Generalized Correlations and Kernel Causality Using R Package generalCorr

Hrishikesh D. Vinod*

June 4, 2016

Abstract

Karl Pearson developed the correlation coefficient r(X,Y) in 1890's. Vinod (2014) develops new generalized correlation coefficients so that when $r^*(Y|X) > r^*(X|Y)$ then X is the "kernel cause" of Y. Vinod (2015a) argues that kernel causality amounts to model selection between two kernel regressions, $E(Y|X) = g_1(X)$ and $E(X|Y) = g_2(Y)$ and reports simulations favoring kernel causality. An R software package called 'generalCorr' (at `www.r-project.org`) computes generalized correlations, partial correlations and plausible causal paths. This paper describes various R functions in the package, using examples to describe them. We are proposing an alternative quantification to extensive causality apparatus of Pearl (2010) and additive-noise type methods in Mooij et al. (2014), who seem to offer no R implementations. My methods applied to certain public benchmark data report a 70-75% success rate. We also describe how to use the package to assess endogeneity of regressors.

Keywords: generalized measure of correlation, non-parametric regression, partial correlation, observational data, endogeneity.

*Vinod: Professor of Economics, Fordham University, Bronx, New York, USA 104 58. E-mail: vinod@fordham.edu. A version of this paper was presented at American Statistical Association's Conference on Statistical Practice (CSP) on Feb. 19, 2016, in San Diego, California, and also at the 11-th Greater New York Metro Area Econometrics Colloquium, Johns Hopkins University, Baltimore, Maryland, on March 5, 2016.

1 Introduction

A new R package in Vinod (2016) called 'generalCorr' provides software tools for computing generalized correlation coefficients and for preliminary determination of causal directions among a set of variables. The package is accessed by R commands (always in the red font for copy and paste):

```
if(!"generalCorr"%in%installed.packages()) {
install.packages("generalCorr",
repos = "http://cran.case.edu/")} ; library(generalCorr)
```

We begin with some background. Elementary statistics teachers wishing to make a distinction between correlation r and causation often use an example where the cause is intuitively known. For example, high crime results in greater deployment of police officers. Some European crime data is included in the 'generalCorr' package. It is accessed by the following commands which summarize the data and plot crime on the horizontal axis and officer deployment on the vertical axis. The output of the code is omitted here for brevity.

```
data(EuroCrime);summary(EuroCrime)
attach(EuroCrime)
cor.test(crim,off)
plot(crim,off)
```

The `cor.test` function used in the code above reports a high correlation coefficient of 0.9900466 which is highly significant (outputs omitted for brevity). The scatterplot shows that there are outliers with both high per capita crime and large number of police officers. However, since these data are genuine, we may not be justified in simply deleting outliers.

Our discussion is intended for practical causal path determination from the type of data illustrated by European crime, while avoiding a purist viewpoint. Holland (1986) and accompanying discussion surveys causality in science and states his motto: "no causation without (experimental) manipulation." Holland criticizes (p. 970) "Granger causality" for economic time series, Granger (1969), as "at bottom indistinguishable from association." We are not alone in rejecting Holland's viewpoint. Why?

In many situations any experimental manipulation of several variable types (e.g., prices, incomes, ethnicity, taxes, wages, weather, education, geography, etc.) can be impossible, unethical, expensive and /or time-consuming.

There are situations where preliminary causal identification is desired to save time and expense, even if experimental manipulations are planned. Hence an interest in some form of preliminary insight into non-experimental causality is strong among all scientists, despite Holland's criticism. In addition to Granger, two recent approaches for causal identification are: (i) the information geometric causal inference (IGCI) method by Daniusis et al. (2012) and Janzing et al. (2014), and (ii) the additive noise model (ANM) by Hoyer et al. (2009) and others.

Causality Assumptions

(A1) Noisy Dependence: If X causes Y, (denoted by the causal path $X \to Y$), X is independently generated (or exogenous) and Y depends on X, where the dependence is subject to random noise.

(A2) Four Causes of Bidirectional Causality: If the data cannot help us choose between two opposite causal paths, $X \to Y$ and $Y \to X$, this can be due to:

1. Intrinsically bi-directional or symmetric causality (illustrated by Boyle's Law: Pressure*Volume =constant) where both X and Y can be exogenous.
2. The presence of confounding variable(s).
3. Linearity of the dependence relation, explained later in Remark 1.
4. Normality of the joint density $f(X, Y)$, also discussed later in Remark 2.

(A3) Transitivity: If X causes Y, and Y causes Z, then X causes Z.

Thus, let us assume an exploratory phase of research where the researcher has observational data and wants to know which causal direction is more plausible. In our European crime example, we choose X is the preferred regressor 'crim' and Y is the intuitively plausible dependent variable 'off' for deployed police officers. We expect the causal direction to show (crim $\to$ off).

Let us write Model 1 as Nadaraya-Watson Kernel regression, (Vinod, 2008, Sec. 8.4), using bandwidths from Hayfield and Racine (2008):

$$Y_t = G_1(X_t) + \epsilon_{1t}, \quad t = 1, \ldots, T, \tag{1}$$

where the functional form of $G_1(.)$ is unknown, except that it is assumed to be a smooth function. Its estimate g_1 by the Nadaraya-Watson Kernel regression, (Vinod, 2008, Sec. 8.4), uses the ratio of a joint density to marginal density. Details are discussed later.

Assuming that (i) $G_1(x) \in \mathcal{G}$, the class of Borel measurable functions, and (ii) $E(Y^2) < \infty$, Li and Racine (2007) prove (p. 59) that $G_1(x)$ is an optimal predictor of Y in mean squared error (MSE). The model can be extended to include additional regressors X_s, which can be control variables. It is convenient to exclude X_s from both models for ease of exposition in much of the following discussion.

The Model 2 regression is:

$$X_t = G_2(Y_t) + \epsilon_{2t}, \quad t = 1, \ldots, T. \tag{2}$$

where $G_2(Y_t)$ is similar to $G_1(X_t)$.

Now we describe how the smooth function G_1 in eq. (1), $Y_t = G_1(X_t) + \epsilon_{1t}$, is estimated by kernel regression methods. Using kernel smoothing to estimate the joint density $f(x, y)$ divided by the marginal density $f(x)$ we write the estimate $g_1(x)$ of the conditional mean function $G_1(x)$ as:

$$g_1(x) = \frac{\sum_{t=1}^{T} Y_t K(\frac{X_t - x}{h})}{\sum_{t=1}^{T} K(\frac{X_t - x}{h})}, \tag{3}$$

where $K(.)$ is the Gaussian kernel function and h is the bandwidth parameter often chosen by leave-one-out cross validation.

Li and Racine (2007) prove (Sec. 2.1) consistent and asymptotically normal (CAN) property of what they call 'local constant estimator' $g_1(x)$ of $G_1(x)$.

Vinod (2014) explains that using superior forecast (or larger R^2) as the criterion for model selection amounts to choosing between generalized correlation coefficients $r^*(Y|X)$ and $r^*(X|Y)$, with details described below. This paper suggests using a variety of statistical model selection tools, many of which are discussed and simulated in Vinod (2013) and further updated in Vinod (2015a) with extensive on-line appendix with R software. The generalized correlations for the crime data are obtained by the R command:

```
options(np.messages=FALSE)
rstar(crim,off)
```

The function `rstar` treats the first variable as X=`crim` and the second as Y=`off`. Its output below reports `corxy` = $r^*(X|Y) = 0.9960$, which is smaller than `coryx` = $r^*(Y|X) = 0.9972$. Thus the model with X=crim is the regressor is superior (better forecasts from larger R^2) implying (crim $\rightarrow$ off). The function also reports the Pearson correlation coefficient as `pearson.r`= $r_{XY} = 0.9900$ and its p-value as zero implying that correlation coefficient is significantly different from zero, (`pv`< 0.0000). All R outputs here use the blue font.

```
$corxy
      cor
0.9960115
$coryx
     cor
0.997196
$pearson.r
      cor
0.9900466
$pv
[1] 1.561488e-24
```

The generalized correlation matrix R^* is obtained by the code where we use `cbind` to create a matrix input needed by the `gmcmtx0` function. Instead of the X, Y notation used above let us denote variables as Xi, Xj for easy matrix generalizations. Let us choose the preferred dependent variable Xi=`off` as the first variable, and Xj=`crim` leading to the matrix elements $R^*(i, j) = r^*(Xi|Xj)$.

```
mtx=cbind(off,crim)
gmcmtx0(mtx)
```

The output matrix is seen to report the "cause" along columns and response along the rows.

```
> gmcmtx0(mtx)
           off     crim
off  1.0000000 0.997196
crim 0.9960115 1.000000
```

Recall that $r^*(X|Y) = r^*(\text{crim}|\text{off}) = 0.9960$ now becomes $R^*(X2|X1) = 0.9960$ and appears at the (2,1) or second row and first column location. Clearly, the inequality $r^*(\text{crim}|\text{off}) < r^*(\text{off}|\text{crim})$ suggests that crim→off is the sensible causal path where officers are deployed in response to high crime rates, not vice versa.

When we have several variables, it becomes difficult to interpret and compare R^*_{ij} with R^*_{ji} as can be seen by invoking the command `gmcmtx0(mtcars)` producing an unwieldy 11×11 matrix for the motor cars engineering specifications data, which is always available in R. We illustrate it with a manageable 3×3 submatrix for the first three columns, for brevity. The R command is:

```
gmcmtx0(mtcars[,1:3])
```

The 3×3 matrix R^* gives along row i and column j the value of $r^*(i|j)$. For example, the value -0.9433 along row=2 and column=1 represents $r^*(\text{cyl}|\text{mpg})$. It is compared to the value -0.8558 of smaller magnitude, in the diagonally opposite location at row=1 and column=2 representing $r^*(\text{mpg}|\text{cyl})$. The causal path is mpg→cyl, implying that desire for better fuel economy reduces the number of cylinders, rather than vice versa.

```
             mpg        cyl       disp
mpg    1.0000000 -0.8557900 -0.9508994
cyl   -0.9433125  1.0000000  0.9759183
disp  -0.8941676  0.9151419  1.0000000
```

Since $|R^*_{13}| > |R^*_{31}|$ the causal path is disp→mpg, implying that engine displacement reduces fuel economy.

Now we illustrate the use of `allPairs` and `somePairs` functions from the package `generalCorr` for causal path identification from data matrices. As we do above, consider only the first three variables out of 11 from the 'mtcars' data for brevity. The first column has 'mpg' or miles per (US) gallon, the second column has 'cyl' representing 'number of cylinders', a categorical variable, and the third column has 'disp' for 'engine displacement in cubic inches'. The 'np' package used here for kernel regressions is already designed to choose suitable bandwidths for categorical variables as well as usual ratio-type variables.

In the terminology of Section 1.1, defined below, this paper considers three criteria. The function `allPairs` implements all three criteria by choosing the `typ` to be the criterion number. The criterion based on R^* matrix is the third

criterion (Cr3). Hence we must call the function `allPairs` with the option `typ=3`. Also, since we want the output to fit here we are choosing the option `dig=4` for displayed digits to be four.

```
attach(mtcars)
options(np.messages=FALSE)
m1=allPairs(cbind(mpg,cyl,disp),typ=3)
m1
```

Since the data can have different number of missing values for different column pairs (as in the European crime data), the function reports the number of non-missing data for each pair. In the 'mtcars' example, there are no missing values with all 32 pairs available. First several lines of output produced by `allPairs` lists the row and column numbers and the corresponding length of non-missing data. The user may ignore these lines, except when different number of lines of data are missing in different data pairs.

```
> m1=allPairs(cbind(mpg,cyl,disp),typ=3)
[1] "n,p,digits" "32"         "3"          "6"
[1] "r* compared"
[1] "no. of pairs, typ " "3"                  "3"
[1] "i,non-missing" "1"             "32"
[1] "i,j,ii" "1"      "2"      "1"
[1] "i,non-missing" "1"             "32"
[1] "i,j,ii" "1"      "3"      "2"
[1] "i,non-missing" "2"             "32"
[1] "i,j,ii" "2"      "3"      "3"
> m1
     X      Y      Cause  r*x|y       r*y|x       r           p-val
[1,] "mpg"  "cyl"  "mpg"  "-0.85579"  "-0.943312" "-0.852162" "0"
[2,] "mpg"  "disp" "disp" "-0.950899" "-0.894168" "-0.847551" "0"
[3,] "cyl"  "disp" "disp" "0.975918"  "0.915142"  "0.902033"  "0"
```

The output object `m1` is in the lower portion of the output above. Since we are choosing a pair of two from a set of three columns of the 'mtcars' data, we must consider (${}^3C_2 = 3$) 3 pairs. The output object `m1` has row [1,], which reports the first pair results with self-explanatory column headings. It shows the causal path `mpg`→`cyl`, because the absolute value of its 'r*y|x'

exceeds 'r*x|y'. It also confirms that the Pearson correlation coefficient `r` has the smallest magnitude along each row. The zero p-value suggests that the estimated `r` is significantly different from zero.

Our results from the `allPairs` function agree with those based on R^* matrix, viz, `mpg`→`cyl` and `disp`→`mpg`. Note that it is possible to get Latex style tabulated output by the commands: `library(xtable)` and `xtable(m1)`.

A study of all possible pairs may be useful to data miners, but often the researcher knows the plausible subset of dependent variable(s) and wishes to assess the causal strength and exogeneity of other variables. In the 'mtcars' data a plausible dependent variable is 'mpg' which may be matched with 'cyl' or 'disp' resulting in only two possible pairs in our abridged (chosen for brevity) illustration with only three variables.

```
m2=somePairs(mtcars[,1:3],typ=3,dig=4)
m2
```

The output is shortened for brevity. Unlike the `allPairs` function above, the first two column headings in the matrix produced by `somePairs` are reversed (Y and X), since `somePairs` is designed for situations where the same dependent variable Y is often fixed for several regressors X on the right hand side, and where we want to know whether the X's are truly exogenous or we need instrumental variables to deal with their endogeneity.

```
> m2
     Y      X      Cause  r*X|Y         r*Y|X        r          p-val
[1,] "mpg"  "cyl"  "mpg"  "-0.943312" "-0.85579"  "-0.85216" "0"
[2,] "mpg"  "disp" "disp" "-0.894168" "-0.950899" "-0.84755" "0"
```

We find that both causal paths: mpg→cyl and disp→mpg based on the matrix R^* above using `gmcmtx0` are correctly summarized by functions `allPairs` and `somePairs`, included for the convenience of the package user.

1.1 Definition of Kernel causality criteria

Assuming A1 to A3, we conclude that variable X kernel causes Y or $X \to Y$, if Model 1 is superior to Model 2 with respect to at least two of the following three inequality criteria based on model-selection.

(Cr1) If X is the cause, Model 1 is more successful than Model 2 in minimizing local kernel regression gradients, or partial derivatives satisfy:

$$|\partial g_1(Y|X)/\partial x| < |\partial g_2(X|Y)/\partial y|, \tag{4}$$

where the inequalities among T local partial derivatives are fuzzy (i.e., may be violated for some subsets). We quantify them with the help of stochastic dominance of four orders (SD1 to SD4) described later.

(Cr2) The estimated Model 1 absolute residuals (i.e., $(|\hat{\epsilon_{1t}}|)$) should be "smaller" than those of Model 2, satisfying the following inequality for each $t = 1, 2, \ldots, T$:

$$(|\hat{\epsilon_{1t}}|) < (|\hat{\epsilon_{2t}}|), \tag{5}$$

where the fuzzy inequalities among T residuals will be summarized by stochastic dominance tools.

(Cr3) The forecasts from Model 1 are "superior" (e.g., the R^2 of Model 1 exceeds the R^2 of Model 2).

Why kernel regressions?

There are at least two advantages:

(a) The kernel regression fits are generally superior to parametric linear or non-linear regressions. For the crime data example, Pearson's correlation is smaller than both: $r_{XY} < r^*(X|Y)$ and $r_{XY} < r^*(Y|X)$. The crime data example illustrates this fact.

(b) Kernel regressions do not place any unnecessary restrictions on the unknown conditional expectations functions, Shaw (2014). For example, the sum of the residuals of parametric regressions is artificially forced to be zero. By contrast, we have conditional expectation 'functions' and estimated errors from (1) and (2) need not sum to zero: $\Sigma(\hat{\epsilon}_{1t} \neq 0)$ and $\Sigma(\hat{\epsilon}_{2t} \neq 0)$. This property is exploited by our second criterion Cr2.

The plan of the remaining paper is as follows. Section 2 discusses the background and confidence intervals for the r^*-based third criterion Cr3. The subection 2.2 uses the crime data example to illustrate statistical inference using the bootstrap, plotting a sampling distribution. Another subection 2.3 discusses a heuristic test for significance of difference between two r^* values avoiding computer intensive bootstraps. A subsection 2.4 briefly mentions the assumptions of alternative approaches avoided here. Section 3 describes details of model selection based on the first two criteria Cr1 and Cr2 with the subsection 3.1 focusing on stochastic dominance. Subsection 3.2 describes a

function summarizing causal assessment by all three criteria, with examples. Subsection 3.3 describes an application to the Klein I model to assess exogeneity of regressors. Section 4 explains an extension to the multivariate case by considering the generalized partial correlation coefficients to assess the effect of X_i on X_j after removing the effect of a set of possibly confounding variable(s) X_k. Section 5 reports a brief summary of results from an application of our methods to 80 data pairs from the benchmark challenge. Section 6 contains a summary and concluding remarks.

2 Background and Inference for the r^*-based Criterion Cr3

Granger (1969) developed causality for time series data based on the criterion of superior forecast and statistical significance of certain coefficients. Vinod (2014) [Sec. 7.1] developed kernel causality by extending Granger's ideas when the data is not necessarily a time series. Since our third criterion (Cr3) based on generalized correlation coefficients r^* is closer to Granger's ideas, it was developed before the other two (Cr1, Cr2) defined in Section 1.1. It is convenient to discuss Cr3 before Cr1 and Cr2, especially because the other two require some familiarity with stochastic dominance borrowed from Financial Economics.

Zheng et al. (2012) define R^2 values as generalized measures of correlation, denoted by GMC(Y|X) or GMC(X|Y), and proved the asymptotic consistency of $\delta = [GMC(X|Y) - GMC(Y|X)]$. Vinod (2014) first claimed that the causal path $X \rightarrow Y$ is plausible when $\delta < 0$.

Definition of Generalized Correlations

Since R^2 is always positive, providing no information regarding the direction of the relation, Vinod (2014) defines:

$$r^*(Y|X) = \text{sign}(r_{XY})\sqrt{\text{GMC(Y|X)}}, \tag{6}$$

where its square root is assigned the sign of the Pearson correlation coefficient. Similarly, $r^*(X|Y) = \text{sign}(r_{XY})\sqrt{\text{GMC(X|Y)}}$, generally distinct from the one in eq. (6). A matrix of generalized correlation coefficients R^* can be computed by the R function `gmcmtx0`. It is illustrated above for the crime data and motor cars data, and is seen to be asymmetric.

2.1 Statistical Inference for Criterion Cr3

The identification of the causal path by Cr3 crucially depends on the asymmetry of the R^* matrix. Statistical inference regarding causal paths from asymmetry is subject to unsolved pitfalls discussed next.

Remark 1 (Linearity pitfall):

When the true functions G_1, G_2 in equations (1) and (2), respectively, are linear, it is well known that the R^2 of both regressions is simply the square of Pearson's standard correlation, $r_{X,Y}$, making $\hat{\delta} \approx 0$. Thus small numerical magnitude of $\hat{\delta}$ is caused by the linearity and does not necessarily imply statistically insignificant kernel causality.

Remark 2 (Normality pitfall):

If the true joint density $f(X, Y)$ is Normal, conditional densities, $f(Y|X)$ and $f(X|Y)$, are also Normal, making g_1, g_2 linear and ultimately making $\hat{\delta} \approx 0$. Again, normality can incorrectly suggest statistically insignificant kernel causality.

In light of Remarks 1 and 2 significance testing of the null hypothesis $\delta = 0$ (implying bi-directional causation) based on the numerical magnitude of $\hat{\delta}$ is problematic whenever the underlying relation is linear or distribution is Normal. Recall that the usual t-test to determine whether the correlation coefficient is significantly different from zero, one can focus on the size of $|r|$. Not so for the causal inference.

If the relations (1) and (2) are N^4 (nonlinear, noisy, non-Normal and non-parametric), possibly involving biological or human agents, we need not worry about this difficulty. The importance of nonlinearity when human agents are involved was mentioned back in 1784 when Kant, the German philosopher, said: "Out of the crooked timber of humanity no straight thing was ever made." Hence causal paths in social sciences may be easier to assess by using Cr3.

Kernel causality based on Cr3 is likely to fail when the relations are bi-directional exact ($E = MC^2$) having no noise components. Hoyer et al. (2009) show that nonlinearities can be a "blessing rather than a curse" in the context of causal identification. Shimizu et al. (2006) show the advantages of non-normality in the same context.

Remark 3 (Wrong cause from R^2):

Consider an example where $X = \epsilon', Y = X^2\eta'$, where ϵ' and η' are independently distributed normal deviates, where $E(X|Y) = E(\epsilon') = 0$, and $E(Y|X) = E(X^2)\,E(\eta') = 0$. Since X needs to be known before we know the corresponding Y, a change in X must "cause" Y to change (through conditional variance). However it can be verified that here the two R^2 values can sometimes suggest incorrect causal direction: $Y \rightarrow X$.

Zhang and Hyvarinen (2009) list two requirements for causal identification in non-linear cases: (i) the assumed causal model should be general enough to approximately reveal the data generating processes (DGP), and (ii) the model should be identifiable, i.e., it is asymmetrical in causes and effects. Since our kernel regressions are flexibly estimated, they are obviously general and their asymmetry is proved in Zheng et al. (2012) under suitable assumptions.

Vinod (2013) reports many favorable simulations and provides tools for statistical inference when the model choice is based on the sign of $\hat{\delta}$, updated in Vinod (2015a).

The maximum entropy bootstrap described in Vinod and López-de-Lacalle (2009) constructs resamples of potentially non-stationary (X, Y) data and estimates $\hat{\delta}_j$ a large number of times, e.g., $j = 1, \ldots, J$ with $J = 999$. These yield an approximate sampling distribution of $\hat{\delta}$.

2.2 Crime Data Bootstrap Inference for $\hat{\delta}$ of Cr3:

We illustrate the sampling distribution of $\hat{\delta}$ in Figure 1 showing (99, 95, 50)% highest density regions. The mode is at -0.0141, a slightly negative value, a desirable sign for the correct causal path `crim` $\rightarrow$`off`.

Vinod (2015a) tackles inference issues (for Cr3 using $\hat{\delta}$) arising from Remarks 2 and 3 by using the asymptotic normality to justify using the maximum entropy bootstrap (R package 'meboot'). The sampling distribution of $\hat{\delta}$ is readily approximated for statistical inference including confidence intervals or 'highest density region' graphics. It uses Hyndman (2008) method where the plots depict three sets of highest density regions for (99, 95 and 50)%, respectively.

Vinod (2015a) defines:

$$\text{P(cause)} = max\{P^*(\delta_j < 0), \quad P^*(\delta_j > 0)\}, \tag{7}$$

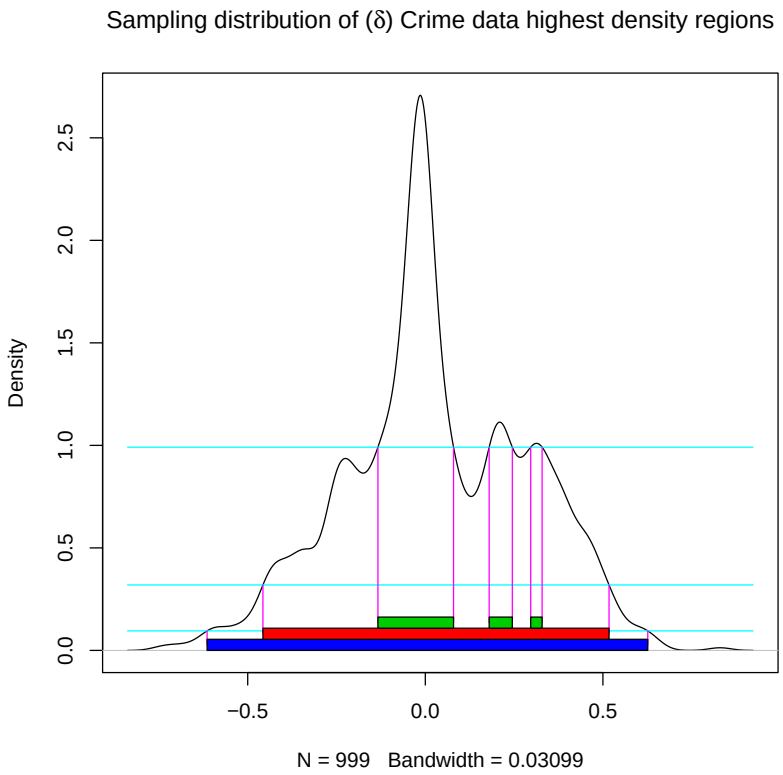


Figure 1: Highest Density Regions for sampling distributions of $\hat{\delta}$ using J=999 resamples of European crime data, where negative values imply correct causal identification

where P(cause) is seen to be the larger of the two rejection probabilities in bootstrap resamples. If we repeat the bootstrap a large number of times (e.g. $L = 1000$) we can numerically approximate the P(Cause) values. A large P(cause) seems to be more desirable because it indicates a larger rejection probability of the null. The `pcause` function in `generalCorr` can be used for the crime data example as follows.

```
options(np.messages=FALSE)
pcause(crim,off,n999=999)
```

The output of the above computer intensive function is next.

```
> pcause(crim,off,n999=999)
[1] 0.5365365
```

The estimate P(cause)=0.5365 for the crime data is only slightly larger than 0.50, suggesting that the observed correct sign might have been due to random variation. A computationally fast and less burdensome heuristic alternative to P(cause) suggested in the Section 2.3 is -0.99, which is also negative, as desired, but statistically insignificant.

2.3 Heuristic test of the difference between two dependent r^* values

If the bootstrap is deemed computationally too demanding, one can formulate the inference problem as one of testing for the difference between two estimates of dependent correlation coefficients, r^*. In 1921 Fisher proposed a variance stabilizing and normalizing transformation for the correlation coefficient, r defined by the formula: $r = \tanh(z)$, involving a hyperbolic tangent. We have:

$$z = \tanh^{-1} r = \frac{1}{2} \log \frac{1+r}{1-r}. \tag{8}$$

An R package 'psych' by Revelle (2014) has references to the literature describing several applications of (8) in tests for difference of two Pearson correlations. One is called the 'paired.r' test for correlations between three variables x, y and z denoted by r(xy), r(xz) and r(yz). Since r(xy)=r(yx) by the symmetry of Pearson's correlations, and since we have only two variables,

the R function `paired.r` obviously cannot be used for our inference problem. However, it is tempting to use it for the following heuristic approximations:

$$\begin{aligned} z\text{stat} &= \text{paired.r}(r^*_{xy}, r^*_{yx}, yz = \text{NULL}, n), \\ t\text{stat} &= \text{paired.r}(r^*_{xy}, r^*_{yx}, yz = \min(|r^*_{xy}|, |r^*_{yx}|), n), \end{aligned} \tag{9}$$

where n is the sample size and where the choice `yz=NULL` yields a standard Normal z-statistic, assuming that the two r^* values are independent. Since we know that they are dependent, and since the R function `paired.r` expects us to specify a numerical estimate of the dependence, one can use the smaller of two absolute values of r^*. Thus the "paired.r" test yields heuristic approximations to t and z test statistics defined in eq. (9). The R function `heurist` of the package 'generalCorr' implements the heuristic t test which requires the r^* correlations and sample size as input arguments, which are obtained as illustrated below.

```
r1=rstar(crim,off)
T=length(crim)
heurist(r1$corxy, r1$coryx,n=T)
```

The output from the above code given below shows that the t-statistic is negative (=0.99) with high p-value (0.33) failing to reject the null hypothesis that either `crim` or `off` can be the cause.

```
Call: paired.r(xy = rxy, xz = ryx, yz = min(rxy, ryx), n = n)
[1] "test of difference between two correlated  correlations"
t = -0.99  With probability =  0.33
```

The heuristic t test is known to be rather conservative.

This completes our discussion of statistical inference for Cr3 based on generalized correlations. In light of the limitations of causal identification based on $\hat{\delta}$, or equivalently on the asymmetric R^* matrix used for Cr3, one needs to use additional model selection criteria Cr1 and Cr2 defined in Section 1.1 and discussed in Section 3 below. They compare the absolute gradients and absolute residuals, respectively, of equation (1) with those of (2). Since our comparisons are somewhat similar to alternative approaches in the literature, they are briefly mentioned in the next subsection.

2.4 Alternative Causality Approaches

Some comments on the alternative approaches from the literature mentioned earlier are included in this subsection. IGCI advocates use information theory to claim that if $X \to Y$, $f(X)$ and $f(Y|X)$ represent independent mechanisms of nature and therefore contain no "information" about each other. Our approach involves a direct comparison of two competing models in equation (1) and (2) without the information theory assumptions.

ANM: The consistency of causal inference under the ANM was established by Kpotufe et al. (2013) who explain (Kolmogorov) complexity measures and kernel regressions similar to our two models. These authors use an equation similar to our eq. (1) upon inserting an explicit requirement that $\epsilon_1 \perp\!\!\!\perp X$, meaning that model 1 errors are orthogonal to X. A key implication of additive noise assumption is that the conditional density $f(Y|X)$ depends on X only through its mean. Since any density has mean, variance, skewness and kurtosis, it is difficult to argue that the ANM assumption is always valid.

The Lemma 4 in Mooij et al. (2014) states the requirement that model 1 errors are orthogonal to X. The ANM will conclude that $X \to Y$ (i.e., choose model 1) if their version of our eq. (2) also satisfies an *absence* of orthogonality for model 2, denoted by $\epsilon_2 \not\perp\!\!\!\perp Y$. Hence we claim that additive-noise type methods also implicitly involve a model choice for their causal identifications.

Mooij et al. (2014) describe various ANM implementations using 88 data pairs. They seem to favor studying the independence of errors with regressors using versions of Hilbert Schmidt Independence Criterion (HSIC) reporting a success rate of over 63%. This paper reports in Section 5 a somewhat better performance (70–75% success) of kernel causality defined above applied to the bivariate 80 of their 88 data pairs.

3 Kernel Regression Model Comparisons for Cr1 and Cr2

This section describes newer tools for overcoming the limitations mentioned in the remarks included in Section 2 by considering additional criteria beyond the goodness-of-fit (R^2) of the two competing kernel regression models, from eq. (1) and eq. (2). This is perhaps a first application of stochastic dominance for model selection.

Standardization:

Any criterion which compares magnitudes estimated by two competing regressions which depend on the units of measurement must be first adjusted to remove such dependence. Hence we often standardize both X, Y separately by subtracting the mean and dividing by the standard deviation.

Since $r^*(Y|X)$ values are not sensitive to units, standardization is not needed for Cr3. However, the absolute values of the gradients of conditional expectation functions g_1, g_2, needed for Cr1 and absolute values of kernel regression residuals needed for Cr2 are generally sensitive to units of measurement. Therefore, we will use standardized data for Cr1 and Cr2.

3.1 Stochastic Dominance for Criteria Cr1 and Cr2

On Wall Street and in Financial Economics a fundamental problem is choosing the best portfolio of stocks and bonds by using past data on returns from such investments. Each portfolio leads to a probability distribution of returns and the portfolio choice is formulated as a problem of choosing the investment offering the best distribution of returns. There is vast and growing published and unpublished literature on this topic.

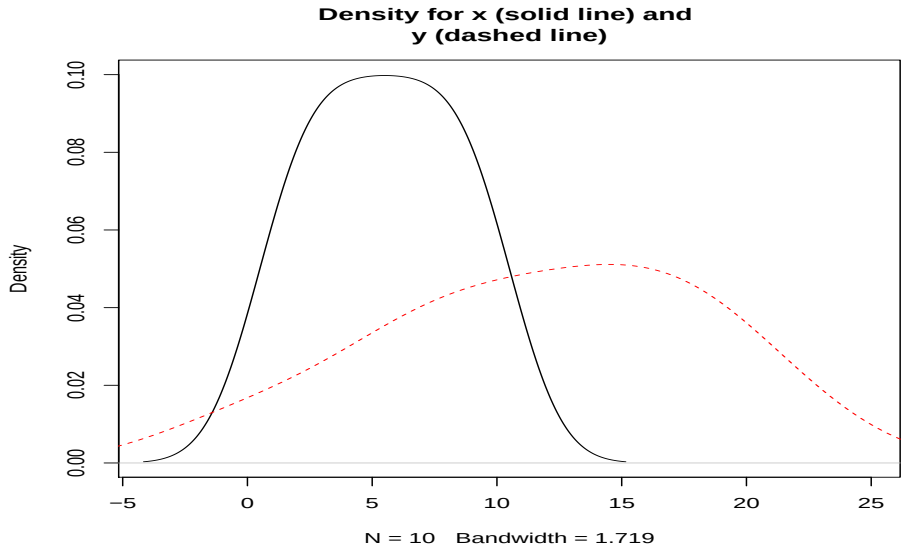
Stochastic dominance (SD) provides well-known comprehensive measures for comparisons of probability distributions, $f(x)$ and $f(y)$, Vinod (2004). We say that $f(x)$ dominates $f(y)$ in the first order (SD1) if their empirical cumulative distribution functions (ecdf) satisfy: $F(x) \leq F(y)$. Why?

A somewhat counter-intuitive sign:

If density $f(X)$ dominates another density $f(Y)$, a density plot for $f(X)$ stays mostly (not everywhere) to the right hand side of the density plot of $f(Y)$. However, the plot of the cumulative density $F(X)$ stays mostly to the *left* hand side of $F(Y)$, or $F(X) - F(Y) \leq 0$, exhibiting a counter-intuitive negativity. See Vinod (2004) (p. 214) illustration using the plots of two beta densities where one is known to dominate. A comparison using artificial data is given in Figure 2.

Let us illustrate stochastic dominance concepts with artificially created small data for $X_t, Y_t, t = 1, 2, \ldots, T = 10$. The dominance of the dependent density $f(Y)$ represented by the dashed line over the density $f(X)$ of the independently generated causal variable is seen in Figure 2.

Figure 2: Smoothed densities for artificial X and Y




```
options(width=65)
set.seed(234);x=sample(1:10);x
y=1+2*x+rnorm(10);y
plot(density(x),main="Density for x (solid line) and
y (dashed line)",xlim=c(-4,25))
lines(density(y),col=2,lty=2)
```

The abridged output of the above code is as follows.

```
> x
 [1]  8 10  1  6  9  4  5  3  2  7
> y
 [1] 17.14013904 21.20918439 -0.03608982 12.51306587 17.91213269
 [6]  9.05785971 12.10397550  6.97438303  5.51484639 15.99005668
```

Let us begin with the Cr3 results based on R^* matrix for these data and note that Cr3 gives the known correct causal path x→y. Although this involves no stochastic dominance at all, it establishes the causal path we seek.

```
gmcmtx0(cbind(x,y))
somePairs(cbind(x,y),typ=3)
```

Recall that `somePairs` labels the first variable in the input matrix as Y and the second variable onward as regressors. However, we get the right label for the 'cause' column in the following output.

```
> gmcmtx0(cbind(x,y))
          x         y
x 1.0000000 0.9907709
y 0.9957212 1.0000000
> somePairs(cbind(x,y),typ=3)
     Y   X   Cause r*X|Y      r*Y|X      r          p-val
[1,] "x" "y" "x"   "0.995721" "0.990771" "0.984616" "0"
```

Note that first order stochastic dominance SD1, summarizes several locally defined central tendencies. Second order dominance (SD2) of $f(x)$ requires their integrals to satisfy: $\int F(x) \leq \int F(y)$, and captures all locally defined dispersions. Similarly, SD3 summarizes several locally defined skewness values and uses $\int \int F(x) \leq \int \int F(y)$. Analogous SD4 for kurtosis requires $\int \int \int F(x) \leq \int \int \int F(y)$.

Computation of SD1 to SD4 using the R software is described in detail in (Vinod, 2008, ch.4). We summarize the basic ideas here. Each ecdf monotonically increases from 0 to 1 with a jump of $1/T$ as the variable increases from its minimum value. A set representing their union is chosen as the new support for the combined random variable conveniently defined as x^j representing cumulated interval widths d_j defined over $2T$ (twice as many) observations. The combined ecdf now assumes that the probability p_i associated with each observation is $1/2T$. The cumulative probability or ecdf is: $F(x^j) = \sum_{i=1}^{j} p_i$, which also monotonically increases from 0 to 1 with a jump of $1/2T$ at each distance d_j for $j = 1, 2, \ldots 2T$.

Premultiplication by a large patterned matrix (I_f), illustrated below, implements the cumulative density computation in Anderson (1996). Now, we illustrate 3×3 representations of I_f allowing the reader to verify that premultiplication by I_f is equivalent to computing a cumulative summation.

$$I_{\mathrm{f}} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

The first order stochastic dominance of the distribution $f(X)$ over $f(Y)$,

(SD1), uses the null hypothesis of no difference between the two as:

$$H_0 : I_f(F(X) - F(Y)) = 0, \text{ against } H_1 : I_f(F(X) - F(Y)) \leq 0. \tag{10}$$

where the alternative hypothesis suggests the dominance of $f(X)$.

Actually SD2 to SD4 require further integrals of these ecdf's. We compute the integrals by using the modified trapezoidal rule (which accommodates unequal widths d_j) according to the formula in Anderson (1996):

$$C(x^j) = \int_0^{x^j} F(z)dz \approx 0.5\left\{F(x^j)d_j + \sum_{i=1}^{j-1}(d_i + d_{i+1})F(x^i)\right\}. \tag{11}$$

Stochastic dominance of order 2 (SD2) uses a similar null hypothesis $H_0 : I_F I_f(F(X) - F(Y)) = 0$, against the alternative $H_1 : I_F I_f(F(X) - F(Y)) \leq 0$, which involves the additional (sparse) matrix I_F needed for implementing the trapezoidal rule and involving distances d_j is illustrated next.

$$I_{\mathrm{F}} = 0.5\begin{bmatrix} d_1 & 0 & 0 \\ d_1 + d_2 & d_2 & 0 \\ d_1 + d_2 & d_2 + d_3 & d_3 \end{bmatrix}.$$

Thus computation of SD2 to SD4 applies eq. (11) repeatedly. Not surprisingly, there are efficient computer programs for this purpose. Novelty here is in using them for model selection for causal determination.

Anderson (1996) describes testing of SD3 obtained by pre-multiplication by an additional I_F. Vinod (2004) extends it to SD4 by pre-multiplication by one more I_F, arguing that it incorporates investor 'prudence' relevant in Finance. Let us denote by ζ the vector of $F(X) - F(Y)$ evaluations at each quantile representing each $(1/2T)$-th segment. Inference for SD1 to SD4 is based on hypotheses regarding $I_f\zeta$, $I_F I_f\zeta$, $I_F I_F I_f\zeta$, and $I_F I_F I_F I_f\zeta$, respectively, similar to eq. (10).

Using the lower case letters to denote the sample values for stochastic dominance of orders 1 to 4, let us define:

$$sd1 = I_f\hat{\zeta}^s,\ sd2 = I_F I_f\hat{\zeta}^s,\ sd3 = I_F I_F I_f\hat{\zeta}^s, \text{ and } sd4 = I_F I_F I_F I_f\hat{\zeta}^s, \tag{12}$$

where the superscript 's' refers to studentized values. Assuming we have T data points for each variable, there are $2T$ estimates of sd's upon bringing two variables on common support.

Now we illustrate the computation of sd1 to sd4 for our artificial example, where the output of `wtdpapb` is used as an input to the function `stochdom2` to compute the stochastic dominance measurement vectors.

```
w1=wtdpapb(x,y) #y should dominate x with mostly positive SDs
print(w1$dj)
stochdom2(w1$dj, w1$wpa, w1$wpb)
```

The following output shows how `dj` are constructed on a common support of both densities. In terms of cumulative densities the T measures of SD1 to SD4 should be positive. Note that we have $T = 10$ estimates of four stochastic dominance measures which need to be summarized. We use their sample means in defining our Av(sd1) to Av(sd4).

```
> w1=wtdpapb(x,y) #y should dominate x with mostly positive SDs
> print(w1$dj)
 [1]  0.000000  1.036090  2.036090  3.036090  4.036090  5.036090
 [7]  5.550936  6.036090  7.010473  7.036090  8.036090  9.036090
[13]  9.093950 10.036090 12.140065 12.549156 16.026147 17.176229
[19] 17.948223 21.245274
> stochdom2(w1$dj, w1$wpa, w1$wpb)
$sd1b
 [1]  0.000000000 -0.002590225  0.002500000  0.025270674
 [5]  0.075721796  0.163853368  0.260994752  0.366626324
 [9]  0.489309599  0.612441171  0.793253192  1.041745662
[13]  1.291829274  1.567821745  1.901673541  2.184029544
[17]  2.464487108  2.679189969  2.813801637  2.866914823

$sd2b
 [1]   0.000000000  -0.001341853  -0.001433705   0.040723425
 [5]   0.244530765   0.847791790   2.026944200   3.921132796
 [9]   6.921390576  10.797399267  16.445542351  24.736149589
[13]  35.346855939  49.696713184  70.756662892  96.392724955
[17] 133.641629392 177.816116713 227.110834524 287.455024005

$sd3b
 [1]  0.000000e+00 -6.951399e-04 -3.520783e-03  5.612278e-02
 [5]  6.317785e-01  3.382296e+00  1.136103e+01  2.931260e+01
 [9]  6.731821e+01  1.296537e+02  2.391171e+02  4.251778e+02
```

```
[13]  6.983737e+02  1.125126e+03  1.856282e+03  2.905074e+03
[17]  4.748356e+03  7.423191e+03  1.105705e+04  1.652310e+04

$sd4b
 [1]  0.000000e+00 -3.601137e-04 -4.652112e-03  7.520008e-02
 [5]  1.463416e+00  1.157104e+01  5.249068e+01  1.752455e+02
 [9]  5.139593e+02  1.206915e+03  2.688653e+03  5.689967e+03
[13]  1.079873e+04  1.994913e+04  3.804638e+04  6.792188e+04
[17]  1.292494e+05  2.337800e+05  3.996237e+05  6.925976e+05
```

The causal direction $X \rightarrow Y$ according to Cr1 requires the inequality of eq. (4), which is $LHS = |\partial g_1(Y|X)/\partial x| < RHS = |\partial g_2(X|Y)/\partial y|$. When the left hand side is 'smaller,' model 1 of (1) has 'smaller' gradients than model 2 of 2). Since we cannot meaningfully compare T inequalities we consider corresponding densities $f(LHS)$ and $f(RHS)$ from absolute values of indicated gradients (apd's).

In the terminology of stochastic dominance by our criterion Cr1, we choose the causal path $X \rightarrow Y$ if $f(LHS)$ is smaller than $f(RHS)$, that is, the RHS density dominates the LHS density by being larger in some overall sense. Equation (12) quantifies dominance orders 1 to 4. We compute $2T$ indexes representing SD1 to SD4 as providing us a comprehensive picture of ranking between two probability distributions. In the R output of the function `stochdom2` above these are denoted as `sd1b` to `sd4b`.

Next, using the central limit theorem we claim that these values are well summarized by simple averages as our sample statistics: Av(sd1) to Av(sd4). The 'generalCorr' package provides convenient functions so that the user need not call `stdpapb` or `stochdom2` functions if the option `typ`=1 or 2 of the function `somePairs` is chosen. Let us recreate the artificial data to illustrate the use of the function `somePairs`.

```
set.seed(234);x=sample(1:10)
y=1+2*x+rnorm(10)
somePairs(cbind(x,y),typ=1,dig=4)
somePairs(cbind(x,y),typ=2,dig=6)
```

Abridged output follows.

```
> somePairs(cbind(x,y),typ=1,dig=4)
     Y    X    Cause SD1apd    SD2apd    SD3apd    SD4apd
```

```
[1,] "x" "y" "y"   "-0.0691" "-0.3365" "-1.2301" "-3.6304"
> somePairs(cbind(x,y),typ=2,dig=6)
     Y   X   Cause SD1res     SD2res     SD3res     SD4res
[1,] "x" "y" "x"   "0.005016" "0.002819" "0.001161" "0.00038"
```

Note that SD1 to SD4 are negative for the gradient based criterion Cr1 (column headings SD1apd to SD4apd) obtained by setting `typ=1`, suggesting the wrong causal path. This illustrates the fact that all criteria do not always suggest the correct causal paths. On the other hand, `typ=2` signs (column headings SD1res to SD4res) are all positive, correctly stating that `x`→`y`, similar to Cr3 noted above. Recall that our definition of kernel causality uses a majority of two out of three criteria.

Now we illustrate the use of functions `allPairs` and `somePairs` using the mtcars data using the options `typ=1,2`. The option `dig`=4 rounds to four digits. The causal directions by these criteria need not agree with those from the R^* matrix.

```
attach(mtcars)
options(np.messages=FALSE)
allPairs(cbind(mpg,cyl,disp),typ=1,dig=4)
somePairs(cbind(mpg,cyl,disp),typ=1,dig=4)
```

Somewhat abridged output for Cr1 is as follows. The signs of Av(sd1) to Av(sd4) depend on the order in which the variables are input. The `allPairs` inputs them as X first and then Y for all possible pairs. Note that with `cbind(mpg,cyl,disp)`, the first variable is not fixed. By contrast, `somePairs` uses the first variable `mpg` as Y which is then paired with the other two. Of course, the signs of Av(sdj) are reversed between the outputs of `allPairs` and `somePairs`, while keeping the correct variable in the 'Cause' column. The results for the choice `typ`=2 based on residuals are omitted for brevity.

```
> allPairs(cbind(mpg,cyl,disp),typ=1,dig=4)
     X      Y      Cause SD1apd      SD2apd      SD3apd    SD4apd
[1,] "mpg"  "cyl"  "cyl" "0.0247"   "0.2691"   "2.1727"    "13.9451"
[2,] "mpg"  "disp" "mpg" "-0.0578" "-0.8837" "-10.4262" "-101.05"
[3,] "cyl"  "disp" "cyl" "-0.0129" "-0.1358" "-1.0766"   "-6.9144"
> somePairs(cbind(mpg,cyl,disp),typ=1,dig=4)
     Y      X      Cause SD1apd      SD2apd      SD3apd   SD4apd
```

```
[1,] "mpg" "cyl"  "cyl" "-0.0247" "-0.2691" "-2.1727" "-13.9451"
[2,] "mpg" "disp" "mpg" "0.0578"  "0.8837"  "10.4262" "101.0538"
```

3.2 Summarizing results of all three criteria

For users' convenience we provide a function `some0Pairs` which reports the results for each of the three criteria and an additional summary matrix with seven columns called `outVote`.

We must first abridge the four numbers produced from Av(sd1) to Av(sd4). We are focusing on their signs defined as (+1 or -1), not their magnitudes. The seven columns produced by this function summarize the signs of Av(sd1) to Av(sd4) stochastic dominance numbers weighted by wt=c(1.2,1.1, 1.05, 1) to compute an overall result for all orders. The weighting is obviously not needed for the third criterion Cr3.

The reason for slightly declining weights on the signs from SD1 to SD4 is simply that the local mean comparisons implicit in SD1 are known to be more reliable than local variance implicit in SD2, local skewness implicit in SD3 and local kurtosis implicit in SD4. The source of slightly declining sampling unreliability of higher moments is the higher power of the deviations from the mean needed in their computations. The summary results for all three criteria are reported in one matrix called `outVote`. Now we illustrate it for the simplest example.

```
some0Pairs(cbind(x,y))
```

Output abridged for brevity is given next.

```
> some0Pairs(cbind(x,y))
$outCr1
     Y   X   Cause SD1apd      SD2apd      SD3apd
[1,] "x" "y" "y"   "-0.069132" "-0.336547" "-1.230067"
     SD4apd
[1,] "-3.630401"

$outCr2
     Y   X   Cause SD1res     SD2res     SD3res     SD4res
[1,] "x" "y" "x"   "0.005016" "0.002819" "0.001161" "0.00038"
```

```
$outCr3
     Y   X   Cause r*x|y      r*y|x      r          p-val
[1,] "x" "y" "x"   "0.995721" "0.990771" "0.984616" "0"

$outVote
     X   Y   Cause Cr1       Cr2      Cr3 sum
[1,] "x" "y" "x"   "-1.0875" "1.0875" "1" "1"
```

These results, based on a very small data set with $T = 10$, do report the correct causal path x→y, based on our eclectic definition. Since Cr1 gives the wrong path, the value of sum=1. When all three criteria are unanimous, value of sum=3.175, which is the case for the crime data described next. It appears that a larger value of sum>1 suggests a stronger determination of the causal path.

```
data(EuroCrime)
attach(EuroCrime)
some0Pairs(cbind(off,crim))
```

Abridged output of the above code has a 'sum' of −3.175 near the end.

```
$outCr1
     Y     X      Cause  SD1apd      SD2apd      SD3apd
[1,] "off" "crim" "crim" "-0.154858" "-3.587453" "-62.985262"
     SD4apd
[1,] "-877.08644"
$outCr2
     Y     X      Cause  SD1res     SD2res      SD3res
[1,] "off" "crim" "crim" "-0.00244" "-0.001206" "-0.000551"
     SD4res
[1,] "-0.000231"
$outCr3
     Y     X      Cause  r*x|y      r*y|x      r          p-val
[1,] "off" "crim" "crim" "0.996012" "0.997196" "0.990047" "0"
$outVote
     X     Y      Cause  Cr1       Cr2       Cr3  sum
[1,] "off" "crim" "crim" "-1.0875" "-1.0875" "-1" "-3.175"
```

It is possible to use the reported number under sum computed by our R function some0Pairs to be attached to the various arrows in directed acyclic

graphs (DAGs) to suggest the direction and strength of causal relation(s). Instead of the `sum`, researcher can choose to attach the value of suitable element of R^* or summaries of SD1 to SD4 on the causal paths. More research is needed to use our tools to supplement the causality apparatus by Pearl (2010).

3.3 Exogeneity in Simultaneous Equation Models

This subsection reports summary results for our three criteria regarding exogeneity of each regressor of the famous Klein I model obtained by using the `someOPairs` function of the 'generalCorr' package.

Klein's specification of the expected consumption equation (stated in terms of fitted coefficients) is:

$$E(\text{cons}) = a_{10} + a_{11}\,\text{coPr} + a_{12}\,\text{coPL} + a_{13}\,\text{wages}. \tag{13}$$

where cons=consumption, coPr=corporate profits, coPL= corporate profits with a lag. Klein data is available in the R package 'systemfit', Henningsen and Hamann (2007). The following code obtains the results to assess the potential endogeneity problem in the first equation of the Klein I model.

According to accepted econometric practice, lagged variables are considered exogenous, because they are pre-determined. Thus the question of endogeneity of lagged variables does not arise, especially since researchers often use lagged variables as 'instrumental variables' to replace potentially endogenous variable. Therefore, our code given below excludes the lagged corporate profits `corpProfLag` from the `cbind` defining the argument to the function `someOPairs`.

```
library(systemfit)
data( "KleinI" )
attach(KleinI)
eqConsump  = cbind(consump, corpProf, wages)
so1=someOPairs(eqConsump)
```

Table 1 reports the summary results using all three criteria. A quick way to assess endogeneity from the results is to focus on the column entitled 'Cause' and look for rows where the dependent variable `consump` also appears as the 'Cause.' The idea is simply that if causal path goes from the dependent variable to the regressor, we have endogeneity issues. Note that this does not

happen along any row of Table 1 allowing us to conclude that there is no endogeneity problem in the first equation of Klein I model. The reader can verify that the same result holds for the remaining two equations also.

Table 1: Table of Summary Results form three criteria applied to the consumption equation of the Klein I model

	X	Y	Cause	Cr1	Cr2	Cr3	sum
1	consump	corpProf	corpProf	-1.0875	-1.0875	-1	-3.175
2	consump	wages	wages	1.0875	-1.0875	-1	-1

4 Causation After Removing Effect of Some Variables

This section considers the generalized correlations between X_i and X_j after removing the effect of a set of variable(s) in X_k. This is an old problem in the context of Pearson correlation coefficients leading to the estimation of partial correlation coefficients. Vinod (2015a) develops generalized partial correlation coefficients starting with the asymmetric matrix R^* of generalized correlation coefficients.

The new R package 'generalCorr' has two functions for this purpose, called `parcor_ijk` and `parcor_ridg`. Let us review the theory behind this generalization before turning to the usage specifics of these R functions in a subsection 4.1.

It is convenient to use a notation similar to Vinod (2015a) to explain the basic concepts. Assume, without loss of generality, that we are interested in comparing treatments X_j for $j = 2, 3, \ldots p$ as they affect X_1. If we assume a linear relation, the $\beta_{1,j;k}$ coefficients measure the specific effect of X_j alone on the dependent variable, X_1, after removing the effects of "all other" variables: $\{k \in [2, \ldots p], k \neq j\}$. In that case one compares the magnitudes of beta coefficients in the multiple regression:

$$E(X_1^s) = \sum_{j=2}^{p} \beta_{1,j;k} X_j^s, \tag{14}$$

where the superscript 's' denotes standardized data, defined to have zero mean: $E(X_i^s) = 0$, and unit variance: $\text{var}(X_i^s = 1)$, for all $i = 1, 2, \ldots p$.

Instead of standardizing data and explicitly computing the above regression, Raveh (1985) suggests a convenient shortcut which exploits the elements of the inverse matrix $R^{-1} = \{r^{jj}\}$. He proves for the first row that:

$$\hat{\beta}_{1,j;k} = -r^{1j}/r^{11}, \tag{15}$$

and a way to link with the partial correlations by inserting the estimated left hand side of eq. (15) on the right hand side, as in:

$$r_{1,j;k} = \hat{\beta}_{1,j;k}/(\sqrt{(r^{jj}/r^{11})}). \tag{16}$$

This extends to an arbitrary i-th row by simply replacing the 1 by i in the above equations. These two equations establish a computationally convenient link between the betas and usual partial correlation coefficients.

If the researcher is choosing between two policy options for influencing X_1, we say that policy option X_2 is superior to X_3 if

$$|\hat{\beta}_{1,2;3}| > |\hat{\beta}_{1,3;2}|, \tag{17}$$

assuming that the magnitude of beta measures "size" of its contribution in standardized regression (14) free from units of measurement.

An alternate to beta is the absolute size of the relevant partial correlation coefficient, which is also unit free. Then the test for superiority of policy option X_2 over X_3 becomes:

$$|r_{1,2;3}| > |r_{1,3;2}|. \tag{18}$$

Since there is no consensus on whether eq. (17) or (18) is the best measure of size of the contribution, let us focus on the partials, which have been generalized in Vinod (2015a) for the case where one replaces linear regressions by kernel regressions. The generalized partial correlations are computed by using equations 15 and 16. Of course, we need to make sure that the inverse matrix R^{*-1} exists. Since the usual symmetric correlation matrix is positive definite, the existence of its inverse is rarely mentioned as a requirement. Since the R^* is asymmetric, we may need to add a ridge-type positive constant (denoted below as `ridgk`) to the diagonal of R^* to ensure that its eigenvalues are all positive real numbers, Vinod (1978).

The generalized partial correlation between (X_1, X_2) after removing the effect of $(X_3, \ldots, X_p)$ is:

$$r^*_{12;3\ldots p} = \frac{R^*_{21}}{\sqrt{R^*_{11}R^*_{22}}}, \tag{19}$$

where R^*_{ij} is the cofactor of R^*. Since the numerator cofactor R^*_{21} is different from R^*_{21}, $r^*_{12;3...p} \neq r^*_{21;3...p}$, implying that the generalized partial correlations will be asymmetric.

In particular, when $p = 3$ we have a new formula:

$$r^*_{12;3} = \frac{r^*_{12} - r^*_{13} r^*_{32}}{\sqrt{(1 - r^*_{13} r^*_{31})}\sqrt{(1 - r^*_{23} r^*_{32})}}. \tag{20}$$

Removing the effect of confounding on kernel causation

Let X_3 be a nuisance variable which might be confounding the causal relationship between X_1 and X_2. Having defined starred partial correlations, Vinod (2015a) defines starred delta as:

$$\delta^*_{1,2;3} = r^{*2}_{1,2;3} - r^{*2}_{2,1;3}. \tag{21}$$

If this delta is negative, we know that $r^{*2}_{2,1;3}$ is the larger of the two, implying that X_1 is the kernel cause of X_2 despite confounding by X_3.

Our code described below denotes `ouij` for $r^*_{i,j;k}$ and `ouji` for $r^*_{j,i;k}$ and checks the sign of the absolute difference, denoted by `rijMrji` for $|r^*_{i,j;k}| - |r^*_{j,i;k}|$. Its interpretation is similar to that of δ^* above, with the negative sign implying that X_i is the more likely kernel cause than X_j.

4.1 R Code for Generalized Partial Correlations

Consider artificial data where we first create `z` independently, then let `x` have an independent component and a component that depends on `z`. Finally we create `y` as dependent on both `x` and `z` with a noise component in its defining relation.

```
set.seed(234)
z=runif(10,2,11);z# z is independently created
x=sample(1:10)+z/10;x  #x is somewhat indep and affected by z
y=1+2*x+3*z+rnorm(10);y #y is caused by both x and z
```

The artificial data are:

```
> z
 [1]  8.710580  9.035412  2.180334  8.984768  2.602191
```

```
 [6]  7.803156 10.364474  8.458780 10.349629  4.558071
> x
 [1]  6.871058  5.903541  9.218033  8.898477  1.260219
 [6]  3.780316  3.036447  7.845878 11.034963  4.455807
> y
 [1] 40.93172 41.01729 25.95145 46.26611 12.31707 32.27355
 [7] 37.23624 42.15213 54.64559 23.60170
```

Now we define a generalized correlation matrix R^* of the three variables.

```
mtx=cbind(x,y,z)
g1=gmcmtx0(mtx);g1
```

We view the R^* matrix before turning to the partial correlation coefficients.

```
> g1
          x         y         z
x 1.0000000 0.9592012 0.2116604
y 0.7127315 1.0000000 0.8636222
z 0.1183994 0.9827227 1.0000000
```

Note that any conclusions based on the asymmetry of the matrix `g1` reported above consider only two variables at a time, ignoring the very presence of any other potentially confounding variable(s). We have created the data deliberately to have such confounding. Now inspecting the matrix `g1` we observe that $r^*(x|y) = 0.9592 > 0.7127 = r^*(y|x)$ incorrectly (because we know how they are created) suggests that `y` kernel causes `x` when we ignore the presence of `z`. On the other hand, $r^*(x|z) = 0.2117 > 0.1184 = r^*(z|x)$ correctly suggests that `z` kernel causes `x` in bivariate comparisons, since `y` is not constructed to vitiate the causal relation between `x` and `z`.

Does the partial correlation between `x` and `y` after removing the effect of `z` correct the incorrect binary result? An application of `parcor_ijk` to matrix `g1` is obtained by the code:

```
parcor_ijk(g1,1,2)
```

The following output shows that the generalized partial correlation between $r^*(x|y)$ denoted as `ouij` and the opposite generalized partial correlation $r^*(y|x)$, denoted as `ouji`, where both are computed after removing the effect of 'z' denoted as `myk` which contains only one variable, which is the third variable in the list input to the matrix.

```
> parcor_ijk(g1,1,2)
$ouij
[1] 1.589513
$ouji
[1] 1.955904
$myk
[1] 3
```

We require that the signs of `ouji` and `ouij` be identical. If not, we cannot rely on the estimated partial correlation coefficients. Since the magnitude `ouji`>`ouij`, the partial correlations satisfy the inequality $r^*(y|x) > r^*(x|y)$. This leads to the correct conclusion that `x`→`y`, after removing the effect of the confounding variable `z`.

An unfortunate fact for the generalized asymmetric R^* matrix is that it need not be positive definite. Hence the partial correlation coefficients need not remain in the correct closed interval: [–1,1]. My R package offers `parcor_ridg` function to solve this problem by using a ridge-type adjustment to the diagonal, mentioned above with a reference to a survey article from the literature. The function is called quite simply as:

```
parcor_ridg(g1,idep=1)
```

The ridge constant needed to make partials to be in the correct range is 4.2367 in the output given next, followed by pairwise generalized partials defined earlier.

```
[1] "final ridgek="    "4.23672976867468"
     nami namj partij   partji    rijMrji
[1,] "x"  "y"  "0.1345" "0.1784"  "-0.0439"
[2,] "x"  "z"  "0.003"  "-0.0105" "-0.0075"
```

Recall that `rijMrji` in the last column represents $|r^*_{i,j;k}| - |r^*_{j,i;k}|$. Its negative sign along the first row correctly implies that $x = X_i$ is the more likely kernel cause than $y = X_j$ after removing the effect of $z = X_k$.

By contrast, the negative sign along the last line (`rijMrji`=–0.0075) referring to the causality between `x` and `z` should be ignored, because the signs of the partials, `ouji`=–0.0105 and `ouij`=0.003 conflict with each other. Note that if X_i and X_j move together, the partial correlation coefficients should

generally satisfy $r^*_{i,j;k} > 0$ and $r^*_{j,i;k} > 0$. If in rare cases the presence of confounding variable(s) X_k reverses the sign, it should reverse both signs, not just one.

The anomaly here should not be surprising in light of the way our data are generated. Recall that `y` is created last and depends on `z` and `x`. Hence, any partial correlation after removing the effect of `y` on the other two variables should not matter. The causal paths `y`→`x` and `y`→`z` are spurious by design. Hence it is gratifying to note that partial correlation coefficient $r^*(x, z; y)$ along the last row of the above output is ambiguous, as it should be.

Similar to `someOPairs`, the function `someCPairs` admits control variable(s) as illustrated below.

```
m3=someOPairs(mtcars[,1:3],dig=4);m3
m4=someCPairs(mtcars[,1:3],ctrl=mtcars[,4],dig=4);m4
m5=someMagPairs(mtcars[,1:3],ctrl=mtcars[,4],dig=4);m5
```

The function `someMagPairs` is intended for use after the causal direction is determined and one wants to have an overall notion of the magnitudes of (the effect of one variable on the other after controlling for `ctrl` variables) relevant partial derivatives, (dy/dx) or (dx/dy). The outputs from the above code are omitted for brevity.

5 Benchmark Application

Mooij et al. (2014) provide a benchmark of 88 sets of observational data where the presumed cause is known. I had technical difficulties in implementing nonparametric kernel regressions on the data in 8 pairs (52 to 55, 71, and 81 to 83), mostly because the data sets are not pairs at all, but have three or more columns. This leaves 80 pairs, some of which are very large to be studied here. Somewhat older results for these data pairs are available at Vinod (2015b). This section reports updated results using the three criteria Cr1 to Cr3 emphasized in this paper and in the R package 'generalCorr'. We apply the function `someOPairs` separately for each data pair, where the implementation is slow, requiring more than a full day of number crunching on my home PC.

Tables 2 and 3 report the summary sign of all four SD measures for Cr1 and Cr2, whereas only (+1, –1) are reported as signs based on Cr3. A column entitled 'sum' further summarizes the overall sign based on all three

criteria. The Table is arranged such that the column entitled 'X' always has the correct 'cause'. Hence the correct sign according to the benchmark website is always positive. This allows us to compute the overall success rate as: (number of positive signs)/80, which equals 55/80 or 68.75%. Some data pairs have $r^*(x|y) \approx r^*(y|x)$, implying potentially bi-directional causality. If we eliminate such data pairs, our success rate increases to over 75%.

Table 2: Summary results for all three criteria using benchmark data, first set of data pairs

	X	Y	Cause	Cr1	Cr2	Cr3	sum
1	ALT	TEMP	TEMP	-1.0875	1.0875	-1	-1
2	ALT	Precip	Precip	-1.0875	0.0625	-1	-2.025
3	Longit	Temp	Longit	1.0875	1.0875	-1	1.175
4	ALT	Sunshine	Sunshine	-1.0875	-1.0875	-1	-3.175
5	RingAg	Len	RingAg	1.0875	1.0875	1	3.175
6	RingAge	ShWt	RingAge	-0.4875	1.0875	1	1.6
7	RingAg	ShDiam	RingAg	1.0875	1.0875	1	3.175
8	RingAg	ShHt	RingAg	1.0875	1.0875	1	3.175
9	RingAg	WholWt	RingAg	-0.4875	0.4875	1	1
10	RingAg	ShWt	RingAg	-0.4875	-0.0625	1	0.45
11	RinAg	VisWt	VisWt	-1.0875	-0.5875	1	-0.675
12	Age	Wage	Age	1.0875	1.0875	1	3.175
13	Disp	mpg	mpg	-1.0875	-1.0875	-1	-3.175
14	HorsP	mpg	mpg	-1.0875	-1.0875	1	-1.175
15	wt	mpg	wt	1.0875	-0.0625	-1	0.025
16	hp	accel	accel	1.0875	-1.0875	-1	-1
17	Age	Divi	Age	1.0875	1.0875	-1	1.175
18	Age	GAG	GAG	-1.0875	1.0875	-1	-1
19	Dur	T2next	T2next	1.0875	-1.0875	-1	-1
20	Lati	temp	temp	-1.0875	1.0875	-1	-1
21	Longi	Preci	Preci	-1.0875	1.0875	-1	-1
22	Age	Wt	Age	-1.0875	1.0875	1	1
23	Age	Ht	Age	-1.0875	1.0875	1	1
24	Age	HrtRat	Age	1.0875	1.0875	1	3.175
25	cement	CoStr	cement	-1.0875	1.0875	1	1
26	Slag	CoStrn	Slag	-1.0875	1.0875	1	1
27	FlyAsh	CoStr	FlyAsh	-1.0875	1.0875	1	1
28	Water	CoStr	Water	-1.0875	1.0875	1	1
29	SuprP	CoStr	SuprP	-1.0875	1.0875	1	1
30	Coars	CoStr	Coars	-1.0875	1.0875	1	1
31	Fine	CoStr	Fine	-1.0875	1.0875	1	1
32	Age	CoStr	Age	1.0875	-1.0875	1	1
33	drnk	mcVol	drnk	1.0875	1.0875	1	3.175
34	drnk	AlkPho	drnk	-1.0875	1.0875	1	1
35	drnk	sgpt	drnk	1.0875	1.0875	-1	1.175
36	drnk	sgot	drnk	1.0875	1.0875	-1	1.175
37	drnk	gama	drnk	1.0875	1.0875	-1	1.175
38	Age	BMI	Age	-1.0875	1.0875	1	1
39	Age	Insu	Age	-1.0875	1.0875	1	1
40	Age	diaBP	Age	-1.0875	1.0875	1	1

Table 3: Summary results for all three criteria using benchmark data, second set of data pairs

	X	Y	Cause	Cr1	Cr2	Cr3	sum
41	AGE	GTT	AGE	1.0875	1.0875	-1	1.175
42	Date	Temp	Date	-1.0875	1.0875	1	1
43	Temp	TmpNxt	Temp	-0.0625	1.0875	1	2.025
44	Presu	PsNxt	Presu	1.0875	-1.0875	1	1
45	Pressu	PsuNxt	PsuNxt	-1.0875	-1.0875	-1	-3.175
46	Humid	HumNxt	Humid	-1.0875	1.0875	1	1
47	WkDay	Cars	Cars	1.0875	-1.0875	-1	-1
48	outd	Indoor	outd	-1.0875	1.0875	1	1
49	Temp	Ozone	Temp	1.0875	1.0875	1	3.175
50	Temp	Ozone	Temp	-1.0875	1.0875	1	1
51	Temp	Ozone	Temp	-1.0875	1.0875	1	1
56	Latit	LifExp	Latit	-1.0875	1.0875	1	1
57	Lati	LifEx	Lati	-1.0875	1.0875	1	1
58	Lati	fLifEx	Lati	-1.0875	1.0875	1	1
59	Lati	fLifE	Lati	-1.0875	1.0875	1	1
60	Lati	LifE	LifE	-1.0875	0.0625	1	-0.025
61	Lati	LifeE	LifeE	-1.0875	-1.0875	1	-1.175
62	Lati	mLifEx	Lati	-1.0875	1.0875	1	1
63	Lati	LifEx	Lati	-1.0875	1.0875	1	1
64	WtrAcc	InfMor	InfMor	-1.0875	-1.0875	1	-1.175
65	HSBret	HSBCrt	HSBret	1.0875	-1.0875	1	1
66	RetHut	ReCKon	ReCKon	-1.0875	-1.0875	-1	-3.175
67	RetCK	ReSHK	RetCK	0.4875	1.0875	1	2.575
68	OpConn	Bytes	OpConn	1.0875	1.0875	-1	1.175
69	OuTemp	InTemp	InTemp	-1.0875	-1.0875	1	-1.175
70	MaleNs	Guess	MaleNs	-1.0875	1.0875	1	1
72	sunSpt	Temp	sunSpt	1.0875	1.0875	-1	1.175
73	EnrUse	CO2	CO2	0.5875	-1.0875	-1	-1.5
74	GNI	LifEx	LifEx	-1.0875	-1.0875	-1	-3.175
75	GNI	YMort	YMort	-1.0875	-1.0875	-1	-3.175
76	PopChg	calChg	PopChg	0.0625	1.0875	-1	0.15
77	SolRad	Temp	Temp	1.0875	-1.0875	-1	-1
78	PhoPFD	EcoPro	PhoPFD	-1.0875	1.0875	1	1
79	PPFdif	NEProd	NEProd	-0.5875	-1.0875	1	-0.675
80	PPFdir	NEProd	PPFdir	-1.0875	1.0875	1	1
84	LnPop	LnEmpl	LnPop	1.0875	1.0875	1	3.175
85	TMeas	Protei	TMeas	-1.0875	1.0875	1	1
86	AptSiz	Rent	Rent	-1.0875	-1.0875	1	-1.175
87	AvTemp	Snow	Snow	-1.0875	1.0875	-1	-1
88	Age	SpineD	Age	1.0875	1.0875	-1	1.175

6 Summary and Concluding Remarks

This paper develops suitable assumptions (A1 to A3) and a practical definition of kernel causality. Since we cannot experimentally vary X to observe its effect on Y, true causality cannot be determined with a 100% success rate from observational data. A less stringent standard for determining "Granger causality" for time series data accepts lower success rates. For cross sectional data we refer to generalized correlation coefficients $r^*(Y|X)$ first defined in Vinod (2014) to define kernel causality as our third criterion Cr3, which is not sensitive to measurement units. The R^* matrix of such coefficients contains suitably signed square roots of Generalized Measures of Correlation (GMCs), recently developed by Zheng et al. (2012).

Vinod (2013) reports extensive simulations where the correct cause is known. It shows that some tools already extant in the literature, including transformations can overcome the problems associated with confounding and non-spherical errors. Vinod (2015a) updates Vinod (2013) to include the computer intensive maximum entropy bootstrap inference and partial correlation coefficients for removing the effect of confounding variables in a multivariate extension.

This paper can be viewed as a vignette describing the detailed steps in the practical use of the R package 'generalCorr.' The package provides European crime data for use as illustrative cross sectional data, which correctly show that high crime rate is the cause of larger police force in European countries, not vice versa. We use the same example to describe statistical inference tools involving the bootstrap and heuristic t tests.

Upon listing certain limitations associated with the R^* matrix, this paper defines an eclectic notion of "kernel causality" based on two out of three criteria (Cr1 to Cr3). We formulate two competing kernel regressions and compare the absolute values of both gradients and residuals along with the goodness of fit. A novelty here is in using stochastic dominance of various orders (SD1 to SD4) for model choice. Our eclectic approach is shown to improve the success rate of causal identification based on 80 pairs of (X, Y) benchmark observational data (presumably issued as a public challenge) where the causal direction is presumably known. Our success rate of about 70-75% for these data is ahead of other attempts known to me.

With additional research it is possible to extend our tools to supplement extensive causality apparatus by Pearl (2010) while relaxing the ANM assumption that the conditional density depends on X only through its mean.

Various R functions relevant for causal path determination and for computation of generalized correlation coefficients and partial correlation coefficients are described with simple numerical examples. These functions and open source nature of R should help in such extensions.

Researchers in various scientific fields and Big Data can benefit from Granger-inspired causality concepts. Our R package 'generalCorr' is convenient, fast and ready for anyone to build upon further. Researchers can save time and resources and propose new research hypotheses by using these tools for *preliminary* identification of causal directions from observational data. Even failed causal directions can help foster disciplined serendipity to scientists and engineers by indicating the presence of confounding variables, missing data and measurement errors.

References

Anderson, G. (1996), "Nonparametric Tests of Stochastic Dominance in Income Distributions," *Econometrica*, 64(5), 1183–1193.

Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K., and Schoelkopf, B. (2012), "Inferring deterministic causal relations," *arxiv*, 1–8, URL `http://arxiv.org/pdf/1203.3475`.

Granger, C. W. J. (1969), "Investigating Causal Relations by Econometric Methods and Cross Spectral Methods," *Econometrica*, 37, 424–438.

Hayfield, T. and Racine, J. S. (2008), "Nonparametric Econometrics: The np Package," *Journal of Statistical Software*, 27, 1–32, URL `http://www.jstatsoft.org/v27/i05/`.

Henningsen, A. and Hamann, J. D. (2007), "systemfit: A Package for Estimating Systems of Simultaneous Equations in R," *Journal of Statistical Software*, 23, 1–40, URL `http://www.jstatsoft.org/v23/i04/`.

Holland, P. W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–970, (includes discussion by many authors).

Hoyer, P., Janzig, D., Mooij, J., Peters, J., and Scholkopf, B. (2009), "Nonlinear causal discovery with additive noise models," *Advances in Neural Information Processing Systems*, 21, NIPS 2008, 689–696.

Hyndman, R. J. (2008), *hdrcde: Highest Density Regions and Conditional Density Estimation*, R package version 2.09, URL `http://CRAN.R-project.org/package=hdrcde`.

Janzing, D., Steudel, B., Shajarisales, N., and Scholkopf, B. (2014), "Justifying information-geometric causal inference," *arxiv*, 1–13, URL `http://arxiv.org/pdf/1402.2499`.

Kpotufe, S., Sgouritsa, E., Janzing, D., and Schoelkopf, B. (2013), "Consistency of causal inference under the additive noise model," in "Proceedings of the 31st International Conference on Machine Learning," , eds. Xing, E. P. and Jebara, T., Journal of Machine Learning Research, vol. 32, pp. 1–9, URL `http://jmlr.org/proceedings/papers/v32/kpotufe14.pdf`.

Li, Q. and Racine, J. S. (2007), *Nonparametric Econometrics*, Princeton University Press.

Mooij, J., Peters, J., , Janzig, D., Zscheischler, J., and Scholkopf, B. (2014), "Distinguishing cause from effect using observational data: methods and benchmarks," *Journal of Machine Learning Research, unpublished archive*, 1–83, URL `http://arxiv.org/abs/1412.3773`.

Pearl, J. (2010), "The Foundations Of Causal Inference," *Sociological Methodology*, 40, 75–149, URL `http://www.jstor.org/stable/41336883`.

Raveh, A. (1985), "On the Use of the Inverse of the Correlation Matrix in Multivariate Data Analysis," *The American Statistician*, 39 (1), 39–42.

Revelle, W. (2014), *psych: Procedures for Psychological, Psychometric, and Personality Research*, Northwestern University, Evanston, Illinois, R package version 1.4.5, URL `http://CRAN.R-project.org/package=psych`.

Shaw, P. (2014), "A nonparametric approach to solving a simple one-sector stochastic growth model," *Economics Letters*, 125, 447–450.

Shimizu, S., Hoyer, P. O., Hyvarinen, A., and Kerminen, A. J. (2006), "A linear non-Gaussian acyclic model for causal discovery," *Journal of Machine Learning Research*, 7, 2003–2030.

Vinod, H. D. (1978), "A survey of ridge regression and related techniques for improvements over ordinary least squares," *Review of Economics and Statistics*, 60, 121–131.

— (2004), "Ranking mutual funds using unconventional utility theory and stochastic dominance," *Journal of Empirical Finance*, 11(3), 353–377.

— (2008), *Hands-on Intermediate Econometrics Using R: Templates for Extending Dozens of Practical Examples*, Hackensack, NJ: World Scientific, ISBN 10-981-281-885-5, URL http://www.worldscibooks.com/economics/6895.html.

— (2013), "Generalized Correlation and Kernel Causality with Applications in Development Economics," *SSRN eLibrary*, URL http://ssrn.com/paper=2350592.

— (2014), "Matrix Algebra Topics in Statistics and Economics Using R," in "Handbook of Statistics: Computational Statistics with R, Vol. 34," , eds. Rao, M. B. and Rao, C. R., New York: North Holland, Elsevier Science, pp. 143–176.

— (2015a), "Generalized Correlation and Kernel Causality with Applications in Development Economics," *Communications in Statistics - Simulation and Computation*, accepted Nov. 10, 2015, URL http://dx.doi.org/10.1080/03610918.2015.1122048.

— (2015b), "Generalized Correlations and Instantaneous Causality for Data Pairs Benchmark," *SSRN eLibrary*, URL http://ssrn.com/abstract=2574891.

— (2016), *generalCorr: Generalized Correlations and Initial Causal Path*, fordham University, New York, R package version 1.0.0, URL https://CRAN.R-project.org/package=generalCorr.

Vinod, H. D. and López-de-Lacalle, J. (2009), "Maximum Entropy Bootstrap for Time Series: The meboot R Package," *Journal of Statistical Software*, 29, 1–19, URL http://www.jstatsoft.org/v29/i05/.

Zhang, K. and Hyvarinen, A. (2009), "On the Identifiability of the Post-Nonlinear Causal Model," *Uncertainty in Artificial Intelligence, UAI*, 647–655, URL http://arxiv.org/pdf/1205.2599.

Zheng, S., Shi, N.-Z., and Zhang, Z. (2012), "Generalized Measures of Correlation for Asymmetry, Nonlinearity, and Beyond," *Journal of the American Statistical Association*, 107, 1239–1252.