1 Introduction

Low-dimensional embedding, manifold learning, clustering, classification, and anomaly detection are among the most important problems in machine learning. The existing methods usually consider the case when each instance has a fixed, finite-dimensional feature representation. We wish to expand the domain of consideration and let each instance correspond to a continuous probability distribution or a function from a nonparametric class in general. At times, these distributions are unknown, but we are given some i.i.d. samples from each distribution. Specifically, we wish to expand the domain of both covariates (inputs) and response (outputs) from real-valued pairs $<X_i, Y_i>$ to distributions.

A common approach to dealing with distributions is to embed the distribution to a Hilbert space, introduce kernels between the distributions, and then use a traditional kernel machine to solve the learning problem. There are both parametric and nonparametric methods proposed in the literature.

Parametric methods usually fit a parametric family (e.g. Gaussians distributions or exponential family) to the densities, and using the fitted parameters they estimate the inner products between the distributions. The problem with parametric approaches, however, is that when the true densities do not belong to the assumed parametric families, then this method introduces some unavoidable bias during the estimation of the inner products between the densities.

A couple of nonparametric approaches exist as well. Since our covariates are represented by finite sets, reproducing kernel Hilbert space (RKHS) based set kernels can be used in these learning problems.

2 Summary

Figure 1: Extension of the standard regression model. Three models: 1) distribution-real (left), distributions $P_1, \ldots, P_m, P_{m+1}$ are unobserved, only the $X_1, \ldots, X_m, X_{m+1}$ sample sets are observable. 2) distribution-distribution (middle), dataset $D = \{(X_i, Y_i)\}_{i=1}^M$ of pairs of sets where $P_i$ and $Q_i = f(P_i)$ are unobserved, instead one observes samples $X_i \sim P_i$ and $Y_i \sim Q_i$. $P_0$ is an unseen query distribution, observed indirectly through $X_0$. We look to estimate the output distribution $Q_0 = f(P_0)$. 3) function-real (right), sparse model where response $Y$ is dependent on a sparse subset of input functions $f_1, \ldots, f_p$. 
We’ve studied distribution to real regression [1], where one aims to regress a mapping \( f \) that takes in a distribution input covariate \( P \subseteq \mathcal{I} \) (for a non-parametric family of distributions \( \mathcal{I} \)) and outputs a real-valued response \( Y = f(P) + \varepsilon \).

We’ve also studied the distribution to distribution regression [2] 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.

Besides the two extensions to the standard regression model, another extension concerns functional covariate.


3 Distribution to Real Regression

We start by laying out the notations and then give the formal definition of the problem. We then present the key results in the paper. Due to space limitation, we leave out the seven assumptions, which can be found in paper [1].

3.1 Notation and Terminology

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m )</td>
<td>number of input distributions in the training set</td>
</tr>
<tr>
<td>( (P_1, Y_1), \ldots, (P_m, Y_m) )</td>
<td>covariate and response pairs, where ( Y_i \in \mathbb{R} )</td>
</tr>
<tr>
<td>( P_i )</td>
<td>a probability distribution on a compact subset ( K \subseteq \mathbb{R}^k )</td>
</tr>
<tr>
<td>( Y_i \sim P_i )</td>
<td>where ( i = 1, \ldots, m ) and ( \mu_i ) is a noise variable with mean 0</td>
</tr>
<tr>
<td>( X_{i1}, \ldots, X_{im_i} \overset{i.i.d.}{\sim} P_i )</td>
<td>samples from the ( i )th training distribution</td>
</tr>
<tr>
<td>( (X_i^1, Y^1_1), \ldots, (X_i^{m_i}, Y_i^{m_i}) )</td>
<td>observed data, where ( X_i^1 = X_{i1}, \ldots, X_{i m_i} )</td>
</tr>
<tr>
<td>( X_{m+1} \subseteq I )</td>
<td>a new batch drawn from a distribution ( P_{m+1} )</td>
</tr>
<tr>
<td>( Y_{m+1} \subseteq \mathbb{R} )</td>
<td>response for ( X_{m+1} )</td>
</tr>
<tr>
<td>( n )</td>
<td>( \min_{1 \leq i \leq m+1} n_i )</td>
</tr>
<tr>
<td>( \hat{f} )</td>
<td>an estimator for the unknown function ( f )</td>
</tr>
<tr>
<td>( \hat{P}_i )</td>
<td>an estimator of ( P_i ) based on ( X_i )</td>
</tr>
<tr>
<td>( \hat{Y}<em>{m+1} = \hat{f} (\hat{P}</em>{m+1}) )</td>
<td>an estimate of ( Y_{m+1} )</td>
</tr>
<tr>
<td>( B, K )</td>
<td>two kernels</td>
</tr>
<tr>
<td>( B(P, h) = { \hat{P} \subseteq \mathbb{D} : D(\hat{P}, P) \leq h } )</td>
<td>( L_1 ) ball of distributions around ( P ) with radius ( h )</td>
</tr>
<tr>
<td>( \Phi_P(h) := \mathbb{P} (B(P, h)) )</td>
<td>small ball probability, where ( P ) is a fixed distribution</td>
</tr>
<tr>
<td>( D(P, Q) )</td>
<td>( L_1 ) distance of the densities of ( P ) and ( Q )</td>
</tr>
</tbody>
</table>

3.2 Problem Definition

We define an estimator \( \hat{f} \) for the unknown function \( f \). Let \( \hat{P}_i \) denote an estimator of \( P_i \) based on \( X_i \), \( (1 \leq i \leq m+1) \).

Given a bandwidth \( h \geq 0 \) and a kernel function \( K \), we define:

\[
\hat{f}(\hat{P}_{m+1}) = \hat{f}(\hat{P}_{m+1}; \hat{P}_1, \ldots, \hat{P}_m) = \left\{ \frac{\sum_{i=1}^{m} Y_i K(\frac{D(P_i, P)}{h})}{\sum_{i=1}^{m} K(\frac{D(P_i, P)}{h})} \right\} \quad \text{if} \sum_{i=1}^{m} K(\frac{D(P_i, \hat{P}_i)}{h}) > 0 \text{ o.w.}
\]

We estimate the density \( \hat{p}_i \) of \( P_i \) with a kernel density estimator, using kernel function \( B \), and bandwidth \( b_i \). The overall goal is to upper bound the risk

\[
R(m, n) = E[\| \hat{f}(\hat{P}; \hat{P}_1, \ldots, \hat{P}_m) - f(P) \|]
\]

3.3 Theorems

Some important assumptions are:

1. \( f \) belongs to the class \( \mathcal{M}(L, \beta, D) \) of Holder continuous functionals on \( \mathbb{D} \); \( 0 < \beta \leq 1 \).
2. Kernel \( K : [0, \infty) \to \mathbb{R} \) is non-negative and Lipschitz continuous. Also, there exist \( 0 < K < 1 \) and \( 0 < r < R < \infty \) such that, for all \( x > 0 \), \( K I_{\{x \in B(0, r)\}} \leq K(x) \leq I_{\{x \in B(0, R)\}} \)
Theorem 1 provides a general upper bound on the risk.

**Theorem 1** Suppose the assumptions hold. Then

\[
R(m, n) \leq \frac{1}{h} \frac{1}{\Phi_p(rh/2)} [C_1 n^{-\frac{1}{2}} + C_2 h^\beta + C_3 \sqrt{\frac{1}{m} \frac{1}{\Phi_p(rh/2)}}] + C_4 \frac{1}{m} \frac{1}{\Phi_p(rh/2)} + (m + 1) e^{-\frac{n}{2} \pi^d}
\]

When the effective dimension of \( P \) as measured by the doubling dimension \( d \) is small, the risk converges to zero. This is summarized in Theorem 2.

**Theorem 2**

\[
R(m, n) \leq \frac{C}{h^{d+1} n^{1/(k+2)}} + C_2 h^\beta + C_3 \frac{1}{m h^d}
\]

The rates depend on whether the third term dominates the first term or not.

\[
R(m, n) = \begin{cases} 
O(m^{-\beta/(2\beta+d)}) & \text{if } \sqrt{\frac{1}{m h^d}} = \Omega(\frac{C_1}{h^{d+1} n^{1/(k+2)}}) \\
O(n^{-\frac{1}{(k+2)(\beta+d)}}) & \text{if } \sqrt{\frac{1}{m h^d}} = O(\frac{1}{n^{(k+1)/(k+2)}})
\end{cases}
\]

### 3.4 Proof outlines

Define the following events

\[
E_0 = \{ \sum_i K_i = 0 \}, \quad E_1 = \{ 0 < \sum_i K_i < K \}, \quad E_2 = \{ K \leq \sum_i K_i \} \\
\hat{E}_0 = \{ \sum_i \hat{K}_i = 0 \}, \quad \hat{E}_1 = \{ 0 < \sum_i \hat{K}_i < K \}, \quad \hat{E}_2 = \{ K \leq \sum_i \hat{K}_i \}
\]

(1)

Use triangle inequality to write

\[
R(m, n) = E[|\hat{f}(\hat{P}; \hat{P}_1, \ldots, \hat{P}_m) - f(P)|] \\
\leq E[|\hat{f}(\hat{P}; \hat{P}_1, \ldots, \hat{P}_m) - \hat{f}(P; P_1, \ldots, P_m)|] + E[|\hat{f}(P; P_1, \ldots, P_m) - f(P)|]
\]

(2)

(3)

Derive the upper bound of equation(2) and equation(3)

#### 3.4.1 Upper bound on equation(2)

The idea is first realizing four possible cases and developing an upper bound for each.

1. \( \sum_i K_i > 0 \) and \( \sum_i \hat{K}_i > 0 \), then \( \Delta \hat{f} = |\frac{Y_i \hat{K}_i}{\sum_i \hat{K}_i} - \frac{Y_i K_i}{\sum_i K_i}| \)

\[
E[\Delta \hat{f} I_{E_2} I_{\hat{E}_2}] \leq C_1 \frac{1}{h} \frac{1}{\Phi_p(rh)} [n^{-\frac{1}{2}} + \sqrt{\frac{1}{m} \frac{1}{\Phi_p(rh)}}]
\]

2. \( \sum_i K_i > 0 \) and \( \sum_i \hat{K}_i = 0 \), then \( \Delta \hat{f} = |\frac{Y_i \hat{K}_i}{\sum_i \hat{K}_i}| \)

\[
E[\Delta \hat{f} I_{E_0} (I_{E_2} + I_{\hat{E}_2})] + E[\Delta \hat{f} I_{\hat{E}_1} (I_{E_2} + I_{\hat{E}_2})] \leq B_Y \zeta + 2B_Y \zeta = 3B_Y \zeta
\]

3. \( \sum_i K_i = 0 \) and \( \sum_i \hat{K}_i > 0 \), then \( \Delta \hat{f} = |\frac{Y_i \hat{K}_i}{\sum_i \hat{K}_i}| \)

\[
E[\Delta \hat{f} I_{E_0} (I_{E_2} + I_{\hat{E}_2})] + E[\Delta \hat{f} I_{E_1} (I_{E_2} + I_{\hat{E}_2})] \leq \frac{2B_Y}{cm} \int \frac{dP(P)}{\Phi_p(rh)} + \frac{B_Y}{cm} \int \frac{dP(P)}{\Phi_p(rh)} = 3B_Y \frac{1}{em} \int \frac{dP(P)}{\Phi_p(rh)}
\]

4. \( \sum_i K_i = 0 \) and \( \sum_i \hat{K}_i = 0 \), then \( \Delta \hat{f} = 0 \)

The upper bound follows by putting the pieces together.
3.4.2 Upper bound on equation

First note that

\[
E[|\hat{f}(P; P_1, \ldots, P_m) - f(p)|] \\
\leq \mathbb{E}\left[\sum_i |f(P_i) - f(P)| K_i |I_{\{\sum_i, K_i > 0\}}\right] + \mathbb{E}\left[\sum_i \mu_i K_i |I_{\{\sum_i, K_i > 0\}}\right] + f_{\text{max}} P(\sum_i K_i = 0)
\]

Proceed with an upper bound on each item

1. 1st item \( \leq L(hR)\beta \)
2. 2nd item \( \leq \mathbb{E}\left[\sum_i \mu_i K_i |I_{\{\sum_i, K_i > 0\}}\right] + \frac{B\psi}{em} \int \frac{dP(p)}{\Phi_P(rh)} \)
3. 3rd item \( \leq f_{\text{max}} \frac{em}{\Phi_P(rh)} \int \frac{dP(p)}{\Phi_P(rh)} \)

3.5 Conclusion

If the number of samples per distribution, \( n \) is large compared to the number \( m \) of distributions, then the learning rate is limited by the number of distributions \( m \) and is in fact precisely the same as the rate of learning a standard \( \beta \)-Holder smooth regression function in \( d \) dimensions. That is, the effect of not knowing the distributions \( P_1, \ldots, P_m \) exactly and only having a finite sample from the distributions is negligible.

When the number of samples per distribution is not large enough, the rate is limited by \( n \), as expected. Notice that the rate gets worse as the dimensionality of each distribution \( k \) grows and as the smoothness \( \beta \) of the regression function deteriorates.

4 Distribution to Distribution Regression

This paper [2] expands paper [1] in the sense that not only the input covariate but also the output response are probability distributions.

4.1 Notation and Terminology


- I: a class of input distributions on \( \Psi^k \subseteq \mathbb{R}^k \)
- O: a class of output distributions on \( \Lambda^l \subseteq \mathbb{R}^l \)
- \( (X_i, Y_i) \): observed i.i.d. samples of input/output distribution from \( P_i \) and \( Q_i \), \( 1 \leq i \leq M \)
- \( Y_{ij} \sim Q_i \)
- \( Q_i \): empirical distribution of \( Y_i \)
- \( \{\phi_i\}_{i \in \mathbb{Z}} \): an orthonormal basis for \( L_2(\Lambda) \)
- \( \hat{Q}_i \): estimator of \( Q_i \)
- \( \hat{X}_0 \): a sample from the query distribution \( P_0 \sim \mathcal{P} \)
- \( \kappa^2(\nu, \sigma) \)
- \( \nu \)
- \( \sigma^{-1} \)

Define weights as

\[
W(\hat{P}_i, \hat{P}_0) = \begin{cases} 
\frac{K(D(\hat{P}_i, \hat{P}_0))}{\sum_{j=1}^M K(D(\hat{P}_j, \hat{P}_0))} & \text{if } \sum_{j=1}^M K(D(\hat{P}_j, \hat{P}_0)) > 0 \\
0 & \text{o.w.}
\end{cases}
\]

where \( D \) is a metric and \( K \) is a kernel function satisfying assumption [2].
4.2 Problem Definition

We define an estimator \( \hat{f}(\hat{p}_0) = \sum_{i=1}^{M} \hat{q}_i W(\hat{P}_i, \hat{P}_0) \) for the pdf of \( f(P_0) \), where the estimate \( \hat{q}_i \) is made using an orthogonal series estimator \( \hat{q}_i(x) = \sum_{\alpha:K_{\alpha}(\nu, \sigma) \leq i} a_{\alpha}(Q_i) \phi_\alpha(x) \)

\[
\varphi_\alpha(x) = \prod_{i=1}^{l} \varphi_{\alpha_i}(x_i), x \in \mathcal{N}, \alpha \in \mathbb{Z}^l
\]

\[ q(x) = \sum_{\alpha \in \mathbb{Z}^l} a_{\alpha}(Q) \varphi_\alpha(x) \]

\[ a_{\alpha}(Q) = \langle \varphi_\alpha, q \rangle = \int_{\mathcal{N}} \varphi_\alpha(z) dQ(z) \in \mathbb{R} \]

\[ \hat{a}_{\alpha} = \sum_{i=1}^{M} a_{\alpha}(Q_i) W(\hat{P}_i, \hat{P}_0) \]

\[
\bar{a}_{\alpha} \equiv \begin{cases} 
\frac{\sum_{i=1}^{M} a_{\alpha}(Q_i) K_{\alpha}(D(P_0, P_i))}{\sum_{j=1}^{M} K_{\alpha}(D(P_0, P_j))} & \text{if } \sum_{j} K(D(P_0, P_j)) > 0 \\
0 & \text{else}
\end{cases}
\]

The goal is to upper bound the \( L_2 \) risk of the estimator \( \hat{f}(\hat{p}_0) \) using the \( L_1 \) metric for \( D \) and kernel density estimation for \( \{\hat{P}_i\}_{i=0}^{M} \).

\[
\mathbb{E}[||\hat{f}(\hat{p}_0) - f(p_0)||_2] \leq \sum_{\alpha \in A_\nu} \mathbb{E}[|\hat{a}_{\alpha} - a_{\alpha}(f(P_0))|] + \mathbb{E}[\sqrt{\sum_{\alpha \in A_\nu} a_{\alpha}^2(f(P_0))}]
\]

Let \( R_\alpha(M, n, m) = \mathbb{E}[|\hat{a}_{\alpha} - a_{\alpha}(f(P_0))|] \). We look to find \( R(M, n, m) \) s.t. \( \forall \alpha \in \mathbb{Z}^l, R_\alpha(M, n, m) \leq R(M, n, m) \).

4.3 Theorems

Theorem 3

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

Let \( \Delta \hat{a}_{\alpha} = |\hat{a}_{\alpha} - \bar{a}_{\alpha}| \).

Bound the first term,

\[
\mathbb{E}[\Delta \hat{a}_{\alpha}] \leq C_1 n^{-\frac{\nu+1}{2}} \mathbb{E}[\frac{1}{\Phi_P(\nu h)}] n^{-\frac{\nu+1}{2}} + C_2 \mathbb{E}[\frac{1}{\Phi_P(\nu h/2)}] + (M + 1) e^{-\frac{1}{2} n^{\frac{\nu+1}{2}}}
\]

Bound the second term

\[
\mathbb{E}[|\hat{a}_{\alpha} - a_{\alpha}(f(P_0))|] \leq C_3 h^{\beta} + C_4 \sqrt{\frac{1}{m M} \mathbb{E}[\frac{1}{\Phi_P(\nu h/2)}]}
\]

\[
+ \frac{C_5}{\sqrt{m M} \mathbb{E}[\frac{1}{\Phi_P(\nu h/2)}]} + \frac{C_6}{M \mathbb{E}[\frac{1}{\Phi_P(\nu h/2)}]}
\]

Theorem 4

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

4.4 Proof outlines of Theorem 3

Define the events same as 1. Use triangle inequality to write

\[
R_\alpha(M, n, m) = \mathbb{E}[|\hat{a}_{\alpha} - a_{\alpha}(f(P_0))|]
\]

\[
\leq \mathbb{E}[|\hat{a}_{\alpha} - \bar{a}_{\alpha}|] + \mathbb{E}[|\bar{a}_{\alpha} - a_{\alpha}(f(P_0))|]
\]

Derive the upper bound of equation 4 and equation 5.
4.4.1 Upper bound on equation 4

The idea is first realizing four possible cases and developing an upper bound for each.

1. $\sum_{i} K_i > 0$ and $\sum_{i} \hat{K}_i > 0$, then
   \[ \mathbb{E}[\Delta \hat{a}_i I_{E_2} I_{E_1}] \leq C_1 \frac{1}{h} \mathbb{E}[\frac{1}{\Phi_P(rh)}] n^{-\frac{3}{2} + \frac{3}{2}} \]

2. $\sum_{i} K_i > 0$ and $\sum_{i} \hat{K}_i = 0$, then
   \[ \mathbb{E}[\Delta \hat{a}_i I_{E_0}(I_{E_1} + I_{E_2})] + \mathbb{E}[\Delta \hat{a}_i I_{E_1}(I_{E_1} + I_{E_2})] \leq \varphi_{max} \zeta(m, M) + 2 \varphi_{max} \zeta(m, M) = 3 \varphi_{max} \zeta(m, M) \]

3. $\sum_{i} K_i = 0$ and $\sum_{i} \hat{K}_i > 0$, then
   \[ \mathbb{E}[\Delta \hat{a}_i I_{E_0}(I_{E_1} + I_{E_2})] + \mathbb{E}[\Delta \hat{a}_i I_{E_1}(I_{E_1} + I_{E_2})] \leq \frac{\varphi_{max}}{eM} \mathbb{E}[\frac{1}{\Phi_P(rh)}] + \frac{2 \varphi_{max}}{eM} \mathbb{E}[\frac{1}{\Phi_P(rh)}] = \frac{3 \varphi_{max}}{eM} \mathbb{E}[\frac{1}{\Phi_P(rh)}] \]

4. $\sum_{i} K_i = 0$ and $\sum_{i} \hat{K}_i = 0$, then $\Delta \hat{a}_i = 0$

The upper bound follows by putting the pieces together.

4.4.2 Upper bound on equation 5

Note that
\[ \mathbb{E}[\hat{a}_i - a_\alpha(f(P_0))] \leq \mathbb{E}[\sum_{i} (a_\alpha(f(P_i)) - a_\alpha(f(P_0))) K_i I_{\{\sum_i K_i > 0\}}] + \mathbb{E}[\sum_{i} \mu_i(i) K_i I_{\{\sum_i K_i > 0\}}] + \mathbb{E}[a_\alpha(f(P_0)) I_{\{\sum_i K_i = 0\}}] \]

Proceed with an upper bound on each term

1. 1st term \( \leq L(hR)^\beta \)
2. 2nd term \( \leq C \sqrt{\frac{1}{mM} \mathbb{E}[\frac{1}{\Phi_P(rh)}]} + \frac{\varphi_{max}}{eM} \mathbb{E}[\frac{1}{\Phi_P(rh)}] \]
3. 3rd term \( \leq \frac{4}{eM} \mathbb{E}[\frac{1}{\Phi_P(rh)}] \)

The result follows by combining the above.

4.5 Proof outlines of Theorem 4

Start by showing
\[ \|\hat{f}(P_0) - f(P_0)\|_2 \leq \sum_{\alpha \in A} |\hat{a}_\alpha - a_\alpha(f(P_0))| + \sqrt{\sum_{\alpha \in A^C} a_\alpha^2(f(P_0))} \]

Thus the bound of the $L_2$ risk is
\[ \mathbb{E}[\|\hat{f}(P_0) - f(P_0)\|_2] \leq |\mathcal{A}| R(M, n, m) + \sqrt{\mathbb{E}[\sum_{\alpha \in A^C} a_\alpha^2(f(P_0))]} \]
\[ \leq CR(M, n, m) t^{\sigma - 1} + \frac{\sqrt{\mathcal{A}}}{t} \]

where $t \sim R(M, n, m)$
4.6 Conclusion

If \( n = \Omega(M^{(\frac{d+d+1}{d+1})^{(k+2)}}) \), then \( M \) is slow growing, and \( R(M, n, m) = O(M^{\frac{-d}{d+1}}) \). If \( M = \Omega(n^{(\frac{d}{d+1})^{(k+2)}}) \), then \( n \) is slow growing, and \( R(M, n, m) = O(n^{\frac{-d}{d+1}}) \). If \( P \) is a doubling dimension, then the rate of convergence for \( E[\|\hat{f}(p_0) - f(p_0)\|_2] \) is polynomial in \( M, n, m \).

5 Nonparametric Divergence Estimation

The goal is to embed distributions into a lower dimensional space without estimating their densities.

\( p, q \)
\( X_{1:n} \equiv (X_1, \ldots, X_N) \) i.i.d samples from \( p \)
\( Y_{1:n} \equiv (Y_1, \ldots, Y_N) \) i.i.d samples from \( q \)
\( \rho_k(i) \) Euclidean distance of the \( k \)th nearest neighbor of \( X_i \) in the sample \( X_{1:n} \setminus x \)
\( \upsilon_k(i) \) Euclidean distance of the \( k \)th nearest neighbor of \( X_i \) in the sample \( Y_{1:m} \setminus x \)

The Rényi divergence is

\[
D_\alpha(p||q) = \int_M \left( \frac{q(x)}{p(x)} \right)^{1-\alpha} p(x) dx
\]

where \( M = \text{supp}(p) \) and \( \alpha \in \mathbb{R} \setminus \{1\} \)

The \( L^2 \) divergence is

\[
L^2(p||q) = \int_M (p(x) - 2q(x) + q(x)^2/p(x))p(x) dx
\]

The estimator for the Rényi divergence is

\[
\hat{D}_\alpha(X_{1:n}||Y_{1:M}) = \frac{1}{N} \sum_{n=1}^N \frac{(N-1)\rho_k^n(X_n)}{M\upsilon_k^n(X_n)} \left( 1 - \alpha \right) B_{k,\alpha}
\]

where \( B_{k,\alpha} = \frac{\Gamma(k)^2}{\Gamma(k-\alpha+1)\Gamma(k+\alpha-1)}, k > |\alpha - 1| \)

For \( k > 2 \), the estimator for the \( L^2 \) divergence is

\[
\hat{L}^2(X_{1:n}||Y_{1:M}) = \frac{1}{N} \sum_{n=1}^N \frac{k - 1}{(N-1)\rho_k^n(X_n)} - \frac{2(k-1)}{M\upsilon_k^n(X_n)} \frac{(N-1)\rho_k^n(X_n)(k-2)(k-1)}{k} / (M\upsilon_k^n(X_n))^2
\]

They claim both estimators are consistent under certain conditions due to the following theorems.

5.1 Rényi divergence estimator : unbiased and consistent

Theorem 5
\[
\lim_{N,M \to \infty} \mathbb{E}[\hat{D}_\alpha] = D_\alpha \quad \lim_{N,M \to \infty} \mathbb{E}[\left( \hat{D}_\alpha - D_\alpha \right)^2] = 0
\]

5.2 \( L^2 \) divergence estimator : unbiased and consistent

Theorem 6
\[
\lim_{N,M \to \infty} \mathbb{E}[\hat{L}^2] = L^2 \quad \lim_{N,M \to \infty} \mathbb{E}[\left( \hat{L}^2 - L^2 \right)^2] = 0
\]

5.3 Proof outlines

Start by defining \( k \)-NN based estimators

\[
\hat{p}_k(x) = \frac{k}{N\rho_k^n(x)} \\
\hat{q}_k(x) = \frac{k}{M\upsilon_k^n(x)}
\]

Use theorems, one can show the estimators are consistent, under the condition \( k(N) \to \infty \). The authors keep \( k \) fixed and prove the consistency of the estimators as follows. They first argue that they do not need to apply consistent
density estimators by realizing equation 6 and 7 have special forms \( \int p(x)p^\gamma(x)q^\beta(x)dx \). In 8 and 9, each of terms is estimated by

\[
\frac{1}{N} \sum_{i=1}^{N} (\hat{p}_k(X_i))^\gamma (\hat{q}_k(X_i))^\beta \quad B_{k,\gamma,\beta}
\]

With the help of two lemmas–Lebesgue and Moments of the Erlang distribution– \( B_{k,\gamma,\beta} \) is a correction factor that ensures asymptotic unbiasedness.

We refer interested readers to the paper [6] for proof detail.

6 Application to Image Classification

The images are represented by unordered, multidimensional, finite sets of feature vectors. The elements of the sets are i.i.d. samples from unknown distributions. Estimate kernel functions between those distributions, then apply kernel classifier for classification.

\[
K(p,q) \quad \text{kernel values between distribution } p \text{ and distribution } q
\]

\[
D_{\alpha,\beta}(p||q) \quad \text{estimator of } K(p,q), \alpha, \beta \in \mathbb{R}
\]

\[
X_{1:n} = (X_1, \ldots, X_n) \quad \text{i.i.d samples from } p
\]

\[
Y_{1:m} = (Y_1, \ldots, Y_m) \quad \text{i.i.d samples from } q
\]

\[
p_k(i) \quad \text{Euclidean distance of the } k\text{th nearest neighbor of } X_i \text{ in the sample } X_{1:n}
\]

\[
v_k(i) \quad \text{Euclidean distance of the } k\text{th nearest neighbor of } X_i \text{ in the sample } Y_{1:m}
\]

6.1 Asymptotic unbiasedness

Theorem 7

\[
\lim_{n,m \to \infty} E[\hat{D}_{\alpha,\beta}(X_{1:n}||Y_{1:m})] = D_{\alpha,\beta}(p||q)
\]

6.2 \( L_2 \) consistency

Theorem 8

\[
\lim_{n,m \to \infty} E[(\hat{D}_{\alpha,\beta}(X_{1:n}||Y_{1:m}) - D_{\alpha,\beta}(p||q))^2] = 0
\]

7 FuSSO

The key idea of FuSSO, a functional analogue to the LASSO, is finding a sparse set of functional input covariates to regress a real-valued response against.

7.1 Estimator for the coefficient

\[
\hat{\beta} = \arg \min_{\beta} \frac{1}{2N} ||Y - \sum_{j=1}^{p} \hat{A}_j \beta_j||^2 + \lambda_N \sum_{j=1}^{p} ||\beta_j||
\]

where \( \hat{A}_j \) is the \( N \times M_n \) with values \( \hat{A}_j(i,m) = \hat{a}_{jm}^{(i)} \) and \( \beta_j = (\beta_{j1}, \ldots, \beta_{jM_n})^T \)

7.2 Sparsistency

The paper goes on proving the key theorem

Theorem 9

\[
\mathcal{P}(\hat{S}_N = S) \to 1
\]

where \( S \) denotes the true set of non-zero functions, and \( \hat{S}_N \) denotes the true set of non-zero functions from \( N \) samples.

We refer interested readers to paper [5] for proof detail.

8 Work Distribution

Together, we read the papers and wrote the report.
References


