A greedy approach to joint distribution learning

Yassine Lahjouji

Supervised by: **Pr. Damir Filipović**

Ecole Polytechnique Fédérale de Lausanne
Mathematics Section
Date : June 2022

A greedy approach to joint distribution learning

Yassine Lahjouji

Abstract

Machine learning methods using Reproducing Kernel Hilbert Spaces (RKHS) have seen a huge increase in interest in the last years for various applications due to the powerful representation power of these spaces. One such application is to embed probability distributions in RKHS's. Different methods exist to formulate such embeddings, but one of the most prominent ones is the low-rank embedding technique proposed by Pr.Filipović, Pr.Multerer and Pr.Schneider in their paper [1]. Thanks to this framework, we can embed joint probability distribution in a tensor product of RKHS's through the Radon-Nikodym derivative and retrieve the true probability distribution by solving a quadratic program. We propose here a new framework designed for joint probability distributions where at least one marginal is finitely discrete. In this case, we consider a greedy approach consisting of finding the optimal local solution for each point of the discrete marginal. Our apporach is theoretically faster.

This thesis is divided into three chapters. The first one being a quick survey of RKHS's and some already existing embeddings of probability distributions. The second chapter will be a reformulation of the low-rank technique . Finally, in the last chapter, we consider our new greedy framework and consider some encouraging numerical results.

The code implementation can be found on my GitHub : *ylahjouji*.

A greedy approach to joint distribution learning

Yassine Lahjouji

Acknowledgements

Firstly, I would like to express my deepest and most heartfelt feelings of appreciation to my supervisor Pr. Damir Filipović. I am extremely thankful for his help throughout this project. His insight and feedback during our meetings were extremely beneficial to me. His direction was cardinal for the realization of this project.

Moreover, I would like to thank all the people that made my journey at EPFL a unique experience that I will cherish my whole life. I also would like to thank EPFL for the opportunity of studying in such an outstanding university.

Finally, I would like to express my profound gratitude for my parents who always found the right words to support me and accompany me through my master.

Contents

Chapter 1

Probability distribution embeddings in RKHS

1.1 Detour into functional analysis

1.1.1 Fundamental results about RKHS

Let $\mathcal{X}$ be a set. We denote by $\mathcal{F}(\mathcal{X}, \mathbb{R})$ the set of functions from $\mathcal{X}$ to $\mathbb{R}$. This set is a vector space under pointwise addition and scalar multiplication.

Definition 1.1. We call a subset $\mathcal{H} \subseteq \mathcal{F}(\mathcal{X}, \mathbb{R})$ a **Reproducing Kernel Hilbert Space**, or more commonly **RKHS** on $\mathcal{X}$, if

- $\mathcal{H}$ is a subspace of $\mathcal{F}(\mathcal{X}, \mathbb{R})$
- There exists an inner product $\langle\ ,\ \rangle_{\mathcal{H}}$ such that $(\mathcal{H}, \langle\ ,\ \rangle_{\mathcal{H}})$ is a Hilbert space.
- $(\forall x \in \mathcal{X})$ The evaluation functional

$$\begin{aligned} E_x \colon \mathcal{H} &\to \mathbb{R} \\ f &\mapsto f(x) \end{aligned}$$

  is a bounded operator, i.e there exists $M_x < +\infty$ such that

$$(\forall f \in \mathcal{H}) \qquad E_x(f) \leq M_x \|f\|_{\mathcal{H}}$$

By using the Riesz-Fréchet representation theorem [2, Chapter 16] on the fucntional E_x, there exists an element of $\mathcal{H}$ which we will denote k_x such that

$$(\forall f \in \mathcal{H}) \qquad f(x) = E_x(f) = \langle f, k_x \rangle_{\mathcal{H}} \tag{1.1}$$

Definition 1.2. The map k_x is called the **reproducing kernel for** x. Considering these maps for all $x \in \mathcal{X}$, we have the map

$$\begin{aligned} k_{\mathcal{X}} \colon \mathcal{X} \times \mathcal{X} &\to \mathbb{R} \\ (x, y) &\mapsto k_{\mathcal{X}}(x, y) = k_y(x) \end{aligned}$$

This map is called the **reproducing kernel for** $\mathcal{H}$.

Notice that we have the following:

$$k_{\mathcal{X}}(x, y) = k_y(x) = \langle k_y, k_x \rangle_{\mathcal{H}} = \langle k_x, k_y \rangle_{\mathcal{H}} = k_{\mathcal{X}}(y, x)$$

Example 1.3. Let $\mathcal{X} = \{1, 2, \dots, n\}$. Considering all the maps $f\colon \mathcal{X} \to \mathbb{R}$, we can identify $\mathcal{F}(\mathcal{X}, \mathbb{R})$ to $\mathbb{R}^n$. Indeed, we could write f as $(f(1), \dots, f(n)) \in \mathbb{R}^n$. Moreover, consider the inner product given by:

$$\langle f, g \rangle = \sum_{i=1}^{n} f(i)g(i)$$

Hence, with this structure, $\mathcal{F}(\mathcal{X}, \mathbb{R})$ is an RKHS on $\mathcal{X}$. A reproducing kernel for $i \in \mathcal{X}$ is given by

$$\begin{aligned} k_i\colon \mathcal{X} &\to \mathbb{R} \\ j &\mapsto k_i(j) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \end{aligned}$$

Hence, the reproducing kernel is given by

$$k_{\mathcal{X}}(i, j) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

Definition 1.4. Let $\mathcal{X}$ be a set and let $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a map. We say that K is a **kernel function** if

$$(\forall n \in \mathbb{N})(\forall x_1, x_2, \dots, x_n \in \mathcal{X}) \qquad (k(x_i, x_j))_{i,j} \quad \text{is a positive semidefinite matrix}$$

Proposition 1.1. *Let $\mathcal{X}$ be a set and let $\mathcal{H}$ be an RKHS on $\mathcal{X}$ with reproducing kernel $k_{\mathcal{X}}$. Then, $k_{\mathcal{X}}$ is a kernel function.*

Proof. This is quite straightforward. Let $x_1, x_2, \dots, x_n \in \mathcal{X}$ and let $v \in \mathbb{R}^n$. Then, we have that:

$$\begin{aligned} v^{\intercal}(k_{\mathcal{X}}(x_i, x_j))_{i,j} v &= \sum_{i,j=1}^{n} k_{\mathcal{X}}(x_i, x_j) v_i v_j \\ &= \sum_{i,j=1}^{n} \langle k_{x_i}, k_{x_j} \rangle_{\mathcal{H}} v_i v_j \\ &= \langle \sum_{i=1}^{n} v_i k_{x_i}, \sum_{j=1}^{n} v_j k_{x_j} \rangle_{\mathcal{H}} \\ &= \| \sum_{i=1}^{n} v_i k_{x_i} \|_{\mathcal{H}} \geq 0 \end{aligned}$$

■

Theorem 1.2 (Moore's theorem)**.** *Let X be a set and let $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function on $\mathcal{X}$. Then, there exists an RKHS $\mathcal{H}$ on $\mathcal{X}$ such that k is a reproducing kernel of $\mathcal{H}$.*

Proof. See [3, Theorem 2.14] ■

Thus, there is a one-to-one correspondence between RKHS and kernel functions. In order to consider the embedding of joint probability distribution, we need to dive into the tensor product of RKHS.

1.1.2 Tensor product of RKHS

Tensor product of Hilbert spaces

Let $\mathcal{H}$ and $\mathcal{K}$ be two Hilbert spaces with respective inner products $\langle\ ,\ \rangle_{\mathcal{H}}$ and $\langle\ ,\ \rangle_{\mathcal{K}}$. For any $h, \hat{h} \in \mathcal{H}$ and $k, \hat{k} \in \mathcal{K}$, we define the following inner product on $\mathcal{H} \otimes \mathcal{K}$:

$$\langle h \otimes k, \hat{h} \otimes \hat{k} \rangle_{\mathcal{H}\otimes\mathcal{K}} = \langle h, \hat{h} \rangle_{\mathcal{H}} \langle k, \hat{k} \rangle_{\mathcal{K}}$$

This is indeed an inner product on the tensor product of $\mathcal{H}$ and $\mathcal{K}$. Taking the completion of this space, we have a Hilbert space which we will also denote $\mathcal{H} \otimes \mathcal{K}$. We refer to the former space as the algebraic tensor product while we refer to the latter as the **tensor product of Hilbert spaces** [4, Section 2.4].

Proposition 1.3. *Let $\mathcal{H}\otimes\mathcal{K}$ be the tensor product of two Hilbert spaces. Let $\{e_x\}_{x\in\mathcal{X}}$ and $\{f_y\}_{y\in\mathcal{Y}}$ be orthonormal bases respectively for $\mathcal{H}$ and $\mathcal{K}$. Then $\{e_x \otimes f_y\}_{x\in\mathcal{X},y\in\mathcal{Y}}$ is an orthonormal basis for $\mathcal{H} \otimes \mathcal{K}$.*

Proof. See [3, Proposition 4.9] ■

Tensor product of RKHS

Let $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ be two RKHS defined respectively on sets $\mathcal{X}$ and $\mathcal{Y}$ with respective reproducing kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$. For any element in the algebraic tensor product $u = \sum_{i=1}^{n} h_i \otimes f_i$, we can identify it to the following map :

$$\begin{aligned} \hat{u}\colon\ & \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \\ & (x, s) \mapsto \hat{u}(x, s) = \sum_{i=1}^{n} h_i(x) f_i(s) \end{aligned}$$

Theorem 1.4. *Consider the kernel function given by*

$$\begin{aligned} k\colon (\mathcal{X} \times \mathcal{Y})\ \times\ & (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R} \\ & ((x_1, y_1), (x_2, y_2)) \mapsto k((x_1, y_1), (x_2, y_2)) = k_{\mathcal{X}}(x_1, x_2) k_{\mathcal{Y}}(y_1, y_2) \end{aligned}$$

and the RKHS $\mathcal{H}(k)$ generated by k. The previous identification $u \mapsto \hat{u}$ can be extendend to a linear isometry from $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ to $\mathcal{H}(k)$.

Proof. See [3] ■

Definition 1.5. K is called the **tensor product of the kernels** $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$, and is denoted $k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$.

Example 1.6. Let $\mathcal{X} = \{1, 2, \ldots, n\}$ and consider the RKHS on $\mathcal{X}$, denoted $\mathcal{H}_{\mathcal{X}}$, given by the reproducing kernel $k_{\mathcal{X}}$ given in (1.3). An orthonormal basis for $\mathcal{H}$ is given by $\{e_i\}_{i\in\mathcal{X}}$ with

$$\begin{aligned} e_i\colon\ & \mathcal{X} \to \mathbb{R} \\ & j \mapsto \delta_{ij} \end{aligned}$$

Let $\mathcal{Y}$ be a set and consider a separable RKHS on $\mathcal{Y}$, denoted $\mathcal{H}_{\mathcal{Y}}$, given by a reproducing kernel $k_{\mathcal{Y}}$. Since $\mathcal{H}_{\mathcal{Y}}$ is separable, there exists a countable orthonormal

basis $\psi_{\mathcal{Y}} := \{\psi_1, \psi_2, \dots, \}$ for $\mathcal{H}_{\mathcal{Y}}$. Thus, by Proposition 1.3, we can write any $h \in \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ as:

$$h = \sum_{i=1}^{D} \sum_{u=1}^{\infty} h_{ui}\, e_i \otimes \psi_u$$

Hence, by considering $\hat{h} \in \mathcal{H}(k)$, where k is the tensor product of the kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$, we have that:

$$\begin{aligned}(\forall (j,s) \in \mathcal{X} \times \mathcal{Y}) \qquad \hat{h}(j,s) &= \sum_{i=1}^{D} \sum_{u=1}^{\infty} h_{ui} e_i(j) \psi_u(s) \\ &= \sum_{i=1}^{D} \delta_{ij} \sum_{u=1}^{\infty} h_{ui} \psi_u(s) \\ &= \sum_{i=1}^{D} e_i(j) \sum_{u=1}^{\infty} h_{ui} \psi_u(s)\end{aligned}$$

For every $i \in \mathcal{X}$, we consider the following maps:

$$h_i = \sum_{u=1}^{\infty} h_{ui} \psi_u \tag{1.2}$$

in $\mathcal{H}_{\mathcal{Y}}$. Thus, we have:

$$\hat{h}(j,s) = \sum_{i=1}^{D} e_i(j) h_i(s)$$

Hence, we have that:

$$\hat{h} = \sum_{i=1}^{D} e_i h_i$$

with e_i being the orthonormal basis of $\mathcal{H}_{\mathcal{X}}$ and h_i maps in $\mathcal{H}_{\mathcal{Y}}$. Thus we can always represent $\hat{h}$ as

$$\hat{h} = (h_1, h_2, \dots, h_D)$$

Moreover, we have that:

$$\hat{h}(j,s) = h_j(s) \tag{1.3}$$

1.2 Traditional embedding

We now move on to consider an important probability distribution embedding defined in [5]. Let $\mathcal{X}$ be a Polish space (complete seperable metric space), and X be a random variable with probability distribution $\mathbb{P}_X$ on $\mathcal{X}$. Let $k_{\mathcal{X}}$ be a reproducing kernel on $\mathcal{X}$. For any $x \in \mathcal{X}$, consider the reproducing kernel for x, i.e. the following map

$$\begin{aligned}k_x : k_{\mathcal{X}}(x, .) \colon \mathcal{X} &\to \mathbb{R} \\ x' &\mapsto k_{\mathcal{X}}(x, x')\end{aligned}$$

Conditioned that this map is L^1-integrable, we can can consider the expectation $\int_{\mathcal{X}} k_{\mathcal{X}}(x,.)\, d\,\mathbb{P}_X$. Thus, our setting induces the following map:

$$\begin{aligned}\mu_k[\mathbb{P}_X]\colon X &\to \mathbb{R} \\ x &\mapsto \int_{\mathcal{X}} k_{\mathcal{X}}(x,.)\, d\,\mathbb{P}_X\end{aligned}$$

Hence, having a certain kernel $k_{\mathcal{X}}$ on $\mathcal{X}$, we can see any probability distribution (as long as it is "smooth" enough) as an element of an RKHS on $\mathcal{X}$. Note that this process determines uniquely the probability distribution $\mathbb{P}_X$ ([5, Theorem 1]). Now, for learning purposes, we only have access to a set of realizations $R_X = \{x_1, x_2, \ldots, , x_n\} \subseteq \mathcal{X}$ of our probability distribution but not the actual distribution. Thus, we don't have access to the map $\mu[\mathbb{P}_X]$. However, we can approximate it by considering the empirical mean on the realisations, i.e. by considering the following map:

$$\mu_k[R_X] = \frac{1}{n}\sum_{i=1}^{n} k_{\mathcal{X}}(x_i,.)$$

This happens to be a good approximation, as the error $\|\mu_k[\mathbb{P}_X] - \mu_k[R_X]\|$ can be bounded by the Rademacher average (a quantity that depends on the reproducing kernel as well as the probability distribution $\mathbb{P}_X$) as explained in [6, Theorem 15]. This embedding has been used for different statistical learning problems as explained below.

1.2.1 Two-sample problem

One of these applications is the two-sample test : check whether two probability distributions are identical by comparing two samples R_X and R_Y realizations of $\mathbb{P}_X$ and $\mathbb{P}_Y$. It is considered as a classical problem in multivariate statistics as the goal is to define a test statistic to either confirm the null hypothesis $\mathcal{H}_0 : \mathbb{P}_X = \mathbb{P}_Y$ or reject it. The use of kernels to address this issue has a long lasting history as shown in [7]. One approach is to use the above embedding to define a test statistic U as follows [8, Lemma 5]:

$$U(X,Y) = \left(\frac{1}{m(m-1)}\sum_{i\neq j} h((x_i,y_i),(x_j,y_j))\right)$$

with

$$h((x_i,y_i),(x_j,y_j)) = (k(x_i,x_j) + k(y_i,y_j)) - (k(x_i,y_j) + k(x_j,y_i))$$

1.2.2 Conditional distribution embedding

Another application is to estimate conditional probabilities. Let $\mathcal{Y}$ be a Polish space and Y a random variable with probability distribution $\mathbb{P}_Y$ on $\mathcal{Y}$. We consider a reproducing kernel on $\mathcal{Y}$ given by $k_{\mathcal{Y}}$. We are interested in the joint distribution of the random variable (X,Y) and especially the conditional distribution of $Y|X = x$ for a certain $x \in \mathcal{X}$. We want to define an embedding for this conditional distribution. To do so we define the following:

Definition 1.7. We call the **cross-covariance operator of X and Y**, denoted by C_{XY}, the following element in the tensor product of RKHS's $\mathcal{F} \otimes \mathcal{G}$ defined by:

$$C_{XY} := \mu_{k_\mathcal{X} \otimes k_\mathcal{Y}}[\mathbb{P}_{(X,Y)}] - \mu_{k_\mathcal{X}}[\mathbb{P}_X] \otimes \mu_{k_\mathcal{Y}}[\mathbb{P}_Y]$$

Notice that it is named as an operator since we can also see it as a map from $\mathcal{F}$ to $\mathcal{G}$. In order to embed the conditional distribution, we want to find an element $\mu_l[\mathbb{P}_{Y|x}]$ in $\mathcal{G}$ everytime that we fix a value x in X. Thus we want to define an operator $\mathcal{U}_{Y|X}$ from $\mathcal{F}$ to $\mathcal{G}$ that verifies the following properties:

$$(\forall x \in \mathcal{X}) \qquad \mu_{k_\mathcal{Y}}[\mathbb{P}_{Y|x}] = \mathbb{E}_{\mathbb{P}_{Y|x}}[k_\mathcal{Y}(Y)|x] = \mathcal{U}_{Y|X}(k_x) \tag{1.4}$$

$$(\forall g \in \mathcal{G}) \qquad \mathbb{E}_{\mathbb{P}_{Y|x}}[g(Y)|x] = \langle g, \mu_{k_\mathcal{Y}}[\mathbb{P}_{Y|x}] \rangle_\mathcal{G} \tag{1.5}$$

Theorem 1.5. *The map $C_{YX} \circ C_{XX}^{-1}$ verifies the properties* (1.4) *and* (1.5)*. Thus $\mathcal{U}_{Y|X}$ is given by this map.*

Proof. See [9, Theorem 4]. ■

From now on, we will refer to the operator $\mathcal{U}_{Y|X}$ as the traditional embedding. Note that given a map $g \in \mathcal{G}$ and $R_{(X,Y)} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ a set of realizations. As showcased in [9, Theorem 5], an estimate of $\mathbb{E}_{\mathbb{P}_{Y|x}}[g(Y)|x]$ is given by

$$\mathbb{E}_{\widehat{\mathbb{P}_{Y|x}}}[g(Y)|x] = [g(y_1) \ldots g_(y_n)]\, M\, [k_x(x_1) \ldots k_x(x_n)]^\intercal \tag{1.6}$$

with

$$M = (K_X + \lambda I)^{-1}$$

Chapter 2

Adaptive joint distribution learning

2.1 Framework

Let X, Y be two random variables taking values respectively in $\mathcal{X}, \mathcal{Y}$. Each r.v. defines a probability measure $\mathbb{P}_X, \mathbb{P}_Y$ on the target spaces. As for the traditional embedding, our goal is to learn the joint distribution $\mathbb{P}_0$ defined by (X, Y). Assume $\mathbb{P}_0 \ll \mathbb{P}_X \otimes \mathbb{P}_Y$, i.e that the true measure is absolutely continuous with respect to the product measure. By the Radon-Nikodym theorem [10, Theorem 7.3], there exists a map $g_0 \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, called the **Radon-Nikodym derivative**, such that:

$$\mathbb{P}_0(A) = \int_A g_0 \, d(\mathbb{P}_X \otimes \mathbb{P}_Y)$$

This Radon-Nikodym derivative verifies positivity and normalization constraints in the following sense

$$g_0 \geq 0 \qquad \mathbb{P}_X \otimes \mathbb{P}_Y \, a.s \tag{2.1}$$

$$\int_Y g_0(., y) \, \mathbb{P}_Y(dy) = 1 \qquad \mathbb{P}_X \, a.s \tag{2.2}$$

$$\int_X g_0(x, .) \, \mathbb{P}_X(dx) = 1 \qquad \mathbb{P}_Y \, a.s \tag{2.3}$$

Considering the linear embeddings

$$\begin{aligned} I_X \colon L^2_{\mathbb{P}_X \otimes \mathbb{P}_Y} &\to L^2_{\mathbb{P}_X} \\ g &\mapsto I_X g \colon = \int_Y g(., y) \, \mathbb{P}_Y(dy) \end{aligned}$$

$$\begin{aligned} I_Y \colon L^2_{\mathbb{P}_X \otimes \mathbb{P}_Y} &\to L^2_{\mathbb{P}_Y} \\ g &\mapsto I_Y g \colon = \int_X g(x, .) \, \mathbb{P}_X(dx) \end{aligned}$$

We can rewrite the conditions (2.2) and (2.3) as

$$I_X g_0 = 1 \ \mathbb{P}_X \, a.s \qquad and \qquad I_Y g_0 = 1 \, \mathbb{P}_Y \ a.s$$

Hence, by this assumption, learning the true distribution is reduced to learning the Radon-Nikodym derivative. We do so by embedding our probability distribution in a tensor product of Reproducing Kernel Hilbert spaces (RKHS).

2.2 Low-rank embedding

2.2.1 Objective function

We consider $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ two separable RKHS respectively on $\mathcal{X}$ and $\mathcal{Y}$ with respective reproducing kernels $k_\mathcal{X}$ and $k_\mathcal{Y}$. From now on, we denote the product space as $Z = \mathcal{X} \times \mathcal{Y}$ and the tensor product of the RKHS's $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ as $\mathcal{H}$ with reproducing kernel the tensor products of the kernels $k_\mathcal{X}$ and $k_\mathcal{Y}$ which we will denote k. Before specifying our embedding, consider the following canonical linear embeddings:

$$J_0 \colon \mathcal{H} \to L^2_{\mathbb{P}_0} \qquad\qquad J \colon \mathcal{H} \to L^2_{\mathbb{P}_X \otimes \mathbb{P}_Y}$$

These operators induce the following Hilbert-Schmidt adjoint operators [11, Chapter 6]:

$$\begin{aligned} J_0^* \colon L^2_{\mathbb{P}_0} &\to \mathcal{H} \\ f &\mapsto J_0^*(f) \end{aligned}$$

with

$$\begin{aligned} J_0^*(f) \colon Z &\to \mathbb{R} \\ z &\mapsto \langle k_z, f \rangle_{L^2_{\mathbb{P}_0}} \end{aligned}$$

Similarly, we define:

$$\begin{aligned} J^* \colon L^2_{\mathbb{P}_X \otimes \mathbb{P}_Y} &\to \mathcal{H} \\ f &\mapsto J^*(f) = \langle k_z, f \rangle_{L^2_{\mathbb{P}_X \otimes \mathbb{P}_Y}} \end{aligned}$$

To learn the Radon-Nikodym derivative g_0, we look for a map

$$g = p + Jh$$

with p being given by a prior distribution that verifies $I_X p = I_Y p = 1$ and $h \in \mathcal{H}$. From now on, w.l.o.g, we will assume that $p = 1$ almost everywhere. This map g should verify the constraints given by (2.1), (2.2) and (2.3). Our objective is to minimize the norm

$$\|g_0 - 1 - Jh\|^2_{L^2_{\mathbb{P}_X \otimes \mathbb{P}_Y}} = \|g_0 - 1\|^2_{L^2_{\mathbb{P}_X \otimes \mathbb{P}_Y}} - 2\langle J_0^* 1 - J^* 1, h \rangle_\mathcal{H} + \langle Jh, Jh \rangle_{L^2_{\mathbb{P}_X \otimes \mathbb{P}_Y}}$$

The first term doesn't depend on h. Thus after considering a ridge-regularization term, our learning objective function becomes

$$\min_{h \in \mathcal{H}} \quad 2\langle J_0^* 1 - J^* 1, h \rangle_\mathcal{H} + \langle Jh, Jh \rangle_{L^2_{\mathbb{P}_X \otimes \mathbb{P}_Y}} + \lambda \|h\|^2_\mathcal{H} \tag{2.4}$$

Rewriting this objective function, we find

$$\min_{h \in \mathcal{H}} \quad -2\langle J_0^* 1 - J^* 1, h \rangle_\mathcal{H} + \langle (J^* J + \lambda) h, h \rangle_\mathcal{H} \tag{2.5}$$

2.2.2 Empirical approximations

Consider a set of realizations $R_{(X,Y)} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ drawn from the true distribution $\mathbb{P}_0$. Our goal is to learn h from this sample by minimizing the objective function. Consider the empirical true distribution

$$\widehat{\mathbb{P}}_0 = \frac{1}{n}\sum_{i=1}^{n} \delta_{(x_i,y_i)}$$

and the marginal ones

$$\widehat{\mathbb{P}}_X = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$
$$\widehat{\mathbb{P}}_Y = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}$$

The empirical product measure is given by:

$$\widehat{\mathbb{P}_X \otimes \mathbb{P}_Y} = \frac{1}{n^2}\sum_{i,j=1}^{n} \delta_{(x_i,y_j)}$$

Notice that every element $f \in L^2_{\widehat{\mathbb{P}_0}}$ can be seen as the vector $(f(x_i, y_i))_{i=1}^n \in \mathbb{R}^n$. Similarly, every element $g \in L^2_{\widehat{\mathbb{P}_X \otimes \mathbb{P}_Y}}$ can be seen as the matrix $(g(x_i, y_j))_{i,j=1}^n$. We denote the elements (x_i, y_j) as $\widehat{Z} = \{z_1, z_2, \ldots, z_{n^2}\}$. In order to perform the computations in the objective function (2.5), we have to consider the empirical linear embedding evaluated at $(x, y) \in Z$. For instance, we have for the embedding $\widehat{J}_0^*$:

$$\begin{aligned}
\widehat{J}_0^* 1(x,y) &= \langle k_{(x,y)}, 1\rangle_{L^2_{\widehat{\mathbb{P}}_0}} \\
&= \int_{X\times Y} k((x,y),z)1(z) \quad \widehat{\mathbb{P}}_0(dz) \\
&= \frac{1}{n}\sum_{i=1}^{n} k((x,y),(x_i,y_i))
\end{aligned}$$

Performing similar computations, we find that:

$$(\widehat{J}\widehat{J}^* g)_{i,j} = \frac{1}{n^2}\sum_{s,t=1}^{n} k((x_i,y_j),(x_s,y_t))g_{s,t}$$

Thus to compute terms in (2.5), we have to build a $n^2 \times n^2$ matrix. This matrix is given by evaluating the feature maps

$$\begin{aligned}
\phi_j = k_{z_j} \colon Z &\to \mathbb{R} \qquad j = 1, 2, \ldots, n^2 \\
z &\mapsto k(z_j, z)
\end{aligned}$$

on $\widehat{Z}$. In particular, performing operations on this matrix such as computing the inverse or eignevalues is $\mathcal{O}((n^2)^3)$. Given the high complexity of this problem, we minimize our objective function on a subspace of $\mathcal{H}$ generated by the feature maps k_{z_i} (the columns of our matrix), where the z_i's form a subsample of the sample $\widehat{Z}$. To do so, we consider a subspace $V = \{\phi_{i_1}, \ldots, \phi_{i_m}\} \subseteq \mathcal{H}$. We show in the following section how to find V through a low-rank approximation of our matrix.

Low-rank approximation

Low-rank approxiamation has been a major field of study in mathematical optimization as many methods have been developed to find such matrices with a reduced rank [12]. In our case, we perform a biorthogonal Chokesky decomposition of the matrix K that gives a low-rank approximation $K \approx LL^\intercal$ with $L \in \mathbb{R}^{N\times m}$, a set of indices $i_1, i_2, \dots, i_m$ corresponding to the chosen columns, as well as a biorthogonal basis B with respect to L, i.e. we have that that $B^\intercal L = I$. This process depends on a threshold parameter ϵ that controls the precision of our approximation. The algorithm is explained in further details in [1, Algortithm 2] and is the reason this embedding is called a low-rank one.

Consider $V = \{\phi_{i_1}, \dots, \phi_{i_m}\}$ where the indices $i_1, \dots, i_m$ are given by our decomposition algorithm and consider the vectors $\boldsymbol{\phi} = [\phi_1\phi_2 \dots \phi_N]$ and $\boldsymbol{\phi_p} = [\phi_{i_1} \dots \phi_{i_m}]$. We can construct a basis $\{\psi_i^{'}\}_{i=1}^m$ which is orthonormal with respect to $\mathcal{H}$ by defining $\psi_i^{'} = \boldsymbol{\phi} b_i$ with $b_1, b_2, \dots b_m$ being the columns of B [1, Theorem 5.2]. Considering the vector $\boldsymbol{\psi}^{'} = [\psi_1^{'} \dots \psi_m^{'}]$ which is composed of the elements $\psi_i^{'}$, we can write it as $\boldsymbol{\psi}^{'} = \boldsymbol{\phi} B$. Now consider U the inverse of the upper triangular factor of the Chokesky decomposition of $K(p,p)$, we have that $L = K(:,p)U$. Moreover, we can derive U from B by the following relation $U = B(p,:)$. Thus, we have that:

$$\boldsymbol{\psi}^{'} = \boldsymbol{\phi} B = \boldsymbol{\phi_p} U$$

Moreover, by performing a spectral decomposition of the matrix $L^\intercal L$, we find that $L^\intercal L = V\Lambda V^\intercal$ where V is composed of the eigenvectors of the decomposition and Λ is the diagonal matrix of eigenvalues. By denoting $v_1, v_2, \dots v_m$ the columns of V, we can construct a basis $\{\psi_i\}_{i=1}^m$ which is orthogonal with respect to both $\mathcal{H}$ and $L^2_{\widehat{\mathbb{P}_X}\otimes\widehat{\mathbb{P}_Y}}$ by defining $\psi_i = \boldsymbol{\psi}^{'} v_i = \boldsymbol{\phi_p} U v_i$. Thus, by denoting the matrix product $UV = Q$ we have that:

$$\boldsymbol{\psi} = \boldsymbol{\phi_p} Q$$

with $\boldsymbol{\psi}$ being an orthogonal basis both in $\mathcal{H}$ and $L^2_{\widehat{\mathbb{P}_X}\otimes\widehat{\mathbb{P}_Y}}$.

Tensor product

In practice, instead of computing the matrix K, we consider the matrices $K_X = [k_{\mathcal{X}}(x_i, x_s)]_{i,s=1}^n$ and $K_Y = [k_{\mathcal{Y}}(y_j, y_t)]_{j,t=1}^n$. Notice that K is nothing but the Kroenecker prouct of the matrices K_X and K_Y where the Kroenecker produt of two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$ is defined as follows:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix},$$

Thus, instead of performing the previous actions on K, we can perform them separately on the matrices K_X and K_Y separately and take the Kroenecker product at each step. More precisely we perform the biorthogonal Chokesky decomposition on the matrices K_X and K_Y yielding respectively the matrices L_X, U_X for K_X and L_Y, U_Y for K_Y. Moreover, performing the spectral decomposition on $L_X^\intercal L_X$ and $L_Y^\intercal L_Y$ yields orthonormal bases $\boldsymbol{\psi_X}$ and $\boldsymbol{\psi_X}$ on respective subspaces V_X and V_Y with $\boldsymbol{\psi_X} = \boldsymbol{\phi_p^X} Q_X$ and $\boldsymbol{\psi_Y} = \boldsymbol{\phi_p^Y} Q_Y$. Thus taking the tensor product of V_X and

V_Y denoted $V = V_X \otimes V_Y$, we have a doubly orthogonal basis induced by K_X and K_Y given by

$$\boldsymbol{\psi} = \boldsymbol{\psi_X} \otimes \boldsymbol{\psi_Y} = (\boldsymbol{\phi_p^X} \otimes \boldsymbol{\phi_p^Y})(Q_X \otimes Q_Y) = \boldsymbol{\phi_p} Q$$

Thus the cost of building this biorthogonal basis is $\mathcal{O}((m_X^2 + m_Y^2)n)$.

2.2.3 Optimization problem

Hence, thanks to the biorthogonal basis, any element $h \in V$ can be written as $h = \boldsymbol{\psi} h' = \boldsymbol{\phi_p} Q h'$ with $h' \in \mathbb{R}^m$ and we have that $h(z) = \boldsymbol{\phi_p}(z) Q h'$. Now, considering our objective function, we have that:

$$\begin{aligned}
-2\langle \widehat{J}_0^* 1 - \widehat{J}^* 1, h\rangle_{\mathcal{H}} &= 2\left(\langle \widehat{J}^* 1, \boldsymbol{\psi}\rangle_{\mathcal{H}} - \langle \widehat{J}_0^* 1, \boldsymbol{\psi}\rangle_{\mathcal{H}}\right) h' \\
&= 2\left(\int_Z \boldsymbol{\phi_p}(z)\, \widehat{\mathbb{P}_X \otimes \mathbb{P}_Y}(dz) - \int_Z \boldsymbol{\phi_p}(z)\, \widehat{\mathbb{P}}_0(dz)\right) Q h' \\
&= 2\,\beta\, Q h' \\
&= 2\alpha h'
\end{aligned}$$

The second integral is computed as follows:

$$\begin{aligned}
\int_Z \boldsymbol{\phi_p}(z)\, \widehat{\mathbb{P}}_0(dz) &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\phi_p}((x_i, y_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n} [k(z_{i_1}, (x_i, y_i)),\ k(z_{i_2}, (x_i, y_i)), \ldots,\ k(z_{i_m}, (x_i, y_i))]
\end{aligned}$$

Similarly, the first integral is computed as follows:

$$\begin{aligned}
\int_Z \boldsymbol{\phi_p}(z)\, \widehat{\mathbb{P}_X \otimes \mathbb{P}_Y}(dz) &= \frac{1}{n^2}\sum_{i,j=1}^{n} \boldsymbol{\phi_p}((x_i, y_j)) \\
&= \frac{1}{n^2}\sum_{i,j=1}^{n} [k(z_{i_1}, (x_i, y_j)),\ k(z_{i_2}, (x_i, y_j)), \ldots,\ k(z_{i_m}, (x_i, y_j))]
\end{aligned}$$

Moreover, since $\boldsymbol{\psi}$ is an orthogonal basis of V with respect to the $\mathcal{H}$-inner product, we have that :

$$\begin{aligned}
\langle Jh, Jh\rangle_{\mathcal{H}} &= \langle \psi h', \psi h'\rangle_{\mathcal{H}} \\
&= h'^{\intercal} I h' \\
&= h'^{\intercal} h'
\end{aligned}$$

And since $\boldsymbol{\psi}$ is also an orthogonal basis of V with respect to the $L^2_{\widehat{\mathbb{P}}_X \otimes \widehat{\mathbb{P}}_Y}$ inner product, we have that :

$$\begin{aligned}
\langle Jh, Jh\rangle_{L^2_{\widehat{\mathbb{P}}_X \otimes \widehat{\mathbb{P}}_Y}} &= \langle \psi h', \psi h'\rangle_{L^2_{\widehat{\mathbb{P}}_X \otimes \widehat{\mathbb{P}}_Y}} \\
&= h'^{\intercal} \Lambda h'
\end{aligned}$$

Thus we can rewrite the objective function as

$$\min_{h' \in \mathbb{R}^m} 2\alpha h' + h'^{\intercal}(\Lambda + \lambda I) h'$$

Notice that the cost of one evaluation of the objective function is $\mathcal{O}(Nm^3)$. Having explicitly written out the terms of our objective function, we now turn to reconsider our constraints in the light of our subspace V.

Normalization constraints

Rewriting the normalization constraints in terms of h, we have that

$$I_X(Jh) = 0 \ \mathbb{P}_Y \, a.s \quad and \quad I_Y(Jh) = 0 \ \mathbb{P}_X \, a.s \tag{2.6}$$

Recall that V is a tensor product given by : $V = V_X \otimes V_Y$. We consider $\boldsymbol{\psi_X} = [\psi_{X,1}, \psi_{X,2}, \ldots, \psi_{X,m_X}]$ and $\boldsymbol{\psi_Y} = [\psi_{Y,1}, \psi_{Y,2}, \ldots, \psi_{Y,m_Y}]$ to be orthonormal bases respectively for V_X and V_Y.
Then $\psi = \psi_X \otimes \psi_Y = [\psi_{X,1} \otimes \psi_{Y,1}, \psi_{X,1} \otimes \psi_{Y,2}, \ldots, \psi_{X,1} \otimes \psi_{Y,m_Y}, \psi_{X,2} \otimes \psi_{Y,1}, \ldots]$ represents an orthonormal basis for V. Thus, for any $h \in V$, we have that

$$\begin{aligned} h &= \sum_{i=1}^{m_X} \sum_{j=1}^{m_Y} h_{j,i} (\psi_{X,i} \otimes \psi_{Y,j}) \\ &= \psi_Y H \psi_X^{\intercal} \\ &= (\psi_X \otimes \psi_Y) \, \mathrm{vec}(H) \\ &= \psi_m h^{'} \end{aligned}$$

with $\mathrm{vec}(H) = h^{'}$ being the vector composed of the concatenation of the columns of H. Thus the constraints written in (2.6) can be written as

$$\begin{aligned} I_X(g) = 1 \quad \mathbb{P}_X \, a.s &\iff I_X(Jh) = 0 \quad \mathbb{P}_X \, a.s \\ &\iff \sum_{i=1}^{m_X} \sum_{j=1}^{m_Y} h_{j,i} I_X(\psi_{X,i} \otimes \psi_{Y,j}) = 0 \, \mathbb{P}_X \, a.s \end{aligned}$$

Moreover, we have that

$$\begin{aligned} I_X(\psi_{X,i} \otimes \psi_{Y,j}) \colon X &\to \mathbb{R} \\ x &\mapsto \int_Y (\psi_{X,i} \otimes \psi_{Y,j})(x, y) \ \mathbb{P}_Y(dy) \end{aligned}$$

And since

$$\begin{aligned} \int_Y (\psi_{X,i} \otimes \psi_{Y,j})(x, y) \ \mathbb{P}_Y(dy) &= \int_Y \psi_{X,i}(x) \psi_{Y,j}(y) \ \mathbb{P}_Y(dy) \\ &= \psi_{X,i}(x) \int_Y \psi_{Y,j}(y) \ \mathbb{P}_Y(dy) \\ &= \psi_{X,i}(x) \, \mathbb{E}_{\mathbb{P}_Y}[\psi_{Y,j}] \end{aligned}$$

By an abuse of notation, we denote $I_X \psi_{Y,j}$ for $\mathbb{E}_{\mathbb{P}_Y}[\psi_{Y,j}]$. Thus we have that

$$I_X(\psi_{X,i} \otimes \psi_{Y,j}) = (I_X \psi_{Y,j}) \psi_{X,i}$$

Hence, we have that:

$$\begin{aligned} I_X(g) = 1 \quad \mathbb{P}_X\, a.s &\iff \sum_{i=1}^{m_X}\sum_{j=1}^{m_Y} h_{j,i}(I_X\psi_{Y,j})\psi_{X,i} = 0 \quad \mathbb{P}_X\, a.s \\ &\iff (I_X\psi_Y)H\psi_X^\intercal = 0 \quad \mathbb{P}_X\, a.s \\ &\iff (I_X\psi_Y)H = 0 \end{aligned}$$

where

$$I_X\psi_Y = (I_X\psi_{Y,1},\ I_X\psi_{Y,2}, \ldots,\ I_X\psi_{Y,m_Y})$$

Consider $\mathbb{1}_s = (1 \ldots 1)$ the vector in $\mathbb{R}^s$ with entries being 1. Then we have that:

$$\begin{aligned} I_X\psi_Y H &\iff (\mathbb{1}_{m_X} \otimes I_X\psi_Y)\,\mathrm{vec}\, H = 0 \\ &\iff \Gamma_X h^{'} = 0 \end{aligned}$$

with $\Gamma_X = \mathbb{1}_{m_X} \otimes I_X\psi_Y$. Similarly, by defining $I_Y\psi_{X,i} = \mathbb{E}_{\mathbb{P}_X}[\psi_{X,i}]$, we have that:

$$\begin{aligned} I_Y(g) = 1 \quad \mathbb{P}_Y\, a.s &\iff H(I_Y\psi_X)^\intercal = 0 \\ &\iff (I_Y\psi_X \otimes \mathbb{1}_{m_Y})\,\mathrm{vec}\, H = 0 \\ &\iff \Gamma_Y h^{'} = 0 \end{aligned}$$

with $\Gamma_Y = I_Y\psi_X \otimes \mathbb{1}_{m_Y}$. Thus normalization constraints become

$$\Gamma_X h^{'} = \Gamma_Y h^{'} = 0 \tag{2.7}$$

Positivity constraint

We now move on to the positivity constraint given in (2.1). We make the restriction to only consider bounded kernels (Gaussian kernel, Laplace RBF kernel, ...) and thus we make the assumption that for any $i \in \{1, 2 \ldots, m\}$, we have that $a_i \leq \psi_i \leq b_i$. Consider the cube $C := \times_{i=1}^{m}[a_i, b_i]$. We are interested in the following set

$$P = \{h^{'} \in \mathbb{R}^m \,/\, \min_{x \in C} x^\intercal h^{'} + 1 \geq 0\}$$

It is clear that we have the following characterization for the set P:

$$h^{'} \in P \iff a^\intercal h_+^{'} - b^\intercal h_-^{'} + 1 \geq 0 \tag{2.8}$$

With these constraints, our optimization problem becomes:

$$\begin{aligned} &\min_{h,h_+,h_- \in \mathbb{R}^m} 2\alpha h + h^\intercal P h \\ &\text{such that } \Gamma_X h = \Gamma_Y h = 0 \\ &\qquad\qquad h = h_+ - h_- \\ &\qquad\qquad a^\intercal h_+ - b^t h_- + 1 \geq 0 \end{aligned} \tag{2.9}$$

with h_+ and h_- respectively the positive and negative parts of h.

Detour into quadratic programming

The problem above is a special case of a quadratic program (QP). Quadratic programming is a field of study where we are interested into the optimization of quadratic functions subject to certain constraints [13, Chapter 2]. The typical quadratic program is written as

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}h^\intercal Ph + q^T h^\intercal \\ \text{subject to} & Gh \leq x \\ & Ah = b \end{array}$$

The first constraint is an inequality constraint, while the second is an equality one. Rewriting the condidtions in (2.9), and considering $\Gamma = (\Gamma_X, \Gamma_Y)$, we have that $\Gamma h = 0$. This represents the equality constraint. However to write the other constraints in the QP framework, we have to delve into conic quadratic optimization. Conic optimization is similar to linear optimization (the objective function is a linear function) with the additional fact that we can impose cone constraints on a subset of the problem variables. More precisely, we can impose that a subset of the problem variables are either in a quadratic cone

$$\mathcal{Q}^n = \left\{ h \in \mathbb{R}^n : h_0 \geq \sqrt{\sum_{j=1}^{n-1} h_j^2} \right\}.$$

or a rotated quadratic cone

$$\mathcal{Q}_r^n = \left\{ h \in \mathbb{R}^n : 2h_0 h_1 \geq \sum_{j=2}^{n-1} h_j^2, \quad h_0 \geq 0, \quad h_1 \geq 0 \right\}.$$

Consider the vector p which represents the diagonal of the matrix P and p_{sqrt} the vector where we take the square root of every element in p (recall that P is a diagonal matrix). Thus we can rewrite our problem as:

$$\begin{aligned} \min_{h,h_+,h_- \in \mathbb{R}_+^m, u \in \mathbb{R}_+} \quad & 2\alpha h + u \\ \text{such that} \quad & \Gamma_X h = \Gamma_Y h = 0 \\ & h = h_+ - h_- \\ & a^\intercal h_+ - b^t h_- + 1 \geq 0 \\ & (0.5, u, h^\intercal p) \in \mathcal{Q}_r^3 \end{aligned} \tag{2.10}$$

2.3 Numerical results

To test the low-rank embedding, we compare it to the traditional one. We draw a sample $S = \{z_1, z_2, \dots z_n\}$ from a joint normal distribution with mean $\mu = [0, 0]$ and covariance matrix $\Sigma = I_2$. Thus, we have two random variables $X, Y \sim \mathcal{N}(\mu, \Sigma)$. We also consider that the correlation between these two random variables is $\rho = 0.5$. Our test is to compute conditional probabilities of Y given X known. We consider Gaussian kernels k_X and k_Y with parameters σ_X, σ_Y in the set $T = \{0.001, 0.01, 0.1\}$. The threshold parameter ϵ for the biorthogonal Chokesky decomposition is also taken in the set T.

First thing we do is to split the data set into three disjoint sets : a training set, denoted S_{tr} on which we will train our data, a validation set S_{val} on which we will tune our hyperparameters $\sigma_X, \sigma_Y, \lambda, \epsilon$ and a testing set S_{test} on which we test our models.

Training

We train both our models on the following data sample $S_{tr} = \{z_1, \dots z_{n_{tr}}\}$. We set the regularization parameter λ to be equal to 0 for the low-rank embedding, and we solve the problem (2.10) using the MOSEK optimization package [14].

Validation

The validation set is used to tune the hyperparameters for both our models ($\sigma_X, \sigma_Y, \epsilon$ for the low-rank embedding and σ_X, λ for the traditional one). To do so, we consider the following loss function:

$$\sum_{i=1}^{n_{val}} \|t(y_i) - \mathbb{E}^{trad,low-rank}[t(Y)|X = x_i]\|_2^2 \tag{2.11}$$

where t is defined as follows

$$\begin{aligned} t\colon Y &\to \mathbb{R}^3 \\ y &\mapsto (\mathbb{1}_{Y\leq -0.3}, \mathbb{1}_{Y\leq -0.2}, \mathbb{1}_{Y\leq -0.1}) \end{aligned}$$

Recall that for the traditional embedding, as seen in (1.6), this expectation is given by:

$$\mathbb{E}^{trad}[t(Y)|X = x_i] = [t(y_1^{train}) \dots t(y_{n_{train}}^{train})]\ M\ [k_X(x_1^{train}, x_i) \dots k_X(x_{n_{train}}^{train}, x_i)]^\intercal$$

On the other hand for the lank-rank embedding approach, after training our model and finding h, it is easy to recover the Radon-Nykodym derivative g. Indeed, it is given by

$$g(z) = 1 + h(z) = 1 + \boldsymbol{\psi}(z)h = 1 + \boldsymbol{\phi_p}(z)Qh$$

Thus, we have for the low-rank embedding:

$$\begin{aligned} \mathbb{E}^{low-rank}[\mathbb{1}_{Y\leq a}|X = x_i] &= \widehat{\mathbb{P}}[Y \leq a|X = x_i] \\ &= \int_{-\infty}^{a} \widehat{\mathbb{P}}_{Y|X}(x_i, dy) \\ &= \int_{-\infty}^{a} g(x_i, y)\, \widehat{\mathbb{P}}_Y(dy) \\ &= \int_{-\infty}^{a} (1 + \boldsymbol{\psi_p}(x_i, y)Qh)\, \widehat{\mathbb{P}}_Y(dy) \end{aligned}$$

We choose the hyperparameters by grid search, i.e. for every combination of hyperparameters we train our model, compute the validation loss function (2.11), and choose the set of hyperparameters which minimize (2.11).

Testing

We test our models by comparing the true expected distribution and the ones computed thanks to the models. More specifically, we compute the following loss function:

$$\sum_{i=1}^{n_{test}} \|\mathbb{E}[t(Y)|X = x_i] - \mathbb{E}^{trad,low-rank}[t(Y)|X = x_i]\|_2^2 \tag{2.12}$$

where the true distribution $Y|X = x_i$ is given by $\mathcal{N}(\rho x_i, (1-\rho^2))$. Indeed, we have that the probability density function of the conditional distribution is given by:

$$f_{Y|X=x_i}(y) = \frac{f_{XY}(x_i, y)}{f_X(x_i)}.$$

Now, we have that the p.d.f. of the joint normal distribution with correleation factor ρ is given by:

$$f_{XY}(x_i, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x_i^2 + y^2 - 2\rho x_i y)\right\}$$

Thus, we have that:

$$\begin{aligned} f_{Y|X=x}(y) &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x_i^2 + y^2 - 2\rho x_i y)\right\} \sqrt{2\pi} \exp\left\{\frac{x_i^2}{2}\right\} \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x_i^2 + y^2 - 2\rho x_i y - x_i^2(1-\rho^2))\right\} \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}(y - \rho x_i)^2\right\} \\ &= f_{\mathcal{N}(\rho x_i, (1-\rho^2))}(y) \end{aligned}$$

We compute our loss function for different numbers of datapoints n. For each number n, we perform 20 runs, and take the average over all the runs.

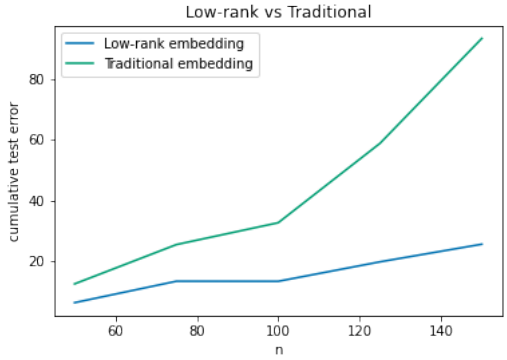

Figure 2.1: Cumulative test error computed over 20 runs for both models on datasets generated by a bivariate joint distribution

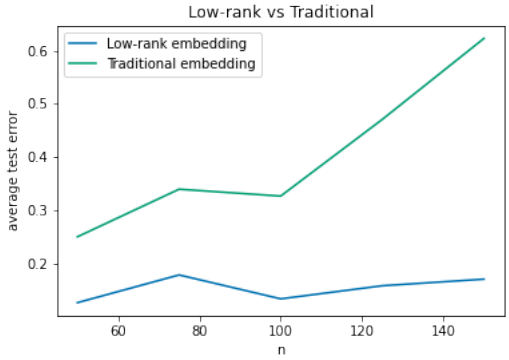

Figure 2.2: Average test error computed over 20 runs for the three models on datasets generated by a bivariate joint distribution

Figures (2.1), (2.2) yields the empirical results for a bivariate gaussian distribution for different number of datapoints n. They show that the low-rank embedding defined in this section performs well better than the traditional one.

Chapter 3

Greedy joint distribution learning

Greedy algorithms have been widely used for different applications in computational mathematics [15][16]. The most prominent ones being Dijkstra's algorithm and Kruskal's algorithm used in graph theory to find minimum paths in a graph. The main idea is to choose the best optimal choice at every stage rather than try to solve the global problem. The perk of this approach is that it generally reduces the computational cost by trying to find optimal solutions locally in a sequential manner.

Let $\mathcal{X}$ be a finite discrete set. W.l.o.g., we consider that $\mathcal{X} = \{1, 2, \dots, D\}$ and consider the RKHS $\mathcal{H}_\mathcal{X}$ with reproducing kernel $k_\mathcal{X}$ as given in (1.3). Let $\mathcal{Y}$ be a set and consider a separable RKHS $\mathcal{H}_\mathcal{Y}$ on $\mathcal{Y}$ with reproducing kernel $k_\mathcal{Y}$. Considering a greedy approach, we try to take advantage of the fact that one of the support sets is finitely discrete. To do so, we show that our problem is equivalent to finding D functions in $\mathcal{H}_\mathcal{Y}$. After that, we solve our problem in a local fashion by finding the best optimum local solution for each h_i.

3.1 Greedy embedding

3.1.1 Objective function

We know from Example (1.6)that for any $h \in \mathcal{H} = \mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}$, we have that:

$$h = \sum_{i=1}^{D} e_i \otimes h_i$$

with $h_i \in \mathcal{H}_\mathcal{Y}$. Now considering the objective function given in (2.4), we can rewrite it as

$$\min_{h \in \mathcal{H}} \quad -2\langle J_0^* 1 - J^* 1, \sum_{i=1}^{D} e_i \otimes h_i \rangle_\mathcal{H} + \langle Jh, Jh \rangle_{L^2_{\mathbb{P}_X \otimes \mathbb{P}_Y}} + \lambda \langle h, h \rangle_\mathcal{H}$$

Moreover, we have that:

$$-2\langle J_0^* 1 - J^* 1, \sum_{i=1}^{D} e_i \otimes h_i \rangle_\mathcal{H} = \sum_{i=1}^{D} -2\langle J_0^* 1 - J^* 1, e_i \otimes h_i \rangle_\mathcal{H}$$

and

$$\langle Jh, Jh \rangle_{L^2_{\mathbb{P}_X \otimes \mathbb{P}_Y}} = \int_{X \times Y} (h(j,s))^2 \, \mathbb{P}_X \times \mathbb{P}_Y(d(j,s)) \tag{3.1}$$

Since

$$\begin{aligned}(h(j,s))^2 &= (\sum_{i=1}^{D} e_i(j)h_i(s))^2 \\ &= h_j(s)^2 \\ &= \sum_{i=1}^{D} e_i(j)(h_i(s))^2\end{aligned}$$

We can rewrite (3.1) as:

$$\begin{aligned}\langle Jh, Jh\rangle_{L^2_{\mathbb{P}_X\otimes\mathbb{P}_Y}} &= \sum_{i=1}^{D}\int_{X\times Y} e_i(j)(h_i(s))^2\,\mathbb{P}_X\times\mathbb{P}_Y(d(j,s)) \\ &= \sum_{i=1}^{D}\int_X e_i(j)\,\mathbb{P}_X(dj)\int_Y h_i(s)^2\,\mathbb{P}_Y(ds) \\ &= \sum_{i=1}^{D}\mathbb{P}_X(i)\langle J_Y h_i, J_Y h_i\rangle_{L^2_{\mathbb{P}_Y}} \\ &= \sum_{i=1}^{D}\mathbb{P}_X(i)\langle J_Y^* J_Y h_i, h_i\rangle_{\mathcal{H}_Y}\end{aligned}$$

Finally, for the last term of the objective function, we have that:

$$\begin{aligned}\langle Jh, Jh\rangle_{\mathcal{H}} &= \langle\sum_{i=1}^{D} e_i\otimes h_i, \sum_{j=1}^{D} e_j\otimes h_j\rangle_{\mathcal{H}} \\ &= \sum_{i,j=1}^{D}\langle e_i\otimes h_i, e_j\otimes h_j\rangle_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\ &= \sum_{i,j=1}^{D}\langle e_i, e_j\rangle_{\mathcal{H}_X}\langle h_i, h_j\rangle_{\mathcal{H}_Y} \\ &= \sum_{i,j=1}^{D}\delta_{ij}\langle h_i, h_j\rangle_{\mathcal{H}_Y} \\ &= \sum_{i=1}^{D}\langle h_i, h_i\rangle_{\mathcal{H}_Y}\end{aligned}$$

Thus, in our setting, the objective function in (2.4) becomes:

$$\min_{h_1,h_2,\ldots,h_D\in\mathcal{H}_Y}\sum_{i=1}^{D} -2\langle J_0^*1 - J^*p, e_i\otimes h_i\rangle_{\mathcal{H}} + \langle(\mathbb{P}_X(i)J_Y^*J_Y + \lambda)h_i, h_i\rangle_{\mathcal{H}_Y} \tag{3.2}$$

Rewriting the constraints (2.6) in this setting, we have:

$$I_Y(Jh) = 0\ \mathbb{P}_X\ a.s$$

Hence, we have that:

$$(\forall\, j\in\mathcal{X})\quad \int_{\mathcal{Y}} h(j,y)\,\mathbb{P}_Y(dy) = 0$$

And since we know by (1.3) that $h(j, y) = h_j(y)$, we have that:

$$(\forall\, j \in \mathcal{X}) \quad \int_{\mathcal{Y}} h_j(y)\, \mathbb{P}_Y(dy) = 0 \tag{3.3}$$

We also have that

$$I_X(g) = 0 \ \mathbb{P}_Y \ a.s$$

Thus, we have $\mathbb{P}_Y$ a.s that :

$$\int_{\mathcal{X}} g(x, y)\, \mathbb{P}_X(dx) = 1$$

We can rewrite it as

$$\sum_{i=1}^{D} (1 + h(x_i, y))\, \mathbb{P}_X(i) = 1\, \mathbb{P}_Y \ a.s$$

Hence, we have that:

$$\sum_{i=1}^{D} (1 + h_i(y))\, \mathbb{P}_X(i) = 1\, \mathbb{P}_Y \ a.s \tag{3.4}$$

Finally, for the positivity constraint:

$$(1 + h) \geq 0 \qquad \mathbb{P}_X \otimes \mathbb{P}_Y \ a.s$$

we have that:

$$(\forall j \in \mathcal{X}) \quad (1 + h_j(y)) \geq 0 \qquad \mathbb{P}_Y \ a.s \tag{3.5}$$

3.1.2 Empirical approximations

Consider a set of realizations $R_{(X,Y)} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ drawn from the true distribution $\mathbb{P}_0$. Our goal, again, is to learn h, or more specifically the h_i's which minimizes the objective function. Now, the difference with the low-rank technique is that h is "composed" of D functions in $\mathcal{H}_{\mathcal{Y}}$. Taking advantage of this fact, we solve the problem locally, i.e. by minimizing the objective function (and thus finding the h_i) sequentially. Thus, we optimize the problem locally for each i, which yields the greedy nature of our approach. Again, we consider the following true empirical distribution:

$$\widehat{\mathbb{P}}_0 = \frac{1}{n} \sum_{i=1}^{n} \delta_{(x_i, y_i)}$$

and the marginal ones:

$$\widehat{\mathbb{P}}_Y = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$$

$$\widehat{\mathbb{P}}_X = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

Notice that in the last sum, elements might be repeated with a strictly positive probability. Suppose we have solved the problem for $h_1, h_2, \ldots, h_{i-1}$, and we now solve the problem for h_i. At this step, the problem is formulated as follows:

$$\min_{h_i \in \mathcal{H}_Y} -2\langle \widehat{J}_0^* 1 - \widehat{J}^* 1, e_i \otimes h_i \rangle_{\mathcal{H}} + \langle (\widehat{\mathbb{P}}_X(i) \widehat{J}_Y^* \widehat{J}_Y + \lambda) h_i, h_i \rangle_{\mathcal{H}_Y} \tag{3.6}$$

As earlier, to perform the computations in this objective function, we will have to compute the kernel matrix K_Y given by

$$K_Y = [k_{\mathcal{Y}}(y_s, s_t)]_{s,t=1}^n$$

Notice that the matrix K_X is given by I_D by the nature of our kernel choice $k_{\mathcal{X}}$. Same as earlier, we perform a biorthogonal Chokesy decompostion on K_Y followed by a spectral decomposition of $L_Y^\intercal L_Y$ to find a biorthogonal basis given by $\boldsymbol{\psi} = \boldsymbol{\phi_p} Q$ and the matrix of eigenvalues Λ_Y. Again, we consider a subspace $V_Y = \text{span}\{\psi_1, \psi_2, \dots \psi_m\}$ of $\mathcal{H}_{\mathcal{Y}}$. Hence, every element $h \in V_Y$ can be written as $h = \boldsymbol{\psi} h^{'} = \boldsymbol{\phi_p} Q h^{'}$. Notice that we only have to perform the biorthogonal Chokesky decomposition of the matrix K_Y once, where we had to do it for both matrices K_X and K_Y for the low-rank approach. Thus, the computational cost of building the biorthogonal basis is $\mathcal{O}(m_Y^2 n)$. Hence, any element $h_i \in V_Y$ can be written as $h_i = \boldsymbol{\psi} h_i^{'} = \boldsymbol{\phi_p} Q h_i^{'}$ with $h^{'} \in \mathbb{R}^{m_Y}$.

With this framework in place, the first term of our objective function becomes

$$\begin{aligned} -2\langle \widehat{J}_0^* 1 - \widehat{J}^* 1, e_i \otimes h_i \rangle_{\mathcal{H}} &= 2\left(\langle \widehat{J}^* 1, e_i \otimes h_i \rangle_{\mathcal{H}} - \langle \widehat{J}_0^* 1, e_i \otimes h_i \rangle_{\mathcal{H}} \right) \\ &= 2\left(\int_Z e_i \otimes h_i \, \widehat{\mathbb{P}_X \otimes \mathbb{P}_Y}(dz) - \int_Z e_i \otimes h_i \, \widehat{\mathbb{P}}_0(dz) \right) \end{aligned}$$

where the second integral can be written as:

$$\begin{aligned} \int_Z e_i \otimes h_i \, \widehat{\mathbb{P}}_0(dz) &= \frac{1}{n} \sum_{j=1}^n (e_i \otimes h_i)(x_j, y_j) \\ &= \frac{1}{n} \sum_{j=1}^n e_i(x_j) h_i(y_j) \\ &= \frac{1}{n} \sum_{j=1}^{n_i} h_i(y_{i_j}) \\ &= \frac{1}{n} \sum_{j=1}^{n_i} \boldsymbol{\phi_p}(y_{i_j}) Q h_i^{'} \end{aligned}$$

with n_i being the number of occurences of i in our data sample and y_{i_j} the y's that correspond to those occurences. Similarly the first integral can be written as :

$$\begin{aligned} \int_Z e_i \otimes h_i \, \widehat{\mathbb{P}_X \otimes \mathbb{P}_Y}(dz) &= \frac{1}{n^2} \sum_{s,j=1}^n (e_i \otimes h_i)(x_s, y_j) \\ &= \frac{1}{n^2} \sum_{s,j=1}^n e_i(x_s) h_i(y_j) \\ &= \frac{1}{n^2} \sum_{j=1}^n h_i(y_j) \\ &= \frac{1}{n^2} \sum_{j=1}^n \boldsymbol{\phi_p}(y_j) Q h_i^{'} \end{aligned}$$

Thus, we can write the first term as

$$\begin{aligned}-2\langle \widehat{J}_0^* 1 - \widehat{J}^* 1, e_i \otimes h_i \rangle_{\mathcal{H}} &= 2\left(\frac{1}{n^2}\sum_{j=1}^{n} \boldsymbol{\phi_p}(y_j) - \frac{1}{n}\sum_{j=1}^{n_i} \boldsymbol{\phi_p}(y_{i_j})\right) Q h_i^{'} \\ &= 2\beta_i Q h_i^{'} \\ &= 2\alpha_i h_i^{'}\end{aligned}$$

while we can write the second term as

$$\langle (\widehat{\mathbb{P}}_X(i)\widehat{J}_Y^*\widehat{J}_Y + \lambda)h_i, h_i\rangle_{\mathcal{H}_Y} = h_i^{'}(\widehat{\mathbb{P}}_X(i)\Lambda_Y + \lambda I)h_i^{'}$$

Hence, the objective function becomes:

$$\min_{h_i^{'} \in \mathbb{R}^{m_Y}} 2\alpha_i h_i{}' + h_i^{'}(\widehat{\mathbb{P}}_X(i)\Lambda_Y + \lambda I)h_i^{'}$$

Notice that the cost of one evaluation of the objective function is $\mathcal{O}(nm_Y^3)$. We now move on to reconsider our constraints.

Positivity constraint

Same as for the low-rank approach, we consider bounded kernels, and thus we make the assumption that for any $j \in \{1, 2, \ldots, m_Y\}$, we have that $a_j \leq \psi_j \leq b_j$. Consider the cube $C := \times_{i=1}^{m_Y}[a_i, b_i]$. Then, under these assumptions, our constraint can be written as:

$$a^\intercal h_{i_+}^{'} - b^\intercal h_{i_-}^{'} + 1 \geq 0 \tag{3.7}$$

Normalization constraints

We have by (3.3) that :

$$I_X h_i = (I_X \boldsymbol{\psi}) h_i^{'} = 0 \tag{3.8}$$

This represents our first constraint. For the second one, since we optimize the problem locally, we have that:

$$\sum_{i'=1}^{i-1}(1 + h_{i'}(y))\,\mathbb{P}_X(i^{'}) + (1 + h_i(y))\,\mathbb{P}_X(i) \leq 1 \qquad \mathbb{P}_Y\, a.s$$

The idea behind this inequality is that having determined $h_1, h_2, \ldots, h_{i-1}$, we can put as much weight as we want on h_i as long as this sum doesn't exceed 1. So we are choosing the optimal h_i for the local problem (3.6) rather than choosing the optimal h_i for the global problem (3.2), hence the greedy nature of our approach. This idea is similar to the one produced in [17, Section 3.2] .Thus, we have that:

$$\sum_{i'=1}^{i-1} h_{i'}(y)\,\mathbb{P}_X(i^{'}) + h_i(y)\,\mathbb{P}_X(i) \leq 1 - \sum_{i'=1}^{i}\mathbb{P}_X(i^{'}) \qquad \mathbb{P}_Y\, a.s$$

Hence, we have that:

$$\sum_{i'=1}^{i-1} \boldsymbol{\psi}(y) h_{i'}^{'}\,\mathbb{P}_X(i^{'}) + \boldsymbol{\psi}(y) h_i^{'}\,\mathbb{P}_X(i) \leq 1 - \sum_{i'=1}^{i}\mathbb{P}_X(i^{'}) \qquad \mathbb{P}_Y\, a.s$$

We can rewrite it as:

$$\boldsymbol{\psi}(y)\left(\sum_{i'=1}^{i-1} h_{i'}^{'}\,\mathbb{P}_X(i^{'}) + h_i^{'}\,\mathbb{P}_X(i)\right) \leq 1 - \sum_{i'=1}^{i}\mathbb{P}_X(i^{'}) \qquad \mathbb{P}_Y\, a.s$$

Thus, we have the following condition:

$$\boldsymbol{\psi}(y)\left(\frac{1}{\mathbb{P}_X(i)}\sum_{i'=1}^{i-1}\mathbb{P}_X(i^{'})h_{i'}^{'} + h_i^{'}\right) \leq \frac{1}{\mathbb{P}_X(i)}\left(1 - \sum_{i'=1}^{i}\mathbb{P}_X(i^{'})\right) \tag{3.9}$$

We are interested in the following set:

$$P_i = \left\{ h_i^{'} \in \mathbb{R}^{m_Y} \,/\, \max_{x\in C} x^{\intercal}\left(\frac{1}{\mathbb{P}_X(i)}\sum_{i'=1}^{i-1}\mathbb{P}_X(i^{'})h_{i'}^{'} + h_i^{'}\right) \leq \frac{1}{\mathbb{P}_X(i)}\left(1 - \sum_{i'=1}^{i}\mathbb{P}_X(i^{'})\right)\right)\right\}$$

We define the vector $v_i \in \mathbb{R}^m$

$$v_i = \frac{1}{\mathbb{P}_X(i)}\sum_{i'=1}^{i-1}\mathbb{P}_X(i^{'})h_{i'}^{'}$$

as well as the set V_i

$$V_i = \{x \in \mathbb{R}^{m_Y} \mid \quad (\forall j = 1, 2, \ldots, m) \qquad x_j \geq -v_i(j)\}$$

Same as for the positivity constraint, the inequality (3.9) is verified if and only if:

$$b^{\intercal}\left(v_i + h_i^u\right) - a^{\intercal}\left(v_i + h_i^l\right) \leq \frac{1}{\mathbb{P}_X(i)}\left(1 - \sum_{i'=1}^{i}\mathbb{P}_X(i^{'})\right)$$

where the j-th coordinate of h_i^u is given by:

$$h_i^u(j) = \begin{cases} h_i^{'}(j) & \text{if} \qquad h_i^{'}(j) \geq -v_i(j) \\ 0 & \text{otherwise} \end{cases}$$

and the j-th coordinate of h_i^l is given by:

$$h_i^l(j) = \begin{cases} 0 & \text{if} \qquad h_i^{'}(j) \geq -v_i(j) \\ -h_i^{'}(j) & \text{otherwise} \end{cases}$$

Hence, we do have that $h_i^{'} = h_i^u - h_i^l$ with $h_i^u, h_i^l \in V_i$. This yield our second normalization constraint which is induced by the greedy approach. Thus the optimization problem to solve to find h_i is given by:

$$\begin{aligned} \min_{h_i, h_i^+, h_i^- \in \mathbb{R}_+^{m_Y}, h_i^u, h_i^l \in V_i, u \in \mathbb{R}_+} \quad & 2\alpha_i h_i + u \\ \text{such that } & (I_X\boldsymbol{\psi})h_i^{'} = 0 \\ & h_i = h_i^+ - h_i^- \\ & a^{\intercal}h_+ - b^{\intercal}h_- + 1 \geq 0 \\ & h_i = h_i^u - h_i^l \\ & b^{\intercal}\left(v_i + h_i^u\right) - a^{\intercal}\left(v_i + h_i^l\right) \leq \frac{1}{\mathbb{P}_X(i)}\left(1 - \sum_{i'=1}^{i}\mathbb{P}_X(i^{'})\right) \\ & (0.5, u, h_i^{\intercal}p) \in \mathcal{Q}_r^3 \end{aligned} \tag{3.10}$$

3.2 Numerical results

We test our greedy embedding against both models previously considered. We consider two frameworks for the joint distribtuions. We first draw a sample from a joint distribution with first marginal $X_1 \sim B(2, 0.5)$ a binomial distribution with parameters 2 and 0.5, and second marginal $Y_1 \sim N(0, 1)$ a normal distribution. For the second framework, we consider a joint distribution with first marginal $X_2 \sim Pois(3)$ a Poisson distribution with rate 3, and second marginal $Y_2 \sim LogNormal(0, 1)$ a lognormal distribution. The Poisson distribution has many applications for financial data, indeed it is commonly used for count variables. On the other hand, the lognormal distribution is one of the most commonly used models for stock prices. For each case, we assume that both marginals are independent. Again our first test will be to compute conditional probabilities, given known values of X. We will consider our hyperparameters in the set T, and split our data into training, validation and testing sets. We consider similar validation and test loss functions as before and report the following results.

3.2.1 Prediction of conditional probabilities

As of earlier, after validation and training, we compare our models by comparing the true expected conditional distribution and the ones computed thanks to the models. For the joint gaussian distribution, the loss function is given by (2.12). However, since the lognormal distribution only takes positive values, we consider a different "dummy" function u instead of t given by:

$$\begin{aligned} u\colon Y &\to \mathbb{R}^3 \\ y &\mapsto (\mathbb{1}_{Y\leq 0.3}, \mathbb{1}_{Y\leq 0.6}, \mathbb{1}_{Y\leq 0.9}) \end{aligned}$$

Thus, the test loss function for the Poisson-Lognormal distribution is given by

$$\sum_{i=1}^{n_{test}} \|\mathbb{E}[u(Y)|X = x_i] - \mathbb{E}^{trad,low-rank,greedy}[u(Y)|X = x_i]\|_2^2 \tag{3.11}$$

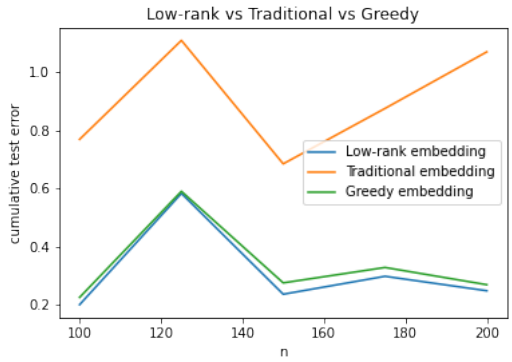

Figure 3.1: Cumulative test error computed over 20 runs for the three models on datasets generated by a Binomial-Gaussian joint distribution

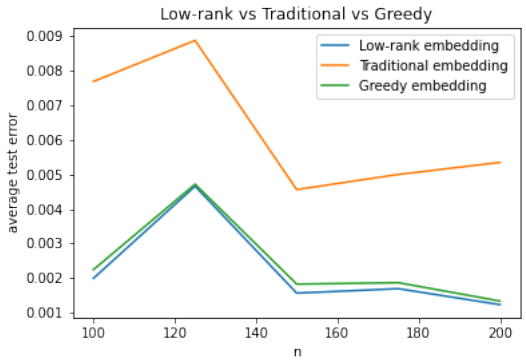

Figure 3.2: Average test error computed over 20 runs for the three models on datasets generated by a Binomial-Gaussian joint distribution

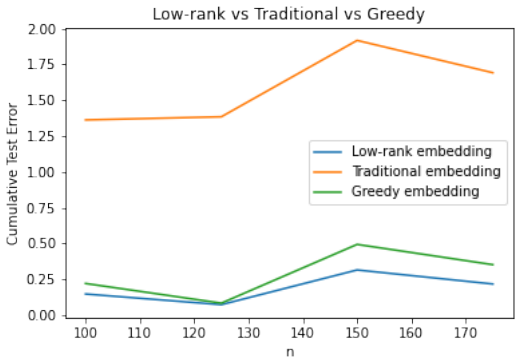

Figure 3.3: Cumulative test error computed over 20 runs for the three models on datasets generated by a Poisson-Lognormal joint distribution

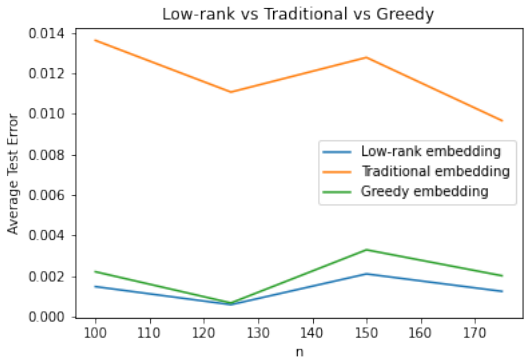

Figure 3.4: Average test error computed over 20 runs for the three models on datasets generated by a Poisson-Lognormal joint distribution

Figures (3.1), (3.2) yield the empirical results for a Binomial-Gaussian distribution for different number of datapoints n. They show that the greedy embedding works considerably well as it quite reproduces the same results as the low-rank embedding. We also notice that both embeddings have considerable better accuracy than the traditional one. Figures (3.3), (3.4) yield the empirical results for a Poisson-Lognormal distribution for different number of datapoints n. Again, for this different joint distribution, both models work considerably better than the traditional embedding. Moreover, the greedy one is not quite off the low-rank embedding.

3.2.2 Classification task

Another important machine learning task is classification. The goal is to assign a label to every class of our data. The simplest form of classification is the binary one. Depending on the value of the features, we have to assign our data point either to class 1 or class 2. For example, for a sample drawn from a Binomial-Gaussian joint distribution, we want to predict the value of $\text{sgn}(Y)|X = x_i$. This value is included in the following set $\{-1, 1\}$. The probability of this r.v. taking value -1 is given by:

$$\begin{aligned} p_i &= \mathbb{P}(\text{sgn}(Y) = -1|X = x_i) \\ &= \mathbb{P}(Y \leq 0|X = x_i) \end{aligned}$$

However, for a sample drawn from a Poisson-Lognormal distribution, we want to predict the value of $\text{sgn}(Y - 1)|X = x_i$. Again, the probability of this r.v. taking value -1 is given by:

$$\begin{aligned} q_i &= \mathbb{P}(\text{sgn}(Y) = -1|X = x_i) \\ &= \mathbb{P}(Y \leq 1|X = x_i) \end{aligned}$$

We use the logistic loss for our loss function, i.e. our loss function for a Binomial-Gaussian distributuin can be written as:

$$-\frac{1}{n_{test}} \sum_{i=1}^{n_t est} 1_{\text{sgn}(y_i)=-1} \ln(p_i) + 1_{\text{sgn}(y_i)=1} \ln(1 - p_i)$$

and for a Poisson-Lognormal distribution:

$$-\frac{1}{n_{test}}\sum_{i=1}^{n_test} 1_{\text{sgn}(y_i-1)=-1}\ln(q_i) + 1_{\text{sgn}(y_i-1)=1}\ln(1-q_i)$$

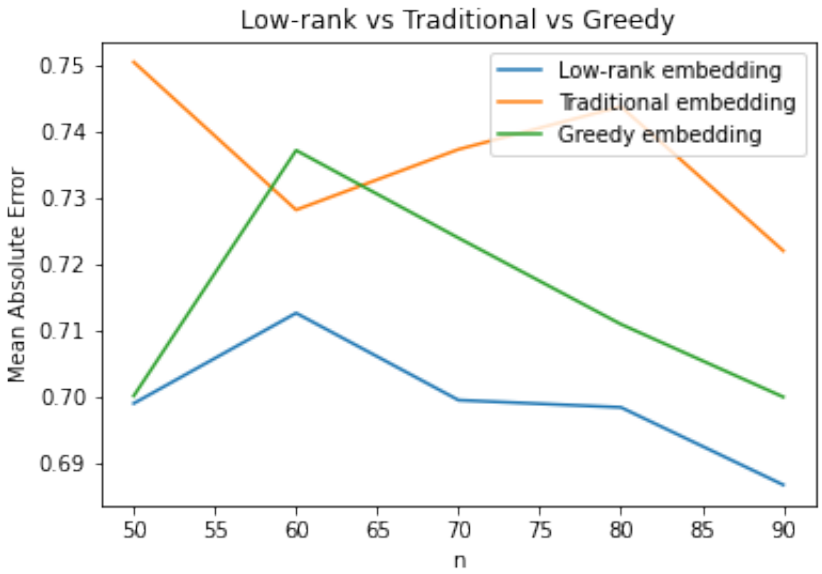

Figure 3.5: Mean absolute error computed over 10 runs for the classification task on datasets generated by a Binomial-Gaussian joint distribution

Figure 3.6: Mean absolute error computed over 10 runs for the classification task on datasets generated by a Poisson-Lognormal joint distribution

Figures (3.5),(3.6) yield the mean absolute error for our three models on a classification task when considering a Binomial-Gaussian distribution and a Poisson-Lognomal one. As expected, the low-rank embedding performs better than the

other two. However, the greedy embedding performs quite well especially for a higher number of datapoints.

3.3 Conclusion

In this work, we introduced a new approach to embed joint probability distributions when one of the marginals is defined on a finite discrete set based on the low-rank embedding introduced in [1]. This approach is based on a greedy framework that permits to find optimum local solutions and thus have a theoretical complexity which is lower than the low-rank embedding. Moreover, this approach also satisfies the positivty and normalization constraints contrary to the traditional embedding introduced in [5]. These findings are accompanied by favorable numerical results.

Bibliography

[1] Damir Filipovic Michael Multerer, Paul Schneider. "Adaptive joint distribution learning". In: (2021). DOI: https://doi.org/10.48550/arXiv.2110.04829.

[2] Clason, Christian. *Introduction to Functional Analysis*. Birkhäuser Cham, 2020. DOI: https://doi.org/10.1007/978-3-030-52784-6.

[3] Paulsen, Vern. *An introduction to the theory of reproducing kernel Hilbert spaces*. Cambridge University Press, 2016.

[4] Michael Reed, Barry Simon. *Methods of modern mathematical physics 1: Functional analysis*. Academic Press, 1972.

[5] Smola, Alex et al. "A Hilbert Space Embedding for Distributions". In: *Algorithmic Learning Theory* (2007), pp. 13–31.

[6] Altun, Yasemin and Smola, Alex. "Unifying Divergence Minimization and Statistical Inference Via Convex Duality". In: (2006), pp. 139–153.

[7] Ferger, Dietmar. "Optimal Tests for the General Two-Sample Problem". In: *Journal of Multivariate Analysis* (2000). DOI: https://doi.org/10.1006/jmva.1999.1879.

[8] Gretton, Arthur et al. "A Kernel Method for the Two-Sample-Problem". In: (2008). DOI: https://doi.org/10.48550/arXiv.0805.2368.

[9] Song, Le et al. "Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems". In: ICML '09 (2009), pp. 961–968. DOI: 10.1145/1553374.1553497. URL: https://doi.org/10.1145/1553374.1553497.

[10] Marek Capiński, Peter Ekkehard Kopp. *Measure, Integral and Probability*. Springer London, 2004.

[11] Retherford, J. R. *Hilbert Space Compact Operators and the Trace Theorem*. Cambridge University Press, 1993.

[12] Kumar, N. Kishore and Schneider, J. "Literature survey on low rank approximation of matrices". In: (2016).

[13] Dostál, Zdenek. *Optimal Quadratic Programming Algorithms*. Springer New York, NY, 2009.

[14] ApS, MOSEK. *MOSEK Fusion API for Python 9.3.20*. 2022. URL: https://docs.mosek.com/latest/pythonfusion/index.html.

[15] Erickson, Jeff. *Algorithms*. 2019.

[16] Curtis, S.A. "The classification of greedy algorithms". In: *Science of Computer Programming* (2003). DOI: https://doi.org/10.1016/j.scico.2003.09.001.

[17] Dadashi, Robert et al. “Primal Wasserstein Imitation Learning”. In: *International Conference on Learning Representations* (2020). DOI: https://doi.org/10.48550/arXiv.2006.04678.