A CONDITIONAL DISTRIBUTION FUNCTION BASED APPROACH TO DESIGN NONPARAMETRIC TESTS OF INDEPENDENCE AND CONDITIONAL INDEPENDENCE

Sohan Seth and José C. Príncipe

Computational NeuroEngineering Laboratory, University of Florida, Gainesville, USA
sohan@cnel.ufl.edu, principe@cnel.ufl.edu

ABSTRACT

Measures of independence and conditional independence are two important statistical concepts that have found profound applications in engineering such as in feature selection and causality detection, respectively. Therefore, designing efficient ways, typically nonparametric, to estimate these measures has been an active research area in the last decade. In this paper, we propose a novel framework to test (conditional) independence, using the concept of conditional distribution function. Although, estimating conditional distribution function is a difficult task on its own, we show that the proposed measures can be estimated efficiently and actually can be expressed as the Frobenius norm of a matrix. We compare the proposed methods with other state-of-the-art techniques and show that they yield very promising results.

Index Terms— Causality, conditional distribution function, conditional independence, estimation, independence, kernel method, nonparametric method.

1. INTRODUCTION

A measure of independence is an important concept that has recently received considerable attention in the area of machine learning due to its potential application in many practical problems such as regression, clustering, feature selection, independent component analysis, etc [1, 2, 3, 4]. Similarly, conditional independence is also an important concept in many different aspects of engineering such as detecting causal relationship between two time series and extracting informative features in a classification or regression setting [2, 5, 6]. Recently, with the advent of kernel based methods, a number of independence tests have been proposed, and have been successfully applied in many practical problems [1, 3]. A few of these tests have been extended for testing conditional independence [2, 5]. However, these tests involve a free parameter, namely the regularization parameter, that affects the performance of the test. Therefore, the search for a more robust test of conditional independence is still active. In this paper, we propose a novel framework for testing conditional independence. The proposed test can be used to test independence as a special case. Although the proposed tests involve a regularization parameter, we show that the performance of these tests are less sensitive to this free parameter. We compare the proposed tests with the state-of-the-art kernel based tests and show that they yield very promising results.

In this paper, we propose a novel framework to design tests of independence and conditional independence using the concept of conditional distribution function. A distribution function based approach has several advantages over a density function based one, since distribution function is always well defined even if the corresponding density function does not exist. Such situations occur, for example, when the random variables are degenerate. Moreover, a distribution function based approach is a better choice when the random variables involve both discrete and continuous values, as, in such cases, estimation of density function breaks down. Inspired by these facts, we propose an approach to design tests of independence and conditional independence based on conditional distribution function. However, a conditional distribution function based approach is rather difficult since the estimation of a conditional statistic is difficult than its unconditional counterpart and conditional distribution is no exception. But, we consider this approach since our objective is not to estimate this statistic precisely. We propose a novel approach to estimate, perhaps not precisely, this quantity and show that it leads to an efficient estimator of conditional independence and dependence. To be more specific, we show that the final estimator is just the Frobenius norm of a matrix. Finally, we apply these methods to a number of synthetic data and show that they perform very well compared to other state-of-the-art techniques and actually are more robust to the choice of free parameters.

The paper is organized as follows. In the next section, we propose measures of independence and conditional independence using conditional distribution function and derive appropriate estimators for these measures. In the following section, we provide rigorous simulation results to corroborate the proposed ideas and compare them with state-of-the-art kernel based techniques. In the final section, we conclude the paper and provide some guidelines for future work.

2. INDEPENDENCE AND CONDITIONAL INDEPENDENCE

Given random variables \((X, Y, Z)\), \((X, Y)\) is said to be conditionally independent given \(Z\) if and only if

\[
P(X < x|Y = y, Z = z) = P(X < x|Z = z)
\]

for all values \((x, y, z)\). Unconditional independence is a special case of conditional independence when the conditioning variable \(Z\) is missing (or, say, is a constant) i.e. \((X, Y)\) are independent if and only if

\[
P(X < x|Y = y) = P(X < x)
\]

for all \((x, y)\).

2.1. Measures of independence and conditional independence

A measure of (conditional) independence is characterized by the property that it attains zero value if and only if the random variables

---

This work is supported by NSF grant ECS 0601271
are (conditionally) independent. Therefore, the following quantities,
\[ M_{CI}^2 = \int (P(X < u | Y = v, Z = w) - P(X < u | Z = w))^2 \]
d\(F_X(u) dF_Y(v, w) \)
and
\[ M_I^2 = \int (P(X < u | Y = v) - P(X < u))^2 dF_X(u) dF_Y(v) \]
are measures of conditional independence and independence, respectively. Note that, we treat \( P(X < u | Z = w) \) and \( P(X < u) \) as functions of three and two variables, respectively; however, these functions do not change in the direction of the missing variable.

### 2.2. Estimators of independence and conditional independence

Let \( g_u(v, w) \) and \( h_u(w) \) be estimates of \( P(X < u | Y = v, Z = w) \) and \( P(X < u | Z = w) \) respectively. Then an estimator of \( M_{CI} \) is given by,
\[ \hat{M}_{CI}^2 = \int (g_u(v, w) - h_u(w))^2 dF_X(u) dF_Y(v, w). \]
Then, using triangle inequality
\[ |M_{CI} - \hat{M}_{CI}| < \]
\[ (\int (P(X < u | Y = v, Z = w) - P(X < u | Z = w))^2 dF_X(u) dF_Y(v, w))^{\frac{1}{2}} \]
\[ + (\int (P(X < u | Y = v, Z = w) - g_u(v))^2 dF_X(u) dF_Y(v, w))^{\frac{1}{2}} \]
This inequality shows that the absolute difference between the actual and estimated \( M_{CI} \) is upper bounded by the distance between the actual and estimated conditional distribution functions. Therefore, we choose the estimators, \( g_u(v, w) \) and \( h_u(w) \), in such a way that the corresponding distances are minimized.

Let us consider the general problem of finding estimate \( g_u(v) \) of \( P(X < u | Y = v) \) such that the distance
\[ J^2 = \int (P(X < u | Y = v) - g_u(v))^2 dF_Y(v) \]
is minimized. Since \( F_X(u) \) is a distribution function, \( J^2 \) is minimized if
\[ J^2(u) = \int (P(X < u | Y = v) - g_u(v))^2 dF_Y(v) \]
is minimized for all \( u \). Note that \( P(X < u) = EI(X < u) \) where \( I \) is the identity function. Using this equality, we expand \( J^2(u) \) to get,
\[ J^2(u) = \int (P(X < u | Y = v) - g_u(v))^2 dF_Y(v) = C - 2 \int P(X < u | Y = v) g_u(v) dF_Y(v) + \int g_u^2(v) dF_Y(v) \]
\[ = C - 2 \int I(u' < u) g_u(v) dF_X(Y(u', v) + \int g_u^2(v) dF_Y(v) \]
where \( C = \int P^2(X < u | Y = v) dF_Y(v) \) is a constant. Assume that we have the following model
\[ g_u(v) = \sum_{i=1}^{m} \alpha_i^u \phi_i(v) \]
where \( \{\phi_i\}_{i=1}^{m} \) is a set of appropriate basis functions and \( \{\alpha_i^u\}_{i=1}^{m} \) is a set of coefficients. Then, using strong law of large numbers,
\[ J^2(u) = C \]
\[ \approx - \frac{2}{n} \sum_{j=1}^{n} \sum_{i=1}^{m} \alpha_i^u \phi_j(y_j) + \frac{1}{n} \sum_{j=1}^{m} \sum_{k=1}^{m} \alpha_i^u \alpha_j^u \phi_i(y_k) \phi_j(y_k) \]
\[ = - \frac{2}{n} \| \Phi \alpha_u + \frac{1}{n} \alpha_u^T \Phi \alpha_u \]
where \( \{x_i, y_i\}_{i=1}^{n} \) are realizations of \( (X, Y) \), \( \{\Phi_{ij}\} \) is a matrix of basis functions evaluated at the sample locations and \( \alpha_i \) and \( \alpha_u \) are vectors of \( I(x_i < u) \) and \( \alpha_u^T \) respectively.

Although this problem is quadratic and theoretically it has a unique solution, in practice, the matrix \( \Phi \) can be illposed depending on the type of basis we are working with. In such cases we need to regularize the solution. The most commonly used regularization technique is the Tikhonov regularization which penalizes the norm of the coefficient vector \( \alpha \). i.e. the regularized cost is given by,
\[ n(J_\alpha^2(u) - C) = \alpha_u^T \Phi I \alpha_u - 2 \alpha_u^T \Phi \alpha_u + \lambda_u \alpha_u^T \alpha_u \]
where \( \lambda_u \) is the regularization parameter. This problem has a unique solution given by
\[ \alpha_u^* = (\Phi^T \Phi + \lambda_u I)^{-1} \Phi^T \beta_u \]
where \( g_u^*(v) = \sum_{i=1}^{n} \alpha_i^*(u) \phi_i(v) \).

Using this solution the estimator of \( M_{CI} \) can be evaluated in the following way,
\[ \hat{M}_{CI}^2 \approx \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (g_{z_i}(y_{j}, z_{j}) - h_{z_j}(z_{j}))^2 \]
\[ = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left( \sum_{k=1}^{m} \alpha_k^u \phi_k((y_{j}, z_{j})) - \sum_{k=1}^{m} \beta_k^u \phi_k((y_{j}, z_{j})) \right)^2 \]
\[ = \frac{1}{n^2} \| \Phi \alpha - \Psi \beta \|^2. \]

where \( \{\phi_i\}_{i=1}^{m} \) and \( \{\psi_i\}_{i=1}^{m} \) are basis functions for \( g \) and \( h \) respectively and \( [\alpha]_i = \alpha_i \) and \( [\beta]_k = \beta_k \) are matrix of coefficients. Replacing the solution of the coefficient vector in the equation and assuming \( \lambda_{z_j} \) is the same for all \( i \), we get,
\[ \hat{M}_{CI}^2 \approx \frac{1}{n^2} \| (\Phi (\Phi^T \Phi + \lambda_u I)^{-1} \Phi^T - \Psi (\Psi^T \Psi + \lambda_u I)^{-1} \Psi^T) \beta \|^2 \]
where \([\beta]_i = [\phi_i(\cdot) \psi] \) is a matrix of 0s and 1s.

Similarly, the estimator of dependence can be evaluated as
\[ \hat{M}_{D}^2 \approx \frac{1}{n^2} \| (\Xi (\Xi^T \Xi + \lambda_x I)^{-1} \Xi^T - n^{-1} J) \beta \|^2 \]
where \([\beta]_{ij} = \xi_j(y_j) \) and \( \{\xi_i\}_{i=1}^{m} \) are basis functions for estimating \( P(X < u | Y = v) \) and \([J]_{ij} = 1 \) is a matrix of ones.
2.3. Choice of basis function and other issues

Since we are trying to find an estimate \( g_v(v) \) of \( P(X < u | Y = v) \) by minimizing the \( L_2 \) distance between these two functions, we need to ensure that the basis functions are rich enough to make this distance arbitrarily small. Therefore, we choose

\[
g_v(v) = \sum_{i=1}^{n} \alpha_i^v \kappa(v, y_i)
\]

where \( \kappa \) is a positive definite kernel, often a Gaussian. We choose this model since it can be shown that under the condition \( n \to \infty \), these function, \( g_v(v) \), can represent any function \( P(X < u | Y = v) \) with arbitrary accuracy in the \( L_2 \) distance sense. To be more specific, we choose \( \kappa_i((y, z)) = \kappa_i(y, y_i)\kappa_i(z, z_i) \) and \( \psi_i(z) = \kappa_i(z, z_i) \) and \( \xi_i = \kappa_i(y, y_i) \). Using these basis functions the matrices \( \Phi, \Psi \) and \( \Xi \) becomes Gram matrices \( K_{Y,Z}[y,z], K_{X,Z} \) and \( K_{Y,Y} \) where \( [K_{L,V}]_{ij} = \kappa(u_i, u_j) \). Then the estimators of conditional independence and dependence, namely \( \mathcal{M}_{CI} \), can be derived from \( \mathcal{M}_{CI} \) by letting the realizations of \( Z \) to be all same such that \( K_{X,Z} = \Xi \).

The measures and estimators derived so far are not symmetric i.e. \( \mathcal{M}_{CI}(X,Y,Z) \neq \mathcal{M}_{CI}(Y,X,Z) \) and same for \( \mathcal{M}_I \). In order to make them symmetric we introduce the following measures,

\[
\mathcal{M}_I^{\beta} = \frac{1}{2}(\mathcal{M}_{CI}(X,Y,Z) + \mathcal{M}_{CI}(X,Z,Y))
\]

and

\[
\mathcal{M}_I^{\beta} = \frac{1}{2}(\mathcal{M}_{CI}(Y,X,Z) + \mathcal{M}_{CI}(X,Y,Z))
\]

Finally, the proposed estimators requires inverting an \( (n \times n) \) dimensional matrix which is \( O(n^3) \) in computation. However, this computation can be reduced substantially by exploiting the fact that the Gram matrices often have a fast decaying eigen structure [1]. Methods such as incomplete Cholesky decomposition can be used for this purpose [7].

3. SIMULATION

In this section we compare the performance of the proposed method with state-of-the-art kernel based methods of conditional independence and dependence, namely \( \text{HSIC} \) and \( \text{NOCCO} \) as described in [5]. We choose these methods since they have recently been shown to work well in many practical problems, both these methods involve two free parameters i.e. the kernel and the regularization parameter and they are computationally equivalent to our methods i.e. they require inverting an \( (n \times n) \) dimensional Gram matrix. We choose to compare the small sample performance of these methods since in many practical problems the number of samples are limited. We use a Gaussian kernel with kernel size \( \sigma \) for all the methods and vary \( \sigma \) with the dimensionality of the problem. Since both the kernel size and the regularization parameter work as smoothing parameter, we fix the kernel size and study the effect of regularization parameter. We observe similar characteristic when the roles of these parameters are exchanged.

![Fig. 1](image)

(a) The figure shows two independent random variables. (b) The figure shows the same random variables when rotated by \( \pi/4 \) in anticlockwise direction. (c) The figure shows the variation of \( M_I^2 \) with angle of rotation. The red curve is the actual data whereas the blue (dotted) curve is the surrogate data. See 3.1.

Table 1. The Table records the number of times \( X' \perp Y' \) has been rejected out of 1000 times. See 3.1.

<table>
<thead>
<tr>
<th>Angle (radian)</th>
<th>0</th>
<th>0.5</th>
<th>1.5</th>
<th>2.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_I^2 (\lambda = 10^{-4}) )</td>
<td>943</td>
<td>936</td>
<td>745</td>
<td>14</td>
</tr>
<tr>
<td>( M_I^2 (\lambda = 10^{-2}) )</td>
<td>938</td>
<td>836</td>
<td>400</td>
<td>0</td>
</tr>
<tr>
<td>( M_I^2 (\lambda = 10^{-3}) )</td>
<td>960</td>
<td>862</td>
<td>292</td>
<td>0</td>
</tr>
<tr>
<td>( M_I^2 (\lambda = 10^{-4}) )</td>
<td>960</td>
<td>818</td>
<td>304</td>
<td>0</td>
</tr>
</tbody>
</table>

3.1. Test of independence

We compare the performance of \( M_I^2 \) and \( \text{NOCCO} \) on a data described in [5]. Consider two independent random variables \( X \) and \( Y \), both having zero mean and unit variance. \( X \) is a uniform random variable whereas \( Y \) is a combination of two uniform random variables each having equal probability of occurrence on disjoint support (See Figure 1). We compare the performance of \( M_I^2 \) and \( \text{NOCCO} \) on a data described in [5]. We choose these methods since they have recently been shown to work well in many practical problems, both these methods involve two free parameters i.e. the kernel and the regularization parameter and they are computationally equivalent to our methods i.e. they require inverting an \( (n \times n) \) dimensional Gram matrix. We choose to compare the small sample performance of these methods since in many practical problems the number of samples are limited. We use a Gaussian kernel with kernel size \( \sigma \) for all the methods and vary \( \sigma \) with the dimensionality of the problem. Since both the kernel size and the regularization parameter work as smoothing parameter, we fix the kernel size and study the effect of regularization parameter. We observe similar characteristic when the roles of these parameters are exchanged.
Table 2. The Table records the number of times $X \perp Y|Z$ has been rejected out of 100 times. See 3.2.

3.2. Test of conditional independence

Next, in order to compare the performance of $I^{COND}$ and $M_{CI}$ we consider the following data generating processes from [8].

1. $Y_t = 0.3Y_{t-1} + \epsilon_t$
2. $Y_t = (-0.5Y_{t-1} + \epsilon_t)I_{Y_{t-1} \leq 1} + (0.4Y_{t-1} + \epsilon_t)I_{Y_{t-1} > 1} + \epsilon_t$
3. $Y_t = 0.5Y_{t-1}$
4. $Y_t = 0.6\Phi(Y_{t-1})Y_{t-1} + \epsilon_t$
5. $Y_t = -0.5Y_{t-1} + 0.5Y_{t-2}(1 + \exp^{-0.5Y_{t-1}})^{-1} + \epsilon_t$
6. $Y_t = 0.1 \log(Y_{t-1}^2) + \sqrt{0.1 + 0.9Y_{t-2}^2} \eta_t$
7. $Y_t = \exp^{-Y_{t-1}^2 + |0.1Y_{t-2}(16 - Y_{t-2})|}/\tau_t$
8. $Y_t = 0.5Y_{t-1} + 0.25Y_{t-2} + 0.125Y_{t-3} + \sqrt{0.3 + |Y_{t-3}|}/\tau_t$
9. $Y_t = \sqrt{0.01 + 0.8Y_{t-1} + 0.64Y_{t-2}^2 + 0.52Y_{t-3}^2}/\tau_t$

$\epsilon_t \sim \mathcal{N}(0, 1)$, $\eta = \sum_{i=1}^{n} \eta_i, \eta_i \sim \mathcal{U}(-0.1, 0.1)$ and $\tau = \sum_{i=1}^{n} \tau_i, \tau_i \sim \mathcal{U}(-1/7, 1/7)$. We test the hypothesis that $Y_t$ is independent of $Y_{t-1}$ given $Y_{t-2}$. It can be seen that this hypothesis is true for DGP1-DGP4 but it is false for DGP5-DGP9.

We again generate 100 sets of 100 samples to first compute an empirical threshold of independence with size 0.05 and generate another 100 sets of 100 samples to compute the empirical rejection rate. We set $\sigma = 2$ for both methods and set $\lambda_1 = \lambda_2$ for $M_{CI}$. In the ideal situation the number of rejections should be 5 for DGP1-DGP4 and 100 for DGP5-DGP9. However, this is hard to achieve from 100 samples. Therefore, we say a method is better if the rejection rate is higher for DGP5-DGP9 and close to 5 for DGP1-DGP4.

We observe that both $M_{CI}$ and $I^{COND}$ show variability with the choice of regularization. However, $M_{CI}$ has shown less variability since it has successfully concluded that DGP1-DGP4 accepts the hypothesis, for all regularization values. $I^{COND}$ on the other hand fails to conclude that over all regularization values. However, $I^{COND}$ exhibits very good rejection rates for DGP5-DGP9 but these values get influenced by the choice of regularization. Finally, we also provide the results reported by [8]. Note that in these results the authors choose the best free parameters by cross validation. We observe that both the proposed method and $I^{COND}$ sometimes provide better result.

4. CONCLUSION

In this paper we propose a novel approach to design tests of conditional independence and independence based on conditional distribution function. We show that the conditional distribution function can be efficiently estimated and the final estimators become the Frobenius norm of a matrix. The proposed methods are computationally effective and we show that they perform well compared to state-of-the-art techniques. However, the performance of these methods can be improved further. First, note that, since our objective is not estimating the conditional distribution function, we have ignored some important features of a conditional distribution function i.e. nonnegativity and monotonicity. These properties can be implemented in the proposed methods by using appropriate constrains. We believe that these constraints will act as a self regularization, eliminating the need of an explicit regularization parameter. Moreover, the problem of estimating the conditional distribution function can be viewed as a regression problem. This can give us a clue about how to choose an appropriate kernel since the kernel selection problem is well studied in the regression literature.

5. REFERENCES


<table>
<thead>
<tr>
<th>Method in [8]</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I^{COND}$ ($\epsilon = 10^{-1}$)</td>
<td>3</td>
<td>17</td>
<td>6</td>
<td>4</td>
<td>65</td>
<td>61</td>
<td>72</td>
<td>75</td>
<td>49</td>
</tr>
<tr>
<td>$I^{COND}$ ($\epsilon = 10^{-2}$)</td>
<td>9</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>21</td>
<td>20</td>
<td>44</td>
<td>37</td>
<td>90</td>
</tr>
<tr>
<td>$I^{COND}$ ($\epsilon = 10^{-3}$)</td>
<td>18</td>
<td>4</td>
<td>16</td>
<td>7</td>
<td>10</td>
<td>5</td>
<td>11</td>
<td>33</td>
<td>72</td>
</tr>
<tr>
<td>$I^{COND}$ ($\epsilon = 10^{-4}$)</td>
<td>11</td>
<td>12</td>
<td>7</td>
<td>6</td>
<td>10</td>
<td>7</td>
<td>7</td>
<td>30</td>
<td>70</td>
</tr>
<tr>
<td>$M_{CI}^{S}$ ($\lambda = 10^{-1}$)</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>42</td>
<td>24</td>
<td>14</td>
<td>61</td>
<td>23</td>
</tr>
<tr>
<td>$M_{CI}^{S}$ ($\lambda = 10^{-2}$)</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>18</td>
<td>14</td>
<td>13</td>
<td>38</td>
<td>27</td>
</tr>
<tr>
<td>$M_{CI}^{S}$ ($\lambda = 10^{-3}$)</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>16</td>
<td>6</td>
<td>8</td>
<td>25</td>
<td>41</td>
</tr>
<tr>
<td>$M_{CI}^{S}$ ($\lambda = 10^{-4}$)</td>
<td>6</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>9</td>
<td>3</td>
<td>3</td>
<td>22</td>
<td>40</td>
</tr>
<tr>
<td>$M_{CI}^{S}$ ($\lambda = 10^{-5}$)</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>11</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>43</td>
</tr>
</tbody>
</table>