the choice of hyperparameters, which, when not chosen appropriately, leads to underfitting or overfitting of the mixture model.

References

[1] Athanassia G. Bacharoglou. Approximation of probability distributions by convex mixtures of gaussian measures. Proceedings of the American Mathematical Society, 138(7):2619–2628, 2010.

[2] Gregory Beylkin and Martin J. Mohlenkamp. Algorithms for numerical analysis in high dimensions. SIAM Journal on Scientific Computing, 26(6):2133–2159, 2005.

[3] Gregory Beylkin, Lucas Monzón, and Xinshuo Yang. Reduction of multivariate mixtures and its applications. Journal of Computational Physics, 383:94–124, 2019.

[4] Erwan Grelier, Anthony Nouy, and Régis Lebrun. Learning high-dimensional probability distributions using tree tensor networks. ArXiv, abs/1912.07913, 2019.

[5] Abolfazl Hashemi, Hayden Schaeffer, Robert Shi, Ufuk Topcu, Giang Tran, and Rachel Ward. Generalization bounds for sparse random feature expansions, 2021.

[6] Johannes Hertrich, Fatima Ba, and Gabriele Steidl. Sparse mixture models inspired by anova decompositions. ETNA - Electronic Transactions on Numerical Analysis, 2021.

[7] Johannes Hertrich, Dang-Phuong-Lan Nguyen, Jean-François Aujol, Dominique Bernard, Yannick Berthomieu, Abdellatif Saadaldin, and Gabriele Steidl. Pca reduced gaussian mixture models with applications in superresolution. Inverse Problems and Imaging, 16(2):341–366, 2022.

[8] Giovanna Jona-Lasinio, Alan Gelfand, and Mattia Jona-Lasinio. Spatial analysis of wave direction data using wrapped gaussian processes. The Annals of Applied Statistics, 6(4), Dec 2012.

[9] Thierry A. Mara and Stefano Tarantola. Variance-based sensitivity indices for models with dependent inputs. Reliability Engineering and System Safety, 107:115–121, 2012. SAMO 2010.
7 Statistical Methods

7.1 Correlation Test

To test whether the features are uncorrelated, we have to verify if their correlation coefficients are zero. Recall that the correlation coefficient of two random variables $X_i$
and \( X_j \) is define as
\[
\text{Cor}(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sigma^2_{X_i} \sigma^2_{X_j}},
\]
where \( \text{Cov}(X_i, X_j) \) represents the covariance of \( X_i \) and \( X_j \) and \( \sigma^2_{X_i}, \sigma^2_{X_j} \) their variance.
If some weighted samples of the random variables are provided, the coefficient are approximated by using their corresponding unbiased weighted samples covariance matrix
\[
\Sigma := \frac{1}{\sum_{n=1}^{N} w_n - 1} \sum_{n=1}^{N} w_n (x^n - \mu)^\top (x^n - \mu),
\]
where
\[
\mu := \frac{1}{\sum_{n=1}^{N} w_n} \sum_{n=1}^{N} w_n x^n
\]
is the weighted samples mean.

### 7.2 Kolmogorov-Smirnov Test

Recall that the Kolmogorov-Smirnov test [12] is a statistical test often used to test if some given samples \( \mathcal{X} \) fit a distribution whose cumulative density function \( F \) is a priori known. The samples fit \( F \) if the Kolmogorov-Smirnov (KS) distance is smaller than a threshold, i.e
\[
D_N(x^n, w_n) = D_N(F, F_N) = \sqrt{\frac{\left(\sum_{n=1}^{N} w_n\right)^2}{\sum_{n=1}^{N} w_n^2} \|F - F_N\|_{\infty} \leq \varepsilon_{KS}},
\]
where \( F_N \) represents the empirical cumulative density function
\[
F_N = \frac{1}{\sum_{n=1}^{N} w_n} \sum_{n=1}^{N} w_n \mathbb{I}[x^n, 1]
\]
of the weighted samples \( \mathcal{X} \). Let
\[
s_n = \frac{\sum_{m=1}^{n} w_m}{\sum_{n=1}^{N} w_n},
\]
if we assume that the samples are ordered increasing then the KS distance becomes
\[
D_N(F, F_N) = \sqrt{\frac{\left(\sum_{n=1}^{N} w_n\right)^2}{\sum_{n=1}^{N} w_n^2} \max_{n=1,\ldots,N} \max \{s_n - x^n, x^n - s_{n-1}\}},
\]
and will be denoted by \( D_N \left( (x^n, w_n)_{n \in [N]} \right) \).
7.3 EM Algorithm

In the following we want to approximate the parameters of the samples distribution under the assumption that the parameters of some fixed mixture components are already known. As already mentioned above, this can be done by the usual Expectation maximization algorithm (EM) (see algorithm 3). To ensure sparse model in the mixing weights, the proximal expectation maximization (Prox-EM) algorithm 4 can be used instead. In the

**Algorithm 3 EM Algorithm for Mixture Models**

1: **Input:** \((x^1, \ldots, x^N) \in \mathbb{T}^{N,d}, (w_1, \ldots, w_N) \in \mathbb{R}^N, \) initial parameters \((\alpha_k^{(0)}, \theta_k^{(0)}) \) \(k \in [K]\).  
2: **Output:** Optimal parameters \((\alpha, \Theta)\)  
3: for \(r = 0, 1, \ldots\) do  
4: **E-Step:** For \(k = 1, \ldots, K\) and \(n = 1, \ldots, N\) compute  
   \[\beta_{n,k}^{(r)} = \frac{\alpha_k^{(r)} p(x^n_{u_k} | \theta_k^{(r)})}{\sum_{j=1}^{K} \alpha_j^{(r)} p(x^n_{u_j} | \theta_j^{(r)})}\]  
5: **M-Step:** For \(k = 1, \ldots, K_1\) compute  
   \[\alpha_k^{(r+1)} = \frac{1}{\sum_{m=1}^{N} w_m \sum_{n=1}^{N} w_n \beta_{n,k}^{(r)}}\]  
   \[\theta_k^{(r+1)} = \arg \max_{\theta_k} \left\{ \sum_{i=1}^{N} w_n \beta_{n,k}^{(r)} \log(p(x^n_{u_k} | \theta_k^{(r)})) \right\}.\]  
6: end for

Prox-EM algorithm, the goal is to minimize the penalized functional

\[L_\gamma(\alpha, \Theta \mid \mathcal{X}) = L(\alpha, \Theta \mid \mathcal{X}) + \gamma \| \alpha \|_0 + \iota \Delta_K(\alpha),\]

where \(\gamma > 0\) represents the learning rate.

If we assume that the samples \(\mathcal{X}\) can be fitted by a \(K\)-components mixture model with parameters \((\alpha_k, \theta_k)_{k \in [K]}\) and for a given \(K_1, |K_1| < K\), the mixtures parameters \((\alpha_k, \theta_k)_{k \in K_1}\) are known, we have to modify the EM-Algorithm to only approximate the parameters \((\alpha_k, \theta_k), k \in K_1^c = [K] \setminus K_1\). Note that this cannot be done directly, by simply fixing the a priori known parameters in the expectation and maximization step. Therefore we will split the mixture components into two groups \(G_1\) and \(G_2\), where \(G_1\) contains components with parameters \((\alpha_k, \theta_k)_{k \in K_1}\) and \(G_2\) those with parameters \((\alpha_k, \theta_k)_{k \in K_1^c}\). Weighting the target distribution samples \(\mathcal{X}\) with the posterior probability  

\[\hat{\beta}_{n,G_2} = \frac{\sum_{k \in K_1^c} \alpha_k p(x_{u_k} \mid \theta_{u_k})}{\sum_{j \in K} \alpha_j p(x_{u_j} \mid \theta_{u_j})},\]

that they belong to \(G_2\) times the ground samples weights, one can use the above described EM-algorithm to find the parameters of the mixture model who fits the best the weighted
samples \( \{(x_n, w_n \beta_{n,G_2})_{n \in [N]}\} \). The output parameters are those of the mixture model from Group \( G_2 \), except to the mixing weights. The mixture components weights has to be rescaled, such that their sum is equal \( \alpha_{G_2} = 1 - \sum_{k \in K_1} \alpha_k \), by multiplying each term with \( \alpha_{G_2} \). To get a better accuracy of the EM-algorithm, we have chosen as initialization parameters the elements of \( \beta \), where the mixing weights has been also rescaled by dividing them with \( \alpha_{G_2} \). For explicit details of the EM-algorithm of the wrapped and the von-Mises distribution see \[6\].

\[ \text{Algorithm 4 Proximal Expectation Maximization (Prox-EM)} \]

1: **Input:** \((x^1, ..., x^N) \in \mathbb{R}^{N,d}, (w_1, ..., w_N) \in \mathbb{R}^N\), initial parameters \((\alpha_k^{(0)}, \theta_{uk}^{(0)})_{k \in [K]}\), learning rate \(\gamma > 0\).

2: **Output:** Optimal parameters \((\alpha, \Theta)\)

3: for \(r = 0, 1, ...\) do

4: Set \(K_c = |\Theta|\)

5: **E-Step:** For \(k = 1, ..., K_c\) and \(n = 1, ..., N\) compute

\[
\beta_{n,k}^{(r)} = \frac{\alpha_k^{(r)} p(x_n \mid \theta_{uk}^{(r)})}{\sum_{j=1}^{K_c} \alpha_j^{(r)} p(x_n \mid \theta_{uj}^{(r)})}
\]

6: **M-Step:** For \(k = 1, ..., K_c\) compute

\[
\alpha_k^{(r+1)} = \frac{1}{\sum_{m=1}^{N} w_m} \sum_{n=1}^{N} w_n \beta_{n,k}^{(r)},
\]

\[
\theta_k^{(r+1)} = \arg \max_{\theta_k} \left\{ \sum_{i=1}^{N} w_n \beta_{n,k}^{(r)} \log(p(x_n \mid \theta_{uk}^{(r)})) \right\}.
\]

7: Order the mixing weights \((\alpha_k)_{k \in K_c}\) and compute the number \(K_0\), where

\[
K_0 = \arg \min_{m=0, ..., K_c-1} \frac{1}{2\gamma} \left( \frac{\sum_{k=1}^{m} \alpha_k}{K_c - m} + \sum_{k=1}^{m} \alpha_k^2 - m \right)
\]

8: Set \(J = [K_c] \setminus [K_0]\) and \(J^c = [K_0]\) such that

\[
\alpha = (\hat{\alpha}_k)_{k \in J}, \quad \text{where} \quad \hat{\alpha}_k = \alpha_k + \frac{1}{|J|} \sum_{k \in J^c} \alpha_k,
\]

\[
\Theta = \{\theta_{uk}\}_{k \in J}
\]

9: end for
7.4 Bayesian Information Criteria (BIC)

The Bayesian Information Criterion (BIC) \[11, 13\] is a statistical method introduced by Schwarz in 1978 for model selection. Given a finite number of models, the BIC is based on the Likelihood the model which given a maximal number of components \(K_{\text{max}}\), we will iteratively train the model with \(K_b = 1, \ldots, K_{\text{max}}\). The optimal number of components \(K_{\text{opt}}\) is the one with the optimal Bayesian Information Criterion (BIC)

\[
\text{BIC}(k) := -2L(\alpha_{tr}, \theta_{tr} | \mathcal{X}_i) + 3k - 1, \tag{7}
\]

where

\[
L(\alpha, \theta | \mathcal{X}_i) = \sum_{n=1}^{N} \tilde{w}^{(i)}_{n,k} \log \left( \sum_{k=1}^{K_b} \alpha_{b,k} p(x_i^n | \theta_{b,k}) \right)
\]

denotes the likelihood of the samples with respect to the trained parameters \(\alpha_{b}, \theta_{b}\) corresponding to \(K_b\) components. Let

\[
\alpha \left( u_{k}^{(i)}, i' \right) \in \Delta_K \left( u_{k}^{(i)}, i' \right), \text{ and } \theta \left( u_{k}^{(i)}, i' \right)
\]

be the estimated parameters associated to the optimal number of mixture component \(K \left( u_{k}^{(i)}, i' \right)\) given by \[7\].

8 Proofs

Before showing Lemma \[3.3\] we will formulate a proposition about the inverse and the determinant of a block matrix

**Proposition 8.1.** \[21\] Let \(\Sigma \in \mathbb{R}^{d \times d}_{+}\) be a positive definite block matrix with the form

\[
\Sigma = \begin{pmatrix} A & B \\ C & D \end{pmatrix}
\]

where \(A \in \mathbb{R}^{n \times n}_{+}, B \in \mathbb{R}^{n \times m}, C \in \mathbb{R}^{m \times n}, D \in \mathbb{R}^{m \times m}_{+}, \) such that \(m + n = d.\) Then the following assertions hold

i. \(\det(\Sigma) = \det(A) \cdot \det(D - CA^{-1}B),\) if \(A\) is invertible,

ii. The matrix \(\Sigma\) is invertible with inverse

\[
\Sigma^{-1} = \begin{pmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{pmatrix},
\]

where

\[
\begin{aligned}
\tilde{A} & := A^{-1} + A^{-1}B \left( D - CA^{-1}B \right)^{-1} CA^{-1}, \\
\tilde{B} & := -A^{-1}B \left( D - CA^{-1}B \right)^{-1}, \\
\tilde{C} & := - \left( D - CA^{-1}B \right)^{-1} CA^{-1}, \\
\tilde{D} & := - \left( D - CA^{-1}B \right)^{-1}.
\end{aligned} \tag{8}
\]
Define vector \( y \) where

\[
\text{Furthermore by proposition 8.1 (8) the inverse covariance matrix yields}
\]

\[
\tilde{A} := (A - BD^{-1}C)^{-1},
\]

\[
\tilde{B} := - (A - BD^{-1}C)^{-1} BD^{-1},
\]

\[
C := -D^{-1}C (A - BD^{-1}C)^{-1},
\]

\[
\tilde{D} := D^{-1} + D^{-1}C (A - BD^{-1}C)^{-1} BD^{-1}.
\]

**Proof.** Let \( n := |u| \). We will only prove the assumption for the marginal of \( f \) with respect to \( X_u \). The marginal pdf with respect to \( X_u \) is given by

\[
f_{X_u}(x_u) = \int_{T^{nc}} f(x) dx_{uc} = \int_{T^{nc}} N(x | \mu, \Sigma) dx_{uc},
\]

\[
= \sum_{l \in \mathbb{Z}^d} \int_{T^{nc}} N(x + l | \mu, \Sigma) dx_{uc},
\]

\[
= \sum_{l \in \mathbb{Z}^d} \int_{T^{nc}} \frac{1}{\sqrt{2\pi} \sigma_{uc}} \exp\left(-\frac{1}{2}((x + l - \mu)\Sigma^{-1}(x + l - \mu))\right) dx_{uc}.
\]

For all \( u \subset [d] \) the matrix \( \Sigma \) can be decomposed into a \( 2 \times 2 \) block matrix

\[
\Sigma = \begin{pmatrix}
\Sigma_{uu} & \Sigma_{uu^c} \\
\Sigma_{u^c u} & \Sigma_{u^c u^c}
\end{pmatrix}.
\]

Proposition 8.1 (ii) implies that

\[
det(\Sigma) = det(\Sigma_{uu}) \cdot det(\Sigma_{u^c u^c} - \Sigma_{u^c u} (\Sigma_{uu})^{-1} \Sigma_{u^c u^c})
\]

Furthermore by proposition 8.1 (ii) the inverse covariance matrix yields

\[
\Sigma^{-1} = \begin{pmatrix}
\tilde{\Sigma}_{uu} & \tilde{\Sigma}_{uu^c} \\
\tilde{\Sigma}_{u^c u} & \tilde{\Sigma}_{u^c u^c}
\end{pmatrix},
\]

where

\[
\tilde{\Sigma}_{uu} = (\Sigma_{uu})^{-1} + (\Sigma_{uu})^{-1} \Sigma_{uu^c} (\Sigma_{u^c u^c} - \Sigma_{u^c u} (\Sigma_{uu})^{-1} \Sigma_{u^c u^c})^{-1} \Sigma_{u^c u} (\Sigma_{uu})^{-1},
\]

\[
\tilde{\Sigma}_{u^c u^c} = -(\Sigma_{uu})^{-1} \Sigma_{u^c u^c} (\Sigma_{uu^c} - \Sigma_{u^c u} (\Sigma_{uu})^{-1} \Sigma_{u^c u^c})^{-1} \Sigma_{u^c u} (\Sigma_{uu})^{-1},
\]

\[
\tilde{\Sigma}_{u^c u} = - (\Sigma_{uu^c} - \Sigma_{u^c u} (\Sigma_{uu})^{-1} \Sigma_{u^c u^c})^{-1} \Sigma_{u^c u} (\Sigma_{uu})^{-1},
\]

\[
\tilde{\Sigma}_{u^c u^c} = (\Sigma_{uu^c} - \Sigma_{u^c u} (\Sigma_{uu})^{-1} \Sigma_{u^c u^c})^{-1}.
\]

Define \( y_l = (x_l - \mu) = (x_l - \mu) \in \mathbb{T}^d \) and \( \mu_u = \Sigma_{uu} (\Sigma_{uu})^{-1} y_u \), since each \( d \)-dimensional vector \( y \) can be decomposed into \( y = (y_u, y_{uc}) \). Thus

\[
(x + l - \mu) (\Sigma^{-1} (x + l - \mu)) = (y_u)^T (\Sigma^{-1}) y_u + (y_{uc})^T (\Sigma^{-1}) y_{uc} + (y_{uc})^T (\Sigma^{-1}) y_{uc} + (y_{uc})^T (\Sigma^{-1}) y_{uc}.
\]

\[
= I + II + III + IV.
\]

27
Replacing the block matrices of the inverse covariance matrix yields

\[ I = \left( y_u^l \right)^\top \left( \Sigma_{uu} \right)^{-1} y_u^l + \left( y_u^l \right)^\top \left( \Sigma_{uu} \right)^{-1} \Sigma_{uu} \left( \Sigma_{uu} \right)^{-1} \Sigma_{uu} \left( \Sigma_{uu} \right)^{-1} \Sigma_{uu} \left( \Sigma_{uu} \right)^{-1} y_u^l, \]

\[ = \left( y_u^l \right)^\top \left( \Sigma_{uu} \right)^{-1} y_u^l + \left( y_u^l \right)^\top \left( \Sigma_{uu} \right)^{-1} \Sigma_{uu} \left( \Sigma_{uu} \right)^{-1} \Sigma_{uu} \left( \Sigma_{uu} \right)^{-1} \Sigma_{uu} \left( \Sigma_{uu} \right)^{-1} y_u^l, \]

\[ = \left( y_u^l \right)^\top \left( \Sigma_{uu} \right)^{-1} y_u^l + \mu_u^l \left( \tilde{\Sigma}_{u\cdot u} \right) \mu_u = I(a) + I(b), \]

and also

\[ II = -\left( y_u^l \right)^\top \left( \Sigma_{uu} \right)^{-1} \left( \Sigma_{u\cdot u} \left( \mu_u \right) \right) - \left( \mu_u \right)^\top \left( \Sigma_{u\cdot u} \right) \left( y_u^l \right), \]

\[ III = -\left( y_u^l \right)^\top \left( \Sigma_{u\cdot u} \left( \mu_u \right) \right) - \left( \mu_u \right)^\top \left( \Sigma_{u\cdot u} \right) \left( y_u^l \right), \]

\[ IV = \left( y_u^l \right)^\top \left( \Sigma_{u\cdot u} \left( \mu_u \right) \right) - \left( \mu_u \right)^\top \left( \Sigma_{u\cdot u} \right) \left( y_u^l \right). \]

Summing up I(b), II, III and IV together yields

\[ I(b) + II + III + IV = \left( y_u^l - \mu_u \right)^\top \left( \tilde{\Sigma}_{u\cdot u} \right) \left( y_u^l - \mu_u \right) = \left( y_u^l - \mu_u \right)^\top \left( \Sigma^{-1} \right) \left( y_u^l - \mu_u \right). \]

This implies that

\[ (x + l - \mu) \Sigma^{-1} (x + l - \mu) = \left( y_u^l \right)^\top \left( \Sigma_{uu} \right)^{-1} y_u^l + \left( y_u^l - \mu_u \right)^\top \left( \Sigma^{-1} \right) \left( y_u^l - \mu_u \right). \]

Equations \[ (10) \] and \[ (9) \] imply that for a fixed \( l \in \mathbb{Z}^d \), the shifted normal density function can be decomposed as

\[ \mathcal{N} \left( x + l \mid \mu, \Sigma \right) = \mathcal{N} \left( \left( x + l \right)_u \mid \mu_u, \Sigma_{uu} \right) \cdot \mathcal{N} \left( \left( x + l \right)_{u^c} \mid \mu_{u^c} - \mu_{u^c}, \tilde{\Sigma}_{u^c u^c} \right) \]

Inserting \[ (11) \] into the definition of the marginal density function \( f_{X_u} \) gives us

\[ f_{X_u} = \sum_{l = (l_u, l_{u^c}) \in \mathbb{Z}^d} \mathcal{N} \left( \left( x + l \right)_u \mid \mu_u, \Sigma_{uu} \right) \cdot \int_{\mathbb{R}^{n_{u^c}}} \mathcal{N} \left( \left( x + l \right)_{u^c} \mid \mu_{u^c} - \mu_{u^c}, \tilde{\Sigma}_{u^c u^c} \right) dx_{u^c} \]

\[ = \sum_{l_u \in \mathbb{Z}^n} \mathcal{N} \left( x_u + l \mid \mu_u, \Sigma_{uu} \right) \cdot \sum_{l_{u^c} \in \mathbb{Z}^{n_{u^c}}} \int_{\mathbb{R}^{n_{u^c}}} \mathcal{N} \left( \left( x + l \right)_{u^c} \mid \mu_{u^c} - \mu_{u^c}, \tilde{\Sigma}_{u^c u^c} \right) dx_{u^c} \]

\[ = \sum_{l_u \in \mathbb{Z}^n} \mathcal{N} \left( x_u + l \mid \mu_u, \Sigma_{uu} \right) \cdot \mathcal{N} \left( z \mid \mu_{u^c} - \mu_{u^c}, \tilde{\Sigma}_{u^c u^c} \right) dz_{u^c}, \]

\[ = \sum_{l_u \in \mathbb{Z}^n} \mathcal{N} \left( x_u + l \mid \mu_u, \Sigma_{uu} \right) \cdot \sum_{l_{u^c} \in \mathbb{R}^{n_{u^c}}} \int_{\mathbb{R}} \mathcal{N} \left( z \mid \mu_{u^c} - \mu_{u^c}, \tilde{\Sigma}_{u^c u^c} \right) dz_{u^c}, \]

\[ = \sum_{l_u \in \mathbb{Z}^n} \mathcal{N} \left( x_u + l \mid \mu_u, \Sigma_{uu} \right) \cdot \int_{\mathbb{R}^{n_{u^c}}} \mathcal{N} \left( z \mid \mu_{u^c} - \mu_{u^c}, \tilde{\Sigma}_{u^c u^c} \right) dz_{u^c}, \]
where \( I_j = [t_j, t_j + 1] \), for all \( j = 1, \ldots, n^c \), and \( l_{u^c} = (t_1, \ldots, t_{n^c}) \). This implies the claim. We can similarly show that the assumption of the marginal density function \( f_{X_{u^c}} \) with respect to \( X_{u^c} \) also holds, by using the second definition of the inverse block matrix from proposition 8.1(ii) for the inverse covariance matrix \( \Sigma^{-1} \) where

\[
\tilde{\Sigma}_{uu} = \left( \Sigma_{uu} - \Sigma_{uu^c} (\Sigma_{u^c u^c})^{-1} \Sigma_{u^c u} \right)^{-1},
\]

\[
\tilde{\Sigma}_{u u^c} = - \left( \Sigma_{uu} - \Sigma_{uu^c} (\Sigma_{u^c u^c})^{-1} \Sigma_{u^c u} \right)^{-1} \Sigma_{uu^c} (\Sigma_{u^c u^c})^{-1},
\]

\[
\tilde{\Sigma}_{u^c u} = - (\Sigma_{u^c u^c})^{-1} \Sigma_{u^c u} \left( \Sigma_{uu} - \Sigma_{uu^c} (\Sigma_{u^c u^c})^{-1} \Sigma_{u^c u} \right)^{-1} \Sigma_{uu^c} (\Sigma_{u^c u^c})^{-1},
\]

and by applying proposition 8.1(iii) to the determinant of the inverse covariance matrix we then get

\[
\det(\Sigma) = 1/\det(\Sigma^{-1}) = 1/\left( \det(\tilde{\Sigma}_{uu}) \cdot \det \left( \tilde{\Sigma}_{u u^c} - \tilde{\Sigma}_{u^c u} \left( \tilde{\Sigma}_{uu^c} \right)^{-1} \tilde{\Sigma}_{u^c u^c} \right) \right),
\]

\[
= 1/\det \left( \left( \Sigma_{uu} - \Sigma_{uu^c} (\Sigma_{u^c u^c})^{-1} \Sigma_{u^c u} \right)^{-1} \right) \cdot \det \left( (\Sigma_{u^c u^c})^{-1} \right),
\]

\[
= \det \left( \Sigma_{uu} - \Sigma_{uu^c} (\Sigma_{u^c u^c})^{-1} \Sigma_{u^c u} \right) \cdot \det (\Sigma_{u^c u^c}),
\]

since the determinant of a matrix is also equal to the inverse of the determinant of its inverse matrix.

Furthermore by the Bayes theorem the conditional density function holds for all \( x \in \mathbb{T}^d \) and \( u \subset [d] \)

\[
f_{X_{u^c} | (X,L)} = \sum_{l_{u^c} \in \mathbb{Z}^{n^c}} f_{X_{u^c} | (X,L)} = \sum_{l_{u^c} \in \mathbb{Z}^{n^c}} \frac{f_{X_{u^c} | (X,L)}}{f_{(X,L)}}.
\]

Hence together with (11) we obtain

\[
f_{(x_{u^c} | (x,l))} = \sum_{l_{u^c} \in \mathbb{Z}^{n^c}} \mathcal{N}(x_{u^c} | x \in [d]) \mathcal{N}(x_{u^c} | \mu_{u^c}, \Sigma_{uu}),
\]

\[
= \sum_{l_{u^c} \in \mathbb{Z}^{n^c}} \mathcal{N}(x_{u^c} | \mu_{u^c}, \Sigma_{uu}) \mathcal{N}(x_{u^c} | \mu_{u^c} - \mu_{l_{u^c}}, \tilde{\Sigma}_{u^c u^c}),
\]

\[
= \sum_{l_{u^c} \in \mathbb{Z}^{n^c}} \mathcal{N}(x_{u^c} | \mu_{u^c} - \mu_{l_{u^c}}, \tilde{\Sigma}_{u^c u^c}).
\]