The Importance of Asymmetry and Monotonicity Constraints in Maximal Correlation Analysis

Elad Domanovitz and Uri Erez

***Abstract*—The maximal correlation coefficient is a well-established generalization of the Pearson correlation coefficient for measuring non-linear dependence between random variables. It is appealing from a theoretical standpoint, satisfying Rényi's axioms for a measure of dependence. It is also attractive from a computational point of view due to the celebrated alternating conditional expectation algorithm, allowing to compute its empirical version directly from observed data. Nevertheless, from the outset, it was recognized that the maximal correlation coefficient suffers from some fundamental deficiencies, limiting its usefulness as an indicator of estimation quality. Another well-known measure of dependence is the correlation ratio which also suffers from some drawbacks. Specifically, the maximal correlation coefficient equals one too easily whereas the correlation ratio equals zero too easily. The present work recounts some attempts that have been made in the past to alter the definition of the maximal correlation coefficient in order to overcome its weaknesses and then proceeds to suggest a natural variant of the maximal correlation coefficient as well as a modified conditional expectation algorithm to compute it. The proposed dependence measure at the same time resolves the major weakness of the correlation ratio measure and may be viewed as a bridge between these two classical measures.**

I. Introduction

Pearson's correlation coefficient is a measure indicating how well one can approximate (estimate in an average least squares sense) a (response) random variable Y as a linear (more precisely affine) function of a (predictor/observed) random variable X, i.e., as $Y = aX + b$.[1] The coefficient is given by

$$\rho(X \leftrightarrow Y) = \frac{\mathrm{Cov(X,Y)}}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}. \tag{1}$$

The coefficient is symmetric in X and Y so it just as well measures how well one can approximate X as a linear function of Y.

The correlation ratio of Y on X, suggested by Pearson (see, e.g., [1]), similarly measures how well one can approximate Y as a general admissible function of X, i.e., as $Y = f(X)$.[2] Specifically, the correlation ratio of Y on X is given by

$$\begin{aligned}\theta(X \to Y) &= \sqrt{\frac{\mathrm{var}(\mathbb{E}[Y|X])}{\mathrm{var}(Y)}} \\ &= \sqrt{1 - \frac{\mathbb{E}[\mathrm{var}(Y|X)]}{\mathrm{var}(Y)}}.\end{aligned} \tag{2}$$

[1]We assume that the random variables X and Y have finite variance.

[2]We define a function $f(\cdot)$ to be admissible w.r.t. the random variable X if it is a Borel-measurable real-valued functions such that $\mathbb{E}[f(X)] = 0$ and $\mathbb{E}[f^2(X)] \leq \infty$.

The correlation ratio can also be expressed as

$$\theta(X \to Y) = \sup_f \rho(f(X) \leftrightarrow Y) \tag{3}$$

where the supremum is taken over all (admissible) functions f (see, e.g., [2]). This measure is naturally nonsymmetric.

We note that one may equivalently say that the correlation ratio measures how well one can approximate Y as $Y = aX' + b$ for some admissible transformation of the random variable $X' = f(X)$. While perhaps seeming superfluous at this point, this view will prove useful when considering different generalizations of the correlation ratio to the case where the observations are a random vector.

Similarly, the Hirschfeld-Gebelein-Rényi maximal correlation coefficient [3]–[5] measures the maximal (Pearson) correlation that can be attained by transforming the pair X, Y into random variables $X' = g(X)$ and $Y' = f(Y)$; that is, how well $X' = aY' + b$ holds in a mean squared error sense for some pair of (admissible) functions f and g. More precisely, the maximum correlation coefficient is defined as the supermum over all (admissible) functions f, g of the correlation between $f(X)$ and $g(Y)$:

$$\rho^{**}_{\max}(X \leftrightarrow Y) = \sup_{g,f} \rho(f(X) \leftrightarrow g(Y)). \tag{4}$$

This measure is again symmetric by definition. We use the superscript "**" to indicate that both functions (applied to the response and the predictor random variables) need not satisfy any restrictions beyond being admissible.

The maximal correlation coefficient has some very pleasing properties. In particular, in [5], Rényi put forth a set of seven axioms deemed natural to require of a measure for dependence between a pair of random variables. He further established that the maximal correlation coefficient satisfies the full set of axioms. It is important to note that one of the axioms requires symmetry. Rényi's seminal work inspired substantial subsequent work aiming to identify other measures of dependence satisfying the set of axioms. We refer the reader to [6] for a survey of some of these.

Another appealing trait of the maximal correlation coefficient, greatly contributing to its popularity, is its relation to the mean square error and hence to a Euclidean geometric framework. In particular, it is readily computable numerically via the alternating conditional expectation (ACE) algorithm of Breiman and Friedman [7]. Moreover, and as recalled in the sequel, the ACE algorithm naturally extends to cover linear estimation of a (transformed) random variable from a component-wise transformed random vector.

Despite its elegance and it being amenable to computation,

the maximal correlation coefficient suffers from some significant deficiencies as was recognized since its inception. It was noted, for instance, that the maximal correlation coefficient equals unity "too easily"; see, e.g. [8] and [9]. Specifically, the maximal correlation coefficient can equal 1 even when the pair of random variables is nearly independent (as also demonstrated below).

Several suggestions were proposed over the years to alter the measure so as to overcome these drawbacks. One important avenue calls for limiting the functions applied to both random variables to be monotone functions [9], [10]. As we observe next, while this restriction indeed results in a more satisfying measure of dependence (in the sense of being a meaningful indicator of the quality of estimation possible), it may well be too harsh of a limitation. Specifically, it is worth quoting the incisive comments of Hastie and Tibshirani in [11]:

"A monotone restriction makes sense for a response transformation because it is necessary to allow predictions of the response from the estimated model. On the other hand, why restrict predictor transformations (such as for displacement and weight in the city gas consumption problem) to be monotone? Instead, why not leave them unrestricted and let the data suggest the shape of the relevant transformation?"

The goal of the present paper is first to reiterate some of the known drawbacks of the maximal correlation coefficient as well as to strengthen the arguments that its definition should be modified. We introduce the notion of ε-monotonicity and we then argue in favor of constraining the transformation only of the response random variable to be ε-monotonic, leading to a proposed semi-ε-monotone maximal correlation measure. We show that this measure does not suffer from the drawbacks of neither the max correlation coefficient nor of those of the correlation ratio. Further, we demonstrate that both the correlation ratio and the suggestion of Hastie and Tibshirani can be viewed as extreme cases of the suggested measure.

In addition, we modify accordingly the ACE algorithm and establish its convergence subject to adding a ε-monotonicity constraint on the transformation applied to the response variable.

II. Shortcomings of the Maximal Correlation Coefficient and a Proposed Modification

As a simple example consider two variables which share only the least significant bit:

$$X = C + \sum_{i=1}^{N} A_i 2^i$$
$$Y = C + \sum_{i=1}^{N} B_i 2^i \tag{5}$$

where A_i, B_i, C are mutually independent random variables, all taking the value 0 or 1 with equal probability. Clearly, applying modulo 2 to both random variables yields a maximal correlation of 1 which seems quite unsatisfactory if our goal is estimation subject to reasonable distortion metrics.

Disconcerted from this behavior of the maximal correlation coefficient, Kimeldorf and Sampson [9] proposed to alter its definition in such a way that it attains the maximal value of 1 only if the pair of random variables is mutually completely dependent.[3] This motivated them to define a modified measure as

$$\rho_{\max}^{mm}(X \leftrightarrow Y) = \sup_{g,f} \rho(f(X) \leftrightarrow g(Y)). \tag{6}$$

where f and g are not only admissible but also monotone functions.

The problem with adopting this definition in the context of estimation is the symmetric constraints it imposes on the two transformations. As the process of estimation/prediction (and more generally inference) is directional, if the goal of the dependence measure is to characterize how well one can achieve the latter tasks, there is no apparent reason to impose any restriction on the transformation applied to the observed data. Several works aiming to modify Rényi's axioms to reflect this asymmetry include [8], [13] and [14].

Remark 1 (Discrete random variables)**.** *While the emphasis in this paper is on continuous random variables, symmetric measures are also generally not appropriate for measuring the dependence between discrete random variables. For instance, a natural measure in this case is the minimal possible probability of error when predicting one from the other. Clearly, this measure is also not symmetric. While minimum error probability is related via universal lower and upper bounds to the conditional entropy and mutual information (the latter being a symmetric measure), as shown in [15] (Equations 5 and 6), the gap between the lower and upper bounds (keeping the probability of error fixed) grows unbounded with the cardinality of the random variables. This is yet another indication that symmetric measures are ill-suited for estimation/prediction purposes.*

Indeed, a natural and quite satisfying directional measure of dependence between random variables is the correlation ratio defined in (3). While Rényi objected to the correlation ratio due to its asymmetric nature, as was noted in [8], when our goal is asymmetric (i.e., estimating Y from X), there is no reason for requiring symmetry from the measure.

Nonetheless, in many cases one does not have strong grounds to assume a particular "parameterization" of the desired (response) random variable Y which is to be estimated. Thus, not allowing to apply any transformation to the response variable, as is the case of the correlation ratio, may in certain cases be too restrictive. In other words, in the absence of a preferred "natural" parametrization of the response variable, one may consider choosing a strictly monotone transformation (change of variables) so as to make it easier to estimate.

Another drawback of the correlation ratio is that it vanishes too easily. Specifically, two dependent random variables can have a correlation ratio of zero.

In light of the considerations discussed, we advocate the following modification to the definition of the maximal correlation coefficient.

[3]Following the work of Lancaster [12], two random variables are said to be mutually completely dependent if they are almost surely invertible functions of one another.

Definition 1. *A function f is said to be ε-increasing if for all $x_2 \geq x_1$.*[4]

$$f(x_2) - f(x_1) \geq \varepsilon(x_2 - x_1). \tag{7}$$

Definition 2. *The semi-ε-monotone maximal correlation measure is defined as*

$$\rho^{*m_\varepsilon}_{\max}(X \to Y) = \sup_{g,f} \rho(f(X) \leftrightarrow g(Y)) \tag{8}$$

where $0 < \varepsilon < 1$ and the supremum is taken over all admissible functions $f(x)$ and ε-increasing (and admissible) functions $g(y)$.

Remark 2. *Limiting g to be an increasing function is done only to simplify notations. Equivalently, g can be limited to be ε-decreasing, which can be defined analogously to Definition 1.*

Remark 3. *Limiting g to be ε-increasing implies that, in particular, it is invertible, which is a natural requirement. The magnitude of ε may be viewed as a means to regularize the reparameterization.*

A. The vector observation case

Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a vector of predictor variables. We may generalize the maximal correlation coefficient as

$$\rho^{**}_{\max}(\mathbf{X} \leftrightarrow Y) = \sup_{g,f} \rho(f(\mathbf{X}) \leftrightarrow g(Y)) \tag{9}$$

where the supremum is over all admissible functions.

Following Breiman and Friedman [7], we may also consider a simplified (quasi-additive) relationship between Y and $\mathbf{X}$ where we seek an optimal linear regression between a transformation of Y and a component-wise non-linear transformation of the predictor random vector $\mathbf{X}$. Denote the fraction of the variance not explained by a regression of $f(Y)$ on $\sum_i f_i(X_i)$ as

$$e^2(g, f_1, \ldots, f_p) = \frac{\mathbb{E}\left[\left(g(Y) - \sum_i f_i(X_i)\right)^2\right]}{\mathbb{E}[g(Y)^2]}. \tag{10}$$

In [7] conditions for the existence of optimal transformations $\{f_i\}, g$ such that the supremum is attained are given, and it is shown that under these conditions the ACE algorithm converges to the optimal transformations.

Going back to the rationale for requiring monotonicity, one may object to the example (5) as being artificial and argue that the maximum correlation coefficient merely captures whatever dependence there is between the random variables. In this respect, it is worthwhile to quote Breiman [16] (commenting on [10]):

"I only know of infrequent cases in which I would insist on monotone transformations. Finding non-monotonicity can lead to interesting scientific discoveries. If the appropriate transformation is monotone, then the fitted spline functions (or ACE transformations) will produce close to a monotonic transformation. So it is hard to see what there is to gain in the imposition of monotonicity."

We demonstrate now that the problematic nature of the maximal correlation coefficient becomes pronounced when considering the multi-variate case and so does the necessity of restricting the transformation of the response variable (only) to be ε-monotone.

Specifically, let us consider again the example of (5). Suppose that Y and X are as defined but that in addition to X, there is another slightly noisy observation of Y, say $\tilde{X} = Y + Z$ where the variance of Z is small with respect to that of Y. Clearly, the maximal correlation coefficient will still equal 1, and the observation $\tilde{X}$ will be discarded even though it could have allowed to estimate Y with small distortion. Thus, in this example, the maximal correlation coefficient is maximized by perfectly estimating the least significant bit while doing away with the more significant bits even though nearly distortion-less reconstruction is possible. See also Section V below for a numerical example.

Ramsay [10] proposed a modification of the ACE algorithm by imposing monotonicity constraints on all transformations. As discussed above, there is no apparent reason to impose such restrictions on the transformations applied to the predictor variables. We formulate an ACE algorithm enforcing ε-monotonicity only on the transformation of the response variable and establish its convergence to a global maximum.

III. Semi-ε-monotone maximal correlation measure and modified Rényi axioms

We follow the approach of Hall [8] in defining an asymmetric variant of the Rényi axioms; more precisely, we adopt a slight variation on the somewhat stronger version formulated by Li [14]. However, unlike both of these works, when it comes to putting forward a candidate dependence measure satisfying the modified axioms, we adopt the approach suggested in [11] and define a new maximal correlation measure where we restrict *only* the function applied to the response variable to be ε-monotone.

Assume $r(X \to Y)$ is to measure the degree of dependence of Y on X. Then it should satisfy the following:

(a) $r(X \to Y)$ is defined for all non-constant random variables X, Y having finite variance.[5]
(b) $r(X \to Y)$ may not be equal to $r(Y \to X)$.
(c) $0 \leq r(X \to Y) \leq 1$.
(d) $r(X \to Y) = 0$ if and only if X, Y are independent.
(e) $r(X \to Y) = 1$ if and only if $Y = f(X)$ almost surely for some admissible function f.
(f) If f is an admissible bijection on $\mathbb{R}$, then $r(f(X) \to Y) = r(X \to Y)$
(g) If X, Y are jointly normal with correlation coefficient ρ, then $r(X \to Y) = |\rho|$.[6]

We next observe that for absolutely continuous (or discrete) distributions, the semi-ε-monotone maximal correlation measure of Definition 2 satisfies the proposed axioms.

[4] This condition may be viewed as the "complement" of the Lipschitz condition.

[5] In [14], the first axiom only requires that $r(X \to Y)$ be defined for *continuous* random variables X, Y.

[6] In [14], the last axiom only requires that if X, Y are jointly normal with correlation coefficient ρ, $r(X \to Y)$ is a strictly increasing function of $|\rho|$.

It is readily verified that axioms (a), (b) and (c) hold. To show that axiom (d) holds, we note that if X, Y are independent, then obviously $\rho^{*m_\varepsilon}_{\max}(X \to Y) = 0$, as so is even $\rho^{**}_{\max}(X \to Y)$. As for the other direction, we first note that it suffices to consider the case where the correlation ratio equals 0 and X, Y are dependent. Since the correlation ratio is 0, it follows from (2) that $\mathbb{E}[Y|X] \equiv const$ (in the mean square sense). We may break the symmetry of $g(y) = y$ by defining, e.g.,

$$g_a(y) = \begin{cases} y & y \geq a \\ \varepsilon y & y < a \end{cases} . \tag{11}$$

Consider two values of x_1 and x_2 for which $p(y|x_i)$ are not identical, as must exist by the assumption of dependence. Let a be a value such that

$$\int^a p(y|x_1) y dy \neq \int^a p(y|x_2) y dy. \tag{12}$$

Without loss of generality, we may assume that the left hand side is smaller than the right hand side (we may rename x_1 and x_2). Recalling that $\varepsilon < 1$, it follows that

$$\int p(y|x_1) g_a(y) dy > \int p(y|x_2) g_a(y) dy \tag{13}$$

Thus,

$$\mathbb{E}[g_a(Y)|X = x_1] \neq \mathbb{E}[g_a(Y)|X = x_2]$$

and hence the correlation ratio between $Y' = g_a(Y)$ and X is non-zero, giving a lower bound to the semi-ε-monotone maximal correlation measure between X and Y.

To show that axiom (e) holds, we note that by definition if $Y = f(X)$ (almost surely), then $\rho^{*m_\varepsilon}_{\max}(X \to Y) = 1$. To show that the opposite direction holds, we recall that if $\rho^{*m_\varepsilon}_{\max}(X \to Y) = 1$, then by the properties of Pearson's correlation coefficient, there is a *perfect* linear regression between $g(Y)$ and $f'(X)$ (g, f' being the maximizing functions of the measure). Hence we have $g(Y) = af'(X) + b$ where g is an increasing function with slope greater than ε. Since g is invertible, we have $Y = g^{-1}(af'(X) + b)$. Denoting $f(X) = g^{-1}(af'(X) + b)$, we note that if f' is admissible, then so is f. Hence, $Y = f(X)$ almost surely.

Axiom (f) trivially holds. To show that axiom (g) holds, we recall that it is well know that when X, Y are jointly normal with correlation coefficient ρ, then $\rho^{**}_{\max}(X \to Y) = |\rho|$ (see, e.g., [17]). Since this implies that the maximal correlation is achieved taking $g(y) = y$ (i.e., a monotone function) and $f(x) = x$ or $f(x) = -x$, it follows that

$$\begin{aligned} \rho^{*m_\varepsilon}_{\max}(X \to Y) &= \rho^{**}_{\max}(X \to Y) \\ &= |\rho|. \end{aligned} \tag{14}$$

Remark 4. *We note that the correlation ratio, defined in* (3)*, satisfies all of the modified axioms except for the "only if" part of axiom (d).*

Remark 5. *We note that one may define other dependence measures satisfying the modified Rényi axioms, most notably via the theory of copulas; see [14]. Nonetheless, we believe that the proposed measure has the advantage of being closely tied to linear regression methods and geometric considerations.*

IV. Modified ACE Algorithm

We now present a modification of the ACE algorithm to compute the semi-ε-monotone maximal correlation measure $\rho^{*m_\varepsilon}_{\max}(X \to Y)$ and show that the algorithm converges to the optimal transformations. We then generalize the algorithm to the quasi-additive multi-variate scenario.

A. Single-variable predictor

Following in the footsteps of [7], recall that the space of all random variables with finite variance is a Hilbert space, which we denote $\mathcal{H}_2$, with the usual definition of the inner product $< X, Y >= \mathbb{E}[XY]$, for $X, Y \in \mathcal{H}_2$. Since applying an admissible function f results in a random variable with finite variance, we may define the subspace $\mathcal{H}_2(X)$ as the set of all random variables that correspond to an admissible function of X.

Similarly, the set of all admissible functions of Y is also a subspace of $\mathcal{H}_2$, which we denote by $\mathcal{H}_2(Y)$. Now, if we limit the functions applied to Y to be ε-increasing, we obtain a closed and convex subset of the Hilbert space $\mathcal{H}_2(Y)$. We denote this set by $\mathcal{M}_\varepsilon(Y)$.

Denoting by $P_{\mathcal{A}}(Y)$ the orthogonal projection of Y onto the closed convex set $\mathcal{A}$,[7] the modified ACE algorithm is described in Algorithm 1.

Algorithm 1 Modified ACE single predictor

```
procedure CalculateSemiMonotoneMeasure
    Set g(Y) = Y/‖Y‖;
    while e²(g, f) decreases do
        f'(X) = P_{H2(X)}(g(Y))
        replace f(X) with f'(X)/‖f'(X)‖
        g'(Y) = P_{M_ε(Y)}(f(X))
        replace g(Y) with g'(Y)/‖g'(Y)‖
    End modified ACE
```

Theorem 1. *The two sequences of functions defined by the ACE algorithm converge to the optimal transformations of the semi-ε-monotone maximal correlation measure.*

Proof. To show convergence to the optimal transformations, we note that the ACE algorithm is an alternating minimization algorithm. This class of algorithms was suggested in [19] and extended in [18]; see also [20].

It is shown in [18] that if P and Q are closed convex subsets of a Hilbert space, alternating minimization converges to the global minimum. As $\mathcal{M}_\varepsilon(Y)$ and $\mathcal{H}_2(X)$ satisfy these conditions, we conclude that the algorithm converges to the optimal transformations. □

B. Multi-variate predictor

In the case of a multi-variate predictor, the ACE algorithm seeks an optimal linear regression between a transformation of Y and a component-wise non-linear transformation of the

[7] Note that $\mathcal{P}_{\mathcal{H}_2(X)}(g(Y)) = \mathbb{E}[g(Y) \mid X]$.

predictor random vector $\mathbf{X}$. The latter transformations are defined by a set of admissible functions $f_1, \ldots, f_p$, each function operating on the corresponding random variable, yielding an estimator of the form $\sum_i f_i(X_i)$.

The modified ACE algorithm for the case of a multi-variate predictor, resticting g to be ε-increasing, is described in Algorithm 2.

Algorithm 2 Modified multi-variate ACE

procedure CALCULATESEMIMONOTONEMEASURE
  Set $g(Y) = Y/\|Y\|$ and $f_1(x_1), \cdots, f_p(x_p) = 0$;
  while $e^2(g, f_1, \ldots, f_p)$ decreases **do**
    while $e^2(g, f_1, \ldots, f_p)$ decreases **do**
      for $k = 1$ to p **do**
        $f_k'(X_k) = \mathcal{P}_{\mathcal{H}_2(X_k)}\left(g(Y) - \sum_{i \neq k} f_i(X_i)\right)$
        replace $f_k(X_k)$ with $f_k'(X_k)/\|f_k'(X_k)\|$
    $g'(Y) = \mathcal{P}_{\mathcal{M}_\varepsilon(Y)}(Y)\left(\sum_i f_i(X_i)\right)$
    replace $g(Y)$ with $g'(Y)/\|g'(Y)\|$
  End modified ACE

V. NUMERICAL EXAMPLES

In this section we give examples in which the semi-ε-monotone maximal correlation measure (evaluated using the modified ACE algorithm) results in a significant improvement over the standard maximal correlation measure (evaluated using the standard ACE algorithm) in the context of estimation of a random variable Y from a random vector $\mathbf{X}$. We further demonstrate its potential for improvement over the correlation ratio.

We begin with a multi-variate example where one of the two observed random variables "masks" the other while the latter is more significant for estimation purposes. In the second example we demonstrate why it is not sufficient to restrict $g(Y)$ to be monotone (or equivalently, to set $\varepsilon = 0$). The third example illustrates why the semi-ε-monotone maximal correlation measure may be advantageous with respect to the correlation ratio.

For simulating ACE, we used the ACE Matlab code provided by the authors of [21]. To limit g to be an ε-monotonic function we used isotonic regression followed by a regularization which is described in more detail below.

A. Example 1 - Multi-variate predictor

Assume that the response variable Y is distributed uniformly over the interval $[0, 1]$. Assume we have two predictor variables

$$X_1 = \mathrm{mod}(\mathrm{Y}, 0.2) + \mathrm{N}_1$$
$$X_2 = Y^3 + N_2 \tag{15}$$

where N_1, N_2 are independent zero-mean Gaussian variables with $\sigma^2_{N_1} = 0.01$ and $\sigma^2_{N_2} = 0.2$.

Calculating the maximal correlation coefficient results in "shadowing" the more significant variable (X_2) for estimation purposes of the response Y. To see this, we start by running the ACE algorithm to evaluate the maximal correlation coefficient between Y and X_1. As can be seen from Figure 1, this results in a very high value. Inspecting the transformations yielding this result, we note that g is not monotonic and hence we cannot recover Y from $g(Y)$.

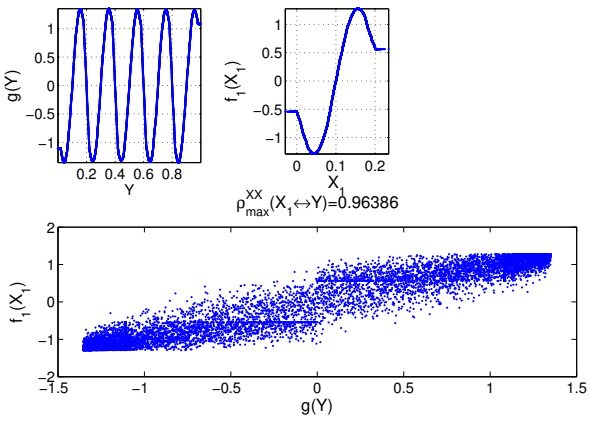

Fig. 1. Example 1: Running ACE on Y and X_1.

Next, we apply the ACE algorithm to calculate the maximal correlation coefficient between Y and X_2. As can be seen from Figure 2, this value is much smaller (than that between Y and X_1) since in this case we have stronger additive noise. Nevertheless, the transformation applied to Y is now monotonic. Therefore, even though the maximal correlation coefficient is smaller, the observation X_2 can better serve for estimation of Y.

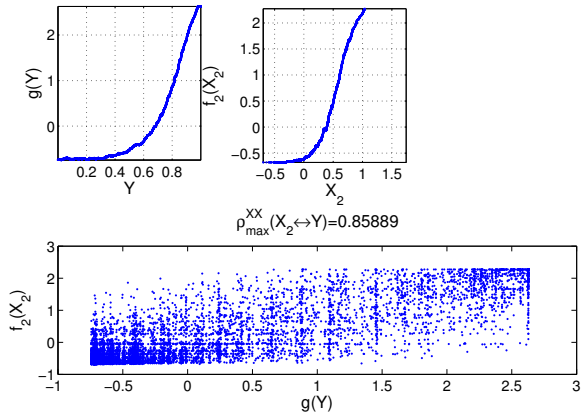

Fig. 2. Example 1: Running ACE on Y and X_2.

Next we apply the ACE algorithm to Y and the vector (X_1, X_2). As can be seen from Figure 3, the ACE algorithm, in order to maximize the correlation, chooses similar functions as in case of running only on Y and X_1, practically choosing to ignore X_2. While, indeed, this maximizes the correlation coefficient, it is far from satisfying from an estimation viewpoint.

We now observe that the modified ACE algorithm does not suffer from this deficiency. To obtain the orthogonal projection

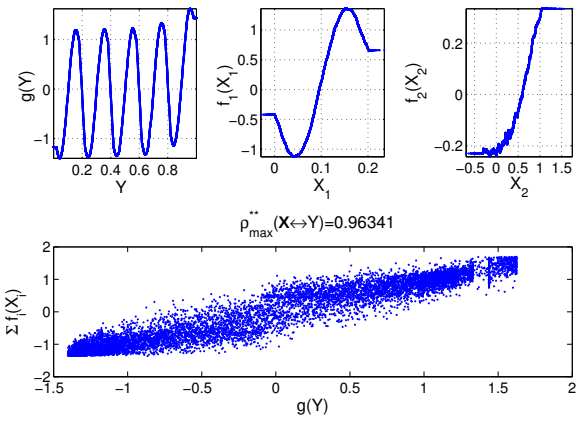

Fig. 3. Example 1: Running ACE on Y, X_1 and X_2.

onto the subset of ε-increasing functions $\mathcal{M}_\varepsilon(Y)$, we use isotonic regression followed by the following regularization[8]

$$g(Y) = g(Y) + \varepsilon \cdot Y. \tag{16}$$

As can be seen from Figure 4, setting $\varepsilon = 0.1$, the resulting value of the semi-ε-monotone maximal correlation measure is very close to the maximal correlation value between Y and X_2. Thus, the algorithm "chooses to ignore" X_1 (even though it suffers from a lower noise level) and bases the estimation on X_2.

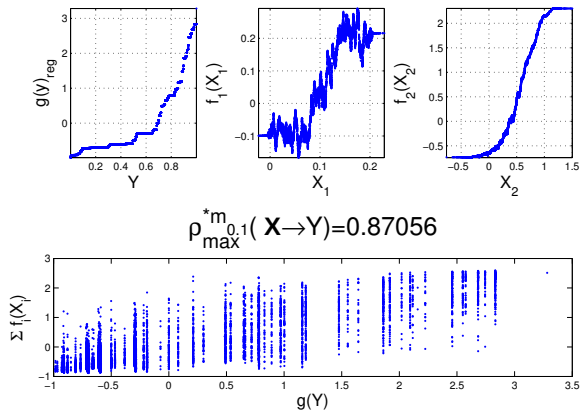

Fig. 4. Example 1: Running modified ACE on Y, X_1 and X_2 with $\varepsilon = 0.1$.

B. Example 2 - Semi-0-monotonicity is insufficient

To illustrate why it does not suffice to limit g to be merely monotone, consider the following example. Assume that the response Y is distributed uniformly over the interval $[-10, 10]$ and that

$$X = \begin{cases} X = Y & Y > 9 \\ X = N_1 & \text{otherwise} \end{cases} \tag{17}$$

where $N_1 \sim \text{Unif}([-1, 1])$ and is independent of Y.

[8]Note that this method of regularization actually forces the minimal slope to be slightly larger than ε.

Limiting g only to be monotone (with no limitation on minimal slope) results in a correlation value of 1 since the optimal solution is to set $g(y) = 0$ in the region it cannot be estimated and $g(y) = y$ otherwise (and then apply normalization). Clearly, the function g is non-invertible as is depicted in Figure 5.

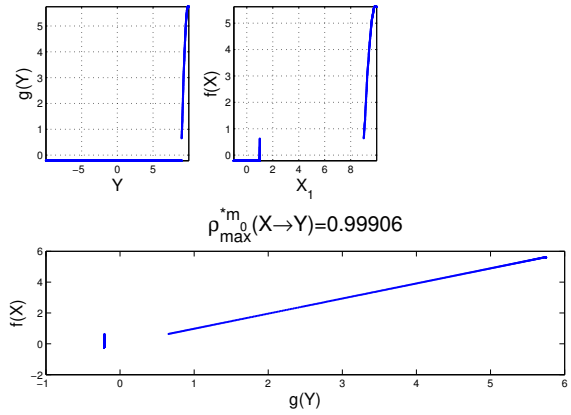

Fig. 5. Example 2: Running modified ACE on Y, X with $\varepsilon = 0$.

Next, we run the modified ACE algorithm, enforcing a minimal slope of $\varepsilon = 0.1$. The results are depicted in Figure 6. This example sheds light on the trade-off that exists when

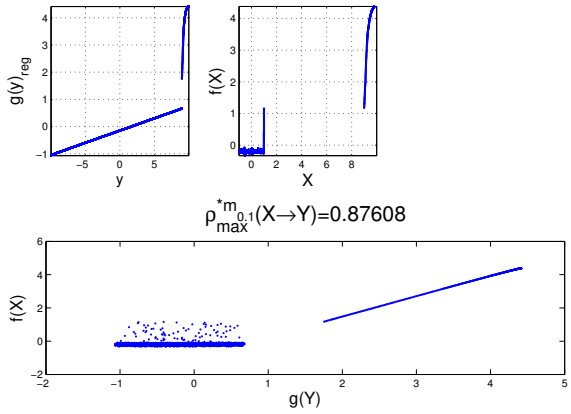

Fig. 6. Example 2: Running modified ACE on Y, X with $\varepsilon = 0.1$

setting the value of ε. Setting ε to be large limits the possible gain over the correlation ratio whereas setting it too low risks overemphasizing regions where the noise is smaller.

C. Example 3 - Comparisons with correlation ratio

The suggested measure can be viewed as a generalization of the correlation ratio (the correlation ratio amounts to setting g to have a constant slope of 1).

We first demonstrate how the semi-ε-monotone maximal correlation measure deals with a well-known example where the correlation ratio equals 0 for a pair of dependent random variables. In this example we assume X and Y are uniformly distributed over a circle with radius 1. The correlation ratio

is 0 as depicted in Figure 7 where we ran ACE enforcing $g(y) = y$.

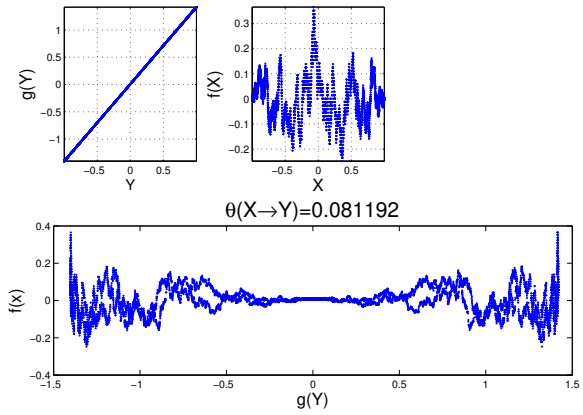

Fig. 7. Example 3a: Optimal transformation corresponding to the correlation ratio.

Applying the semi-ε-monotone maximal correlation measure with $\varepsilon = 0.1$ yields a much larger correlation. Thus, it manages to capture the dependence between X and Y. This is depicted in Figure 8.

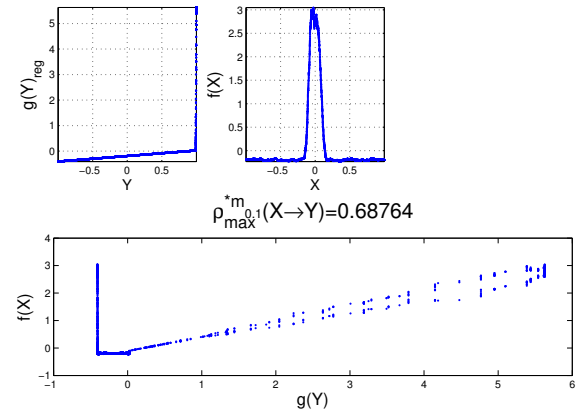

Fig. 8. Example 3b: Optimal transformations corresponding to the semi-ε-monotone maximal correlation measure with $\varepsilon = 0.1$.

The next example demonstrates another potential advantage over the correlation ratio. As was already noted, there are cases where there is no a priori preferred (natural) parameterization for the response variable and thus choosing one that maximizes the correlation may be a reasonable approach as we now demonstrate.

Assume that the response variable Y is distributed uniformly over the interval $[0, 10]$ and that the predictor variable X is

$$X = \log(Y) + N, \tag{18}$$

where N is a zero-mean Gaussian (and independent of Y) with unit variance. Comparing the correlation ratio (Figure 9) to the semi-0.1-monotone correlation measure (Figure 10) reveals that the correlation of the latter is significantly higher.

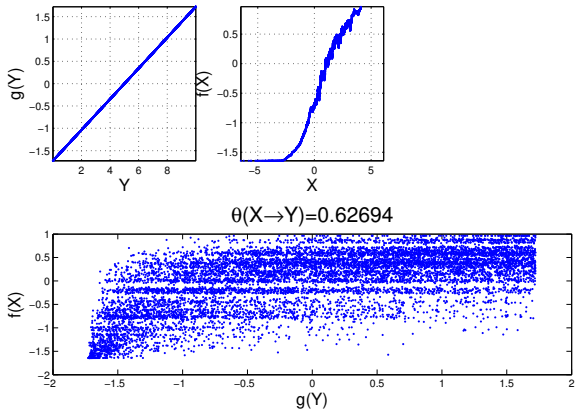

Fig. 9. Example 3b: Optimal transformations corresponding to the correlation ratio.

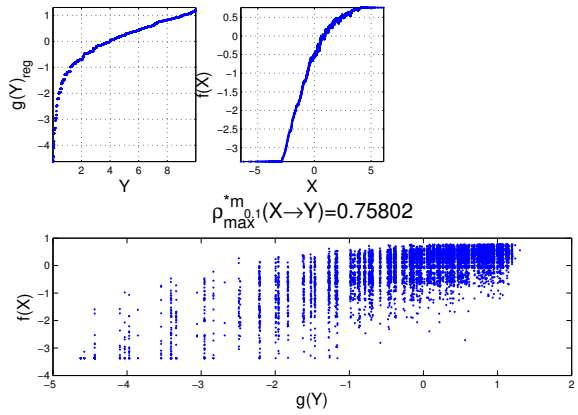

Fig. 10. Example 3b: Optimal transformations corresponding to the semi-0.1-monotone correlation measure.

REFERENCES

[1] H. Cramér, *Mathematical methods of statistics (PMS-9)*. Princeton University Press, 2016, vol. 9.

[2] A. Rényi, "New version of the probabilistic generalization of the large sieve," *Acta Mathematica Hungarica*, vol. 10, no. 1-2, pp. 217–226, 1959.

[3] H. O. Hirschfeld, "A connection between correlation and contingency," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 31, no. 4. Cambridge University Press, 1935, pp. 520–524.

[4] H. Gebelein, "Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung," *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 21, no. 6, pp. 364–379, 1941.

[5] A. Rényi, "On measures of dependence," *Acta Mathematica Hungarica*, vol. 10, no. 3-4, pp. 441–451, 1959.

[6] D. Drouet-Mari and S. Kotz, *Correlation and dependence*. World Scientific, 2001.

[7] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *Journal of the American Statistical Association*, vol. 80, no. 391, pp. 580–598, 1985.

[8] W. Hall, *On characterizing dependence in joint distributions*. University of North Carolina, Department of Statistics, 1967.

[9] G. Kimeldorf and A. R. Sampson, "Monotone dependence," *The Annals of Statistics*, pp. 895–903, 1978.

[10] J. O. Ramsay, "Monotone regression splines in action," *Statistical Science*, vol. 3, no. 4, pp. 425–441, 1988.

[11] T. Hastie and R. Tibshirani, "[monotone regression splines in action]: Comment," *Statistical Science*, vol. 3, no. 4, pp. 450–456, 1988.

[12] H. Lancaster, "Correlation and complete dependence of random variables," *The Annals of Mathematical Statistics*, vol. 34, no. 4, pp. 1315–1321, 1963.

[13] K. Joag-Dev, "Measures of dependence," *Handbook of Statistics*, vol. 4, pp. 79–88, 1984.

[14] H. Li, "A true measure of dependence," University Library of Munich, Germany, Tech. Rep., 2016.

[15] D. Tebbe and S. Dwyer, "Uncertainty and the probability of error (corresp.)," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 516–518, 1968.

[16] L. Breiman, "[monotone regression splines in action]: Comment," *Statistical Science*, vol. 3, no. 4, pp. 442–445, 1988.

[17] H. O. Lancaster, "Some properties of the bivariate normal distribution considered in the form of a contingency table," *Biometrika*, vol. 44, no. 1/2, pp. 289–292, 1957.

[18] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Statistics and Decisions*, vol. 1, pp. 205–237, 1984.

[19] W. Cheney and A. A. Goldstein, "Proximity maps for convex sets," *Proceedings of the American Mathematical Society*, vol. 10, no. 3, pp. 448–450, 1959.

[20] C. L. Byrne, "Alternating minimization and alternating projection algorithms: A tutorial," *Sciences New York*, pp. 1–41, 2011.

[21] H. Voss and J. Kurths, "Reconstruction of non-linear time delay models from data by the use of optimal transformations," *Physics Letters A*, vol. 234, no. 5, pp. 336–344, 1997.