Distribution to Distribution Regression

Junier B. Oliva JOLIVA@CS.CMU.EDU
Barnabás Póczos BAPOCZOS@CS.CMU.EDU
Jeff Schneider JEFF.SCHNEIDER@CS.CMU.EDU
Machine Learning Department, School of Computer Science, Carnegie Mellon University
5000 Forbes Avenue Pittsburgh PA 15213 USA

Abstract

We analyze 'Distribution to Distribution regression' where one is regressing a mapping where both the covariate (inputs) and response (outputs) are distributions. No parameters on the input or output distributions are assumed, nor are any strong assumptions made on the measure from which input distributions are drawn from. We develop an estimator and derive an upper bound for the $L2$ risk; also, we show that when the effective dimension is small enough (as measured by the doubling dimension), then the risk converges to zero with a polynomial rate.

1. Introduction

In standard regression analysis, one is concerned with inferring a mapping of a real valued vector of covariates (features) $X \in \mathbb{R}^d$ to a real valued vector response $Y \in \mathbb{R}^k$. While such a model encompasses many real-world problems, the restriction of finite dimensions on input and output domains excludes the regression of more complex objects. For example, in functional analysis one considers regression where the input domain are functions (Ferraty & Vieu, 2006).

Probability distributions are another infinite dimensional domain of interest for regression. Recently, (Poczos et al., 2012) considered regressing a mapping of probability distributions to a real valued response. Instead, in this paper we study *distribution to distribution regression*, where both the input covariate and the output response are probability distributions. Furthermore, we take a nonparametric approach, making as few and as weak assumptions as possible on the nature of input/output distributions and the mapping between them.

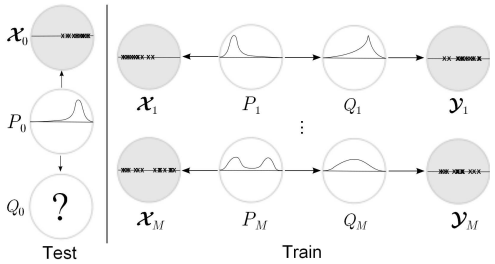


Figure 1. Model of dataset $\mathcal{D} = \{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^M$ of pairs of sets where P_i and $Q_i = f(P_i)$ are unobserved, instead one observes samples $\mathcal{X}_i \sim P_i$ and $\mathcal{Y}_i \sim Q_i$. P_0 is an unseen query distribution, observed indirectly through $\mathcal{X}_0$. We look to estimate the output distribution $Q_0 = f(P_0)$.

This framework is quite general and applicable to many real-world problems. For instance, consider the case where one collects samples at evenly spaced times $t_1, \ldots, t_M$, then given the probability distribution at time t_i it is natural to try to predict the distribution at time t_{i+1}. Furthermore, there are many more domain-specific uses of distribution to distribution regression. For example, in business one may consider the mapping of the distribution of some weather features to the distribution of some shipping route features. Also, in finance one may be interested in the mapping of the distribution of one sector's prices to the distribution of prices for another sector.

Our main contributions are as follows. First, we develop a nonparametric estimator for distribution to distribution regression. Second, with weak assumptions on the nature of input/output distributions and the measure that input distributions are sampled from, we derive an upperbound on the rate of convergence for the $L2$ risk. Lastly, we show that if the measure that input distributions are sampled from is a doubling measure, then the rate of convergence for the $L2$ risk is polynomial.

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

2. Model

Let $\mathcal{I}$ be a class of input distributions on $\Psi^k \subseteq \mathbb{R}^k$ that have a density with respect to the Lebesgue measure. Similarly, let $\mathcal{O}$ be a class of output distributions on $\Lambda^l \subseteq \mathbb{R}^l$ that have a density with respect to the Lebesgue measure. We regress a functional $f : \mathcal{I} \mapsto \mathcal{O}$.

Consider a set $\{(P_1, Q_1), \ldots, (P_M, Q_M)\}$ where $P_i \in \mathcal{I}$, $Q_i \in \mathcal{O}$, and $Q_i = f(P_i)$. Under the usual regression setting one would observe pairs of variables directly, however, we shall not since typically the true distribution of a real-world sample is unknown. Instead, we will consider the case where one only indirectly observes input/output distributions through i.i.d. samples $(\mathcal{X}_i, \mathcal{Y}_i)$ from P_i and Q_i respectively. That is, one has a dataset of input/output samples $\mathcal{D} = \{(\mathcal{X}_1, \mathcal{Y}_1), \ldots, (\mathcal{X}_M, \mathcal{Y}_M)\}$ where $\mathcal{X}_i = \{X_{i1}, \ldots, X_{in_i}\}$ with each $X_{ij} \overset{iid}{\sim} P_i$ and $\mathcal{Y}_i = \{Y_{i1}, \ldots, Y_{im_i}\}$ with $Y_{ij} \overset{iid}{\sim} Q_i$ (see Figure 1). Furthermore, each input distribution is taken to be sampled i.i.d. from a measure $\mathcal{P}$ on $\mathcal{I}$: $P_1, \ldots, P_M \overset{iid}{\sim} \mathcal{P}$.

Let $\{p_i\}_{i=1}^M \cup \{q_i\}_{i=1}^M$ be the pdfs corresponding to distributions $\{P_i\}_{i=1}^M \cup \{Q_i\}_{i=1}^M$. Note, for ease of notation, we shall use $Q_i = f(P_i)$ and $q_i = f(p_i)$, P_i and p_i interchangeably depending on whether speaking of a distribution or its density. Using nonparametric density estimation, one may estimate the input/output pdfs as $\{\tilde{p}_i\}_{i=1}^M \cup \{\tilde{q}_i\}_{i=1}^M$ where $\tilde{p}_i$ is estimated from $\mathcal{X}_i$ and $\tilde{q}_i$ is estimated from $\mathcal{Y}_i$. We shall denote the distributions correspond to pdfs $\{\tilde{p}_i\}_{i=1}^M \cup \{\tilde{q}_i\}_{i=1}^M$ as $\{\tilde{P}_i\}_{i=1}^M \cup \{\tilde{Q}_i\}_{i=1}^M$; that is $\tilde{P}_i(A) = \int_A \tilde{p}_i(x)\mathrm{d}x$, and $\tilde{Q}_i(A) = \int_A \tilde{q}_i(x)\mathrm{d}x$.

Often, nonparametric estimators for the real-valued regression setting take the form of some linear smoother. That is, to get an estimate at a new input query point $x_0 \in \mathbb{R}^d$, the function $g(x_0)$ is estimated as $\hat{g}(x_0) = \sum_i Y_i W(X_i, x_0)$ where $W(X_i, x_0) \in \mathbb{R}$ are weights depending on input observation $X_i \in \mathbb{R}^d$, and $Y_i \in \mathbb{R}$ is the (typically noisy) output observation of $g(X_i)$. Here, instead of a query point x_0, we have a query distribution $P_0 \sim \mathcal{P}$; however, again we consider the case when we are only given P_0 indirectly through a sample $\mathcal{X}_0 = \{X_{01}, \ldots, X_{0n_0}\} \overset{iid}{\sim} P_0$. In order to construct an estimate of $f(p_0)$, the pdf corresponding to the output distribution for P_0, we will apply a linear smoother using estimates of pdfs obtained using the observed input/output samples. That is our estimate for the pdf of $f(P_0)$ will have the form:

$$\hat{f}(\tilde{p}_0) = \sum_{i=1}^{M} \tilde{q}_i W(\tilde{P}_i, \tilde{P}_0), \tag{1}$$

where $\tilde{P}_0$ is the estimator of P_0 estimated from $\mathcal{X}_0$.

We will use orthogonal series estimators for output densities q_i, kernel smoothers for weights $W(\tilde{P}_i, \tilde{P}_0)$, and kernel density estimators for input densities p_i. Clearly, other regression methods, and density estimators may be used; we chose these methods primarily for ease of analysis.

3. Related Work

Distribution to distribution regression is related to the aforementioned functional analysis. However, the objects this model works over–distributions and their densities–are inferred through datasets of samples drawn from the objects, with varying finite sizes. In functional analysis, the objects–functions–are inferred through datasets of (X, Y) pairs that are often taken to be arbitrarily dense in the domain of the objects. For a comprehensive background in functional analysis see (Ferraty & Vieu, 2006) and (Ramsay & Silverman, 2002).

A common approach to working with distributions in ML tasks is to embed the distributions in a Hilbert space, then using kernels and kernel machines, solve a learning problem. The most straight forward of these methods is to fit a parametric model to distributions for estimating inner products (Jebara et al., 2004; Jaakkola et al., 1999; Moreno et al., 2003). Kernels have also been developed using nonparametric methods over distributions. For example, since distributions are observed only through finite sets, set kernels may be applied (Smola et al., 2007). Moreover, the representer theorem was recently generalized for the space of probability distributions (Muandet et al., 2012). Futhermore, kernels based nonparametric estimators of divergences have also been explored (Póczos et al., 2012a;b).

Recently, a non-Hilbert space approach was taken for regression with distribution covariates and real-valued response (Poczos et al., 2012), where an upper-bound was provided in hopes of better understanding the effect of sample sizes and number of input/output pairs on estimation risk. In this paper, we aim to provide a similar understanding to an even richer model, which spans a wider range of problems.

4. Orthogonal Series Estimator for Output Density

As previously mentioned, we do not directly observe $q_i = f(p_i)$, instead we are only given a finite sample drawn from q_i. In order to provide a linearly smoothed

estimate of $f(p)$ for unseen p as (1), we must first make an estimate of q_i. We will consider the case where the estimate $\tilde{q}_i$ is made using an orthogonal series estimator (see e.g. Tsybakov (2008)).

Suppose that $\Lambda^l \subseteq \mathbb{R}^l$, the domain of output densities is compact s.t. $\Lambda = [a, b]$. Let $\{\varphi_i\}_{i\in\mathbb{Z}}$ be an orthonormal basis for $L_2(\Lambda)$. Then, the tensor product of $\{\varphi_i\}_{i\in\mathbb{Z}}$ serves as an orthonormal basis for $L_2(\Lambda^l)$; that is,

$$\{\varphi_\alpha\}_{\alpha\in\mathbb{Z}^l} \quad \text{where} \quad \varphi_\alpha(x) = \prod_{i=1}^{l} \varphi_{\alpha_i}(x_i),\ x \in \Lambda^l$$

serves as an orthonormal basis (so we have $\forall \alpha, \gamma \in \mathbb{Z}^l,\ \langle\varphi_\alpha, \varphi_\gamma\rangle = I_{\{\alpha=\gamma\}}$).

Let $Q \in \mathcal{O}$, then

$$q(x) = \sum_{\alpha\in\mathbb{Z}^l} a_\alpha(Q)\varphi_\alpha(x) \quad \text{where} \tag{2}$$

$$a_\alpha(Q) = \langle\varphi_\alpha, q\rangle = \int_{\Lambda^l} \varphi_\alpha(z)\mathrm{d}Q(z) \ \in \mathbb{R}.$$

We make an anisotropic Sobolev ellipsoid type assumption about the projection coefficients $a(Q) = \{a_\alpha(Q)\}_{\alpha\in\mathbb{Z}^l}$:

$$\mathcal{O} = \{Q : a(Q) \in \Theta(\nu, \sigma, \bar{A})\} \quad \text{where} \tag{3}$$

$$\Theta(\nu, \sigma, \bar{A}) = \left\{ \{a_\alpha\}_{\alpha\in\mathbb{Z}^l} : \sum_{\alpha\in\mathbb{Z}^l} a_\alpha^2 \kappa_\alpha^2(\nu, \sigma) < \bar{A} \right\},$$

$$\kappa_\alpha^2(\nu, \sigma) = \sum_{i=1}^{l} (\nu_i|\alpha_i|)^{2\sigma_i} \text{ for } \nu_i, \sigma_i, \bar{A} > 0.$$

See (Ingster & Stepanova, 2011; Laurent, 1996) for other analysis with this type of assumption. The assumption in (3) will control the tail-behavior of projection coefficients and allow us to effectively estimate $Q \in \mathcal{O}$ using a finite number of projection coefficients on the empirical distribution of a sample.

Given a sample $\mathcal{Y}_i = \{Y_{i1}, \ldots, Y_{im_i}\}$ where $Y_{ij} \overset{iid}{\sim} Q_i \in \mathcal{O}$ and $Q_i = f(P_i)$, let $\widehat{Q}_i$ be the empirical distribution of $\mathcal{Y}_i$; i.e. $\widehat{Q}_i(Y = Y_{ij}) = \frac{1}{m_i}$. Our estimator for q_i will be:

$$\tilde{q}_i(x) = \sum_{\alpha\ :\ \kappa_\alpha(\nu,\sigma)\le t_i} a_\alpha(\widehat{Q}_i)\varphi_\alpha(x) \quad \text{where} \tag{4}$$

$$a_\alpha(\widehat{Q}_i) = \int_{\Lambda^k} \varphi_\alpha(z)\mathrm{d}\widehat{Q}_i(z) = \frac{1}{m_i}\sum_{j=1}^{m_i} \varphi_\alpha(Y_{ij}). \tag{5}$$

Choosing t_i optimally can be shown to lead to $\mathbb{E}[\|\tilde{q}_i - q_i\|_2^2] = O(m_i^{-\frac{2}{2+\sigma^{-1}}})$, where $\sigma^{-1} = \sum_{j=1}^{l} \sigma_j^{-1}$, $m_i \to \infty$ (Nussbaum, 1983).

5. Estimator of Output Density for Unseen Input Density

As previously mentioned, the estimator of $q = f(p_0)$ that we'll use to estimate the output query distribution when given a sample from a new input query distribution P_0 is (1). Let

$$\mathcal{A}_t = \{\alpha \in \mathbb{Z}^l \ :\ \kappa_\alpha(\nu, \sigma) \le t\} \tag{6}$$

Our estimator $\hat{q}_0 = \hat{f}(\tilde{p}_0)$ will be as follows. Let all $\tilde{q}_i$ (4) have coefficients $\alpha \in \mathcal{A}_t$, (1) may be written as:

$$\begin{aligned}
\hat{q}_0(x) = [\hat{f}(\tilde{p}_0)](x) &= \sum_{i=1}^{M} \tilde{q}_i(x) W(\tilde{P}_i, \tilde{P}_0) \\
&= \sum_{i=1}^{M} \left(\sum_{\alpha\in\mathcal{A}_t} a_\alpha(\widehat{Q}_i)\varphi_\alpha(x) \right) W(\tilde{P}_i, \tilde{P}_0) \\
&= \sum_{\alpha\in\mathcal{A}_t} \left(\sum_{i=1}^{M} a_\alpha(\widehat{Q}_i) W(\tilde{P}_i, \tilde{P}_0) \right) \varphi_\alpha(x) \\
&= \sum_{\alpha\in\mathcal{A}_t} \hat{a}_\alpha \varphi_\alpha(x)
\end{aligned} \tag{7}$$

where $\hat{a}_\alpha = \sum_{i=1}^{M} a_\alpha(\widehat{Q}_i) W(\tilde{P}_i, \tilde{P}_0)$. That is, our estimator can be interpreted as smoothing the projection density estimates at each output density or, equivalently, as building a new projection density estimate using smoothed projection coefficients from each output density.

With (7), (2) we can upperbound the L_2 loss of $\hat{f}(\tilde{p}_0)$:

$$\begin{aligned}
&\|\hat{f}(\tilde{p}_0) - f(p_0)\|_2 = \\
&= \left(\int_{\Lambda^k} \left(\sum_{\alpha\in\mathcal{A}_t} (\hat{a}_\alpha - a_\alpha(f(P_0)))\varphi_\alpha(x) \right.\right. \\
&\quad \left.\left. - \sum_{\alpha\in\mathcal{A}_t^c} a_\alpha(f(P_0))\varphi_\alpha(x) \right)^2 \mathrm{d}x \right)^{1/2} \\
&= \sqrt{\sum_{\alpha\in\mathcal{A}_t} (\hat{a}_\alpha - a_\alpha(f(P_0)))^2 + \sum_{\alpha\in\mathcal{A}_t^c} a_\alpha^2(f(P_0))} \qquad (8) \\
&\le \sum_{\alpha\in\mathcal{A}_t} |\hat{a}_\alpha - a_\alpha(f(P_0))| + \left(\sum_{\alpha\in\mathcal{A}_t^c} a_\alpha^2(f(P_0)) \right)^{1/2}
\end{aligned}$$

where (8) follows from orthonormality. Thus,

$$\mathbb{E}[\|\hat{f}(\tilde{p}_0) - f(p_0)\|_2] \le \sum_{\alpha\in\mathcal{A}_t} \mathbb{E}\left[|\hat{a}_\alpha - a_\alpha(f(P_0))|\right] \tag{9}$$

$$+ \mathbb{E}\left[\sqrt{\sum_{\alpha\in\mathcal{A}_t^c} a_\alpha^2(f(P_0))} \right]. \tag{10}$$

Thus, we may upperbound the absolute risk for each projection coefficient in (9), and control (10) using (3). Note, the risk for each projection coefficient is akin to the distribution covariate/ real output problem studied in (Poczos et al., 2012). In fact, using a trivial rewrite: $a_\alpha(\widehat{Q}_i) = a_\alpha(Q_i) + \mu_\alpha^{(i)}$ with, $\mu_\alpha^{(i)} = a_\alpha(\widehat{Q}_i) - a_\alpha(Q_i)$ where $\mathbb{E}[\mu_\alpha^{(i)}] = 0$, but unlike in (Poczos et al., 2012) sample sizes will now play a role in the nature of the "noise" $\mu_\alpha^{(i)}$.

In order to make weights $W(\tilde{p}_i, \tilde{p}_0)$ we will use kernel smoothing. That is, $W(\tilde{P}_i, \tilde{P}_0) =$

$$\begin{cases} \frac{K\left(\frac{D(\tilde{P}_i,\tilde{P}_0)}{h}\right)}{\sum_{j=1}^{M} K\left(\frac{D(\tilde{P}_j,\tilde{P}_0)}{h}\right)} & \text{if } \sum_{j=1}^{M} K\left(\frac{D(\tilde{P}_j,\tilde{P}_0)}{h}\right) > 0 \\ 0 & \text{else} \end{cases} \tag{11}$$

where D is a metric and K is a kernel function satisfying assumption $\mathfrak{A}_2$ below.

6. L_2 Risk Analysis

We will analyze the L_2 risk of the estimator $\hat{f}(\tilde{p}_0)$ using the L_1 metric for D and kernel density estimation for $\{\tilde{p}_i\}_{i=0}^M$. That is, suppose that a kernel density estimator is used to estimate input densities:

$$\tilde{p}_i(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{1}{b_i^k} B\left(\frac{\|x - X_{ij}\|}{b_i}\right), \tag{12}$$

where $b_i > 0$ is a bandwidth parameter, B is an appropriate kernel function (see e.g. Tsybakov (2008)), and $\|\cdot\|$ is the Euclidean norm. Furthermore, suppose that $D(\tilde{P}_0, \tilde{P}_i) = \|\tilde{p}_0 - \tilde{p}_i\|_1$ is the L_1 distance $\int |\tilde{p}_0(x) - \tilde{p}_i(x)| \mathrm{d}x$.

6.1. Assumptions

We shall assume $\mathfrak{A}_1$ through $\mathfrak{A}_6$:

- ($\mathfrak{A}_1$) *L_∞ Hölder continuous functional.* The unknown functional f is in a class $\mathcal{M} = \mathcal{M}(L, \beta)$:

$$\mathcal{M} = \Big\{ f : \forall P, P' \in \mathcal{I}, \max_{\alpha \in \mathbb{Z}^k} |a_\alpha(f(P)) - a_\alpha(f(P'))| \leq L D(P, P')^\beta \Big\}. \tag{13}$$

- ($\mathfrak{A}_2$) *Asymmetric boxed and Lipschitz kernel.* The kernel K satisfies the following properties: $K : [0, \infty] \to \mathbb{R}]$ is nonnegative and Lipschitz continuous with Lipschitz constant L_K. Furthermore, there exist constants $0 < \underline{K} < 1$ and $0 < r < R < \infty$ such that:

$$\forall x > 0,\ \underline{K} I_{\{x \leq r\}} \leq K(x) \leq I_{\{x \leq R\}}$$

- ($\mathfrak{A}_3$) *Class of input/output distributions.* The class $\mathcal{I}$ is the set of distribution $\mathcal{H}_k(1)$ with densities that are 1-smooth Hölder functions, as in (Rigollet & Vert, 2009). Furthermore, the domain of distributions in $\mathcal{I}$, $\Psi^k \subseteq \mathbb{R}^k$, is assumed to be compact. Also, it is assumed $\mathcal{O}$ is as (3) and Section 4.

- ($\mathfrak{A}_4$) *Bounded basis.* Assume $\max_{\alpha \in \mathbb{Z}} \|\varphi_\alpha\|_\infty < \varphi_{\max}$ for some $0 < \varphi_{\max} < \infty$. Furthermore, clearly $\forall i \in \{1, \ldots, M\}$, $\forall \alpha \in \mathbb{Z}^l$, $a_\alpha(\widehat{Q}_i) < \varphi_{\max}$. Moreover, it is assumed that n_i and m_i are independent of P_i. By (3), we know that $\forall \alpha \neq 0$, $a_\alpha(Q_i) < \bar{A}$; assume further that $a_0(Q_i) < \bar{A}$.

- ($\mathfrak{A}_5$) *Lower bound on sample sizes.* Assume that $\min\left(\{n_i\}_{i=0}^M\right) = n$, $\min\left(\{m_i\}_{i=1}^M\right) = m$ and $e^{n^{\frac{1}{2+k}}}/M \to \infty$ as $M \to \infty$.

- ($\mathfrak{A}_6$) *Relation between n and h.* Assume that $C_* n^{-\frac{1}{2+k}} \leq rh/4$

6.2. Lemmas

Before deriving upper bounds, we state several Lemmas. For proofs, please refer to the Appendix section. Let $\mathcal{B}_D(P, h) \equiv \{P' \in \mathcal{I} : D(P, P') \leq h\}$, $\Phi_P(h) \equiv \mathcal{P}(\mathcal{B}_D(P, h))$, where P is a fixed distribution. Also let $K_j \equiv K\left(\frac{D(P_0,P_j)}{h}\right)$ and $\tilde{K}_j \equiv K\left(\frac{D(\tilde{P}_0,\tilde{P}_j)}{h}\right)$.

Lemma 1 $\mathbb{P}\left(\sum_{i=1}^M K_i = 0\right) \leq \mathbb{P}\left(\sum_{i=1}^M K_i \leq \underline{K}\right) \leq \frac{1}{eM}\mathbb{E}\left[\frac{1}{\Phi_P(rh)}\right]$.

Let $\zeta(n, M) \equiv \frac{1}{eM}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right] + (M+1)e^{-\frac{1}{2}n^{\frac{k}{2+k}}}$.

Lemma 2 $\mathbb{P}\left(\sum_{i=1}^M \tilde{K}_i = 0\right) \leq \mathbb{P}\left(\sum_{i=1}^M \tilde{K}_i \leq \underline{K}\right) \leq \zeta(n, M)$.

Lemma 3 $\mathbb{E}\left[\frac{I_{\{\sum_i K_i \leq K\}}}{\sum_i K_i}\right] \leq \frac{1+1/K}{M\underline{K}}\mathbb{E}\left[\frac{1}{\Phi_P(rh)}\right]$.

6.3. Upper bound

We look to analyze the L2 risk of our estimator (9),(10). As previously mentioned, (10) can be upperbounded using (3).

Let $R_\alpha(M, n, m) \equiv \mathbb{E}\left[|\hat{a}_\alpha - a_\alpha(f(P_0))|\right]$, where arguments M, n, m emphasize the dependence on the respective sample size bounds. We look to find $R(M, n, m)$ s.t. $\forall \alpha \in \mathbb{Z}^l$ $R_\alpha(M, n, m) \leq R(M, n, m)$.

Hence, using (9) and (10): $\mathbb{E}[\|\hat{f}(\tilde{p}_0) - f(p_0)\|_2] \leq$

$$|\mathcal{A}_t| R(M, n, m) + \sqrt{\mathbb{E}\left[\sum_{\alpha \in \mathcal{A}_t^c} a_\alpha^2(f(P_0))\right]} \tag{14}$$

where $|\mathcal{A}_t|$ is the cardinality of set $\mathcal{A}_t$ as in (6). In order to derive a bound for $R(M,n,m)$, we will use some similar arguments in (Poczos et al., 2012). For analysis purposes let $\bar{a}_\alpha \equiv$

$$\begin{cases} \frac{\sum_{i=1}^{M} a_\alpha(\widehat{Q}_i) K\left(\frac{D(P_0,P_i)}{h}\right)}{\sum_{j=1}^{M} K\left(\frac{D(P_0,P_i)}{h}\right)} & \text{if } \sum_j K\left(\frac{D(P_0,P_i)}{h}\right) > 0, \\ 0 & \text{else} \end{cases}. \tag{15}$$

That is, $\bar{a}_\alpha$ is the kernel regression estimator using the true input distribution $\{P_i\}_{i=0}^M$ instead of the estimated distributions $\{\tilde{P}_i\}_{i=0}^M$. We upper-bound the pointwise risk as:

$$R_\alpha(M,n,m) \le \mathbb{E}[|\hat{a}_\alpha - \bar{a}_\alpha|] \tag{16}$$
$$+ \mathbb{E}[|\bar{a}_\alpha - a_\alpha(f(P_0))|] \tag{17}$$

For our bounds, we shall take $b_i = n^{-\frac{1}{2+k}}$ the asymptotically optimal bandwidths (up to constants) for kernel density estimation under MSE loss.

6.3.1. Bound on Eq. 16

Let $\Delta\hat{a}_\alpha = |\hat{a}_\alpha - \bar{a}_\alpha|$, we look to upperbound $\mathbb{E}[\Delta\hat{a}_\alpha]$.

Define the following six events E_0, E_1, E_2 $\tilde{E}_0, \tilde{E}_1, \tilde{E}_2$ as: $E_0 = \{\sum_i K_i = 0\}$, $E_1 = \{0 < \sum_i K_i \le \underline{K}\}$, $E_2 = \{\underline{K} < \sum_i K_i\}$, $\tilde{E}_0 = \{\sum_i \tilde{K}_i = 0\}$, $\tilde{E}_1 = \{0 < \sum_i \tilde{K}_i \le \underline{K}\}$, and $\tilde{E}_2 = \{\underline{K} < \sum_i \tilde{K}_i\}$. Clearly, $\mathbb{E}[\Delta\hat{a}_\alpha] = \sum_{i=0}^2 \sum_{j=0}^2 \mathbb{E}[\Delta\hat{a}_\alpha I_{E_i} I_{\tilde{E}_j}]$.

If $0 < \sum_i K_i$, then clearly $\forall\alpha$ $|\bar{a}_\alpha| = \left|\sum_i \frac{a_\alpha(\widehat{Q}_i)K_i}{\sum_i K_i}\right| < \varphi_{\max}$. Thus, $\mathbb{E}\left[\left|\sum_i \frac{a_\alpha(\widehat{Q}_i)K_i}{\sum_i K_i}\right| I_{\tilde{E}_0}(I_{E_1} + I_{E_2})\right]$

$$\le \varphi_{\max}\mathbb{E}\left[I_{\{\sum_i K_i > 0 \wedge \sum_i \tilde{K}_i = 0\}}\right]$$
$$= \varphi_{\max}\mathbb{P}\left(\sum_i K_i > 0, \sum_i \tilde{K}_i = 0\right)$$
$$\le \varphi_{\max}\mathbb{P}\left(\sum_i \tilde{K}_i = 0\right) \le \varphi_{\max}\zeta(m,M), \tag{18}$$

with (18) following from Lemma 2.

Similarly, $\mathbb{E}\left[\left|\frac{\sum_i a_\alpha(\widehat{Q}_i)\tilde{K}_i}{\sum_i \tilde{K}_i}\right| I_{E_0}(I_{\tilde{E}_1} + I_{\tilde{E}_2})\right] \le \frac{\varphi_{\max}}{eM}\mathbb{E}\left[\frac{1}{\Phi_P(rh)}\right]$, by Lemma 1.

Furthermore, $\mathbb{E}\left[\Delta\hat{a}_\alpha I_{E_1}(I_{\tilde{E}_1} + I_{\tilde{E}_2})\right]$

$$\le \mathbb{E}\left[\left(\left|\frac{\sum_i a_\alpha(\widehat{Q}_i)K_i}{\sum_i K_i}\right| + \left|\frac{\sum_i a_\alpha(\widehat{Q}_i)\tilde{K}_i}{\sum_i \tilde{K}_i}\right|\right) \times I_{E_1}(I_{\tilde{E}_1} + I_{\tilde{E}_2})\right]$$
$$\le 2\varphi_{\max}\mathbb{E}\left[I_{E_1}(I_{\tilde{E}_1} + I_{\tilde{E}_2})\right] \le 2\varphi_{\max}\mathbb{E}\left[I_{E_1}\right]$$
$$= 2\varphi_{\max}\mathbb{P}(0 < \sum_i K_i \le \underline{K}) \le \frac{2\varphi_{\max}}{eM}\mathbb{E}\left[\frac{1}{\Phi_P(rh)}\right]$$

by Lemma 1; likewise $\mathbb{E}\left[\Delta\hat{a}_\alpha I_{\tilde{E}_1}(I_{E_1} + I_{E_2})\right]$

$$\le 2\varphi_{\max}\mathbb{P}(0 < \sum_i \tilde{K}_i \le \underline{K}) \le 2\varphi_{\max}\zeta(m,M).$$

by Lemma 2.

Lemma 4 $\mathbb{E}\left[\Delta\hat{a}_\alpha I_{\tilde{E}_2} I_{E_2}\right] \le \frac{C_1}{h}\mathbb{E}\left[\frac{1}{\Phi_P(rh)}\right] n^{-1/(2+k)}$ (for $C_1 > 0$ specified in proof, see Appendix).

Hence, combining: $\mathbb{E}[\Delta\hat{a}_\alpha] \le C_1 \frac{n^{-\frac{1}{2+k}}}{h}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right] + \frac{C_2}{M}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right] + (M+1)e^{-\frac{1}{2}n^{\frac{k}{2+k}}}$, with $C_i > 0$.

6.3.2. Bound on Eq. 17

Note that $\mathbb{E}|\bar{a}_\alpha - a_\alpha(f(P_0))| =$

$$\mathbb{E}\left|\frac{\sum_i a_\alpha(\widehat{Q}_i)K_i}{\sum_i K_i} I_{\{\sum_i K_i > 0\}} - a_\alpha(f(P_0))\right|$$
$$= \mathbb{E}\left|\frac{\sum_i (a_\alpha(Q_i) + \mu_\alpha^{(i)})K_i}{\sum_i K_i} I_{\{\sum_i K_i > 0\}} - a_\alpha(f(P_0))\right|$$
$$\le \mathbb{E}\left|\frac{\sum_i (a_\alpha(f(P_i)) - a_\alpha(f(P_0)))K_i}{\sum_i K_i} I_{\{\sum_i K_i > 0\}}\right| \tag{19}$$
$$+ \mathbb{E}\left|\frac{\sum_i \mu_\alpha^{(i)} K_i}{\sum_i K_i} I_{\{\sum_i K_i > 0\}}\right| \tag{20}$$
$$+ \mathbb{E}\left|a_\alpha(f(P_0)) I_{\{\sum_i K_i = 0\}}\right|, \tag{21}$$

We bound the three terms in (19), (20), and (21).

To bound (19), let $I = I_{\{\sum_i K_i > 0\}}$. Using $\mathfrak{A}\mathbf{1}$,

$$\mathbb{E}\left[\frac{\sum_i |a_\alpha(f(P_i)) - a_\alpha(f(P_0))|K_i}{\sum_i K_i} I\right]$$
$$\le \mathbb{E}\left[\frac{\sum_i L D(P_i,P_0)^\beta K_i}{\sum_i K_i} I\right] \le L(hR)^\beta,$$

since by $\mathfrak{A}\mathbf{2}$ $\text{supp}(K) \subseteq [0,R]$ so $LD(P_i,P_0)^\beta K_i \le L(hR)^\beta K_i$. To bound (20), first we bound $\mathbb{E}[|\mu_\alpha^{(i)}|]$ and $\sqrt{\mathbb{E}[|\mu_\alpha^{(i)}|^2]}$.

Lemma 5[1] $\mathbb{E}[|\mu_\alpha^{(i)}|] \le \sqrt{\mathbb{E}[|\mu_\alpha^{(i)}|^2]} \le C m^{-\frac{1}{2}}$

Let $\underline{I} = I_{\{\underline{K} > \sum_i K_i > 0\}}$ and $\bar{I} = I_{\{\sum_i K_i > \underline{K}\}}$, hence

[1] see Appendix, $C > 0$.

$I = \bar{I} + \underline{I}$. To bound (20):

$$\mathbb{E}\left|\frac{\sum_i \mu_\alpha^{(i)} K_i}{\sum_i K_i} I\right| = \mathbb{E}\left|\frac{\sum_i \mu_\alpha^{(i)} K_i}{\sum_i K_i}(\bar{I} + \underline{I})\right|$$
$$\leq \mathbb{E}\left[\left|\frac{\sum_i \mu_\alpha^{(i)} K_i}{\sum_i K_i}\bar{I}\right|\right] + \mathbb{E}\left[\left|\frac{\sum_i \mu_\alpha^{(i)} K_i}{\sum_i K_i}\underline{I}\right|\right]$$
$$\leq \mathbb{E}\left[\left|\frac{\sum_i \mu_\alpha^{(i)} K_i}{\sum_i K_i}\right| \bar{I}\right] + c_1 m^{-\frac{1}{2}} \mathbb{P}\left(\sum_i K_i < \underline{K}\right)$$
$$\leq \mathbb{E}\left[\left|\frac{\sum_i \mu_\alpha^{(i)} K_i}{\sum_i K_i}\right| \bar{I}\right] + \frac{c_1 m^{-\frac{1}{2}}}{eM}\mathbb{E}\left[\frac{1}{\Phi_P(rh)}\right],$$

second last line by Lemma 5, and last by Lemma 1.

Lemma 6[1] $\mathbb{E}\left[\left|\frac{\sum_i \mu_\alpha^{(i)} K_i}{\sum_i K_i}\right| \bar{I}\right] \leq C\sqrt{\frac{1}{mM}\mathbb{E}\left[\frac{1}{\Phi_P(rh)}\right]}$

Lastly, using Lemma 1, a bound on (21):

$$\mathbb{E}\left|a_\alpha(f(P_0)) I_{\{\sum_i K_i = 0\}}\right| \leq \frac{\bar{A}}{eM}\mathbb{E}\left[\frac{1}{\Phi_P(rh)}\right]$$

Combining the above we have that $\mathbb{E}|\bar{a}_\alpha - a_\alpha(f(P_0))|$

$$\leq C_3 h^\beta + C_4\sqrt{\frac{1}{mM}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right]}$$
$$+ \frac{C_5}{\sqrt{m}M}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right] + \frac{C_6}{M}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right].$$

6.3.3. Projection Coefficient Regression Convergence Rate

Synthesizing, $\forall \alpha \in \mathbb{Z}^l$ $R_\alpha(M,n,m) \leq R(M,n,m)$, where: $R(M,n,m) \equiv$

$$C_1\frac{n^{-\frac{1}{2+k}}}{h}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right] + \frac{C_2}{M}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right]$$
$$+ C_3 h^\beta + C_4\sqrt{\frac{1}{mM}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right]}$$
$$+ \frac{C_5}{\sqrt{m}M}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right] + (M+1)e^{-\frac{1}{2}n^{\frac{k}{2+k}}}. \quad (22)$$

6.3.4. Convergence Rate for Distribution to Distribution Regression

Note that $\sqrt{\sum_{\alpha\in\mathcal{A}_t^c} a_\alpha^2(f(P_0))}$

$$= \frac{1}{t}\sqrt{\sum_{\alpha\in\mathcal{A}_t^c} t^2 a_\alpha^2(f(p_0))} \leq \frac{1}{t}\sqrt{\sum_{\alpha\in\mathcal{A}_t^c} \kappa_\alpha^2(\nu,\sigma) a_\alpha^2(f(P_0))}$$
$$\leq \frac{1}{t}\sqrt{\sum_{\alpha\in\mathbb{Z}^l} \kappa_\alpha^2(\nu,\sigma) a_\alpha^2(f(P_0))} \leq \frac{\sqrt{\bar{A}}}{t}.$$

Furthermore, note that if we have a bound $\forall \alpha \in \mathcal{A}_t$, $c \geq |\alpha_i|$ then $(2c+1)^l \geq |\mathcal{A}_t|$, by a simple counting argument. Let $\lambda = \text{argmin}_i \nu_i^{2\sigma_i}$. We have $\alpha \in \mathcal{A}_t$ iff $\frac{t^2}{\nu_\lambda^{2\sigma_\lambda}} \geq \frac{1}{\nu_\lambda^{2\sigma_\lambda}}\sum_{i=1}^l (\nu_i|\alpha_i|)^{2\sigma_i} \geq \sum_{i=1}^l |\alpha_i|^{2\sigma_i}$, so $\nu_\lambda^{-\frac{\sigma_\lambda}{\sigma_i}} t^{\frac{1}{\sigma_i}} \geq |\alpha_i|$. Thus, $|\mathcal{A}_t| \leq \prod_{i=1}^l (2\nu_\lambda^{-\frac{\sigma_\lambda}{\sigma_i}} t^{\frac{1}{\sigma_i}} + 1)$. Thus, asymptotically $|\mathcal{A}_t| = O(t^{\sigma^{-1}})$ where $\sigma^{-1} = \sum_{j=1}^l \sigma_j^{-1}$; and, as $M,n,m \to \infty$ and with appropriate h, by (14) we have:

$$\mathbb{E}[\|\hat{f}(\tilde{p}_0) - f(p_0)\|_2] \leq CR(M,n,m)t^{\sigma^{-1}} + \frac{\sqrt{\bar{A}}}{t}.$$

Choosing $t \sim R(M,n,m)^{-1/(\sigma^{-1}+1)}$ leads to our first major result, a bound on the $L2$ risk.

Theorem 7

$$\mathbb{E}[\|\hat{f}(\tilde{p}_0) - f(p_0)\|_2] \leq C' R(M,n,m)^{1/(\sigma^{-1}+1)} \quad (23)$$

As a corollary, if $\sigma_i = \rho > 0$, then clearly:

$$\mathbb{E}[\|\hat{f}(\tilde{p}_0) - f(p_0)\|_2] \leq C' R(M,n,m)^{\rho/(k+\rho)}$$

6.3.5. Doubling Dimension

Clearly, the bounds on $R(M,n,m)$ and $\mathbb{E}[\|\hat{f}(\tilde{p}_0) - f(p_0)\|_2]$ depend on the quantity $\mathbb{E}\left[(\Phi_P(rh/2))^{-1}\right]$. It can be shown that without further assumptions the quantity can be relatively large, leading to slow rates. However, here we will focus on the case when we may control the effective dimension of the support of $\mathcal{P}$.

Following (Kpotufe, 2011), we look to control the effective dimension through the doubling dimension. We say that $\mathcal{P}$ is a doubling measure, with effective dimension d, if $\exists c > 0$ for all $u > 0$ and $1 > \epsilon > 0$: $\frac{\mathcal{P}(\mathcal{B}_D(S,u))}{\mathcal{P}(\mathcal{B}_D(S,\epsilon u))} < \left(\frac{c}{\epsilon}\right)^d$. Hence, $E\left[\frac{1}{\Phi_P(rh/2)}\right]$

$$= \mathbb{E}\left[\frac{\Phi_P(1)}{\Phi_P(1)}\frac{1}{\Phi_P(\frac{rh}{2})}\right] \leq (\tfrac{rh}{2})^{-d} C\mathbb{E}\left[\frac{1}{\Phi_P(1)}\right] \leq C' h^{-d}$$

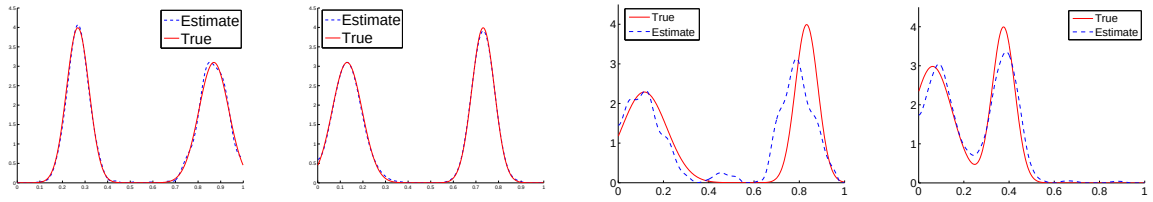
Clearly, $\frac{1}{M} = \Omega(\frac{1}{\sqrt{mM}})$, as $M,m \to \infty$. To further simplify, assume that $n = \Theta(m)$. Then, $R(M,n,m) \leq$

$$R(M,n) \equiv C_1\frac{n^{-\frac{1}{2+k}}}{h^{d+1}} + \frac{C_2}{Mh^d} + C_3 h^\beta + C_4\sqrt{\frac{1}{nMh^d}}.$$

We will analyze $R(M,n)$ depending on the dominating term, and choosing the bandwidth h optimally. Furthermore, in order to assure that the $(M+1)e^{-\frac{1}{2}n^{\frac{k}{2+k}}}$ term in (22) does not dominate we slightly extend assumption $\mathfrak{A}_5$ as follows: $M = O(n^{-\frac{\beta}{(k+2)(\beta+d+1)}} e^{n^{\frac{k}{k+2}}})$.

Lemma 8[2] (*Case 1*) If $\frac{1}{Mh^d} = \Omega(\sqrt{\frac{1}{nMh^d}})$ and $\frac{1}{Mh^d} =$

[2] see Appendix.



(a) Input Distribution $M, \eta = 10000$ (b) Output Distribution $M, \eta = 10000$ (c) Output Distribution $M, \eta = 100$ (d) Output Distribution $M, \eta = 1000$

Figure 2. (a) Example unseen input distribution p_0 shown in solid red, $\tilde{p}_0$ in dashed blue using $M, \eta = 10000$. (b) Corresponding output distribution, $f(p_0)$ shown in solid red, $\hat{f}(\tilde{p}_0)$ in dashed blue using $M, \eta = 10000$. (c),(d) Different true and estimated output distributions using other M, η (corresponding input distributions not shown).

$\Omega(\frac{n^{-\frac{1}{2+k}}}{h^{d+1}})$, then $R(M,n) = O(h^\beta + \frac{1}{Mh^d})$ and choosing h optimally leads to $R(M,n) = O(M^{-\frac{\beta}{\beta+d}})$.

Hence, $R(M,n,m) = O(M^{\frac{-\beta}{\beta+d}})$. Note this case implies $n = \Omega(M^{\frac{(\beta+d+1)(k+2)}{\beta+d}})$; since M is slow growing, it makes sense that the rate be driven by it.

Lemma 9[2] (*Case 2*) If $\frac{n^{-\frac{1}{2+k}}}{h^{d+1}} = \Omega(\sqrt{\frac{1}{nMh^d}})$ and $\frac{n^{-\frac{1}{2+k}}}{h^{d+1}} = \Omega(\frac{1}{Mh^d})$, then $R(M,n) = O(h^\beta + n^{\frac{-1}{2+k}}h^{-(d+1)})$ and choosing h optimally leads to $R(M,n) = O(n^{-\frac{\beta}{(k+2)(\beta+d+1)}})$.

Thus, $R(M,n,m) = O(n^{-\frac{\beta}{(k+2)(\beta+d+1)}})$. This case implies $M = \Omega(n^{\frac{\beta+d}{(k+2)(\beta+d+1)}})$; thus, the rate is again intuitive since n is slow growing in this case.

Lastly, it can be shown that if one choose h optimally then $(nMh^d)^{-1/2}$ cannot dominate.

Lemma 10[2] If one chooses h optimally it can not be that $\sqrt{\frac{1}{nMh^d}} = \Omega(\frac{n^{-\frac{1}{2+k}}}{h^{d+1}})$ and $\sqrt{\frac{1}{nMh^d}} = \Omega(\frac{1}{Mh^d})$.

Hence we have that in any case, if $\mathcal{P}$ is a doubling measure then we have a polynomial rate on $R(M,n,m)$. Hence by (23), we have our second major result:

Theorem 11 If $\mathcal{P}$ is a doubling dimension, then the rate of convergence for $\mathbb{E}[\|\hat{f}(\tilde{p}_0) - f(p_0)\|_2]$ is polynomial in M, n, m.

6.3.6. Different Distances, Estimators

We note that a very similar analysis may be employed when using the $L2$ metric as the distance D, and using projection series estimators for $\{\tilde{p}_i\}_{i=0}^M$. On a practical note, in this case the $L2$ distance among the estimated input distribution and the estimated query distribution is just the $\ell 2$ distance of their projection coefficients. Due to space constraints, further details are omitted.

7. Experiments

In order to assess the empirical performance of distribution to distribution estimation, we performed experiments on both synthetic and real data. In both cases, the dataset we are operating on is of the form $\{(\mathcal{X}_1, \mathcal{Y}_1), \ldots, (\mathcal{X}_M, \mathcal{Y}_M)\}$, where $(\mathcal{X}_i, \mathcal{Y}_i)$ is a pair of samples drawn from input/output distributions: $\mathcal{X}_i \overset{iid}{\sim} P_i$, $\mathcal{Y}_i \overset{iid}{\sim} Q_i$. The estimated input/output densities p_i, q_i, were estimated using projection series estimators with the cosine basis: $\psi_0(x) \equiv 1$, $\psi_j(x) = \sqrt{2}\cos(j\pi x)$, $j \geq 1$. The kernel used for the regression weights was the triangle kernel: $(1-|x|)_+$. Parameters were selected by cross validating log likelihoods.

7.1. Synthetic Dataset

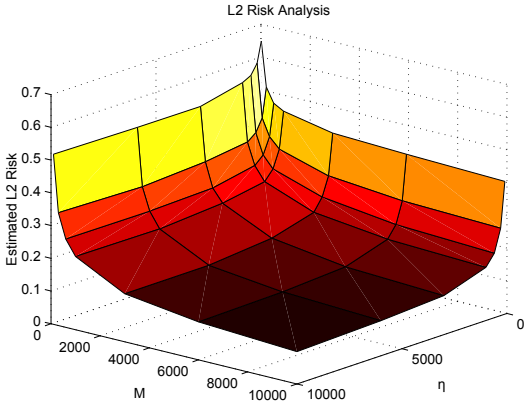
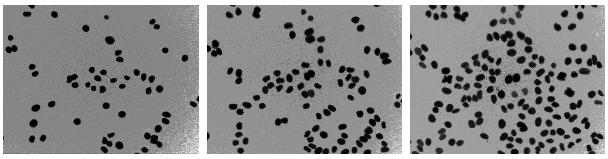


Figure 3. Grid of estimated $L2$ risks for varying values of M, η.

To better understand the effects of sample size quantities n, m and M, we generated synthetic datasets with varying sizes and tested the effectiveness of our estimator. The input/output distribution were created as follows: First we draw $\mu_1, \mu_2 \sim \text{Unif}[0,1]$ and $\sigma_1, \sigma_2 \sim \text{Unif}[.05, .1]$, then pdfs are $p(x) = \frac{1}{2}g(x;\mu_1,\sigma_1) + \frac{1}{2}g(x;\mu_2,\sigma_2)$, $q(x) = \frac{1}{2}g(x;1-\mu_1,\sigma_1) +$



(a) Frame 1 (b) Frame 50 (c) Frame 100

Figure 4. Cell images at different time frames.

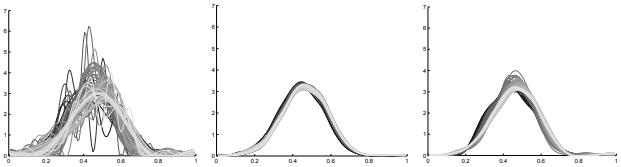
$\frac{1}{2}g(x; 1-\mu_2, \sigma_2)$ where g is the truncated normal pdf on $[0,1]$: $g(x;\mu,\sigma) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})/(\Phi(\frac{1-\mu}{\sigma}) - \Phi(\frac{-\mu}{\sigma}))$ with ϕ and Φ being the standard normal pdf and cdf (see Figures 2(a) and 2(b)).

A grid was populated evaluating our estimator by generating M pairs of (p_i, q_i) input/output densities as described above. Then from each (p_i, q_i), η points are drawn: $|\mathcal{X}_i| = \eta$, and $|\mathcal{Y}_i| = \eta$. That is, $m_i = n_i = \eta$. M and η were chosen to be in $\{100, 300, 600, 1000, 3000, 6000, 10000\}$. For each configuration of M, η the $L2$-risk is reported as the average $L2$-loss calculated for a separate test set of 5000 input/output sample pairs of η points (Figure 3). Not surprisingly the fastest direction to decrease the $L2$ risk is by increasing M, η simultaneously. It is also interesting to note that holding either M or η fixed and increasing the other decreases the risk, but eventually levels off. This is exactly predicted by our theory, since once either size M or η get much bigger than the other size, the smaller size drives the rate. Furthermore, one may see in Figures 2(c) and 2(d), that even for smaller sample sizes, the estimator still produces useful estimates.

7.2. Cell Dataset

Next, we used a dataset of a time-series of images of HeLa cells (Buck et al., 2009). A total of 100 time-frames were used, containing from 53 to 149 cells each. In each time-frame 49 image nuclear features were extracted from each cell. All features were rescaled to lie within $[0,1]$ (see Figure 4). At each time-frame, we look to regress the distribution of one of the nuclear features (e.g. short-axis length) when given the distribution of another nuclear feature in the time-frame (e.g. long-axis length). That is, we use $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^{100}$, where $(\mathcal{X}_i, \mathcal{Y}_i)$ are samples of the input/output feature at the i^{th} time-frame.

In addition to a distribution to distribution estimator (DDE), one may use a conditional distribution based estimator (CDE). That is, estimate the output distribution as follows: estimate the conditional distribution of a cell's output feature given the input feature, then when given a frame, estimate the output distribution as the average of the conditional distribution given the input feature values for each cell in the time-frame. Note that this CDE requires much more knowledge and special conditions than the DDE since a one to one mapping between input/output samples must exist[3], and one needs to know the mapping.



(a) Short-axis $\{\tilde{q}_i\}$ (b) CDE $\{\hat{f}(\tilde{p}_i)\}$ (c) DDE $\{\hat{f}(\tilde{p}_i)\}$

Figure 5. Densities at different time frames, lighter lines correspond to later time-frames. (a) Estimated output densities for a short-axis length feature with projection series estimators on each $\mathcal{Y}_i$. (b) CDE estimated output densities for frame i, $\hat{f}(\tilde{p}_i)$, holding out $(\mathcal{X}_i, \mathcal{Y}_i)$. (c) DDE estimated output densities for frame i holding out $(\mathcal{X}_i, \mathcal{Y}_i)$.

Since now we are estimating over a real-world dataset, no longer do we know the true density for samples. Instead, we compare the cross-validated log-likelihoods (CVLL) (using a holdout input/output sample pair) of the CDE to the DDE. We regress the mapping of the distribution of cell long-axis length to the distribution of cell short-axis length, where CDE yields a CVLL 6657.09 and DDE yields 6714.95 (see Figure 5). We note that although the CDE uses much more information, the DDE yields a better likelihood for estimated output distributions for unseen input distributions. Also, the DDE is able to capture change in distributions than CDE, which stays much more stationary. It may be of scientific interest to consider conditions under which one expects DDE to outperform CDE.

8. Discussion and Conclusion

In conclusion, we have provided an estimator for performing regression when both covariates and responses are distributions; also, upper bounds were derived for the risk of the estimator. No parametric assumptions were made on the input/output distribution, nor on the measure from which input distributions are drawn from in the estimator or upper bound results. Furthermore, if an assumption is made on the doubling dimension of the measure of input distributions $\mathcal{P}$, then we show that the $L2$ risk of the estimated output density converges at a polynomial rate. In future work we will derive lower bounds for the risk. Furthermore, we will test the performance of the estimator on other real-world datasets.

[3] E.g., this was not the case with the synthetic dataset.

References

Buck, T.E., Rao, A., Coelho, L.P., Fuhrman, M.H., Jarvik, J.W., Berget, P.B., and Murphy, R.F. Cell cycle dependence of protein subcellular location inferred from static, asynchronous images. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pp. 1016–1019. IEEE, 2009.

Ferraty, F. and Vieu, P. *Nonparametric functional data analysis: theory and practice*. Springer, 2006.

Ingster, Y. and Stepanova, N. Estimation and detection of functions from anisotropic sobolev classes. *Electronic Journal of Statistics*, 5:484–506, 2011.

Jaakkola, T.S., Haussler, D., et al. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pp. 487–493, 1999.

Jebara, T., Kondor, R., and Howard, A. Probability product kernels. *The Journal of Machine Learning Research*, 5:819–844, 2004.

Kpotufe, S. k-nn regression adapts to local intrinsic dimension. *arXiv preprint arXiv:1110.4300*, 2011.

Laurent, B. Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24(2):659–681, 1996.

Moreno, P.J., Ho, P., and Vasconcelos, N. A kullback-leibler divergence based kernel for svm classification in multimedia applications. *Advances in Neural Information Processing Systems*, 16:1385–1393, 2003.

Muandet, K., Schölkopf, B., Fukumizu, K., and Dinuzzo, F. Learning from distributions via support measure machines. *arXiv preprint arXiv:1202.6504*, 2012.

Nussbaum, M. On optimal filtering of a function of many variables in white gaussian noise. *Problemy Peredachi Informatsii*, 19(2):23–29, 1983.

Poczos, B., Rinaldo, A., Singh, A., and Wasserman, L. Distribution-Free Distribution Regression. *AISTATS 2013, arXiv preprint arXiv:1302.0082*, 2012.

Póczos, B., Xiong, L., and Schneider, J. Nonparametric divergence estimation with applications to machine learning on distributions. *arXiv preprint arXiv:1202.3758*, 2012a.

Póczos, B., Xiong, L., Sutherland, D.J., and Schneider, J. Nonparametric kernel estimators for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2989–2996. IEEE, 2012b.

Ramsay, J.O. and Silverman, B.W. *Applied functional data analysis: methods and case studies*, volume 77. Springer New York:, 2002.

Rigollet, P. and Vert, R. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15(4): 1154–1178, 2009.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pp. 13–31. Springer, 2007.

Tsybakov, Alexandre B. *Introduction to nonparametric estimation*. Springer, 2008.