Statistical Inference for Paths of Stochastic Processes

Philip Ernst, Ph.D.
Associate Professor of Statistics
Department of Statistics, Rice University

Support from NSF DMS-1811936 and ONR-N00014-18-1-2192
gratefully acknowledged

ONR Review
October 15, 2019

What is the problem that you are trying to solve?

Objective

We wish to build the first demonstrably correct statistical tests for testing independence (or dependence) of pairs of paths of stochastic processes. As part of a broad effort, we will focus on tests to detect independence for the following pairs of paths of Gaussian processes: Wiener processes, Ornstein-Uhlenbeck (OU) processes, fractional Ornstein-Uhlenbeck (fOU), and fractional Brownian motion (fBm).

Why is it important?

A fundamental statistical question

The mathematical importance of our work emerges via a fundamental yet crucial question for practitioners of statistics: given a pair of sequences of random variables $\{X_k, Y_k\}$ $(k = 1, 2, \ldots, n)$, how can we measure the strength of the dependence of the sequences X and the sequence Y?

An often misleading statistic

The classical Pearson correlation coefficient offers a solution that is standard and often powerful. However, it is also widely used even in situations where little is known about its empirical properties, e.g. when the sequence of random variables are not identically distributed or independent. The Pearson correlation is often calculated between two time series.

After ninety years, we finally have mathematical proof that it is misleading!

Let $(X_k)_k$ and $(Y_k)_k$ be two independent random walks (partial sums of two independent sets of i.i.d. variables). The empirical correlation is defined as

$$\theta_n = \frac{\sum_{k=1}^n X_k Y_k - \frac{1}{n}(\sum_{k=1}^n X_k)(\sum_{k=1}^n Y_k)}{\sqrt{\sum_{k=1}^n X_k^2 - \frac{1}{n}(\sum_{k=1}^n X_k)^2}\sqrt{\sum_{k=1}^n Y_k^2 - \frac{1}{n}(\sum_{k=1}^n Y_k)^2}}. \quad (1)$$

The recent work of Ernst, Shepp, Wyner (*The Annals of Statistics*, 2017) closes a 90-year old open problem posed by Udny Yule (Yule, 1926) regarding the dispersion of so-called "nonsense" correlation. The work shows that statistical tests based on Pearson correlation using *partial sums of random variables* are indeed **mathematically erroneous**.

Yule's "nonsense correlation"

Let us quickly review the work of Ernst, Shepp, and Wyner (2017) Yule's "Nonsense Correlation" Solved! *The Annals of Statistics*, **45**: 1789-1809. The problem is to mathematically prove Yule's 1926 empirical finding of "nonsense correlation."

Yule (1926)

"We sometimes obtain between quantities varying with time (time-variables) quite high correlations to which we cannot attach any physical significance whatever, although under the ordinary test the correlation would be held to be certainly 'significant'. The occurrence of such 'nonsense correlations' makes one mistrust the serious arguments that are sometimes put forward on the basis of correlations between time series." (*Journal of the Royal Statistical Society*, 89(1)).

What Phenomena Was Yule Addressing?

Two DISTINCT but often confused phenomena.

- "Spurious correlation": Correlation observed when two time series are themselves dependent on an unobserved third time series.
- "Nonsense correlation": Correlation observed between two independent time series without any regard to a third time series.

Formal Mathematical Setup of the Problem

Consider two sequences of i.i.d. random variables, $X \equiv \{X_n\}_{n\in\mathbb{N}}$ and $Y \equiv \{Y_n\}_{n\in\mathbb{N}}$ with finite variances. Let $S_n = X_1 + ... + X_n$ and $T_n = Y_1 + ... + Y_n$. Consider the following statistic:

$$\theta_n^{(1)} = \frac{\sum_{i=1}^n S_i T_i - \frac{1}{n}\sum_{i=1}^n S_i \sum_{j=1}^n T_j}{\sqrt{\sum_{i=1}^n S_i^2 - \frac{1}{n}(\sum_{i=1}^n S_i)^2}\sqrt{\sum_{i=1}^n T_i^2 - \frac{1}{n}(\sum_{i=1}^n T_i)^2}}.$$

"Nonsense Correlation"

"Nonsense correlation": Correlation observed between two independent time series, without any regard to a third time series.

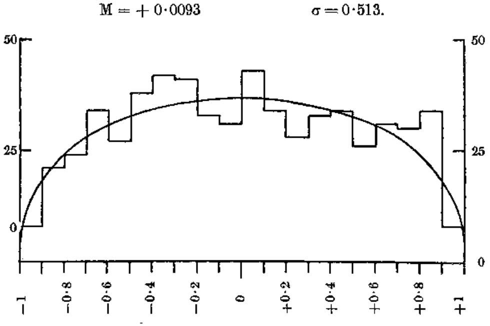

Figure: Yule (1926): 600 correlations between samples of 10 observations from from two independent and identically distributed random walks.

Is $\sigma \approx .5$????

A Simulation: "Volatile" Correlation

In lieu of "nonsense correlation," Ed George conceived of the term "volatile" correlation; the correlation is "volatile" in the sense that its distribution is heavily dispersed and is frequently large in absolute value.

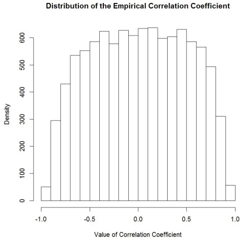

Figure: Consider S_i and T_i to be two independent symmetric random walks. The histogram reports the empirical correlation of partial sums of two independent symmetric random walks with $n = 10,000$ for each random walk, calculated by $\theta_n^{(2)}$. The simulation is repeated 10,000 times.

Simulation Data

The histogram reports that the middle 95% of the observed correlation coefficients fall in the interval [-.83, .83]. The lesson to be learned is that Pearson correlation is not always a useful statistic; in the case of two independent random walks, the observed correlation coefficient has a very different distribution than the nominal t-distribution. This was recently brought to light by McShane and Wyner (2011) in *Annals of Applied Statistics*, critique of efforts to reconstruct the earth's historical temperatures using correlated time series. **Providing solid mathematical justification for avoiding these types of methodological errors is a major motivation for our work!**

Moments of θ- A Simulation

The result of averaging 10,000 simple Monte Carlo iterations of the first ten moments of θ can be found below in Table 1.

$\mathbb{E}\left[\theta^0\right]$	$\mathbb{E}\left[\theta^1\right]$	$\mathbb{E}\left[\theta^2\right]$	$\mathbb{E}\left[\theta^3\right]$	$\mathbb{E}\left[\theta^4\right]$	$\mathbb{E}\left[\theta^5\right]$
1	-.00116634	.235057	-.000524066	.109276	-.000283548
$\mathbb{E}\left[\theta^6\right]$	$\mathbb{E}\left[\theta^7\right]$	$\mathbb{E}\left[\theta^8\right]$	$\mathbb{E}\left[\theta^9\right]$	$\mathbb{E}\left[\theta^{10}\right]$	
.0609591	-.00016797	.0378654	-.000105611	.0251693	

Table: Moments of θ obtained from 10,000 Monte Carlo iterations.

Note that odd moments should all be zero, since θ is symmetric. But recall that the goal of our work is to analytically compute the moments...a question that remained open since 1926!

Back to Statistics: The Statistic $\theta_n^{(2)}$

Consider two sequences of i.i.d. random variables, $X \equiv \{X_n\}_{n\in\mathbb{N}}$ and $Y \equiv \{Y_n\}_{n\in\mathbb{N}}$ with finite variances. Let $S_n = X_1 + ... + X_n$ and $T_n = Y_1 + ... + Y_n$. Consider the following statistic:

$$\theta_n^{(2)} = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i - \frac{1}{n}\sum_{i=1}^n X_i \frac{1}{n}\sum_{j=1}^n Y_j}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2 - (\frac{1}{n}\sum_{i=1}^n X_i)^2}\right)\left(\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2 - (\frac{1}{n}\sum_{i=1}^n Y_i)^2}\right)}$$

It is easy to see that when X and Y are independent, $\theta_n^{(2)} \to 0$ almost surely (by the law of large numbers). Thus, $\theta_n^{(2)}$ could (at least in principle) be used to test the independence of the two sequences.

The Statistic $\theta_n^{(1)}$

$$\theta_n^{(1)} = \frac{\sum_{i=1}^n S_i T_i - \frac{1}{n}\sum_{i=1}^n S_i \sum_{j=1}^n T_j}{\sqrt{\sum_{i=1}^n S_i^2 - \frac{1}{n}(\sum_{i=1}^n S_i)^2}\sqrt{\sum_{i=1}^n T_i^2 - \frac{1}{n}(\sum_{i=1}^n T_i)^2}}.$$

The same is NOT the case for $\theta_n^{(1)}$. When X and Y are independent, $\theta_n^{(1)}$ is asymptotically non-zero. This is because the partial sums S_n and S_m are are dependent while only X_n and X_m are independent. Thus, the two statistics $\theta_n^{(1)}$ and $\theta_n^{(2)}$ are only *formally similar* but with *different natures*.

A Formal Setup

Let $W_1(t)$ and $W_2(t)$ denote two independent Wiener processes on the interval [0,1]. We will prove Yule's phenomenon by analytically determine the second moment of the empirical correlation coefficient, the latter which is given by the following statistic:

$$\theta := \frac{\int_0^1 W_1(t)W_2(t)dt - \int_0^1 W_1(t)dt \int_0^1 W_2(t)dt}{\sqrt{\int_0^1 W_1^2(t)dt - \left(\int_0^1 W_1(t)dt\right)^2}\sqrt{\int_0^1 W_2^2(t)dt - \left(\int_0^1 W_2(t)dt\right)^2}}, \quad (2)$$

NOTE that θ is the continuous time analog of $\theta_n^{(1)}$, which we recall as

$$\theta_n^{(2)} = \frac{\sum_{i=1}^n S_i T_i - \frac{1}{n}\sum_{i=1}^n S_i \sum_{j=1}^n T_j}{\sqrt{\sum_{i=1}^n S_i^2 - \frac{1}{n}(\sum_{i=1}^n S_i)^2}\sqrt{\sum_{i=1}^n T_i^2 - \frac{1}{n}(\sum_{i=1}^n T_i)^2}}.$$

It was shown in Phillips (1986) that $\theta_n^{(1)}$ converges weakly to θ.

Vision for Proof of Yule's "Nonsense Correlation" and Preview of Main Result

Part 1: Rewrite θ in an alternate form that involves stochastic integrals.

Part 2: We will introduce the function F, which is well suited to calculating the moments of θ. We will proceed to explicitly calculate F. This part uses tools from classical functional anaylsis, Mercer's theorem, etc.

Part 3: We will eventually obtain the variance of θ as

$$\int_0^\infty du_1 \int_0^{u_1} \sqrt{\frac{u_1 u_2}{\sinh u_1 \sinh u_2}} \frac{u_1 u_2}{u_1 + u_2} \left(\frac{\frac{1}{u_1^2}\left(1 - \frac{u_1}{\sinh u_1}\cosh u_1\right) - \frac{1}{u_2^2}\left(1 - \frac{u_2}{\sinh u_2}\cosh u_2\right)}{u_1 - u_2} \right) du_2 .$$

Although it is not possible to calculate the double integral above in elementary terms, it has removable singularities and shockingly converges very nicely at all points where any of u_1, u_2, or $u_1 - u_2$ vanishes. In addition to providing an explicit formula for the second moment of θ, we offer implicit formulas for higher moments of θ.

The Moments

After enormous efforts, we obtain the following theorem:

Theorem

$$\mathbb{E}\left[\theta^{2n}\right] = \binom{2n}{n}\frac{2n}{2^{2n}}\sum_{r=n}^{\infty}\int_0^1 \binom{r-1}{n-1}(1-v^2)^{n-1}v^{2(r-n)}dv \int_0^{\infty} 2S(u)T(u)s_r(u)du, \tag{3}$$

where $S(u) = \sqrt{\frac{u}{\sinh u}}$ *and* $T(c) = \frac{1}{c}\frac{S'(c)}{S(c)}$.

The second moment of θ

The proposition below gives an explicit formula for the second moment of θ:

Proposition

The second moment of θ, corresponding to $n = 1$, can be calculated explicitly as

$$\int_0^\infty du_1 \int_0^{u_1} \sqrt{\frac{u_1 u_2}{\sinh u_1 \sinh u_2}} \frac{u_1 u_2}{u_1 + u_2} \left(\frac{\frac{1}{u_1^2}\left(1 - \frac{u_1}{\sinh u_1}\cosh u_1\right) - \frac{1}{u_2^2}\left(1 - \frac{u_2}{\sinh u_2}\cosh u_2\right)}{u_1 - u_2} \right) du_2 .$$

The final result

Using the above expression for the second moment of θ we numerically obtain a value of .240522. Recall that the Monte Carlo simulation in Table 1 reported a value of .235057. One must proceed numerically to calculate higher order moments.

How is it solved today, and what are the limitations of current solutions?

The work of Ernst et al. (2017) analytically confirms that this classical t-test is unsuitable for even the basic case of two i.i.d. random walks. The only currently available test to interpret this correlation is the classical Pearson correlation test, with a test statistic t^* given by $t^* := \theta_n \sqrt{n-2}/\sqrt{1-(\theta_n)^2}$. There are of course (mostly for discrete-time data) cross-covariance and distance metrics, as well as copula methods, but none of these methods are formulated for tests of independence of paths of stochastic processes.

Enter Larry Brown: Email exchanges in October 2017 (with approximate paraphrasing)

Larry: "Now that you've solved Yule, what about thinking about a test that would do better."
Philip: "Are there not such pathwise tests for inference for *pairs* of stochastic processes?"
Larry: "I don't think so... perhaps statistical inference and stochastic processes have not quite met. I checked with a few experts (who would know if such a question were indeed open) and it seems open indeed. Check Horvath *JMVA* 2013 for a start..."

Horvath et al. (2013)

Horvath et. al (2013) tests for independence in a set of n square-integrable random functions under a stationarity assumption. It considers the finite-dimensional Box-Ljung-Pierce approach and is based on sample auto-covariance functions. Under the null hypothesis that the random functions are independent, the auto-covariances' $L^2\left([0,1]^2\right)$-norm for fixed "lags" h should be close to 0 for all possible pairs of functions. The test statistic sums over all lags up to $h \leq H(n)$.

Horvath et al. (2013)

The alternative in Horvath et al. as in any infinite-dimensional context, has to be quite specific, to avoid a low test power. As $n \to \infty$, by tailoring $H(n)$, the authors show asymptotic normality of their test statistic under the null, and explosion under the alternative, thus achieving high power.

This being said, Horvath et al. does not answer the critical question of interest to us, since in the context of the Pearson correlation questions for stochastic processes, one is primarily interested in pairs of paths, i.e. the case $n = 2$.

What About A Common Sense Idea?

A “common-sense” idea would be to use the increments of W in θ_n, since this would get us back to the case of i.i.d. data when starting with true random walks. This idea defeats the purpose of trying to exploit the fact that data is presented to us as a time series or process path, and more importantly, it is not directly applicable for any process other than the Wiener process.

Long Memory!!

However, for processes with long memory such as fractional Brownian motion (fBm) B^H, passing to its increment process (fractional Gaussian noise, or fGn) may present no advantage whatsoever.

Enter Larry Brown: Email exchanges in December 2017 (with approximate paraphrasing)

Larry: "Philip, you should start with moment based tests. But for this you need to know the law of θ. You should know it for every Gaussian process imaginable! Besides, this would solve the remaining ninety year old open problem!

A more general set-up achieved in Ernst, L.C.G. Rogers, and Q. Zhou (2019)

For two continuous-time Gaussian processes $X_1(t), X_2(t)$ on the interval $[0, T]$, Pearson's correlation statistic is

$$\rho(T) = \frac{Y_{12}(T)}{\sqrt{Y_{11}(T)Y_{22}(T)}},$$

where the random variables $Y_{ij}(T)\,(i,j=1,2)$ are defined as

$$Y_{ij}(T) = \int_0^T X_i(u)X_j(u)du - T\,\bar{X}_i\bar{X}_j, \quad \bar{X}_i = \frac{1}{T}\int_0^T X_i(u)du.$$

First observation

$$\rho(T) = \frac{Y_{12}(T)}{\sqrt{Y_{11}(T)Y_{22}(T)}}.$$

ρ is the ratio of three (dependent) random variables. If we can compute the joint moment generating function (MGF), then the moments of ρ of any order can be evaluated (in principle).

First observation

From MGF to moments

Let $M_Z(t) = E(e^{tZ})$ be the MGF of a random variable such that $M_Z(t)$ exists in a neighborhood of zero. Then, for $k = 1, 2, \ldots$

$$E[Z^k] = M_Z^{(k)}(0),$$
$$E[Z^{-k}] = \frac{1}{\Gamma(k)} \int_0^\infty t^{b-1} M_Z(-t) dt.$$

First observation

From MGF to moments

Let $M_Z(t) = E(e^{tZ})$ be the MGF of a random variable such that $M_Z(t)$ exists in a neighborhood of zero. Then, for $k = 1, 2, \ldots$

$$E[Z^k] = M_Z^{(k)}(0),$$
$$E[Z^{-k}] = \frac{1}{\Gamma(k)} \int_0^\infty t^{b-1} M_Z(-t) dt.$$

The second equation follows from

$$\frac{z^k}{\Gamma(k)} \int_0^\infty t^{k-1} e^{-tz} dt = 1.$$

So the formula for $E[Z^{-k}]$ also applies to non-integer k!

First observation

From MGF to moments

Let Z, W be two random variables, where $W \geq 0$, and

$$M_{Z,W}(t_1, t_2) = E[e^{t_1 Z + t_2 W}]$$

be the (joint) MGF which exists in a neighborhood of 0. Then,

$$M_{Z,W}^{(k,0)}(0, -t) = E[Z^k e^{-tW}]$$

for $k = 1, 2, \ldots$. For $m > 0$,

$$E\left[\frac{Z^k}{W^m}\right] = \frac{1}{\Gamma(m)} \int_0^\infty t^{m-1} M_{Z,W}^{(k,0)}(0, -t) dt.$$

First observation

$$\rho(T) = \frac{Y_{12}(T)}{\sqrt{Y_{11}(T)Y_{22}(T)}}.$$

For our problem, define the MGF of (Y_{11}, Y_{12}, Y_{22}) by

$$\begin{aligned}\phi(S) &= E\left[\exp\left\{-\frac{1}{2}(s_{11}Y_{11} + 2s_{12}Y_{12} + s_{22}Y_{22})\right\}\right] \\ &= E\left[\exp\left\{-\frac{1}{2}\int_0^T (X(u)-\bar{X})^\top S\,(X(u)-\bar{X})du\right\}\right],\end{aligned}$$

where $X(t) = (X_1(t), X_2(t))$, $\bar{X} = (\bar{X}_1, \bar{X}_2)$ and S is a 2×2 positive-definite matrix with entries s_{ij}.

First observation

From MGF to moments

$$\rho(T) = \frac{Y_{12}(T)}{\sqrt{Y_{11}(T)Y_{22}(T)}}.$$

$$\phi(S) = E\left[\exp\left\{-\frac{1}{2}(s_{11}Y_{11} + 2s_{12}Y_{12} + s_{22}Y_{22})\right\}\right].$$

$$E\rho^k = \frac{(-1)^k}{2^k\Gamma(k/2)^2}\int_0^\infty\int_0^\infty s_{11}^{k/2-1}s_{22}^{k/2-1}\frac{\partial^k\phi}{\partial s_{12}^k}(s_{11},0,s_{22})\,ds_{11}\,ds_{22}.$$

Second observation

If we can compute

$$\psi(S,z) = E\left[\exp\left\{-\frac{1}{2}\int_0^T (X(u)+z)^\top S\,(X(u)+z)du\right\}\right], z \in \mathbb{R}^2,$$

then $\phi(S)$ can be obtained by

$$\begin{aligned}\phi(S) &= E\left[\exp\left\{-\frac{1}{2}\int_0^T (X(u)-\bar{X})^\top S\,(X(u)-\bar{X})du\right\}\right] \\ &= \int_{\mathbb{R}^2} \frac{T(\det S)^{1/2}\psi(S,z)}{2\pi}\,dz.\end{aligned}$$

Evaluating $\psi(S, z)$

Hence, we only need to figure out how to compute

$$\psi(S, z) = E\left[\exp\left\{-\frac{1}{2}\int_0^T (X(u) + z)^\top S\,(X(u) + z)du\right\}\right].$$

There are many existing results and tools for evaluating such quadratic functionals of Gaussian processes, in particular Donati-Martin and Yor (1997).

Evaluating $\psi(S, z)$

$$\psi(S,z) = E\left[\exp\left\{-\frac{1}{2}\int_0^T (X(u)+z)^\top S\,(X(u)+z)du\right\}\right].$$

Reducing to one-dimensional problems

In particular, if

- X_1, X_2 are independent,
- for any rotation matrix R, the distribution of RX is the same as X,

then the calculation of $\psi(S, z)$ can be further reduced to that of its one-dimensional counterparts.

Wiener processes

If X_1, X_2 are two independent Wiener processes, then

$$\phi_{\mathrm{Bm}}(S) = \left(\frac{\theta_1 \theta_2 T^2}{\sinh \theta_1 T \sinh \theta_2 T} \right)^{1/2},$$

from which we get the formula provided in Ernst et al. (2017).

$$E\rho^2 = \int_0^\infty \int_0^v \frac{uv\sqrt{uv}}{(v^2 - u^2)\sqrt{\sinh u \sinh v}} \left(\frac{1}{u \tanh u} - \frac{1}{v \tanh v} - \frac{1}{u^2} + \frac{1}{v^2} \right) du\, dv.$$

Wiener processes

k	2	4	6	8
$E\rho^k$	0.240523	0.109177	0.060862	0.037788
k	10	12	14	16
$E\rho^k$	0.025114	0.017504	0.012641	0.009385

Table: Numerical values of the moments of Yule's nonsense correlation for two independent Wiener processes.

Wiener processes

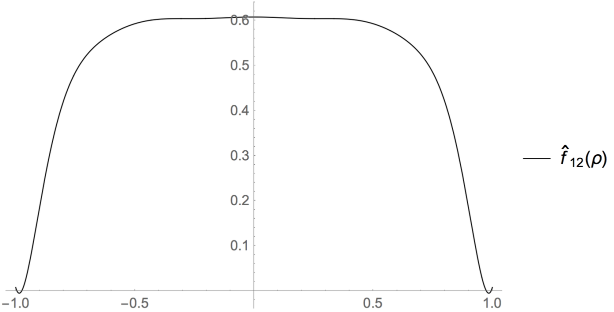

Figure: The 12th-order polynomial approximation to the probability density function of ρ for two independent Wiener processes.

Ornstein-Uhlenbeck processes

If $X = (X_1, X_2)$ follows the stochastic differential equation

$$dX(t) = -rX(t)dt + dW(t), \qquad r \in (0, \infty),$$

then,

$$\phi_{\mathrm{OU}}(S; r) = g(\theta_1^2; r)g(\theta_2^2; r),$$

where θ_1^2, θ_2^2 are the eigenvalues of S and

$$g(\theta^2) = \sqrt{T}e^{rT/2}\left\{\frac{\theta^2}{\eta^4}[2r(\cosh\eta T - 1) + \eta\sinh\eta T] + \frac{r^2 T}{\eta^3}[\eta\cosh\eta T + r\sinh\eta T]\right\}^{-1}$$

Ornstein-Uhlenbeck processes

r	0.1	0.2	0.3	0.4	0.5	1
$E\rho^2$	0.23209	0.22438	0.21734	0.21091	0.20504	0.18231
r	2	5	10	20	50	100
$E\rho^2$	0.15583	0.11454	0.07627	0.04404	0.01907	0.00971

Table: Numerical values of the second moment of Yule's nonsense correlation for two independent Ornstein-Uhlenbeck processes with mean reversion parameter r and $T = 1$.

Brownian bridges

If $X = (X_1, X_2)$ follows the stochastic differential equation

$$dX(t) = -\frac{X(t)}{1-t}dt + dW(t), \qquad t \in [0,1],$$

then

$$\phi_{\mathrm{Bb}}(S) = \frac{\theta_1\theta_2}{4\sinh(\theta_1/2)\sinh(\theta_2/2)},$$

where θ_1^2, θ_2^2 are the eigenvalues of S.

Brownian bridges

k	2	4	6	8
$E\rho^k$	0.149001	0.047864	0.0201829	0.009876

Table: Numerical values of the moments of Yule's nonsense correlation for two independent Brownian bridges.

Correlated Brownian motions

Let $X_1(t)$, $X_2(t)$ be two Brownian motions with constant correlation c, represented by the following SDE

$$dX_1(t) = dW_1(t), \qquad dX_2(t) = c\,dW_1(t) + \sqrt{1-c^2}\,dW_2(t).$$

Then the process $MX(t)$ is a two-dimensional Brownian motion with independent coordinates where

$$M = M(c) = \begin{bmatrix} 1 & 0 \\ -c(1-c^2)^{-1/2} & (1-c^2)^{-1/2} \end{bmatrix}.$$

Hence,

$$\phi_{\mathrm{cBm}}(S) = \phi_{\mathrm{Bm}}((M^{-1})^\top S M^{-1}).$$

Correlated Brownian motions

c	0	0.1	0.2	0.3	0.4
$E\rho$	0	0.08873	0.17792	0.26804	0.35963
$E\rho^2$	0.24052	0.24550	0.26061	0.28636	0.32368
$\mathrm{Var}(\rho)$	0.2405	0.2376	0.2290	0.2145	0.1943
c	0.5	0.6	0.7	0.8	0.9
$E\rho$	0.45338	0.55004	0.65071	0.75698	0.87151
$E\rho^2$	0.37407	0.43986	0.52477	0.63509	0.78298
$\mathrm{Var}(\rho)$	0.1685	0.1373	0.1013	0.0621	0.0235

Table: Numerical values of the moments of Yule's nonsense correlation for two correlated Brownian motions with correlation coefficient c ($T = 1$).

Asymptotics of $\rho(T)$

In all previous examples, we have assumed $T = 1$.

Asymptotics of $\rho(T)$

In all previous examples, we have assumed $T = 1$.

What if we let T go to ∞?

Will the distribution of $\rho(T)$ change?

Will $\rho(T)$ goes to zero?

Wiener processes

By the self-similarity of Wiener process, $W(tT)$ has the same distribution as $\sqrt{T}\tilde{W}(t)$ where $\tilde{W}$ is another Wiener process. This implies that the distribution of $\rho(T)$ does not depend on T.

Ornstein-Uhlenbeck processes

Recall that an stationary OU processes with mean-reverting parameter $r > 0$ has $E[X(t)] = 0$ and $\mathrm{Var}[X(t)] = 1/2r$ for every t.

Central limit theorem for $\rho(T)$

For two independent OU processes with mean-reverting parameter $r > 0$, we can prove that

$$\sqrt{T}\rho(T) \xrightarrow{D} N\left(0, \frac{1}{2r}\right).$$

Ornstein-Uhlenbeck processes

Central limit theorem for $\rho(T)$

Recall that $\rho(T) = Y_{12}(T)/\sqrt{Y_{11}(T)Y_{22}(T)}$. To prove the CLT, Since $Y_{11}(T)/T \to 1/2r$ a.s. by ergodic theorem, we only need to show the weak convergence of $Y_{12}/\sqrt{T}$. Recall

$$Y_{12}(T) = \int_0^T X_1(u)X_2(u)du - T\bar{X}_1(T)\bar{X}_1(T).$$

- Show that $\sqrt{T}\bar{X}_1(T)\bar{X}_1(T)$ converges to 0 in L^2.
- Show the normal convergence of $T^{-1/2}\int_0^T X_1(u)X_2(u)du$ by computing its characteristic function.

Limitations

The above methodology does not work when X is not a solution to a linear SDE!

A Strategy Going Forward: Joint work with Frederi Viens, Michigan State University

Our proposed tests' originality is their major departure from both t-tests and tests based on asymptotic normality. We know that for Wiener and fBm paths, the Pearson correlation θ's distribution is constant for increasing horizon $T \to \infty$, and is not in the least related to a CLT (unlike the t-distribution!). We can use this remarkable stationarity at the level of a test statistic to our advantage! We will rely heavily on the theory of **Wiener chaos** and the optimal fourth moment theorem of Nourdin and Peccati (2015).

References

1. Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Statistical Science* **10** 364-376.
2. Boas, R. (1954). Entire Functions. Academic Press: New York.
3. Donati-Martin, C. and Yor, M. (1997). Some Brownian functionals and their laws. *The Annals of Probability* **25** 1011-1058.
4. Ernst, P., Shepp, L., and Wyner, A. (2017). Yule's 'nonsense correlation' solved! *The Annals of Statistics*, **45**:1789-1809.
5. Ernst, P., Rogers, L.C.G., and Zhou, Q. (2019). The distribution of Yule's 'nonsense correlation' *https://arxiv.org/abs/1909.02546*.
6. Horvath, L., Huskova, M., and Rice, G. (2013). Test of independence for functional data. *Journal of Multivariate Analysis*, **117**: 100-119.

References (continued)

1. McShane, B.B. and Wyner, A.J. (2011). A statistical analysis of multiple temperature proxies: are reconstructions of surface temperatures over the last 1000 years reliable? *Annals of Applied Statistics Statistics* **5** 5-44.

2. Nourdin, I. and Peccati, G. (2015). The optimal fourth moment theorem. *Proceedings of the American Mathematical Society*, **143** 3123-3133.

3. Phillips, P.C.B. (1986). Understanding spurious regressions in econometrics. *Journal of Econometrics* **33** 311-340.

4. Yule, G.U. (1926). Why do we sometimes get nonsense correlations between time series? A study in sampling and the nature of time series. *Journal of the Royal Statistical Society* **89** 1-63.

Part 1

Proposition

We have the equality

$$\theta = \frac{X_{1,2}}{\sqrt{X_{1,1}X_{2,2}}}, \tag{4}$$

where

$$X_{i,j} = \int_0^1 \int_0^1 (\min(s_1, s_2) - s_1 s_2)\, dW_i(s_1) dW_j(s_2), \tag{5}$$

and where θ *is, as before,*

$$\theta := \frac{\int_0^1 W_1(t)W_2(t)dt - \int_0^1 W_1(t)dt \int_0^1 W_2(t)dt}{\sqrt{\int_0^1 W_1^2(t)dt - \left(\int_0^1 W_1(t)dt\right)^2}\sqrt{\int_0^1 W_2^2(t)dt - \left(\int_0^1 W_2(t)dt\right)^2}}.$$

"Wishful Thinking" is Usually Working Behind the Scenes

QUESTION: Why should I want the term $\min(s_1, s_2) - s_1 s_2$??

"Wishful Thinking" is Usually Working Behind the Scenes

QUESTION: Why should I want the term $\min(s_1, s_2) - s_1 s_2$??
ANSWER: It is the covariance of a pinned Wiener process on [0,1].

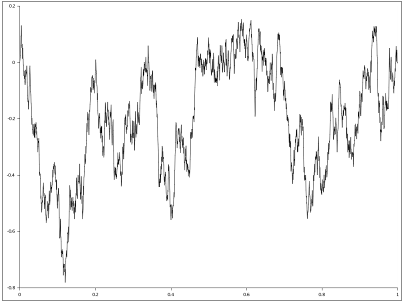

Figure: Pinned Wienier process on [0,1].

Part 2: Defining and Calculating F

We will need to appeal to **Mercer's theorem**, which gives a representation of a symmetric positive-definite function on a square as a sum of a convergent sequence of product functions.

We define for $|a| \leq 1, \beta_i \geq 0, i = 1, 2,$ the integral

$$F(\beta_1, \beta_2, a) = \mathbb{E}\left[e^{a\beta_1\beta_2 X_{1,2} - \frac{\beta_1^2}{2} X_{1,1} - \frac{\beta_2^2}{2} X_{2,2}}\right], \tag{6}$$

where the $X_{i,j}$ are as defined in Equation (5). Under the above conditions, the expectation is finite and thus F is well defined. This is because $X_{1,2} = \theta\sqrt{X_{1,1}X_{2,2}}$ and $|\theta| \leq 1$ so the exponent is at most

$$-\frac{1}{2}\left(\beta_1\sqrt{X_{1,1}} - \beta_2\sqrt{X_{2,2}}\right)^2 \leq 0. \tag{7}$$

Thus the expectand is bounded by unity, and so, for this range, $F(\beta_1, \beta_2, a) \leq 1$.

Key Theorem

Here is the key theorem.

Theorem

$$F(\beta_1, \beta_2, a) = \frac{1}{\sqrt{\frac{\sinh c^+}{c^+} \frac{\sinh c^-}{c^-}}}, \tag{8}$$

where

$$c^\pm = c^\pm(\beta_1, \beta_2, a) = \sqrt{\frac{(\beta_1^2 + \beta_2^2) \pm \sqrt{(\beta_1^2 - \beta_2^2)^2 + 4a^2\beta_1^2\beta_2^2}}{2}}. \tag{9}$$

Intuition: the functions M and K

Motivated by the definition of $X_{i,j}$ in equation (5), which we recall as:

$$X_{i,j} = \int_0^1 \int_0^1 (\min(s_1, s_2) - s_1 s_2)\, dW_i(s_1) dW_j(s_2). \tag{10}$$

we define M by

$$M(s_1, s_2) = \min(s_1, s_2) - s_1 s_2, \ \ s_1, s_2 \in [0, 1], \tag{11}$$

which is the covariance of pinned Brownian motion on $[0, 1]$.For $i_1, i_2 \in \{1, 2\}$ and $s_1, s_2 \in [0, 1]$, we define the kernel function K_{i_1,i_2} by

$$K_{i,i}(s_1, s_2) = -\beta_i^2 M(s_1, s_2) \tag{12}$$

and

$$K_{1,2}(s_1, s_2) = K_{2,1}(s_1, s_2) = a\beta_1\beta_2 M(s_1, s_2).$$

Let

$$K = \begin{bmatrix} K_{1,1} & K_{1,2} \\ K_{2,1} & K_{2,2} \end{bmatrix}.$$

Eigenvalues of T_K

The eigenvalues of T_K can be calculated in the form of $\frac{1}{\pi^2 n^2}$. This allows us to use the following product formula (Boas, 1954)

$$\frac{\sin(z)}{z} = \prod_{n=1}^{\infty}\left(1 - \frac{z^2}{\pi^2 n^2}\right). \tag{13}$$

Note that for any complex number z,

$$\frac{\sin(z)}{z} = \frac{\sinh(-iz)}{-iz}.$$

Since $z^{\pm}$ are purely imaginary, we observe $z^{\pm} = ic^{\mp}$ and the $c^{\mp}$ are nonnegative real. In particular, since $c^{\mp} = -iz^{\pm}$, we have that

$$\frac{\sin(z^+)}{z^+}\frac{\sin(z^-)}{z^-} = \frac{\sinh(-iz^+)}{-iz^+}\frac{\sinh(-iz^-)}{-iz^-} = \frac{\sinh c^+}{c^+}\frac{\sinh c^-}{c^-}.$$

Part 3: Extracting the Moments

Denote

$$F_i(\beta_1, \beta_2, z) = \frac{\partial}{\partial \beta_i} F(\beta_1, \beta_2, z), i = 1, 2;$$

$$F_3(\beta_1, \beta_2, z) = F'(\beta_1, \beta_2, z) = \frac{\partial}{\partial z} F(\beta_1, \beta_2, z).$$

The goal of this section is to prove the following:

Theorem

The moments of θ, where θ is defined in Equation (2), satisfy

$$\sum_{n=1}^{\infty} \frac{z^{2n}}{2n} \mathbb{E}\left[\theta^{2n}\right] \frac{(n!)^2 2^{2n}}{(2n)!} = \int_0^{\infty} \frac{d\beta_1}{\beta_1} \int_0^{\infty} \frac{d\beta_2}{\beta_2} z F'(\beta_1, \beta_2, z).$$