Supplemental Information

June 1, 2016

Contents

I Datasets 2

II Constant Volatility Transformations for the Wright-Fisher Process (WFP) 3

III Modification of KS-test to account for dependence among CVTs 5

IV Simulation of Surrogate Data 7
1 WFP simulation 7
2 Parameter Estimation 8
3 Simulating Surrogate Data for Male Tongue 9

V PERMNOS 10
Part I
Datasets

Microbiome datasets were obtained from

http://www.genomebiology.com/2011/12/5/R50/additional

The collection of these data is described in Caporaso et al. 2011 [1]. All sequence count data were converted to relative abundances. The data were not collected at uniform increments, so the actual date stamps were used to calculate $\Delta t$ in the data. $\Delta t$ ranged from 1 to 18 days, with most $\Delta t$ being 1 day.

Breeding Bird Survey data was downloaded from

https://www.pwrc.usgs.gov/bbs/rawdata/[2] and all points along all routes in the United States were pooled, while points from Canada were ignored.

Market share and Market capitalization data were obtained by collecting share prices and share volumes from the US stock database ©2014 Center for Research in Security Prices (CRSP), The University of Chicago Booth School of Business, accessed through the Wharton Research Data Services. Out of over 500 companies in the S&P 500 from January 1, 2000 to January 1, 2005, only 459 companies were in the index continuously during that interval. One company, PERMNO 76656, recorded a negative day-end share price that appears to be a typo and was corrected - a string of day-end prices in the text file from CRSP was: 26.3200 26.6400 24.9900 -24.9300 27.7000 27.6600 27.9100. The -24.9300 was changed to 24.9300. Another company, PERMNO 10324, had a “NaN” entry in the string of share prices: 28.9100 31.5800 NaN 32.6100 35.1100; this company was removed from analysis. 6 remaining PERMNOs, 88313, 88319, 88587, 89463, 89813, and 90435, had zero entries - these PERMNOs were removed from the data. The PERMNOs for the remaining 451 S&P500 companies used in our analysis are listed at the bottom of this supplemental information file.

Share prices were converted to relative abundances to obtain market shares, and each company’s capitalization (the product of share price and share volume) was converted to relative abundances to obtain the market capitalization. The dates of the day-end prices were noted and time-increments were defined as the number of days between the dates of the day-end prices (e.g. $\Delta t = 1$ for most weekdays, $\Delta t = 3$ for most weekends, etc.).

For a comparison of relative-abundance-distribution based methods of assessing neutrality, a time point was taken from each of the datasets, the counts were rarefied to 5000 for comparison across datasets (for the market capitalization, the “unit” was taken as the market share of the smallest company), and species-abundance distributions were fit using the R package ’sads‘. Many of the datasets appear neutral from this perspective, and all of the datasets whose
RADs are well-fit by neutral theory’s expectations are in fact rejected using our test.

Figure 1: Empirical rank-abundance distributions (circles) plotted against the best-fit from neutral theory (blue line). Tests based on these static features may provide useful comparisons between models, but time-series data are ultimately needed to distinguish between models with similar or identical stationary properties.

Part II
Constant Volatility
Transformations for the Wright-Fisher Process (WFP)

The WFP is defined by the Itô stochastic differential equation defining fluctuations in relative abundances in an $n$-species community, $X_t = (X_1^t, \ldots, X_n^t)$,

$$dX_t = \lambda (\rho - X_t) dt + \sigma (X_t) dW_t$$  \hspace{1cm} (1)

where $\lambda > 0$ and $\rho, X_t \in \Delta^n$ the $n$-dimensional simplex, i.e. $X_i^t > 0$ for all $i$ and $t$, and $X_1^1 + \ldots + X_n^t = 1$ for all $t$. The quadratic covariation between the relative abundances of different species is given by the elements of $\Sigma (X_t) = \frac{1}{2} \sigma \sigma^T$, where

$$\Sigma_{i,j} = \begin{cases} X_i^t (1 - X_i^t) & i = j \\ -X_i^t X_j^t & i \neq j \end{cases}$$  \hspace{1cm} (2)

For systems with state-dependent noise, the volatility can’t be measured directly as doing so would require multiple samples from each point, allowing calculation of the covariance matrices at each point. An alternative, suitable for
testing if the volatility comes from our particular model, is to find a function, \( f \), which has constant volatility when its input is a WFP and non-constant volatility otherwise. Thus, we seek a function \( f : \Delta^n \mapsto \mathbb{R} \) such that

\[
\lim_{\Delta t \to 0} \text{var} \left[ \frac{f(X_t+\Delta t) - f(X_t)}{\sqrt{\Delta t}} \right] = C
\]

(3)

for some constant, \( C \). Applying Itô’s lemma to equation 1, we have that

\[
\lim_{\Delta t \to 0} \text{var} \left[ \frac{\Delta f}{\sqrt{\Delta t}} \right] = \nabla f^T \sigma \sigma^T \nabla f.
\]

(4)

If we let our constant be \( C = 2 \) a constant volatility function will be a function \( f \) such that

\[
\nabla f^T \Sigma \nabla f = 1
\]

(5)

noting the trivial case with \( C = 0 \) can always be obtained by any function \( f \left( X_1^t + \ldots + X_n^t \right) = f(1) \). Simplifying notation by denoting \( X_i^t = x_i \), equation (5) can be expanded using equation (2):

\[
\sum_{i=1}^{n-1} \sum_{j>i}^n x_i x_j \left( \frac{\partial f}{\partial x_i} - \frac{\partial f}{\partial x_j} \right)^2 = 1.
\]

(6)

If \( n = 3 \), and we denote \( x_1 = x, x_2 = y \) and \( x_3 = z \), we obtain the partial differential equation

\[
x y \left( \frac{\partial f}{\partial x} - \frac{\partial f}{\partial y} \right)^2 + x z \left( \frac{\partial f}{\partial x} - \frac{\partial f}{\partial z} \right)^2 + y z \left( \frac{\partial f}{\partial y} - \frac{\partial f}{\partial z} \right)^2 = 1
\]

(7)

defining constant-volatility transformations for a 3-species WFP.

When \( n = 2 \), the PDE in equation 6 is simply

\[
x y \left( \frac{\partial f}{\partial x} - \frac{\partial f}{\partial y} \right)^2 = 1
\]

(8)

which can be solved by substituting \( u = x + y \) and \( v = x - y \), simplifying to

\[
(u^2 - v^2) \left( \frac{\partial f}{\partial v} \right)^2 = 1.
\]

(9)

Since \( X_i \in \Delta^2 \) and \( u = 1 \) the solution is simply

\[
f = \arcsin(v).
\]

(10)

For \( n > 2 \), a set of solutions can be intuited by using the grouping invariance of the WFP, i.e. that higher dimensional Wright-Fisher Processes can be collapsed into lower-dimensional Wright-Fisher processes by simply grouping species. For example, where \( n = 3 \), we can set \( w = x + y \) and the resulting process \( Z_t = (w, z) \) is a 2-species WFP. Thus, \( f = \arcsin(x + y - z) \) is a CVT
for \( n = 3 \). Since any re-grouping of species in a WFP yields a WFP, and since we always know the CVTs for the 2-species case, we can define a set of CVTs for an \( n \)-dimensional WFP:

\[
fa(X_t) = \arcsin \left( \sum_{i=1}^{n} a_i X_i^t \right)
\]

where \( a_i = \pm 1 \). All \( 2^n \) of the these CVTs solve equation 6 and have constant-volatility when operating on a WFP.

Two details about this test should be noted. First, one might consider constructing a test on the magnitude - and not just the constancy - of the volatility. Doing so will provide a stronger test of the particular WFP written in equation 1, but will have a higher false-positive rate as a test of the neutrality of competition. This is clear by considering the process

\[
dY_t = \lambda (\rho - Y_t) dt + \alpha \sigma (Y_t) dW_t
\]

formed by rescaling the volatility of a standard WFP by a scalar, \( \alpha \). \( Y_t \) has quadratic covariation \( d\left[Y_t^i, Y_t^j\right] = \alpha^2 Y_t^i Y_t^j \) that is still bilinear in \( Y_t^i \) and \( Y_t^j \) and thus the species can still be grouped to form lower-dimensional WFPs - this process is still neutral in the sense that species identities do not affect their dynamics, but its volatility is a different magnitude than the WFP of equation (1). Tests for the constancy, and not the magnitude, of the volatility of \( f_a \) will provide more robust tests of the neutral symmetry - grouping invariance.

The second detail is that the \( 2^n \) constant volatility tests are redundant and some are uninformative. For instance, the sample volatilities of \( f_a \) and \( f_{-a} \) are equal, and the volatilities for \( a = (1, 1, ..., 1) \) are zero since \( f_{\pm 1} = \arcsin (\pm 1) \). Thus, there are only \( 2^{n-1} - 1 \) informative choices of CVT.

Part III
Modification of KS-test to account for dependence among CVTs

The P-values arising from CVT tests are dependent as two CVTs may overlap greatly in their grouping of species in the community. Consequently, a standard KS-test on the distribution of P-values will have a high false-positive rate (FIGURE S1).

Correcting the high false positive rate of a standard KS-test requires accounting for the dependence among the CVTs. Chicherportiche et al. [3] developed a modified goodness of fit test for dependent data that requires calculation of copulas. However, rather than calculating copulas explicitly and using the methods developed by Chicherportiche et al., we resort to simulation of the null distribution of the KS statistic for CVTs operating on a WFP.
A standard Kolmogorov-Smirnov test on the uniformity of P-values from multiple constant-volatility tests will yield a high false-positive rate due to the dependence structure of the CVTs. (left sub-panel) KS tests from the 840,000 WFP simulations described here yield a high false-positive rate. (right sub-panel) To correct for this, we analyze distribution of KS statistics for the 840,000 simulations to parametrize the scaling of critical values of the KS statistic corresponding to \( P = 0.05 \) and \( P = 0.005 \) as functions of the number of species, \( n \), the number of time points, \( T \), and the number of randomly drawn CVTs, \( n_a \).

For all simulations, we used \( T_{\text{sim}} = 1 \) and \( \lambda = 50 \). The number of species, \( n \), took 7 equally spaced values from \( n = 12 \) to \( n = 24 \). The number of time-samples, \( T \), within the simulation window, \( T_{\text{sim}} \), took 10 equally spaced values from \( T = 125 \) to \( T = 1250 \). The number of CVTs, \( n_a \), took values 6 in \( n_a \in \{50, 100, 200, 400, 800, 1600\} \).

Stable, numerical integration of the WFP is not trivial, especially for large numbers of species, \( n \). The full details of the numerical integration are discussed below under “Simulation of Surrogate Data” and covered in greater detail in the thesis of Washburne (2015) [4]. For each of the 420 combinations of \( \{n, T, n_a\} \), 2,000 WFP trajectories were simulated. Using the methods from Washburne (2015), simulating the 840,000 trajectories described above resulted in only 3 simulations leaving the bounded domain, \( \Delta^n \).

We wished to understand how the \( P = 0.05 \) and \( P = 0.005 \) critical values of the KS statistic scaled with the number of species, the number of CVTs, and the number of time points. Plots of the \( P = 0.05 \) critical values, \( K_{0.05} \), as a function of \( n \), \( T \), and \( n_a \) are shown in figure S2. For the range of \( n \), \( T \), and \( n_a \) considered, \( K_{0.05} \) is roughly constant in \( T \), decays as a power law for \( n_a \), and decays slower than a power law in \( n \) (figure S2). Our values of \( n \) range from \( n = 335 \) to \( n = 1295 \), well beyond the values of \( n \) simulated. For a conservative estimate of \( K_{0.05} \) for values of \( n \) in our analysis, we assume a power law decay in \( n_a \) and fit the critical values as a power law in \( n_a \), \( K_\alpha = c n_a^b \). We obtain \((c, b) = (0.9286, -0.4128)\) for \( \alpha = 0.95 \) and \((c, b) = (1.0942, -0.3968)\) for \( \alpha = 0.995 \).
Figure 3: (top-left) For large numbers of species, the $P = 0.95$ critical value of the KS statistic, $K_{95}$, in neutral simulations follows a power law in the number of CVTs, $n$. (top-right) $K_{95}$ does not change significantly with $T$. (bottom-left) For large $n$, $K_{95}$ will decrease with the number of species, $n$. (bottom-right) Zooming in on the plot of $n$ versus $K_{95}$ for high $n$. The decay in $K_{95}$ appears to be leveling off faster than a power-law in $n$.

A full perturbation analysis including varying $\lambda$ and $T_{sim}$, or an explicit calculation of the copulas, will improve the accuracy of this method for testing the neutrality of competitive systems.

Part IV
Simulation of Surrogate Data

Simulating the WFP is not trivial due to the bounded state space, $\Delta^n$. Some tools for the numerical simulation of the WFP were developed in the thesis of Washburne (2015) [4]. The relevant tools are discussed here.

1 WFP simulation

We used Washburne’s (2015) methods for simulating the WFP, namely simulating a log-ratio transformation,

$$g^i(X_i) = \log \left( \frac{X_i}{1 - X_i} \right)$$  (13)
for $i = 1, \ldots, n$. Trajectories for the log-ratio transformation of the WFP were computed by integrating the SDE,
\[ dg_i = \left( e^{g_i} + e^{-g_i} \right)^2 \left[ \frac{\lambda p_i - 1/2}{1 + e^{-g_i}} - \frac{1}{2} \left( \frac{n - 2}{n} \right) \right] dt + \nabla g_i^T \sigma(g_i) dW_t \]
obtained by Ito’s lemma and the addition of a dimensionality correction to the drift, $-\frac{1}{2} \left( \frac{n-2}{n} \right) \left( e^{g_i/2} + e^{-g_i/2} \right)^2 dt$, shown to keep $X_t \in \Delta^n$ without affecting the bilinear quadratic covariation of $X_t$ [4].

We also use Washburne’s method for factoring the covariance matrix, $\Sigma$, into a zero-sum matrix and a competition matrix,
\[ \Sigma = PV P^T \]
where $P$ is the matrix,
\[
P = \begin{pmatrix}
1^T_{n-1} & 0^T_{n-2} & \cdots & 0 \\
1^T_{n-2} & \ddots & \ddots & \ddots \\
-1_{n-1} & \ddots & 0 & \cdots \\
-1_{n-2} & \ddots & 1 & \cdots & -1
\end{pmatrix}.
\]
Here, $1_k$ is the $k \times 1$ vector containing all ones, $0_k$ is the $k \times 1$ vector containing all zeroes, and $I_k$ is the $k \times k$ identity matrix. $V$ is a diagonal matrix whose $i$th diagonal is the $i$th element of the vectorized, upper-triangular region of $\Sigma$. For example, if $n = 4$ and $X_1^1 = x$, $X_1^2 = y$, $X_1^3 = z$ and $X_1^4 = w$, we have the factorization
\[
\Sigma = \begin{pmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & -1 & 0 & -1 & 0 & 1 & 0 \\
0 & 0 & -1 & 0 & -1 & -1 & 1
\end{pmatrix} \begin{pmatrix}
1 & 1 & 0 & 0 \\
0 & 1 & -1 & 0 \\
1 & 0 & 0 & -1 \\
0 & 1 & -1 & 0
\end{pmatrix}.
\]

All WFP trajectories simulated here were simulated by Euler-Maruyama integration of the dimension-corrected Ito SDE of the log-ratio transform, $g$, using the covariance matrix factorization described above. The trajectories of $g_t$ were then mapped through the inverse of $g_t$ to obtain the WFP $X_t$.

## 2 Parameter Estimation

Surrogate datasets are constructed by numerically integrating equation (1) over a time interval, $[0, T_{sim}]$. To match at WFP to the data, we need to estimate parameters in the model, $\lambda$, $\rho$, and $T$. First, we estimate $\lambda$ and $\rho$. Second, since
The parameters $\lambda$ and $\rho$ can be estimated by their relation to the stationary mean & variance of the WFP. Consider a two-species WFP, $n = 2$, and for simplicity let’s use the notation $X_t^1 = X_t$ and $X_t^2 = Y_t$ and $\rho_1 = \rho$. Denoting the first and second moments of the WFP $m_t = E[X_t]$ and $s_t = E[(X_t)^2]$, we can use Itô’s lemma and Itô’s isometry to yield the system of ODEs

\begin{align}
\frac{dm_t}{dt} &= \lambda (\rho - m_t) \\
\frac{ds_t}{dt} &= 2 (1 + \lambda \rho) m_t - 2 (1 + \lambda) s_t
\end{align}

which yields the stationary mean, $m_\infty = \lim_{t \to \infty} m_t$ and second moment $s_\infty = \lim_{t \to \infty} s_t$

\begin{align}
m_\infty &= \rho \\
s_\infty &= \frac{1 + \lambda \rho}{1 + \lambda} m_\infty.
\end{align}

Thus, the stationary mean is $\mu = \rho$ and the stationary variance is given by $v^2 = s_\infty - m_\infty^2$, yielding $v^2 = \rho \frac{(1 - \rho)}{1 + \lambda}$. These two equations give one way of estimating $\lambda$ and $\rho$. Given a two-species system, we estimate the sample mean, $\hat{\mu}$, and sample variance, $\hat{v}^2$, and from those obtain our estimates of $\lambda$ and $\rho$:

\begin{align}
\hat{\rho} &= \frac{\hat{\mu} (1 - \hat{\mu})}{\hat{v}^2} - 1 \\
\hat{\lambda} &= \frac{\hat{\mu} (1 - \hat{\mu})}{\hat{v}^2} - 1
\end{align}

The last parameter that remains to be estimated is the length of time, $T_{\text{sim}}$, to simulate our surrogate datasets. An estimate, $\hat{T}_{\text{sim}}$, can be obtained by noting that the true autocorrelation time of the WFP is $\tau_c = \frac{1}{\lambda}$, i.e. the autocorrelation $\text{Cor}(X_t, X_{t+\tau_c}) = e^{-1}$. For a dataset with $M$ time points, the autocorrelation function can be interpolated to find the time lag, $\tau$ with $0 \leq \tau \leq M$, where $C(X_t, X_{t+\tau}) = e^{-1}$. We want a surrogate dataset to be simulated for the same number of autocorrelation times as our empirical data, i.e. $\frac{T_{\text{sim}}}{\tau_c} = \frac{M}{\tau}$, thus the time-length for the surrogate dataset will be

\[\hat{T}_{\text{sim}} = \frac{M}{\lambda \tau_c}.
\]

3 Simulating Surrogate Data for Male Tongue

To produce the overlayed scatter plots of $\nu_t$ versus $f_t$ for real & surrogate data, the male tongue was grouped into two groups: the first 200 species (as ordered in the link http://www.genomebiology.com/2011/12/5/R50/additional) formed group 1 and the remaining 173 species formed group 2. The relative abundances of group 1 was $\hat{\rho} = 0.5652$, and the rate of mean reversion estimated at
\[ \lambda = 18.0063. \] The estimated timescale of the dataset for simulation of the WFP was \( T_{\text{sim}} = 13.4872. \) 4,000 2-species WFP trajectories were simulated by initializing the community at \( X_0 = \hat{\rho}, \) integrating with \( \Delta t = 10^{-6}, \) and sampling 332 points in \([0, T_{\text{sim}}]\) such that the relative distance between the samples matched those of the 332 time points in the data. Of the 4,000 trajectories simulated, none left the bounded domain.

To analyze the distribution of P-values from constant-volatility tests, we generated 16,000 surrogate trajectories of \( f_t. \) To generate these surrogate data, 16,000 random groupings, \( a, \) were chosen by randomly assigning \( a_i = \pm 1, \) \( \hat{\rho} \) was estimated as average relative abundance of one group, \( \hat{\lambda} \) was kept at \( \hat{\lambda} = 18.0063 \) from the previous estimation, and one WFP trajectory was simulated and subsampled as above. Generating 16,000 trajectories required discarding 11 trajectories which left the bounded domain. The mean distance from the boundary, \( \delta = 0.5 - |\hat{\rho} - 0.5|, \) was significantly lower in the 11 discarded trajectories compared to trajectories included in our analysis; trajectories included in our analysis had \( \delta = 0.3598, \) compared to \( \delta = 0.0367 \) for the 11 trajectories discarded. Consequently, these surrogate data had a very slight bias against extremely lop-sided groupings.

Part V
PERMNOS

10078 10104 10107 10108 10137 10138 10145 10147 10225 10401 10516 10562
10874 10890 10942 11081 11308 13688 13856 13928 14008 14277 14322 14323
11850 11896 11955 11970 11976 12052 12060 12079 12140 12431 12490 12503
12570 13100 13119 13688 13856 13928 14008 14277 14322 14323 14541 14593 14656
15720 16424 16432 16600 16678 17005 17144 17478 17750 17806 17830 18016
18092 18163 18401 18411 18542 18729 19300 19502 19561 20053 20220 20482
20862 21178 21866 21207 21371 21573 21776 21792 21936 21979 22032
22103 22111 22293 22309 22517 22692 22779 22840 22947 23036 23077
23114 23229 23309 23318 23473 23579 23712 23819 23915 23931 23990 24101
24046 24109 24205 24221 24272 24360 24459 24563 24643 24766 24942 24985
25013 25081 25320 25419 25487 25769 25778 25785 25953 26112 26382 26403
26518 26710 26825 27430 27633 27713 27756 27828 27887 27959 27983 27991
28222 28484 29946 30382 30681 32379 32791 34032 34746 34817 34833 35044
33051 36397 36468 36499 37584 38156 38682 38703 39087 39490 39538 39642
39917 40061 40125 40272 40416 40539 41080 41355 41443 42024 42083 42200
42534 42796 42906 43123 43350 43449 43553 43772 44061 44644 44951 45536
45495 45671 45751 45794 46378 46674 46886 47175 47466 47896 47941
48267 48274 48485 48506 48633 48725 48960 49015 49154 49373 49429 49656
49630 49905 50032 50227 50876 50906 51043 51263 51369 51377 51596 51706
52038 52337 52476 52695 52919 52978 53065 53479 53613 53831 54148 54181
References

[1] Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, et al. Moving pictures of the human microbiome. Genome Biol. 2011;12(5):R50.

[2] Partridge K, Ziolkowski Jr D, Hudson M. North american breeding bird survey dataset 1966-2013, version 2013.0. US Geological Survey, Patuxent Wildlife Research Center<www.pwrc.usgs.gov/BBS/RawData. 2014;

[3] Chicheportiche R, Bouchaud JP. Goodness-of-fit tests with dependent observations. Journal of Statistical Mechanics: Theory and Experiment. 2011;2011(09):P09003.

[4] Washburne A. Competition and Coexistence in an Unpredictable World; 2015.