Causal Learning

Tubingen University

Ahmad Ehyaei

XICOR

Definition 0.1: Chatterjee Correlation

iven random variables X, Y, where is Y is not a constant, Chatterjee correlation ξ is defined as

$$\xi(X,Y) = \frac{\int Var(\mathbb{E}(1_{\{Y\geq t\}}|X))d\mu(t)}{\int Var(1_{\{Y\geq t\}})d\mu(t)},$$

where μ is the law of Y.

Let $\{(X_i, Y_i)\}_{i=1}^n$ be i.i.d. pairs following the same distribution as (X, Y). Rearrange the data as $(X_{(1)}, Y_{(1)}), \dots, (X_{(n)}, Y_{(n)})$ such that $X_{(1)} < \cdots < X_{(n)}$. Let r_i be the rank of $Y_{(i)}$, i.e. the number of j such that $Y_{(j)} \leq Y_{(i)}$. Then the correlation coefficient ξ_n is defined to be

$$\xi_n(X,Y) := 1 - \frac{3\sum_{i=1}^{n-1}|r_{i+1} - r_i|}{n^2 - 1}.$$

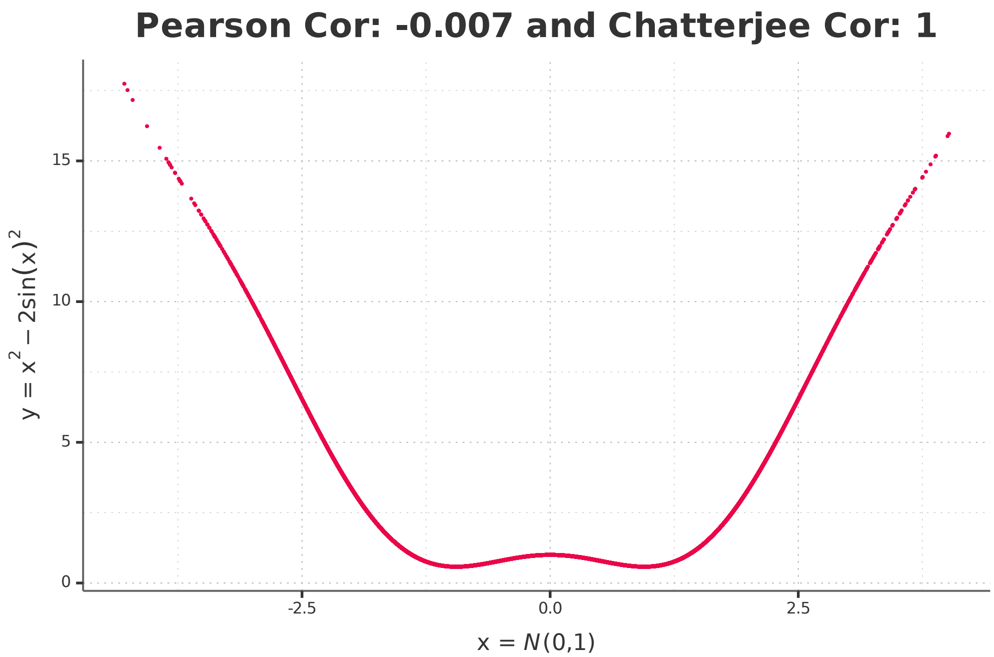
The properties of Chatterjee's correlation coefficient are:

- $\xi(X,Y) \in [0,1]$
- $\xi(X,Y) = 0$ if and only if X and Y are independent.
- $\xi(X,Y) = 1$ if and only if atleast one of X and Y is a measurable function of the other.
- ξ is not symmetric in X, Y. This is intentional and useful as we might want to study if Y is a measurable function of X, or X is a measurable function of Y. To get a symmetric coefficient, it suffices to consider max$(\xi(X,Y), \xi(Y,X))$.
- ξ_n is based on ranks, and for the same reason, it can be computed in $O(n \log n)$.

Theorem 0.1

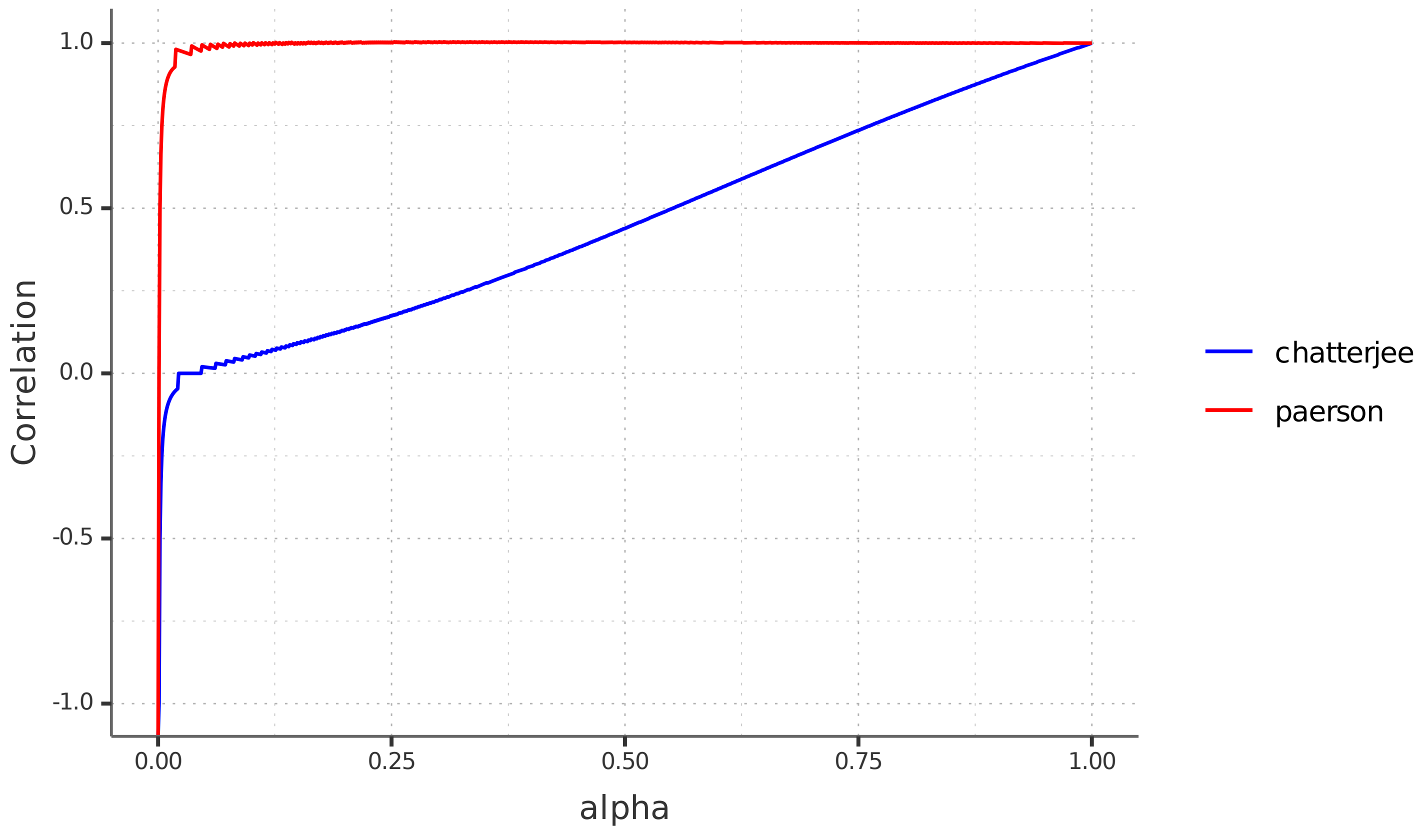
f Y is not almost surely a constant, then as $n \to \infty$, $\xi_n(X,Y)$ converges almost surely to $\xi(X,Y)$.

```
n = 100000
x <- rnorm(n)
z <- rnorm(n)
y = x^2-2*sin(x)^2+1
chatterjee = xicor(x, y, pvalue=TRUE)$x %>% round(3)
paerson = cor(x, y) %>% round(3)
```



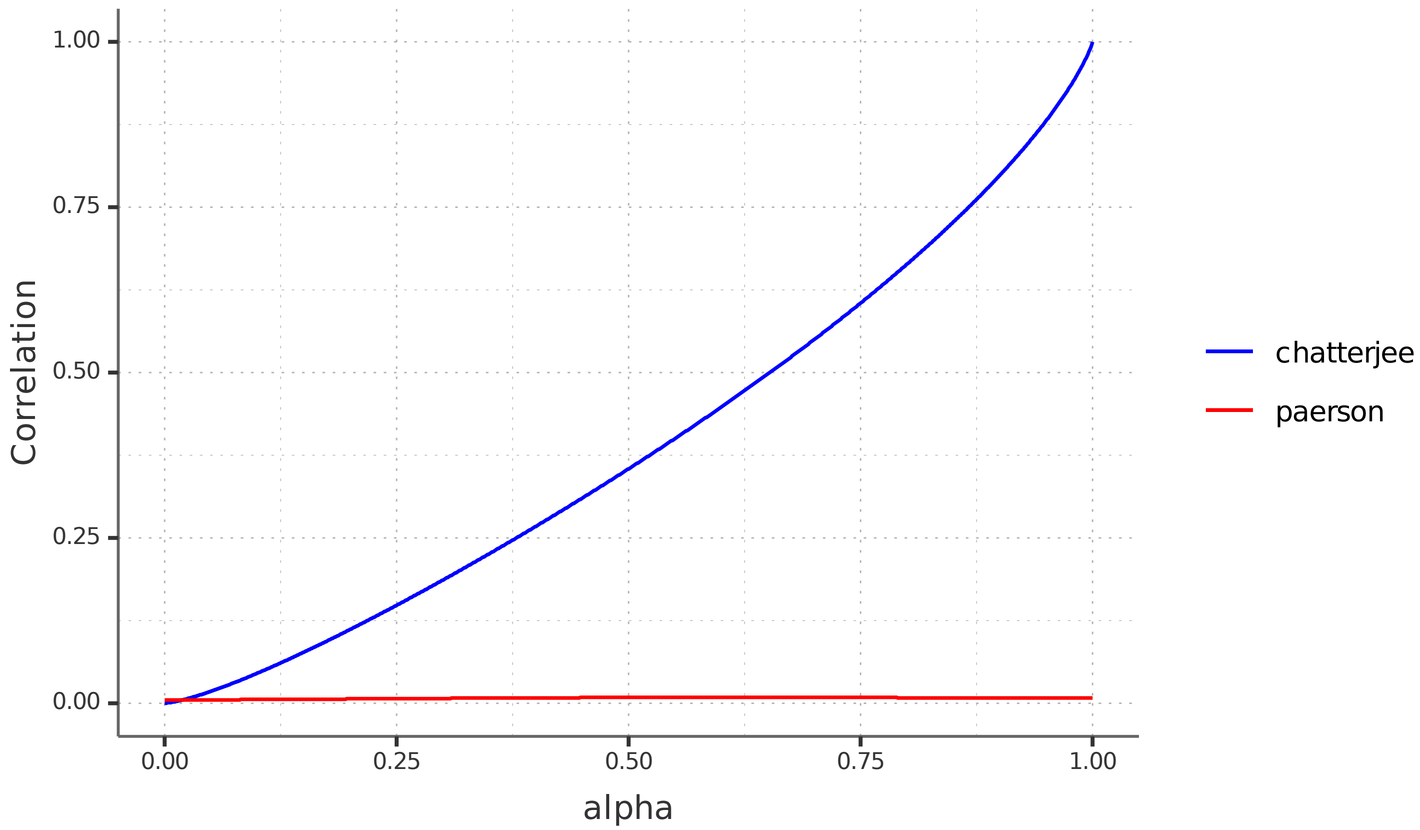

```
alpha = seq(0,1,0.001)
cor_table <- data.frame(alpha = alpha, chatterjee = 0, paerson = 0)
for(i in 1:length(alpha)){
y = alpha[i]*x + (1-alpha[i])*z
cor_table$chatterjee[i] = xicor(x, y, pvalue=TRUE)$x %>% round(3)
cor_table$paerson[i] = cor(x, y) %>% round(3)
}
cor_table$correction_term = alpha/(sqrt(alpha^2 + (1-alpha)^2))
write_rds(cor_table,"data/cor_table.rds")
```

```
sim = read_rds("data/cor_table.rds")
colors <- c("chatterjee" = "blue", "paerson" = "red")

ggplot()+
  geom_line(data = sim, aes(x = alpha, y = chatterjee/correction_term, color = "chatterjee"))+
  geom_line(data = sim, aes(x = alpha, y = paerson/correction_term, color = "paerson"))+
  theme_scientific()+
  labs(y = "Correlation", color = NULL)+
  scale_color_manual(values = colors)
```



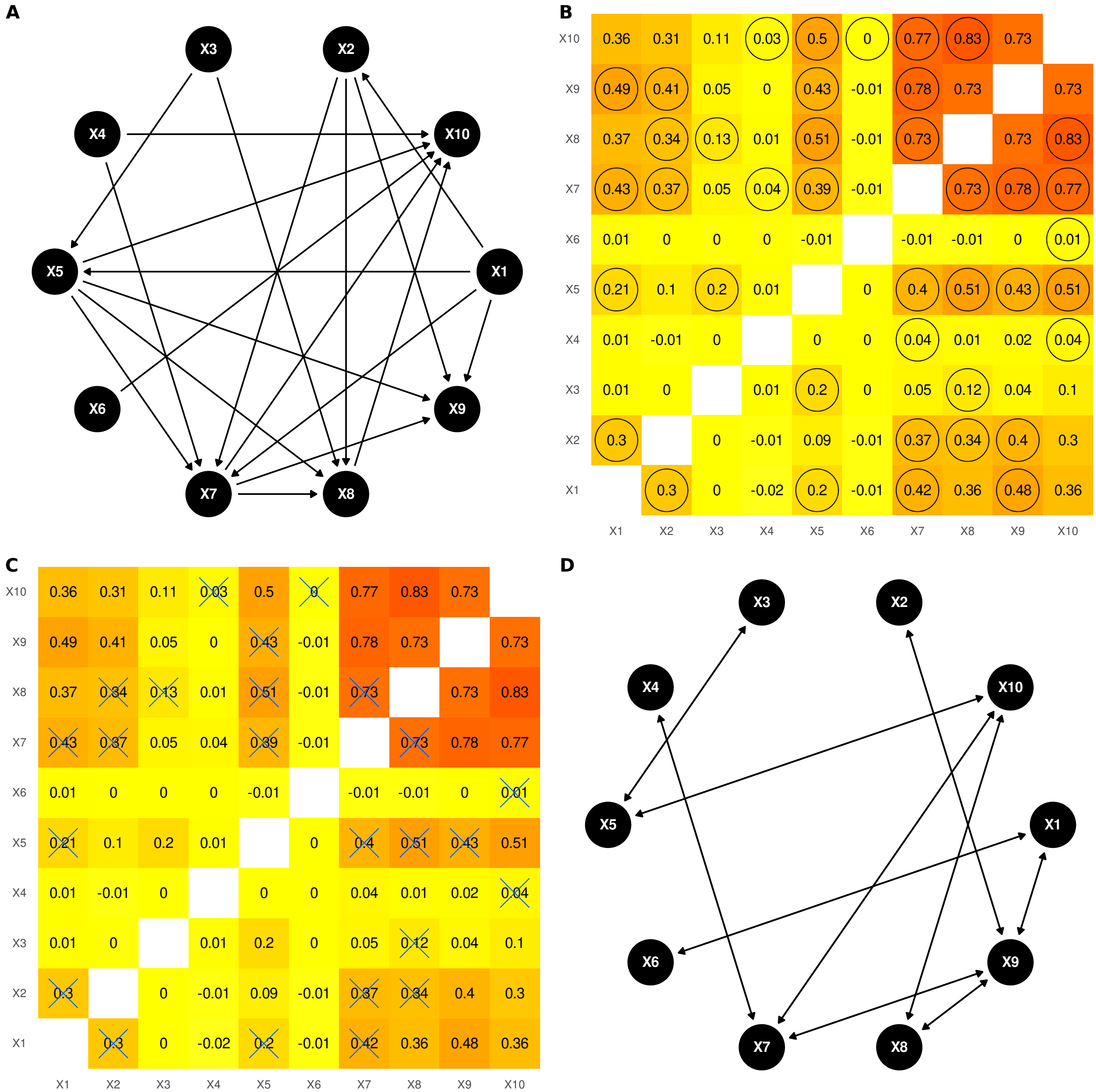

```
alpha = seq(0,1,0.001)
cor_table <- data.frame(alpha = alpha, chatterjee = 0, paerson = 0)

for(i in 1:length(alpha)){
y = alpha[i]*x^2 + (1-alpha[i])*z^2
cor_table$chatterjee[i] = xicor(x, y, pvalue=TRUE)$x %>% round(3)
cor_table$paerson[i] = cor(x, y) %>% round(3)
}
cor_table$correction_term = alpha/(sqrt(alpha^2 + (1-alpha)^2))
write_rds(cor_table,"data/cor_table_2.rds")
```

```
sim = read_rds("data/cor_table_2.rds")
colors <- c("chatterjee" = "blue", "paerson" = "red")

ggplot()+
  geom_line(data = sim, aes(x = alpha, y = chatterjee, color = "chatterjee"))+
  geom_line(data = sim, aes(x = alpha, y = paerson, color = "paerson"))+
  theme_scientific()+
  labs(y = "Correlation", color = NULL)+
  scale_color_manual(values = colors)
```



ESTIMATING LARGE CAUSAL POLYTREE SKELETONS FROM SMALL SAMPLES

```r
w = bernouli_dag(10, 0.4)

# 2. Plot Dag Related To Adjacency Matrix
dag <- adjacency_to_dag(w)
p_graph <- ggdag(dag, layout = "circle") + theme_dag()

# 3. Generate SCM corresponding to Noise and Adjacency Matrix
scm <- lin_scm(w = w, noise = normal_noise, n = 10000)

# 4. Estimate Tree by Chatterjee Algorithms
xi <- cor_mat(normal_mat(scm))

# 5. plot the correlation matrix VS Adjacency
p_true <- cor_adj_plot(xi, w)

# 6. Estimate Tree
e <- estimate_tree(xi)

e_dag <- adjacency_to_dag(e)
e_graph <- ggdag(e_dag, layout = "circle") + theme_dag()

# 7. find the difference between true and estimated graph
p_fault <- cor_fault_plot(xi, w, e)


p <- ggarrange(p_graph, p_true, p_fault,e_graph,
          labels = c("A", "B", "C","D"),
          ncol = 2, nrow = 2)
ggsave("plots/graph.pdf", p, device = cairo_pdf,width = 30, height = 30, unit ="cm")
```




```
[1] 0.2980773
```

```
[1] 0.3261963
```

```
[1] 0.1393531
```

```
[1] 0.1720412
```

#https://igraph.org/r/doc/isomorphic.html

In the below work the corelation extend to multivariate r.v. and use the nearest neighbor Azadkia-Chatterjee's correlation coefficient adapts to manifold data Fang Han and Zhihan Huang