Nonparametric estimation of a multivariate density under Kullback-Leibler loss with ISDE

Louis Pujol

Nonparametric estimation of a multivariate density under Kullback-Leibler loss with ISDE

Louis Pujol*

Abstract

In this paper, we propose a theoretical analysis of the algorithm ISDE, introduced in previous work. From a dataset, ISDE learns a density written as a product of marginal density estimators over a partition of the features. We show that under some hypotheses, the Kullback-Leibler loss between the proper density and the output of ISDE is a bias term plus the sum of two terms which goes to zero as the number of samples goes to infinity. The rate of convergence indicates that ISDE tackles the curse of dimensionality by reducing the dimension from the one of the ambient space to the one of the biggest blocks in the partition. The constants reflect a combinatorial complexity reduction linked to the design of ISDE.

Keywords— Multivariate Density Estimation, Independence Structure, Nonparametric Density Estimation

*Université Paris-Saclay, CNRS, Inria, Laboratoire de Mathématiques d'Orsay, 91405, Orsay, France. louis.pujol@universite-paris-saclay.fr

1 NOTATIONS

Let f be a density function (a nonnegative real function whose integral is equal to 1) over $\mathbb{R}^d$. If we think of f from a statistical viewpoint, it is natural to refer to the indices $\{1, \ldots, d\}$ as the features.

Let $S \subset \{1, \ldots, d\}$, we denote by f_S the marginal density of f over S. For all $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$

$$f_S(x) = \int f(x) \prod_{i \notin S} dx_i. \tag{1.1}$$

With a slight abuse of notation, to highlight the fact that $f_S(x)$ does not depend on $(x_i)_{i \notin S}$, we write $f_S(x_S)$ instead of $f_S(x)$.

Let k be an positive integer not greater than d. We denote by Set_d^k the set of all subsets of $\{1, \ldots, d\}$ with cardinal not greater than k and by Part_d^k the collection of all partitions of $\{1, \ldots, d\}$ constructed with blocks in Set_d^k. We also use the shortcuts $\mathrm{Set}_d = \mathrm{Set}_d^d$ and $\mathrm{Part}_d = \mathrm{Part}_d^d$.

2 INTRODUCTION

In a previous work ([8]), we have introduced ISDE (Independence Structure Density Estimation). ISDE estimates a density f from a set of iid realizations $X_1, \ldots, X_N$ considering the Independence Structure (IS) hypothesis. This paper is devoted to a theoretical analysis of this algorithm. In this introduction, we review existing theory about IS, introduce ISDE, and set the goals of the present work.

2.1 Curse of dimensionality and independence Structure

Minimax Risk Let $X_1, \ldots, X_N$ be *iid* realizations of a random variable in $\mathbb{R}^d$ admitting a density f. The goal of density estimation is to construct an estimator $\hat{f}$ of the density. We can measure the hardness of such an estimation task using

the minimax framework. Assume that the true density belongs to some known model $\mathcal{F}$ and let D be a (pseudo)distance on $\mathcal{F}$, the minimax risk is defined as follows:

$$\mathcal{R}(D, \mathcal{F}) := \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}\left[D(f, \hat{f})\right] \tag{2.1}$$

where the inf is taken over all measurable functions from the data to $\mathcal{F}$. More specifically, a great part of the literature on the topic deals with the asymptotic regime of $\mathcal{R}(D, \mathcal{F})$ with respect to N.

Hölder Balls Let $\mathcal{U}$ be an open subset of $\mathbb{R}^d$ and $g : \mathcal{U} \to \mathbb{R}$ a function. Let $\gamma = (\gamma_1, \ldots, \gamma_d) \in \mathbb{N}^d$ be a multiindex and let $|\gamma| = \sum_{i=1}^d \gamma_i$ be its order. The partial differentiate operator D^γ is defined as follows

$$D^\gamma g = \frac{\partial^{|\gamma|} g}{\partial_1^{\gamma_1} \ldots \partial_d^{\gamma_d}}. \tag{2.2}$$

For a positive number β, if we denote by s the larger integer strictly lower than β and let $\delta = \beta - s \in (0, 1]$, g belongs to the Hölder ball $\mathcal{H}(\beta, L)$ where L is a positive real number if both following conditions are fulfilled

$$\begin{cases} \max\limits_{|\gamma| \leq s} \sup\limits_{x \in \mathcal{U}} \|D^\gamma g(x)\| \leq L \\ \max\limits_{|\gamma| = s} \sup\limits_{x, y \in \mathcal{U}} |D^\gamma g(x) - D^\gamma g_i(y)| \leq L \|x - y\|^\delta. \end{cases} \tag{2.3}$$

If g is defined on a close subset $\mathcal{C}$ of $\mathbb{R}^d$, we say that $g \in \mathcal{H}(\beta, L)$ if the restriction of g to the interior of $\mathcal{C}$.

Minimax Risk over Hölder Balls In [3], the minimax rate of this family of functions was studied considering L_p distances. In particular, the result with the squared L_2 distance is the following

$$\mathcal{R}\left(\|.\|_2^2, \mathcal{H}^\beta(d, H)\right) \sim N^{-\frac{2\beta}{2\beta + d}}. \tag{2.4}$$

We can interpret this bound as a manifestation of the curse of dimensionality because of its dependence on d. A solution is to consider the IS model introduced in [5].

Independence Structure For $k \leq d$, we define a family of functions:

$$\mathcal{D}_d^k = \left\{ f : \mathbb{R}^d \to \mathbb{R} |\ \exists \mathcal{P} \in \text{Part}_d^k : f(x) = \prod_{S \in \mathcal{P}} f_S(x_S) \right\}. \tag{2.5}$$

In probabilistic terms, a density f over $\mathbb{R}^d$ belongs to $\mathcal{D}_d^k$ if we can group these features into independent blocks. Another viewpoint is that the random variable characterized by f admits a graphical model, a collection of disjoint fully connected cliques of size not greater than k. It was showed in [9] that

$$\text{R}\left(\|.\|_2^2, \mathcal{H}^\beta(d, H) \cap \mathcal{D}_d^k\right) \sim N^{-\frac{2\beta}{2\beta+k}}. \tag{2.6}$$

The striking fact here is that the hardness of the estimation problem is no longer related to the ambient dimension but instead to the size of the biggest block of the partition on which the density function is decomposable.

2.2 ISDE

As explained in [8], the density estimation problem under squared L_2 loss does not lead to a feasible algorithm. This is why we change the loss function to the Kullback-Leibler (KL) divergence. If $\hat{f}$ is an estimator of f, the KL loss between f and $\hat{f}$ is defined as

$$\text{KL}\left(f \| \hat{f}\right) = \int \log\left(\frac{f}{\hat{f}}\right) f. \tag{2.7}$$

This formulation is well suited to IS as it involves log-densities, and a log of a product of marginal densities is a sum of log of marginal densities. From an algorithmic viewpoint, this formulation allowed us to implement an algorithm with reasonable running time and memory usage (see [8] for details). ISDE operates as follows.

1. Two independent datasets $W_1, \ldots, W_m$ and $Z_1, \ldots Z_n$ are extracted from $X_1, \ldots, X_N$.

2. $W_1, \ldots, W_m$ is used to compute marginal density estimators $(\hat{f}_S)_{S \in \mathrm{Set}_d^k}$. Any multidimensional density estimation procedure can be used for this step.

3. $Z_1, \ldots, Z_n$ is used to compute $(\ell_n(S))_{S \in \mathrm{Set}_d^k}$ where

$$\ell_n(S) = \frac{1}{n} \sum_{i=1}^{n} \log \hat{f}_S((Z_i)_S). \tag{2.8}$$

4. The optimization problem

$$\max_{\mathcal{P} \in \mathrm{Part}_d^k} \sum_{S \in \mathcal{P}} \ell_n(S) \tag{2.9}$$

is solved with an exact integer programming optimization procedure (branch-and-bound).

The output is a partition $\hat{\mathcal{P}}$ and an estimator of f taking the form $\hat{f}_{\hat{\mathcal{P}}} = \prod_{S \in \hat{\mathcal{P}}} \hat{f}_S$. The formulation with log-densities leads to a combinatorial complexity reduction. At first glance, the problem of density estimation under IS necessitates manipulating data structures of size P_d^k while ISDE only lies on data structures of size S_d^k.

2.3 Goal of this work

This paper is intended to provide a theoretical analysis of ISDE by upper-bounding the quantity $\mathrm{KL}\left(f \| \hat{f}\right)$. In particular, we will show that the introduction of IS tackles the curse of dimensionality and that the constants in the upper-bound reflect the combinatorial complexity reduction implemented in ISDE.

2.4 Organization of the paper

In section 3 we establish a first decomposition on the risk involving oracle partitions. In section 4 we introduce the regularity conditions on the proper density

and establish that an upper-bound for the uniform loss between marginal densities of f and marginal estimators is sufficient to obtain a convergence result for ISDE. In section 5 we show that it is possible to obtain an upper-bound for uniform estimation of marginal densities for a particular estimator. Then in section 6 we state the desired upper-bound for the estimator outputted by ISDE.

3 KULLBACK-LEIBLER RISK DECOMPOSITION

In this section, we show that the KL loss between f and $\hat{f}_{\hat{\mathcal{P}}}$, the estimator outputted by ISDE, decomposes as the sum of three terms with a clear interpretation.

3.1 Oracles partitions

We denote by $\hat{\mathcal{P}}$ the partition outputted by ISDE. Let $P_n[.]$ denotes the empirical measure associated with the sample $Z_1, \ldots, Z_n$ and $P[.]$ the measure associated with the true density f. $\hat{\mathcal{P}}$ is solution of the following optimisation problem :

$$\hat{\mathcal{P}} \in \underset{\mathcal{P} \in \mathrm{Part}_d^k}{\arg\min}\ P_n\left(-\log(\hat{f}_{\mathcal{P}})\right) \tag{3.1}$$

The partition $\hat{\mathcal{P}}$ is random depending on both W and Z. Let us define two other meaningful partitions.

$$\tilde{\mathcal{P}} \in \underset{\mathcal{P} \in \mathrm{Part}_d^k}{\arg\min}\ P\left(-\log(\hat{f}_{\mathcal{P}})\right) = \underset{\mathcal{P} \in \mathrm{Part}_d^k}{\arg\min}\ \mathrm{KL}\left(f \| \hat{f}_{\mathcal{P}}\right) \tag{3.2}$$

and

$$\mathcal{P}_* \in \underset{\mathcal{P} \in \mathrm{Part}_d^k}{\arg\min}\ P\left(-\log(f_{\mathcal{P}})\right) = \underset{\mathcal{P} \in \mathrm{Part}_d^k}{\arg\min}\ \mathrm{KL}\left(f \| f_{\mathcal{P}}\right). \tag{3.3}$$

$\tilde{\mathcal{P}}$ is a random partition depending on W but not on Z. it is the best combination of the estimators $(\hat{f}_S)_{S\in \mathrm{Set}_d^k}$ if we consider that the quantities $\left(P(-\log \hat{f}_S)\right)_{S\in \mathrm{Set}_d^k}$ are known.

$\mathcal{P}_*$ is not random. It is only a function of k. $f_{\mathcal{P}_*}$ can be interpreted as the Kullback-Leibler projection of f on the model $\mathcal{D}_d^k$ thanks to the following property.

Proposition 3.1

$$f_{\mathcal{P}_*} \in \arg\min_{g\in\mathcal{D}_d^k} \mathrm{KL}\left(f\|g\) \tag{3.4}$$

Proof. Let $g \in \mathcal{D}_d^k$ and denote by $\mathcal{P}_g$ a partition such that $g = \prod_{S\in\mathcal{P}_g} g_S$. We have

$$\begin{aligned}
\mathrm{KL}\left(f\|g\right) &= \int \log\left(\frac{f}{g}\right) f && (3.5)\\
&= \int \log\left(\frac{f}{f_{\mathcal{P}_g}}\right) f + \int \log\left(\frac{f_{\mathcal{P}_g}}{g}\right) f && (3.6)\\
&= \mathrm{KL}\left(f\|f_{\mathcal{P}_g}\right) + \sum_{S\in\mathcal{P}_g} \mathrm{KL}\left(f_S\|g_S\right) && (3.7)\\
&\leq \mathrm{KL}\left(f\|f_{\mathcal{P}_g}\right) && (3.8)
\end{aligned}$$

with equality if $g = f_{\mathcal{P}_g}$. Then

$$\begin{aligned}
\arg\min_{g\in\mathcal{D}_d^k} \mathrm{KL}\left(f\|g\right) &= \arg\min_{\mathcal{P}\in\mathrm{Part}_d^k} \min_{g\in\mathcal{D}_d^k} \mathrm{KL}\left(f\|g\right) && (3.9)\\
&= \arg\min_{\mathcal{P}\in\mathrm{Part}_d^k} \mathrm{KL}\left(f\|f_{\mathcal{P}}\right) && (3.10)
\end{aligned}$$

□

3.2 Kullback-Leibler risk upper-bound

We are now in a position to establish a control of the Kullback-Leibler risk for $\hat{f}_{\hat{P}}$ involving the oracles partitions.

Lemma 3.2: Kullback-Leibler risk control

$$\mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right) \leq \mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right) + \sum_{S_*\in\mathcal{P}_*} \mathrm{KL}\left(f_{S_*}\|\hat{f}_{S_*}\right) + (P-P_n)(\log\hat{f}_{\tilde{\mathcal{P}}} - \log\hat{f}_{\hat{\mathcal{P}}}) \tag{3.11}$$

Proof. We start by decomposing $\mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right)$ as follows

$$\begin{aligned}
\mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right) &= \mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right) && (3.12)\\
&+ \mathrm{KL}\left(f\|\hat{f}_{\mathcal{P}_*}\right) - \mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right) && (3.13)\\
&+ \mathrm{KL}\left(f\|\hat{f}_{\tilde{\mathcal{P}}}\right) - \mathrm{KL}\left(f\|\hat{f}_{\mathcal{P}_*}\right) && (3.14)\\
&+ \mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right) - \mathrm{KL}\left(f\|\hat{f}_{\tilde{\mathcal{P}}}\right). && (3.15)
\end{aligned}$$

Then, as $\mathrm{KL}\left(f\|\hat{f}_{\tilde{\mathcal{P}}}\right) \leq \mathrm{KL}\left(f\|\hat{f}_{\mathcal{P}_*}\right)$,

$$\begin{aligned}
\mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right) &\leq \mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right) && (3.16)\\
&+ \mathrm{KL}\left(f\|\hat{f}_{\mathcal{P}_*}\right) - \mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right) \qquad \text{(i)} && (3.17)\\
&+ \mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right) - \mathrm{KL}\left(f\|\hat{f}_{\tilde{\mathcal{P}}}\right) \qquad \text{(ii)}. && (3.18)
\end{aligned}$$

Now, we rewrite (i)

$$\begin{aligned}
\mathrm{KL}\left(f\|\hat{f}_{\mathcal{P}_*}\right) - \mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right) &= \int \log\left(\frac{f(x)}{\hat{f}_{\mathcal{P}_*}(x)}\right) f(x)dx - \int \log\left(\frac{f(x)}{f_{\mathcal{P}_*}(x)}\right) f(x)dx && (3.19)\\
&= \int \log\left(\frac{f_{\mathcal{P}_*}(x)}{\hat{f}_{\mathcal{P}_*}(x)}\right) f(x)dx && (3.20)\\
&= \sum_{S_*\in\mathcal{P}_*} \int \log\left(\frac{f_{S_*}(x)}{\hat{f}_{S_*}(x)}\right) f_{S_*}(x)dx && (3.21)\\
&= \sum_{S_*\in\mathcal{P}_*} \mathrm{KL}\left(f_{S_*}\|\hat{f}_{S_*}\right). && (3.22)
\end{aligned}$$

And we upper-bound (ii). As $P_n\left[\log \hat{f}_{\hat{\mathcal{P}}} - \log \hat{f}_{\tilde{\mathcal{P}}}\right] \geq 0$:

$$\begin{aligned}
\mathrm{KL}\left(f \| \hat{f}_{\hat{\mathcal{P}}}\right) - \mathrm{KL}\left(f \| \hat{f}_{\tilde{\mathcal{P}}}\right) &= P\left[-\log \hat{f}_{\tilde{\mathcal{P}}}\right] - P\left[-\log \hat{f}_{\hat{\mathcal{P}}}\right] && (3.23)\\
&\leq P\left[\log \hat{f}_{\tilde{\mathcal{P}}} - \log \hat{f}_{\hat{\mathcal{P}}}\right] + P_n\left[\log \hat{f}_{\hat{\mathcal{P}}} - \log \hat{f}_{\tilde{\mathcal{P}}}\right] && (3.24)\\
&= (P - P_n)(\log \hat{f}_{\tilde{\mathcal{P}}} - \log \hat{f}_{\hat{\mathcal{P}}}). && (3.25)
\end{aligned}$$

□

Three terms appear in the upper bound, and they can be easily interpreted.

- $\mathrm{KL}\left(f \| f_{\mathcal{P}_*}\right)$ is a bias term. It is the intrinsic error of the model $\mathcal{D}_d^k$ and can be thought of as a distance from f to $\mathcal{D}_d^k$ thanks to proposition 3.1.
- $\sum_{S_* \in \mathcal{P}_*} \mathrm{KL}\left(f_{S_*} \| \hat{f}_{S_*}\right)$ is an approximation term. It is a random quantity depending on the sample W and represents the error made when $f_{\mathcal{P}_*}$ is estimated with $\hat{f}_{\mathcal{P}_*}$. Conditionally to W, it depends on how the estimation of log-likelihoods made thanks to Z is accurate and quantifies our ability to output the optimal partition.

In the sequel of the paper, we will focus on upper-bounding the approximation and selection terms as they are the random quantities of interest in our problem. We treat the bias term as a structural error, and we focus on upper-bounding the quantity

$$\mathrm{KL}\left(f \| \hat{f}_{\hat{\mathcal{P}}}\right) - \mathrm{KL}\left(f \| f_{\mathcal{P}_*}\right). \tag{3.26}$$

A study of the bias in a multivariate Gaussian framework can be found in appendix A.

4 CONDITIONS AND OBJECTIVE

4.1 Regularity conditions

Bounding condition Density estimation under Kullback-Leibler loss is known to be a challenging problem. One work by [2] has studied the asymptotic convergence rates for kernel estimators in a one-dimensional setting. It was shown that the tails of the kernel must be chosen appropriately regarding the tails property of the proper density to have convergent estimators. In this work, we restrict our attention to densities that are lower and upper bounded by some positive quantities. This is done to avoid hardly tractable tail behavior issues. In the sequel, we consider that the following bounding condition is valid for all $S \in \text{Set}_d^k$

$$e^{-A|S|} \leq f_S \leq e^{A|S|} \; \forall S \in \mathcal{P}_* \tag{BC}$$

Note that if we impose a positive lower bound on the marginal densities, we must consider that f is compactly supported. In the sequel, we will suppose that the support of f is $[0,1]^d$.

Hölder Regularity We will consider in the sequel that it exists $\beta \in (0,2]$ and $L > 0$ such that $f_S \in \mathcal{H}(\beta, L)$ for all $S \in \text{Set}_d^k$. We will use the following approximation property for functions in Hölder balls.

Lemma 4.1: Approximation

Let us consider that $g \in \mathcal{H}(\beta, H)$ with $\beta \in (0,2)$ and the domain of g is $\mathcal{U} \subset \mathbb{R}^d$, then for all $x \in \mathcal{U}$ and u such that $x + u \in \mathcal{U}$. If $\beta \in (0,1]$ then

$$|g(x) - g(x+u)| \leq L\|u\|^\beta. \tag{4.1}$$

If $\beta \in (1,2]$ then

$$\left| g(x) - g(x+u) - \sum_{k=1}^{d} \partial_k u_k g(x) \right| \leq L\|u\|^\beta. \tag{4.2}$$

4.2 Objective

Our goal is to propose an estimation procedure for the collection of marginal densities $(f_S)_{S\in \mathrm{Set}_d^k}$. If we are able to ensure, simultaneously for all $S \in \mathrm{Set}_d^k$ a uniform control

$$\|\hat{f}_S - f_S\|_\infty \leq \epsilon_S < e^{-A|S|}(1 - e^{-A|S|}). \tag{UC}$$

Then we can upper-bound the approximation term and the selection term thanks to the following proposition.

Proposition 4.2: Consequences of uniform control

If the uniform control (UC) is satisfied and (BC) is true, then

1. A bounding condition is satisfied by all the estimators $(\hat{f}_S)_{S\in \mathrm{Set}_d^k}$

$$e^{-2A|S|} \leq \hat{f}_S \leq e^{2A|S|}. \tag{$\widehat{\mathrm{BC}}$}$$

2. For all $S \in \mathrm{Set}_d^k$, the Kullback-Leibler divergence between f_S and $\hat{f}_S$ can be upper-bounded

$$\mathrm{KL}\left(f_S \| \hat{f}_S\right) \leq e^{2A|S|}\epsilon_S \tag{4.3}$$

3. Conditionnaly on W, the selection term can be upper-bounded with high probability. More precisely if $\delta_n \in (0, 1)$ we have

$$\mathbb{P}\left[\left|(P - P_n)(\log \hat{f}_{\tilde{\mathcal{P}}} - \log \hat{f}_{\hat{\mathcal{P}}})\right| \geq 2d\sqrt{\frac{2Ak}{n}}\sqrt{\log\left(\frac{2S_d^k}{\delta}\right)}\,\middle|\, W\right] \leq \delta_n. \tag{4.4}$$

Proof. Proof of 1: Under (BC) we have

$$e^{-A|S|} - \|\hat{f}_S - f_S\|_\infty \leq \hat{f}_S \leq e^{A|S|} + \|\hat{f}_S - f_S\|_\infty. \tag{4.5}$$

Now, under (UC)

$$e^{-A|S|} - \|\hat{f}_S - f_S\|_\infty \geq e^{-A|S|} - e^{-A|S|}(1 - e^{-A|S|}) = e^{-2A|S|} \tag{4.6}$$

and

$$e^{A|S|} + \|\hat{f}_S - f_S\|_\infty \leq e^{A|S|} + e^{-A|S|}(1 - e^{-A|S|}) \tag{4.7}$$
$$\leq e^{A|S|} + e^{3A|S|}\left(e^{-A|S|}(1 - e^{-A|S|})\right) \tag{4.8}$$
$$= e^{A|S|} + e^{2A|S|}(1 - e^{-A|S|}) = e^{2A|S|}. \tag{4.9}$$

Proof of 2: Let us compute

$$\mathrm{KL}\left(f_S \| \hat{f}_S\right) = \int \log\left(\frac{f_S}{\hat{f}_S}\right) f_S \tag{4.10}$$
$$\leq \int \left(\frac{f_S - \hat{f}_S}{\hat{f}_S}\right) f_S \tag{4.11}$$

Using ($\widehat{\mathrm{BC}}$), 1 / $\hat{f}_S \leq e^{2A|S|}$

$$\leq e^{2A|S|}\|f_S - \hat{f}_{S,h_m}\|_\infty \tag{4.12}$$
$$\leq e^{2A|S|}\epsilon_S \tag{4.13}$$

Proof of 3: Let $S \in \mathrm{Set}_d^k$, under ($\widehat{\mathrm{BC}}$) we have $\log \hat{f}_S \in [-2A|S|, 2A|S|]$. Using Hoeffding inequality, we obtain

$$\mathbb{P}\left[\left|(P - P_n)\log \hat{f}_S\right| \geq \sqrt{\frac{2A|S|}{n}}\sqrt{\log\frac{2S_d^k}{\delta}}|W\right] \leq \frac{\delta}{S_d^k}. \tag{4.14}$$

Now, by union bound :

$$\mathbb{P}\left[\sup_{S\in\mathrm{Set}_d^k}\left|(P - P_n)\log \hat{f}_S\right| \geq \sqrt{\frac{2A|S|}{n}}\sqrt{\log\frac{2S_d^k}{\delta}}|W\right] \leq \delta. \tag{4.15}$$

This leads to :

$$\mathbb{P}\left[2d\sup_{S\in\mathrm{Set}_d^k}\left|(P - P_n)\log \hat{f}_S\right| \leq 2d\sqrt{\frac{2Ak}{n}}\sqrt{\log\frac{2S_d^k}{\delta}}|W\right] \geq 1 - \delta. \tag{4.16}$$

Now, we remark that

$$|(P-P_n)\log\hat{f}_{\tilde{\mathcal{P}}}-\log\hat{f}_{\hat{\mathcal{P}}}| = \left|\sum_{S\in\tilde{\mathcal{P}}}(P-P_n)\log\hat{f}_S-\sum_{S\in\hat{\mathcal{P}}}(P-P_n)\log\hat{f}_S\right| \tag{4.17}$$

$$\leq \sum_{S\in\tilde{\mathcal{P}}}\left|(P-P_n)\log\hat{f}_S\right|+\sum_{S\in\hat{\mathcal{P}}}\left|(P-P_n)\log\hat{f}_S\right| \tag{4.18}$$

$$\leq 2d\sup_{S\in\mathrm{Set}_d^k}\left|(P-P_n)\log\hat{f}_S\right|. \tag{4.19}$$

Then, we have

$$\mathbb{P}\left[|(P-P_n)\log\hat{f}_{\tilde{\mathcal{P}}}-\log\hat{f}_{\hat{\mathcal{P}}}|\geq 2d\sqrt{\frac{2Ak}{n}}\sqrt{\log\frac{2S_d^k}{\delta}}|W\right]\leq\delta. \tag{4.20}$$

□

5 UNIFORM DENSITY ESTIMATION FOR MARGINAL DENSITIES

5.1 For a fixed S

In this subsection, we fix a subset of variables $S\in\mathrm{Set}_d^k$ and we study the problem of constructing an estimator $\hat{f}_S$ based on the sample $W_1,\ldots,W_m$ giving a control of $\|f_S-\hat{f}_S\|_\infty$ in order to verify (UC). We decompose the error as a sum of a bias and a variance term as follows

$$\|f_S-\hat{f}_S\|_\infty\leq\underbrace{\left\|f_S-\mathbb{E}\left[\hat{f}_S\right]\right\|_\infty}_{\text{Bias}}+\underbrace{\left\|\mathbb{E}\left[\hat{f}_S\right]-\hat{f}_S\right\|_\infty}_{\text{Variance}}. \tag{5.1}$$

5.1.1 Bias upper-bound

Choice of the kernel function In the following we will use density estimator based on an ancillary function K called kernel. K is a nonnegative integrable function on $\mathbb{R}$ such that $\int K(x)dx = 1$, we consider the following assumptions on K:

$$\begin{cases} \forall x \in \mathbb{R}, K(-x) = K(x) \\ \text{Supp}(K) \in [-1, 1] \\ \int xK(x)dx = 0 \\ \|K\|_\infty < \infty \end{cases} \tag{A.K}$$

We will also assume that, if $K^S_{h,x} : u \mapsto \frac{1}{h^{|S|}} \prod_{k \in S} K\left(\frac{x_k - u_k}{h}\right)$, the family of function

$$\mathcal{F}_S = \left\{K^S_{h,x}, h > 0, x \in \mathbb{R}^{|S|}\right\} \tag{5.2}$$

is a bounded VC class of functions. It means that it exists positive numbers A and ν such that for any probability measure P over $\mathbb{R}^{|S|}$ and any $\tau \in (0, 1)$ we have

$$\mathcal{N}\left(\mathcal{F}_S, L_2(P), \tau\right) \leq \left(\frac{A\|K\|_\infty}{\tau}\right)^\nu \tag{5.3}$$

where $\mathcal{N}\left(\mathcal{F}_S, L_2(P), \tau\right)$ is the τ-covering number of $\mathcal{F}_S$ for the $L_2(P)$ distance. As proved in [1] this condition is met for almost all classical kernels. An example of kernel function K satisfying all the assumptions is the Epanechnikov kernel K_{Epa}

$$K_{\text{Epa}}(x) = \frac{3}{4}(1 - x^2)\mathbb{1}_{[-1,1]}(x). \tag{5.4}$$

Boundary issue One must be aware of the issue induced by the fact that f_S is supported on $[0, 1]^{|S|}$. We define the usual kernel density estimator (KDE) as follows. Let h be a positive real number. The KDE for the marginal density f_S associated with the kernel K, the bandwidth $h > 0$, and the sample W is defined as

$$\hat{f}^{\text{KDE}}_{h,S}(x) = \frac{1}{mh^{|S|}} \sum_{i=1}^{m} \prod_{k \in S} K\left(\frac{(W_i)_k - x_k}{h}\right) \tag{5.5}$$

We remark that even in the samples $W_1, \dots, W_m$ belong to $[0,1]^d$, there is no reason to have $\hat{f}^{\text{KDE}}_{h,S}$ supported in $[0,1]^{|S|}$.

In this setting, the bias of the classical KDE does not go to zero as $h \to 0$, illustrating the boundary issue induced by estimating a compactly supported density.

Proposition 5.1: Boundary issue

Let $f_S \in \mathcal{H}(2, L)$, the bias

$$\left\| \mathbb{E}\left[\hat{f}^{\text{KDE}}_{h,S}\right] - f_S \right\|_\infty \tag{5.6}$$

does not tend to 0 as $h \to 0$.

Proof.

$$\mathbb{E}\left[\hat{f}^{\text{KDE}}_{h,S}\right](0) - f_S(0) = \frac{1}{h^{|S|}} \int_{[0,1]^{|S|}} f_S(t) \prod_{k \in S} K\left(\frac{t_k}{h}\right) dt - f_S(0) \tag{5.7}$$

$$= \int_{[-1,1]^{|S|}} [f_S(hu) - f_S(0)] \prod_{k \in S} K(u_k) du_k \tag{5.8}$$

As $\int xK(x) = 0$

$$= \int_{[-1,1]^{|S|}} \left[f_S(hu) - f_S(0) - h \sum_{k \in S} u_k \partial_k f_S(0) \right] \prod_{k \in S} K(u_k) du_k \tag{5.9}$$

Now as $f_S(x) = 0$ for $x \notin [0,1]^{|S|}$

$$\begin{aligned} &= \int_{[0,1]^{|S|}} \left[f_S(hu) - f_S(0) - h \sum_{k \in S} u_k \partial_k f_S(0) \right] \prod_{k \in S} K(u_k) du_k \\ &\quad - f_S(0) \int_{[-1,1]^{|S|} \setminus [0,1]^{|S|}} \prod_{k \in S} K(u_k) du_k \\ &\quad - h \sum_{k \in S} \partial_k f_S(0) \int_{[-1,1]^{|S|} \setminus [0,1]^{|S|}} u_k \prod_{k \in S} K(u_k) du_k \end{aligned} \tag{5.10}$$

The third term in the final sum tends to 0 with h. The same is true for the first term as

$$\begin{aligned}
&\Big| \int_{[0,1]^{|S|}} \left[f_S(hu) - f_S(0) - h \sum_{k \in S} u_k \partial_k f_S(0) \right] \prod_{k \in S} K(u_k) du_k \Big| \\
&\qquad \leq \int_{[0,1]^{|S|}} \left| f_S(hu) - f_S(0) - h \sum_{k \in S} u_k \partial_k f_S(0) \right| \prod_{k \in S} K(u_k) du_k \qquad (5.11) \\
&\qquad \leq L h^2 \sum_{k \in S} \int_{[0,1]^{|S|}} u_K^2 \prod_{k \in S} K(u_k) du_k \qquad (5.12) \\
&\qquad \leq L \sigma_K^2 h^2. \qquad (5.13)
\end{aligned}$$

Now, as $\int_{[-1,1]^{|S|} \setminus [0,1]^{|S|}} \prod_{k \in S} K(u_k) du_k = \frac{2^{|S|}-1}{2^{|S|}}$, we conclude that

$$\lim_{h \to 0} \mathbb{E}\left[\hat{f}_{h,S}^{\mathrm{KDE}} \right](0) = f_S(0) \frac{2^{|S|}-1}{2^{|S|}} \geq e^{-A|S|} \frac{2^{|S|}-1}{2^{|S|}} > 0. \qquad (5.14)$$

□

Mirror-Image KDE To correct the boundary bias previously introduced, a solution is to add a correction to the estimator $\hat{f}_{h,S}^{\mathrm{KDE}}$ near the boundary of the domain of definition. Let us define three mirroring operations for a number $x \in [0,1]$

$$M^{-1}(x) = -x; \; M^0(x) = x; \; M^1(x) = 2 - x. \qquad (5.15)$$

We define the mirror-image KDE as a KDE constructed over the sample W augmented with mirror reflections of each point over all axis. An illustration of this operation in dimension 2 is given by fig. 1.

This estimator is an extension to every dimension of the one proposed in [6]. The formal definition is

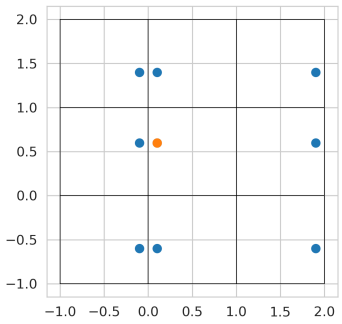
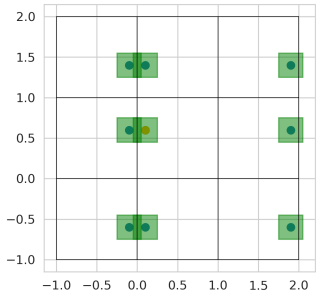
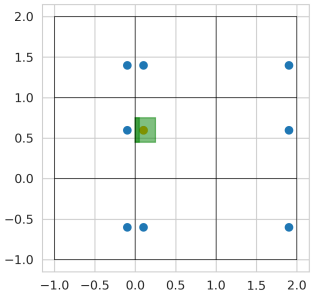


Figure 1: A datapoint in $[0,1]^d$ (in orange) and his 8 mirror-images (in blue)



(a) A kernel is fitted over all points of the augmented dataset



(b) Restriction to the unit hypercube

Figure 2: Construction of the mirror-image KDE

$$\hat{f}^{\mathrm{MI}}_{m,S}(x) = \mathbb{1}_{[0,1]^{|S|}}(x)\frac{1}{mh^{|S|}}\sum_{i=1}^{m}\sum_{a\in\{-1,0,1\}^{|S|}}\prod_{k\in S}K\left(\frac{M^{a_k}(W_i)_k - x_k}{h}\right). \tag{5.16}$$

It consists in summing multidimensional kernel over the points of the augmented samples and restrict the domain of the obtained function to $[0,1]^{|S|}$ as illustrated in fig. 2. Roughly speaking it consists in flipping the part of $\hat{f}^{\mathrm{KDE}}_{h,S}$ that fall outside $[0,1]^{|S|}$ inside it. $\hat{f}^{\mathrm{MI}}_{m,S}$ is supported on $[0,1]^{|S|}$ and $\int_{[0,1]^{|S|}}\hat{f}^{\mathrm{MI}}_{m,S}(x)dx = 1$.

Bias for mirror-image KDE Under an ad-hoc condition on the partial derivatives of f_S at the boundary of $[0,1]^{|S|}$ it is possible to bound the bias for the mirror-image KDE. Our result is an extension of the lemma 3.1 in [6] to every dimension and every $\beta \in (0,2]$ while the analysis in the original paper was restricted to bi-dimensional densities and $\beta = 2$. With our proof strategy, we find a better constant in the upper-bound for $|S| = 2$ and $\beta = 2$.

Proposition 5.2: Bias for mirror-image KDE

Let us assume that for all sequence $(x_n)_{n\in\mathbb{N}}$ in $[0,1]^{|S|}$, if x_n converges to a boundary point of $[0,1]^{|S|}$, then for all $k \in S$ $\lim_{n\to\infty} \partial_k f_S(x_n) = 0$. Then

$$\left\| f_S - \mathbb{E}\left[\hat{f}^{\mathrm{MI}}_{m,S}\right] \right\|_\infty \leq C_1 h^\beta \tag{5.17}$$

where $C_1 = L|S|^{\beta/2} \left(2\|K\|_\infty\right)^{|S|}$ if $\beta < 2$ and $C_1 = L|S|$ if $\beta = 2$

Proof. We define f^{MI}_S as the function defined over $[-1,2]^{|S|}$ such that for all $x \in [0,1]^{|S|}$ and $a \in \{-1,0,1\}^{|S|}$

$$f^{\mathrm{MI}}_S(M^a(x)) = f_S(X) \tag{5.18}$$

where $M^a(x) = \left(M^{a_k}(x_k)\right)_{k\in S}$. The property that the partial derivatives of f_S vanish near the boundary of $[0,1]^{|S|}$ ensures that $\partial_k f^{\mathrm{MI}}_S$ is continuous on $(-1,2)^{|S|}$ and so $f^{\mathrm{MI}}_S \in \mathcal{H}(2,L)$.

Let $x \in [0,1]^{|S|}$, we want to bound $\left| f_S(x) - \mathbb{E}\left[\hat{f}^{\mathrm{MI}}_{m,S}(x)\right] \right|$. Assume first that $x \in [0,1/2]^{|S|}$ and denote by $\mathcal{A}$ the set $\{k \in S : x_k < h\}$. We start by considering the situation where $|\mathcal{A}| \geq 1$. For all $k \in \mathcal{A}$ and all $t \in [0,1]$, $K\left(\frac{t-(2-x_k)}{h}\right) = 0$ because the support of K is $[-1,1]$, $h \leq 1/2$ and $x_k < h$. For all $k \in S \setminus \mathcal{A}$ and all $t \in [0,1]$, $K\left(\frac{t-(2-x_k)}{h}\right) = 0$ and $K\left(\frac{t-(-x_k)}{h}\right) = 0$. Then the expected value of $\hat{f}^{\mathrm{MI}}_{m,S}$ at the point x can be written as

$$\mathbb{E}\left[\hat{f}^{\mathrm{MI}}_{m,S}\right](x) = \sum_{\mathcal{B}\subset\mathcal{A}} \frac{1}{h^{|S|}} \int_{[0,1]^{|S|}} \prod_{k\in\mathcal{B}} K\left(\frac{t_k + x_k}{h}\right) \prod_{k\in S\setminus\mathcal{B}} K\left(\frac{t_k - x_k}{h}\right) f_S(t)dt \tag{5.19}$$

Now, for $\mathcal{B} \in \mathcal{A}$, we denote $x_{\mathcal{B}}$ the vector such that $(x_{\mathcal{B}})_k = x_k$ if $k \notin \mathcal{B}$ and $(x_{\mathcal{B}})_k = -x_k$ if $k \in \mathcal{B}$. We have

$$\mathbb{E}\left[\hat{f}^{\mathrm{MI}}_{m,S}\right](x) = \sum_{\mathcal{B}\subset\mathcal{A}} \int_{\chi^S_{\mathcal{B}}} \prod_{k\in S} K(u_k)\, f_S(x_{\mathcal{B}} + uh)du \tag{5.20}$$

where $\chi^S_{\mathcal{B}} = \left\{u \in [-1,1)^{|S|} : x_{\mathcal{B}} + uh \in [0,1]\right\}$. We see that $\chi^S_{\mathcal{B}} = \prod_{k\in S}[\underline{u}_k, \bar{u}_k)$ where $\underline{u}_k = -x_k/h$ if $k \in \mathcal{B}$, -1 otherwise and $\bar{u}_k = -x_k/h$ if $k \in \mathcal{A}\setminus\mathcal{B}$, 1 otherwise. What is more, as $f^{\mathrm{MI}}_S = f_S$ on $[0,1]^{|S|}$ we have

$$\mathbb{E}\left[\hat{f}^{\mathrm{MI}}_{m,S}\right](x) = \sum_{\mathcal{B}\subset\mathcal{A}} \int_{\chi^S_{\mathcal{B}}} \prod_{k\in S} K(u_k)\, f^{\mathrm{MI}}_S(x_{\mathcal{B}} + uh)du. \tag{5.21}$$

Now, as $(\chi^S_{\mathcal{B}})_{\mathcal{B}\subset\mathcal{A}}$ forms a partition of $[-1,1)^{|S|}$, we have

$$f_S(x) = \sum_{\mathcal{B}\subset\mathcal{A}} \int_{\chi^S_{\mathcal{B}}} \prod_{k\in S} K(u_k)\, f_S(x)du \tag{5.22}$$

$$= \sum_{\mathcal{B}\subset\mathcal{A}} \int_{\chi^S_{\mathcal{B}}} \prod_{k\in S} K(u_k)\, f^{\mathrm{MI}}_S(x_{\mathcal{B}})du \tag{5.23}$$

We denote by $\delta_{\mathcal{B}}(u,\beta)$ the quantity

$$\begin{cases} f^{\mathrm{MI}}_S(x_{\mathcal{B}} + uh) - f^{\mathrm{MI}}_S(x_{\mathcal{B}}) & \text{if } \beta \in (0,1] \\ f^{\mathrm{MI}}_S(x_{\mathcal{B}} + uh) - f^{\mathrm{MI}}_S(x_{\mathcal{B}}) - h\sum_{k\in S} u_k \partial_k f^{\mathrm{MI}}_S(x_{\mathcal{B}}) & \text{if } \beta \in (1,2] \end{cases} \tag{5.24}$$

From lemma 4.1 we have

$$|\delta_{\mathcal{B}}(u,\beta)| \leq Lh^{\beta}\|u\|^{\beta} \tag{5.25}$$

And, as $\int xK(x) = 0$, we have

$$\left| f_S(x) - \mathbb{E}\left[\hat{f}^{\mathrm{MI}}_{m,S}\right](x) \right| = \left| \sum_{\mathcal{B}\subset\mathcal{A}} \int_{\chi^S_{\mathcal{B}}} \prod_{k\in S} K(u_k)\, \delta_{\mathcal{B}}(u) du \right| \tag{5.26}$$

$$\leq Lh^\beta \int_{[-1,1]^{|S|}} \prod_{k\in S} K(u_k) \|u\|^\beta du \tag{5.27}$$

If $\beta = 2$

$$\int_{[-1,1]^{|S|}} \prod_{k\in S} K(u_k)\|u\|^\beta du = \int_{[-1,1]^{|S|}} \prod_{k\in S} K(u_k) \sum_{k\in S} u_k^2 du \tag{5.28}$$

$$= \sum_{k\in S} \int_{-1}^{1} K(u) u^2 du \tag{5.29}$$

$$\leq \sum_{k\in S} \int_{-1}^{1} K(u) du = |S|. \tag{5.30}$$

If $\beta < 2$

$$\int_{[-1,1]^{|S|}} \prod_{k\in S} K(u_k)\|u\|^\beta du \leq \|K\|_\infty^{|S|} \int_{[-1,1]^{|S|}} \|u\|^\beta du \tag{5.31}$$

$$\leq \|_\infty^{|S|} \sqrt{|S|}^\beta \int_{[-1,1]^{|S|}} du \tag{5.32}$$

$$= |S|^{\beta/2} \left(2\|K\|_\infty\right)^{|S|}. \tag{5.33}$$

Then $\sup_{x\in[0,1/2]^{|S|}} \left| f_S(x) - \mathbb{E}\left[\hat{f}^{\mathrm{MI}}_{m,S}\right](x) \right| \leq C_1 h^\beta$. By symmetry the same inequality is true when the sup is taken over $[0,1]^{|S|}$.

□

Then considering the mirror-image KDE leads to a correction of the boundary issue previously mentioned.

5.1.2 Variance upper-bound

To upper-bound the variance of the mirror-image KDE, we will use corollary 15 of [4]. Our setting is not the same as in this paper as we deal with mirror-image KDE. Then in order to obtain the same result, we must ensure that the family of functions

$$\mathcal{F}_S^{\mathrm{MI}} = \left\{ K_{x,h}^{\mathrm{MI}} | x \in [0,1]^{|S|}, h \in (0, 1/2) \right\} \tag{5.34}$$

where

$$K_{x,h}^{\mathrm{MI}} : u \mapsto \frac{1}{h^{|S|}} \mathbb{1}_{[0,1]^{|S|}} \sum_{a \in \{-1,0,1\}^{|S|}} \prod_{k \in S} K\left(\frac{M^{a_k}(u_k) - x_k}{h} \right) \tag{5.35}$$

is a bounded VC class of function. We know that $\mathcal{F}_S$ is a bounded VC class of function. The results of section 2.6 of [11] indicate that a family of functions is a bounded VC class if and only if the associated collection of sublevels is a VC class of sets. Now, we remark that the sublevels of functions in $\mathcal{F}_S^{\mathrm{MI}}$ can be written as intersections of sublevels of functions in $\mathcal{F}_S$ intersected with $[0,1]^{|S|}$. Then, as intersections preserve the VC class property for collection of sets (see [10]), $\mathcal{F}_S^{\mathrm{MI}}$ is a bounded VC class of functions, and the corollary 15 of [4] applies, leading to the following result.

Proposition 5.3: Variance

Let $h_{m,S}$ be a bandwidth in $(0, 1/2)$ and $\delta_m \in (0,1)$. With probability $1 - \delta_m$

$$\left\| \hat{f}_{S,h_{m,S}} - \mathbb{E}\left[\hat{f}_{m,h_{m,S}} \right] \right\|_\infty \leq C_2 \sqrt{\frac{\log\left(1/h_{m,S}\right) + \log(2/\delta_m)}{m h_{m,S}^{|S|}}}. \tag{5.36}$$

The constant C_2 depends on $|S|$, on $\|K\|_\infty$ and on $\|K'\|_\infty$.

5.1.3 Conclusion

Now, as we have for a control of the bias and the variance term for every bandwidth $h_{m,S} \in (0, 1/2)$, by choosing appropriately $h_{m,S}$ it is possible to bound $\|f_S - \hat{f}_{S,h_{m,S}}\|_\infty$.

Proposition 5.4: Convergence

Choosing $h_{m,S} \asymp (1/m)^{\frac{1}{2\beta+|S|}}$, it exists a constant C_S such that with probability at least $1 - \delta_m$

$$\|f_S - \hat{f}_{S,h_{m,S}}\|_\infty \leq C_S \sqrt{\log m + 2\log(2/\delta_m)} \left(\frac{1}{m}\right)^{\frac{\beta}{2\beta+|S|}}. \tag{5.37}$$

5.2 Uniformity over Set_d^k

We have just established a control in high probability for the quantity $\|f_S - \hat{f}_{S,h_{m,S}}\|_\infty$ for a given S. Our objective is to have such a control uniformly over Set_d^k. Applying a union-bound, we obtain the following result.

Proposition 5.5: Uniform control of uniform error over all subsets

Let us denote $C_k = \max_{S \in \mathrm{Set}_d^k} C_S$. We have, with probability at least $1 - S_d^k \delta_m$

$$\sup_{S \in \mathrm{Set}_d^k} \left\| \hat{f}_{S,h_m^S} - f_S \right\|_\infty \leq C_k \sqrt{\log m + 2\log(2/\delta_m)} \left(\frac{1}{m}\right)^{\frac{2}{4+k}}. \tag{5.38}$$

In particular with the choice $\delta_m = \frac{2}{m S_d^k}$, we have that with probability $1 - \frac{2}{m}$

$$\sup_{S \in \mathrm{Set}_d^k} \left\| \hat{f}_{S,h_m^S} - f_S \right\|_\infty \leq C_k \sqrt{3\log m + 2\log S_d^k} \left(\frac{1}{m}\right)^{\frac{2}{4+k}}. \tag{5.39}$$

6 THEOREM

Let now m_0 be the smallest integer m such that

$$C_k\sqrt{\log m + 2\log\left(2/\delta_m\right)}\left(\frac{1}{m}\right)^{\frac{2}{4+k}} \leq e^{-A|S|}\left(1-e^{-A|S|}\right). \tag{6.1}$$

If $m \geq m_0$, we know that on an event $\mathcal{A}_m^k$ of probability at least $1 - S_d^k\delta_m$

$$\|f_S - \hat{f}_{S,h_m}\|_\infty \leq C_k\sqrt{\log m + 2\log\left(2/\delta_m\right)}\left(\frac{1}{m}\right)^{\frac{\beta}{2\beta+k}}. \tag{6.2}$$

Then, on $\mathcal{A}_m^k$ (UC) is satisfied with $\epsilon_S = C_k\sqrt{\log m + 2\log\left(2/\delta_m\right)}\left(\frac{1}{m}\right)^{\frac{\beta}{2\beta+k}}$ for all $S \in \mathrm{Set}_d^k$. As a consequence, using proposition 4.2, we have

$$\sum_{S\in\mathcal{P}_*} \mathrm{KL}\left(f_S \| \hat{f}_{S,h_m^s}\right) \leq e^{2Ak}|\mathcal{P}_*|C_k\sqrt{\log m + 2\log\left(2/\delta_m\right)}\left(\frac{1}{m}\right)^{\frac{2}{4+k}}. \tag{6.3}$$

And, on $\mathcal{A}_m^k$, for $\delta_n \in (0, 1/S_d^k)$ with probability $1 - S_d^k\delta_n$

$$(P - P_n)(\log \hat{f}_{\tilde{\mathcal{P}}} - \log \hat{f}_{\hat{\mathcal{P}}}) \leq 2d\sqrt{\frac{Ak}{n}}\sqrt{\log(2/\delta_n)}. \tag{6.4}$$

Now, with the choice $\delta_m = 1/(2S_d^k m)$ and $\delta_n = 1/(2S_d^k n)$ we obtain the following result

Theorem 6.1: Final bound

If for all $S \in \mathrm{Set}_d^k$, $f \in \mathcal{H}(2, L)$, and for all sequence $(x_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} x_n = x^*$ where x^* belongs to the boundary of $[0,1]^{|S|}$ and for all $k \in S$ $\lim_{n\to\infty} \partial_k f_S(x_n) = 0$. With the choice $\hat{f}_S = \hat{f}^{\mathrm{MI}}_{h_{m,S},S}$ where $h_{m,S} \approx \left(\frac{1}{m}\right)^{\frac{1}{2+|S|}}$, we have with probability at least $(1-1/m)(1-1/n)$

$$\begin{aligned}\mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right) - \mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right) \leq & e^{2Ak}\sqrt{2}|\mathcal{P}_*|C_k\sqrt{\log m + \log\left(S_d^k\right)}\left(\frac{1}{m}\right)^{\frac{\beta}{2\beta+k}} \\ & + 2d\sqrt{\log n + \log\left(S_d^k\right)}\sqrt{\frac{Ak}{n}} \end{aligned} \tag{6.5}$$

Ignoring logarithmic factors, the rate of convergence of the approximation term is $\left(\frac{1}{m}\right)^{\frac{\beta}{2\beta+k}}$. The dependence of this quantity in k illustrates that ISDE tackles the curse of dimensionality for the density estimation problem under KL loss in the same spirit that [9] showed that his estimator does for the squared L_2 loss.

Ignoring logarithmic factors again, the rate of convergence of the selection term is $\frac{1}{\sqrt{n}}$. This is a classical rate of convergence for hold-out procedures with bounded loss (see corollary 8.8 in [7]).

The term $\log(S_d^k)$ in the upper-bound illustrates the combinatorial complexity reduction operated by ISDE. The presence of the log of the number of hypotheses is classical for hold-out procedures with bounded loss (see again corollary 8.8 in [7]). In our context, we have reduced the combinatorial complexity from the number of partitions P_d^k to the number of subsets S_d^k.

7 CONCLUSION

In this paper, we have studied the convergence properties of ISDE. In particular, we have shown that under suitable assumptions on the true density and for the mirror-image KDE as marginal density estimator, we can provide an upper-bound

valid with high-probability of the quantity

$$\mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right) - \mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right). \tag{7.1}$$

This bound highlights how ISDE tackles the curse of dimensionality and reduces the combinatorial complexity of the density estimation problem under IS compared to a brute-force approach. These results offer a theoretical validation of the empirical observations presented in [8].

To complete the study, it let to study how the bias term $\mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right)$ behaves. It is hard to give a precise statement on this quantity in a general setting. One simple situation is when $f \in \mathcal{D}_k^d$. In this case $\mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right) = 0$. This bias can also be explicitly evaluated in some multivariate Gaussian frameworks, see appendix A.

Acknowledgement This work was supported by the program Paris Region Ph.D. of DIM Mathinnov and was partly supported by the French ANR Chair in Artificial Intelligence TopAI - ANR-19-CHIA-0001. The author is thankful to Marc Glisse and Pascal Massart for their constructive remarks on this work.

References

[1] Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002.

[2] Peter Hall. On kullback-leibler loss and density estimation. *The Annals of Statistics*, pages 1491–1519, 1987.

[3] Rafael Hasminskii, Ildar Ibragimov, et al. On density estimation in the view of kolmogorov's ideas in approximation theory. *The Annals of Statistics*, 18(3):999–1010, 1990.

[4] Jisu Kim, Jaehyeok Shin, Alessandro Rinaldo, and Larry Wasserman. Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume

dimension. In *International Conference on Machine Learning*, pages 3398–3407. PMLR, 2019.

[5] Oleg Lepski. Multivariate density estimation under sup-norm loss: oracle approach, adaptation and independence structure. *Annals of Statistics*, 41(2):1005–1034, 2013.

[6] Han Liu, Larry Wasserman, and John Lafferty. Exponential concentration for mutual information estimation with application to forests. *Advances in Neural Information Processing Systems*, 25, 2012.

[7] Pascal Massart. Concentration inequalities and model selection. 2007.

[8] Louis Pujol. Isde: Independence structure density estimation. *arXiv preprint arXiv:2203.09783*, 2022.

[9] Gilles Rebelles et al. Lp adaptive estimation of an anisotropic density under independence hypothesis. *Electronic journal of statistics*, 9(1):106–134, 2015.

[10] Aad Van Der Vaart and Jon A Wellner. A note on bounds for vc dimensions. *Institute of Mathematical Statistics collections*, 5:103, 2009.

[11] AW van der Vaart, A.W. van der Vaart, A. van der Vaart, and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996.

A Bias term in the multivariate Gaussian framework

In this appendix, we study the bias $\mathrm{KL}\left(f \| f_{\mathcal{P}_*}\right)$ in a multivariate Gaussian framework where exact computations are possibles.

A.1 Model and notations

Model If Σ denotes a $d \times d$ definite positive matrix, we denote by f_Σ the density of a multivariate centered Gaussian random variable with covariance Σ and by $\Sigma_{\mathcal{P}}$ the matrix defined as follows.

$$\Sigma_{\mathcal{P}}(i,j) = \begin{cases} \Sigma(i,j) & \text{if } i \text{ and } j \text{ belongs to the same bloc in } \mathcal{P} \\ 0 & \text{else.} \end{cases} \tag{A.1}$$

If S_1 and S_2 are subsets of $\{1,\dots,d\}$, we denote by $\Sigma(S_1,S_2)$ the $|S_1| \times |S_2|$ submatrix matrix of Σ where we keep the intersection of rows in S_1 and columns in S_2, to keep notations compact, we write $\Sigma(S)$ instead of $\Sigma(S,S)$

For a multivariate Gaussian random variable with covariance Σ, $f_\Sigma \in \mathcal{D}_d^k$ is equivalent to the fact that it exists a permutation σ of $\{1,\dots,d\}$ such that $P_\sigma \Sigma P_\sigma^{-1}$ is block-diagonal with blocks of size smaller than $k \times k$. For clarity, in what follows, we will always consider that this property is met with $\sigma = \mathrm{id}$, meaning that we restrict ourselves to partitions in which each block is made of consecutive features. This does not imply a loss of generality.

We now consider that $f = f_\Sigma$ and

$$\Sigma = \Sigma_{\mathcal{P}} + \epsilon \tag{A.2}$$

where $\Sigma_{\mathcal{P}}$ is a block-diagonal covariance matrix corresponding to the independence structure $\mathcal{P}$ and ϵ is a "small" (in a sense to be defined later) definite positive matrix. The question is how this perturbation influences the bias term. In order to answer it, we must control $\mathrm{KL}\left(f \| f_{\mathcal{P}}\right)$ for all $\mathcal{P}$ in Part_d^k.

A.2 Some useful lemmas

Computation of KL losses The first useful result is an explicit computation of $\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_\mathcal{P}}\right)$ for any $\mathcal{P}$ in Part_d^k.

Lemma A.1: Exact computation of KL between two centered multivariate Gaussian

For every $\mathcal{P} \in \mathrm{Part}_d^k$

$$\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_\mathcal{P}}\right) = \frac{1}{2}\left(\sum_{S\in\mathcal{P}} \log\det\Sigma(S) - \log\det\Sigma\right) \tag{A.3}$$

Or if $\lambda_1 \leq \cdots \leq \lambda_d$ are the eigenvalues of Σ and $\lambda_1^\mathcal{P} \leq \cdots \leq \lambda_d^\mathcal{P}$ the eigenvalues of $\Sigma_\mathcal{P}$

$$\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_\mathcal{P}}\right) = \frac{1}{2}\sum_{i=1}^{d} \log\left(\frac{\lambda_i^\mathcal{P}}{\lambda_i}\right) \tag{A.4}$$

Proof. The density f_Σ has the following expression.

$$f_\Sigma(x) = \frac{1}{(2\pi)^{d/2}(\det\Sigma)^{1/2}} \exp\left(-\frac{1}{2}x^{\mathrm{T}}\Sigma^{-1}x\right). \tag{A.5}$$

We compute the KL divergence between f_Σ and $f_{\Sigma_\mathcal{P}}$

$$\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_\mathcal{P}}\right) = \int \log\left(\frac{f_\Sigma(x)}{f_{\Sigma_\mathcal{P}}(x)}\right) f_\Sigma(x)dx \tag{A.6}$$

$$= \frac{1}{2}\log\frac{\det\Sigma_\mathcal{P}}{\det\Sigma}\underbrace{\int f_\Sigma(x)dx}_{=1}$$

$$+ \frac{1}{2}\underbrace{\int x^{\mathrm{T}}\Sigma_\mathcal{P}^{-1}x f_\Sigma(x)dx}_{=\mathrm{Tr}(\Sigma_\mathcal{P}^{-1}\Sigma)}$$

$$+ \frac{1}{2}\underbrace{\int x^{\mathrm{T}}\Sigma^{-1}x f_\Sigma(x)dx}_{=\mathrm{Tr}(\Sigma^{-1}\Sigma)=d} \tag{A.7}$$

$$= \frac{1}{2}\left(\log\det\Sigma_\mathcal{P} - \log\det\Sigma + \mathrm{Tr}\left(\Sigma_\mathcal{P}^{-1}\Sigma\right) - d\right) \tag{A.8}$$

Now, we define a permutation $\sigma_\mathcal{P}$ of $\{1,\dots,d\}$ such that :

$$\Sigma_\mathcal{P} = P_{\sigma_\mathcal{P}}\left(\begin{array}{c|c|c|c} \Sigma(S_1) & \mathbf{0} & \dots & \mathbf{0} \\ \hline \mathbf{0} & \Sigma(S_2) & \ddots & \vdots \\ \hline \vdots & \ddots & \ddots & \mathbf{0} \\ \hline \mathbf{0} & \dots & \mathbf{0} & \Sigma(S_M) \end{array}\right) P_{\sigma_\mathcal{P}}^{-1} \tag{A.9}$$

where $\{S_1,\dots,S_M\}$ denotes the blocks of $\mathcal{P}$. It is then clear that $\log\det\Sigma_\mathcal{P} = \sum_{S\in\mathcal{P}}\log\det\Sigma(S)$. We also have

$$\Sigma = P_{\sigma_\mathcal{P}}\left(\begin{array}{c|c|c|c} \Sigma(S_1) & \Sigma(S_1,S_2) & \dots & \Sigma(S_1,S_M) \\ \hline \Sigma(S_2,S_1) & \Sigma(S_2) & \ddots & \vdots \\ \hline \vdots & \ddots & \ddots & \Sigma(S_{M-1},S_M) \\ \hline \Sigma(S_M,S_1) & \dots & \Sigma(S_M,S_{M-1}) & \Sigma(S_M) \end{array}\right) P_{\sigma_\mathcal{P}}^{-1}. \tag{A.10}$$

Then

$$\Sigma_\mathcal{P}^{-1}\Sigma = P_{\sigma_\mathcal{P}}\left(\begin{array}{c|c|c|c} I_{|S_1|} & X_{1,2} & \dots & X_{1,M} \\ \hline X_{2,1} & I_{|S_2|} & \ddots & \vdots \\ \hline \vdots & \ddots & \ddots & X_{M-1,M} \\ \hline X_{M,1} & \dots & \Sigma(S_M,S_{M-1}) & I_{|S_M|} \end{array}\right) P_{\sigma_\mathcal{P}}^{-1} \tag{A.11}$$

where for $i\neq j$, $X_{i,j}$ is a $|S_i|\times|S_j|$ matrix. Then $\mathrm{Tr}\left(\Sigma_\mathcal{P}^{-1}\Sigma\right) = d$.

The formulation of the result involving the eigenvalues comes from the fact that $\det \Sigma = \prod_{i=1}^{d} \lambda_i$ and $\det \Sigma_{\mathcal{P}} = \prod_{i=1}^{d} \lambda_i^{\mathcal{P}}$. □

Some computation of determinants We define the $p \times p$ matrix

$$A_\sigma^p = \begin{pmatrix} 1 & \sigma & \dots & \sigma \\ \sigma & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma \\ \sigma & \dots & \sigma & 1 \end{pmatrix}. \tag{A.12}$$

If k divides d we define

$$\Sigma_\sigma^{(d,k)} = \left(\begin{array}{c|c|c|c} A_\sigma^k & \mathbf{0} & \dots & \mathbf{0} \\ \hline \mathbf{0} & \ddots & \ddots & \vdots \\ \hline \vdots & \ddots & \ddots & \mathbf{0} \\ \hline \mathbf{0} & \dots & \mathbf{0} & A_\sigma^k \end{array}\right) \tag{A.13}$$

For $\epsilon > 0$ we define

$$\Sigma_{\sigma,\epsilon}^{(d,k)} = \left(\begin{array}{c|c|c|c} A_\sigma^k & \epsilon & \dots & \epsilon \\ \hline \epsilon & \ddots & \ddots & \vdots \\ \hline \vdots & \ddots & \ddots & \epsilon \\ \hline \epsilon & \dots & \epsilon & A_\sigma^k \end{array}\right) \tag{A.14}$$

Lemma A.2: Property for block matrices

i $\det A_\sigma^p = [1-\sigma]^{p-1}\,[1+(p-1)\sigma]$

ii $\det \Sigma_\sigma^{(d,k)} = [1-\sigma]^{\frac{d}{k}(k-1)}[1+(k-1)\sigma]^{\frac{d}{k}}$

iii $\det \Sigma_{\sigma,\epsilon}^{(d,k)} = [1-\sigma]^{\frac{d}{k}(k-1)}[1+(k-1)\sigma+(d-k)\epsilon][1+(k-1)\sigma-k\epsilon]^{\frac{d}{k}-1}$

Proof. (i) We start by computing the eigenvalues of A^p_σ. We remark that

$$A - (1-\sigma)I = \begin{pmatrix} \sigma & \dots & \sigma \\ \vdots & & \vdots \\ \sigma & \dots & \sigma \end{pmatrix}. \tag{A.15}$$

Then it is clear that $x \in \mathbb{R}^p \in \ker\left(A^p_\sigma - (1-\sigma)I\right) \Leftrightarrow x \in \{y \in \mathbb{R}^p : \sum_{i=1}^p y_i = 0\}$, which is a linear subspace of $\mathbb{R}^p$ of dimension $p-1$.

Then we remark that if $\mathbb{1}_p = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ then

$$A^p_\sigma \mathbb{1}_p = (1 + (p-1)\sigma)\,\mathbb{1}_p \tag{A.16}$$

Then $1+(p-1)\sigma$ is an eigenvalue of A and it could not be of multiplicity greater than 1 as we have just proven than $(1-\sigma)$ has a multiplicity of $p-1$. Using the fact that the determinant is the product of the eigenvalues, we obtain

$$\det A^p_\sigma = (1-\sigma)^{p-1}\,(1 + (p-1)\sigma) \tag{A.17}$$

The proof of (ii) follows immediately as the determinant of a block-diagonal matrix is the product of the derminants of the diagonal blocks.

(iii) We determine the eigenvalues of $\Sigma^{(d,k)}_{\sigma,\epsilon}$. To this end we will find a set of d linearly independent eigenvectors. We remark that

$$\Sigma^{(d,k)}_{\sigma,\epsilon}\mathbb{1}_d = (1 + (k-1)\sigma + (d-k)\epsilon)\mathbb{1}_d. \tag{A.18}$$

Then $1+(k-1)\sigma+(d-k)\epsilon$ is an eigenvalue of $\Sigma^{(d,k)}_{\sigma,\epsilon}$ with multiplicity at least 1. Now, we remark that for all integer $0 \le i \le \frac{d}{k} - 1$ and $2 \le j \le k$ we have

$$\Sigma^{(d,k)}_{\sigma,\epsilon}(e_{ik+1} - e_{ik+j}) = (1-\sigma)(e_{ik+1} - e_{ik+j}). \tag{A.19}$$

Then $(1-\sigma)$ is an eigenvalue of $\Sigma^{(d,k)}_{\sigma,\epsilon}$ with multiplicity at least $\frac{d}{k}(k-1)$. Finally, if for $i < j$ we denote by $\mathbb{1}^i_j = \sum_{k=i}^j e_k$, we remark that for all integer $1 \le i \le \frac{d}{k} - 1$

$$\Sigma^{(d,k)}_{\sigma,\epsilon}\left(\mathbb{1}^0_k - \mathbb{1}^{ik+1}_{(i+1)k}\right) = (1 + (k-1)\sigma - k\epsilon)\left(\mathbb{1}^0_k - \mathbb{1}^{ik+1}_{(i+1)k}\right). \tag{A.20}$$

Then $1 + (k-1)\sigma + (d-k)\epsilon$ is an eigenvalue of $\Sigma_{\sigma,\epsilon}^{(d,k)}$ with multiplicity at least $\frac{d}{k} - 1$.

As $1 + \frac{d}{k}(k-1) + \left(\frac{d}{k} - 1\right) = d$, we now that the eigenvalues of $\Sigma_{\sigma,\epsilon}^{(d,k)}$ are $1 + (k-1)\sigma + (d-k)\epsilon)$, $1 - \sigma$ and $1 + (k-1)\sigma + (d-k)\epsilon$ with multiplicity 1, $\frac{d}{k}(k-1)$ and $\frac{d}{k} - 1$.

□

A.3 Control of the bias

Almost independence structure the following property precise the KL loss between $f_{\Sigma_{\sigma,\epsilon}^{(d,k)}}$ and $f_{\Sigma_{\sigma}^{(d,k)}}$, with a particular look at the situation where $\epsilon \to 0$.

Proposition A.3: Almost independence

$$\begin{aligned} \mathrm{KL}\left(f_{\Sigma_{\sigma,\epsilon}^{(d,k)}} \| f_{\Sigma_{\sigma}^{(d,k)}}\right) = &-\frac{1}{2}\log\left(1 + \frac{d-k}{1+(k-1)\sigma}\epsilon\right) \\ &-\frac{1}{2}\left(\frac{d}{k} - 1\right)\log\left(1 - \frac{k}{1+(k-1)\sigma}\epsilon\right) \end{aligned} \tag{A.21}$$

At the limit $\epsilon \to 0$

$$\mathrm{KL}\left(f_{\Sigma_{\sigma,\epsilon}^{(d,k)}} \| f_{\Sigma_{\sigma}^{(d,k)}}\right) \underset{\epsilon\to 0}{=} \frac{d(d-k)}{4(1+(k-1)\sigma)^2}\epsilon^2 + o(\epsilon^2) \tag{A.22}$$

Proof. As $\Sigma_{\sigma}^{(d,k)}$ is a block-diagonal submatrix of $\Sigma_{\sigma,\epsilon}^{(d,k)}$, using lemma A.1 we have

$$\mathrm{KL}\left(\Sigma_{\sigma,\epsilon}^{(d,k)} \| \Sigma_{\sigma}^{(d,k)}\right) = \frac{1}{2}\log\left(\frac{\det \Sigma_{\sigma}^{(d,k)}}{\det \Sigma_{\sigma,\epsilon}^{(d,k)}}\right) \tag{A.23}$$

Let $\beta = 1 + (k-1)\sigma$ Now, using lemma A.2, we have

$$\mathrm{KL}\left(\Sigma_{\sigma,\epsilon}^{(d,k)} \| \Sigma_{\sigma}^{(d,k)}\right) = \frac{1}{2}\left[\log\left(\frac{\beta^{\frac{d}{k}}}{[\beta + (d-k)\epsilon][\beta - k\epsilon]^{\frac{d}{k}-1}}\right)\right] \tag{A.24}$$

$$= \frac{1}{2}\left[\frac{d}{k}\log\beta - \log(\beta + (d-k)\epsilon) - \left(\frac{d}{k} - 1\right)\log\left(\beta - k\epsilon\right)\right] \tag{A.25}$$

$$= \frac{d}{2k}\log\beta - \frac{\log\beta}{2} - \frac{1}{2}\log\left(1 + \frac{d-k}{\beta}\epsilon\right)$$

$$- \frac{1}{2}\left(\frac{d}{k} - 1\right)\log\beta - \frac{1}{2}\left(\frac{d}{k} - 1\right)\log\left(1 - \frac{k}{\beta}\epsilon\right) \tag{A.26}$$

$$= -\frac{1}{2}\log\left(1 + \frac{d-k}{\beta}\epsilon\right) - \frac{1}{2}\left(\frac{d}{k} - 1\right)\log\left(1 - \frac{k}{\beta}\epsilon\right) \tag{A.27}$$

Now, we use that

$$\log(1 + \frac{d-k}{\beta}\epsilon) \underset{\epsilon\to 0}{=} \frac{d-k}{\beta}\epsilon - \frac{(d-k)^2}{2\beta^2}\epsilon^2 + o(\epsilon^2) \tag{A.28}$$

and

$$\log(1 - \frac{k}{\beta}\epsilon) \underset{\epsilon\to 0}{=} -\frac{k}{\beta}\epsilon - \frac{k^2}{2\beta^2}\epsilon^2 + o(\epsilon^2). \tag{A.29}$$

And we obtain

$$\mathrm{KL}\left(\Sigma_{\sigma,\epsilon}^{(d,k)} \| \Sigma_{\sigma}^{(d,k)}\right) \underset{\epsilon\to 0}{=} -\frac{d-k}{2\beta}\epsilon + \frac{(d-k)^2}{4\beta^2}\epsilon^2$$

$$+ \frac{\left(\frac{d}{k} - 1\right)k}{2\beta}\epsilon + \frac{\left(\frac{d}{k} - 1\right)k^2}{4\beta^2}\epsilon^2 + o(\epsilon^2) \tag{A.30}$$

$$\underset{\epsilon\to 0}{=} \frac{(d-k)^2 + kd - k^2}{4\beta^2}\epsilon^2 + o(\epsilon^2) \tag{A.31}$$

$$\underset{\epsilon\to 0}{=} \frac{d^2 - 2kd + k^2 + kd - k^2}{4\beta^2}\epsilon^2 + o(\epsilon^2) \tag{A.32}$$

$$\underset{\epsilon\to 0}{=} \frac{d(d-k)}{4\beta^2}\epsilon^2 + o(\epsilon^2) \tag{A.33}$$

□

Optimal structure for small k The following proposition establish that if $\Sigma = A_\sigma^d$ and $k < d$, $\mathcal{P}_*$ is composed of a maximum number of blocks of size k.

Proposition A.4: Optimal Structure

Suppose that $\Sigma = A_\sigma^d$. A structure $s = (s_i)_{i=1}^M$ is a list of positive integer with $\sum_{i=1}^M s_i = d$. To a structure is associated a partition with blocks of consecutive features with size $s_1, \dots, s_M$. For any structure s we have

$$\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_s}\right) = \frac{1}{2}\left(\sum_{i=1}^{M} \log\left(\frac{1+(s_i-1)\sigma}{1-\sigma}\right) - \log\left(\frac{1+(d-1)\sigma}{1-\sigma}\right)\right). \tag{A.34}$$

If we denote by p and r the only integers such that $d = pk + r$ where $r < k$, we have

$$s^* = (\underbrace{k, \dots, k}_{p \text{ times}}, r) \tag{A.35}$$

Proof. We combine lemma A.1 and lemma A.2 to obtain

$$\begin{aligned}
\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_s}\right) &= \frac{1}{2}\left(\sum_{S\in\mathcal{P}} \log\det A_\sigma^{s_i} - \log\det A_\sigma^d\right) && \text{(A.36)}\\
&= \frac{1}{2}[\sum_{S\in\mathcal{P}} (s_i - 1)\log(1-\sigma) + \log(1+(s_i-1)\sigma) \\
&\quad - (d-1)\log(1-\sigma) - \log(1+(d-1)\sigma)] && \text{(A.37)}\\
&= \frac{1}{2}\left(\underbrace{\sum_{i=1}^{M}(s_i-1)}_{=d-M} - (d-1)\right)\log(1-\sigma) \\
&\quad + \frac{1}{2}\left(\sum_{i=1}^{M}\log\left(1+(s_i-1)\sigma\right) - \log(1+(d-1)\sigma)\right) && \text{(A.38)}\\
&= \frac{1}{2}\left(\sum_{i=1}^{M}\log\left(\frac{1+(s_i-1)\sigma}{1-\sigma}\right) - \log\left(\frac{1+(d-1)\sigma}{1-\sigma}\right)\right) && \text{(A.39)}
\end{aligned}$$

Now we want to prove that the structure minimizing $\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_s}\right)$ is $(\underbrace{k,\ldots,k}_{p \text{ times}},r)$. To do so we start by remarking that for any $s=(s_i)_{i=1}^M \neq (k,\ldots,k,r)$ it exists $i\neq j$ such that $s_i \neq k$ and $s_j \neq k$. We will prove that, for our minimization problem it is always possible to find a better structure $\tilde{s}$ with the following

i if $s_i+s_j \leq k$, $\tilde{S}=(s_k)_{k\notin\{i,j\}} \cup (s_i+s_j)$

ii if $\exists l>0 : s_i+s_j=k+l$, $\tilde{S}=(s_k)_{k\notin\{i,j\}} \cup (k,l)$

To prove (i), we start from

$$2\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_s}\right) = \sum_{k=1}^{M} \log\left(\frac{1+(s_k-1)\sigma}{1-\sigma}\right) - \log\left(\frac{1+(d-1)\sigma}{1-\sigma}\right) \tag{A.40}$$

$$\begin{aligned} 2\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_{\tilde{s}}}\right) = &\sum_{k=1,\ldots,M,k\notin\{i,j\}} \log\left(\frac{1+(s_k-1)\sigma}{1-\sigma}\right) + \log\left(\frac{1+(s_i+s_j-1)\sigma}{1-\sigma}\right) \\ &- \log\left(\frac{1+(d-1)\sigma}{1-\sigma}\right). \end{aligned} \tag{A.41}$$

Then to prove that $\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_s}\right) > \mathrm{KL}\left(f_\Sigma \| f_{\Sigma_{\tilde{s}}}\right)$ it is sufficient to prove that for all $a,b\geq 1$, $g(a,b)>0$ where

$$g(a,b) = \log\left(\frac{1+(a-1)\sigma}{1-\sigma}\right) + \log\left(\frac{1+(b-1)\sigma}{1-\sigma}\right) - \log\left(\frac{1+(a+b-1)\sigma}{1-\sigma}\right). \tag{A.42}$$

Let us start by computing $\partial_1 g(a,b)$

$$\partial_1 g(a,b) = \frac{\sigma}{1+(a-1)\sigma} - \frac{\sigma}{1+(a+b-1)\sigma} \tag{A.43}$$

$$= \frac{\sigma}{\left(1+(a-1)\sigma\right)\left(1+(a+b-1)\sigma\right)}\left(1+(a+b-1)\sigma-1-(a-1)\sigma\right) \tag{A.44}$$

$$= \frac{b\sigma^2}{\left(1+(a-1)\sigma\right)\left(1+(a+b-1)\sigma\right)} \geq 0. \tag{A.45}$$

Then for any $b\geq 1$, $g(a,b)$ is nondecreasing in a. As a and b play similar roles in $g(a,b)$, we have that for any $a\geq 1$, $g(a,b)$ is nondecreasing in b. To prove that

$g(a,b) \geq 0$ it is sufficient to show that $g(1,1) > 0$.

$$g(1,1) = \log\left(\frac{1}{1-\sigma}\right) + \log\left(\frac{1}{1-\sigma}\right) - \log\left(\frac{1+\sigma}{1-\sigma}\right) \tag{A.46}$$

$$= -\log\left((1-\sigma)(1+\sigma)\right). \tag{A.47}$$

Now, as $\sigma \in (0,1)$, $(1-\sigma)(1+\sigma) \in (0,1)$. Then $\log\left((1-\sigma)(1+\sigma)\right) < 0$, implying $g(1,1) > 0$.

To prove (ii) we start from

$$2\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_s}\right) = \sum_{k=1}^{M} \log\left(\frac{1+(s_k-1)\sigma}{1-\sigma}\right) - \log\left(\frac{1+(d-1)\sigma}{1-\sigma}\right) \tag{A.48}$$

$$2\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_{\tilde{s}}}\right) = \sum_{k=1,\ldots,M, k\notin\{i,j\}} \log\left(\frac{1+(s_k-1)\sigma}{1-\sigma}\right) + \log\left(\frac{1+(k-1)\sigma}{1-\sigma}\right)$$
$$+ \log\left(\frac{1+(l-1)\sigma}{1-\sigma}\right) - \log\left(\frac{1+(d-1)\sigma}{1-\sigma}\right). \tag{A.49}$$

Then to prove that $\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_s}\right) > \mathrm{KL}\left(f_\Sigma \| f_{\Sigma_{\tilde{s}}}\right)$ it is sufficient to prove that for all $x \in [l,k]$, $h(x)$ attains its minimum at l or k where

$$h(x) = \log\left(1+(x-1)\sigma\right) + \log\left(1+((k+l)-x-1)\sigma\right). \tag{A.50}$$

Let us start by computing $h'(x)$

$$h'(x) = \sigma\left(\frac{1}{1+(x-1)\sigma} - \frac{1}{1+((k+l)-x-1)\sigma}\right) \tag{A.51}$$

$$= \frac{\sigma\left(1+((k+l)-x-1)\sigma - 1 - (x-1)\sigma\right)}{(1+(x-1)\sigma)(1+((k+l)-x-1)\sigma)} \tag{A.52}$$

$$= \frac{\sigma^2}{(1+(x-1)\sigma)(1+((k+l)-x-1)\sigma)}((k+l)-2x). \tag{A.53}$$

Then h increases from l to $(k+l)/2$ and decreases from $(k+l)/2$ to k and $h(l) = h(k)$ the minimum of h is attained on l and k.

□

Conclusion We finish this appendix by establishing a general upper bound of $\mathrm{KL}\left(f_\Sigma \| f_{\mathcal{P}_*}\right)$ where $\Sigma = \Sigma_{\sigma,\epsilon}^{(d,k^*)}$.

Theorem A.5: Upper-bound for the bias in a multivariate Gaussian framework

If $\Sigma = \Sigma_{\sigma,\epsilon}^{(d,k^*)}$ and if $k < k^*$. Let (p, r) be the unique couple of integer with $0 \leq r < k$ such that $k^* = pk + r$, we have

$$\begin{aligned}\mathrm{KL}\left(f_\Sigma \| f_{\mathcal{P}_*}\right) \leq \mathrm{KL}\left(f_{\Sigma_{\sigma,\epsilon}^{(d,k^*)}} \| f_{\Sigma_\sigma^{(d,k^*)}}\right) + \frac{dp}{2k^*}\log\left(\frac{1+(k-1)\sigma}{1-\sigma}\right) \\ + \frac{d}{2k^*}\log\left(\frac{1+(r-1)\sigma}{1-\sigma}\right) - \frac{d}{2k^*}\log\left(\frac{1+(k^*-1)\sigma}{1-\sigma}\right)\end{aligned} \tag{A.54}$$

with

$$\mathrm{KL}\left(f_{\Sigma_{\sigma,\epsilon}^{(d,k)}} \| f_{\Sigma_\sigma^{(d,k)}}\right) \underset{\epsilon\to 0}{=} \frac{d(d-k)}{4(1+(k-1)\sigma)^2}\epsilon^2 + o(\epsilon^2) \tag{A.55}$$

Proof. Let us consider

- the structure $\tilde{s} = (\underbrace{k, \ldots, k}_{p \text{ times}}, r)$, and $\mathcal{P}_{\tilde{s}}$ the associated partition of k^* features,
- the structure $s = (\underbrace{\tilde{s}, \ldots, \tilde{s}}_{d/k^* \text{ times}})$, and $\mathcal{P}_s$ the associated partition of d features,
- the structure $s_0 = (\underbrace{k^*, \ldots, k^*}_{d/k^* \text{ times}})$ and $\mathcal{P}_0$ the associated partition of d features.

We can upper-bound the bias term $\mathrm{KL}\left(f_\Sigma \| f_{\mathcal{P}_*}\right)$ as follows

$$\mathrm{KL}\left(f_\Sigma \| f_{\mathcal{P}_*}\right) \leq \mathrm{KL}\left(f_\Sigma \| f_{\mathcal{P}_s}\right) \tag{A.56}$$

$$= \int \log\left(\frac{f_\Sigma}{f_{\mathcal{P}_s}}\right) f_\Sigma \tag{A.57}$$

$$= \int \log\left(\frac{f_\Sigma}{f_{\mathcal{P}_0}}\right) f_\Sigma + \int \log\left(\frac{f_{\mathcal{P}_0}}{f_{\mathcal{P}_s}}\right) f_\Sigma \tag{A.58}$$

$$= \mathrm{KL}\left(\Sigma_{\sigma,\epsilon}^{(d,k^*)} \| \Sigma_\sigma^{(d,k^*)}\right) + \int \log\left(\frac{f_{\mathcal{P}_0}}{f_{\mathcal{P}_s}}\right) f_\Sigma \tag{A.59}$$

The blocks of the partition $\mathcal{P}_s$ are subsets of blocks of the partition $\mathcal{P}_0$, then

$$\int \log\left(\frac{f_{\mathcal{P}_0}}{f_{\mathcal{P}_s}}\right) f_\Sigma = \int \log\left(\frac{\prod_{S\in\mathcal{P}_0} f_S}{\prod_{S\in\mathcal{P}_0}(f_{\mathcal{P}_s})_S}\right) f_\Sigma \tag{A.60}$$

$$= \sum_{S\in\mathcal{P}_0} \int \log\left(\frac{f_S}{(f_{\mathcal{P}_s})_S}\right) f_S \tag{A.61}$$

$$= \sum_{S\in\mathcal{P}_0} \mathrm{KL}\left(f_{A^d_\sigma} \| f_{(A^d_\sigma)_{\mathcal{P}_{\tilde{s}}}}\right) \tag{A.62}$$

$$= \frac{d}{k^*}\mathrm{KL}\left(f_{A^d_\sigma} \| f_{(A^d_\sigma)_{\mathcal{P}_{\tilde{s}}}}\right) \tag{A.63}$$

Now, using proposition A.3

$$\mathrm{KL}\left(f_{A^d_\sigma} \| f_{(A^d_\sigma)_{\mathcal{P}_{\tilde{s}}}}\right) = \frac{1}{2}\Bigg(p\log\left(\frac{1+(k-1)\sigma}{1-\sigma}\right) + \log\left(\frac{1+(r-1)\sigma}{1-\sigma}\right) - \log\left(\frac{1+(k^*-1)\sigma}{1-\sigma}\right)\Bigg) \tag{A.64}$$

And, using proposition A.4, we have that

$$\mathrm{KL}\left(f_{\Sigma^{(d,k)}_{\sigma,\epsilon}} \| f_{\Sigma^{(d,k)}_{\sigma}}\right) \underset{\epsilon\to 0}{=} \frac{d(d-k)}{4(1+(k-1)\sigma)^2}\epsilon^2 + o(\epsilon^2). \tag{A.65}$$

Then we have proven the desired upper-bound.

□