Statistical mechanics of time series

Riccardo Marcaccioli[1,*] and Giacomo Livan[1,2,†]

[1] *Department of Computer Science, University College London, 66-72 Gower Street, London WC1E 6EA, UK*
[2] *Systemic Risk Centre, London School of Economics and Political Sciences, Houghton Street, London WC2A 2AE, UK*

Countless natural and social multivariate systems are studied through sets of simultaneous and time-spaced measurements of the observables that drive their dynamics, i.e., through sets of time series [1, 2]. Typically, this is done via hypothesis testing: the statistical properties of the empirical time series are tested against those expected under a suitable null hypothesis. This is a very challenging task in complex interacting systems, where statistical stability is often poor due to lack of stationarity and ergodicity. Here, we describe an unsupervised, data-driven framework to perform hypothesis testing in such situations. This consists of a statistical mechanical theory - derived from first principles - for ensembles of time series designed to preserve, on average, some of the statistical properties observed on an empirical set of time series. We showcase its possible applications on a set of stock market returns from the NYSE.

Hypothesis testing lies at the very core of the scientific method. In its general formulation, it hinges upon contrasting the observed statistical properties of a system with those expected under a null hypothesis. In particular, hypothesis testing allows to discard potential models of a system when empirical measurement that would be exceedingly unlikely under them are made.

However, there is often no theory to guide the investigation of a system's dynamics. What is worse, in many practical situations one may be given a single - and possibly unreproducible - set of experimental data. This is indeed the case when dealing with most complex systems whose collective dynamics often are neither stationary nor ergodic, ranging from climate [3, 4] to brain activity [5] and financial markets [6–8]. This, in turn, makes hypothesis testing in complex systems a very challenging task, that potentially prevents from assessing which properties observed in a given data sample are "untypical", i.e, unlikely to be observed again in a sample collected at a different point in time.

This issue is usually tackled by constructing ensembles of artificial time series sharing some characteristics with those generated by the dynamics of the system under study. This can be done either via modelling or in a purely data-driven way. In the latter case, the technique most frequently used by both researchers and practitioners is bootstrapping [9, 10], which amounts to generating partially randomised versions of the available data via resampling, that can then be used as a null benchmark to perform hypothesis testing. Depending on its specificities, bootstrapping can account for autocorrelations and cross-correlations in time series sampled from multivariate systems. However, it relies on assumptions, such as sample independence and some form of stationarity [11], which limit its power when dealing with complex systems.

As far as model-driven approaches are concerned, the literature is extremely vast [12]. Broadly speaking, modelling approaches rely on a priori structural assumptions for the system's dynamics, and on identifying the parameters values that best explain the available set of observations within a certain class of models (e.g., via Maximum Likelihood [12]). A widely used class for multivariate time series is that of autoregressive models, such as VAR [13], ARMA [14], and GARCH [15], which indeed were originally introduced, among other reasons, to perform hypothesis testing [14]. In such models, the future values of each time series are given by a linear combination of past values of one or more time series, each characterised by their own idiosyncratic noise to capture the fluctuations of individual variables. Such a structure is most often dictated by its simplicity rather than by first principles. As a consequence, once calibrated, autoregressive models produce rather constrained ensembles of time series that do not allow to explore scenarios that differ substantially from those observed empirically.

Here we propose a novel formalism to perform hypothesis testing on sets of time series based on the ensemble theory of statistical mechanics. Starting from first principles, we will introduce a (gran)canonical ensemble of correlated time series subject to constraints based on the properties of an empirically observed set of measurements. This, in turn, will result in a multivariate probability distribution which allows to unbiasedly sample values centred on such measurements. The theory we propose in the following shares some similarities with the canonical ensemble of complex networks [16–18], and, as we will show, inherits its powerful calibration method based on Likelihood maximization [19].

Let $\mathcal{W}$ be the set of all real-valued sets of N time series of length T (i.e., the set of real-valued matrices of size $N \times T$), and let $\overline{W} \in \mathcal{W}$ be the empirical set of data we want to perform hypothesis testing on (i.e., $\overline{W}_{it}$ stores the value of the i-th variable experimentally measured at time t). Our aim is to define a probability density function $P(W)$ on $\mathcal{W}$ such that the expectation values $\langle \mathcal{O}_\ell(W) \rangle$ of a set of observables ($\ell = 1, \ldots, L$) coincide with the value $\overline{O}_\ell$ of the corresponding quantity as empirically measured from $\overline{W}$.

Following Boltzmann and Gibbs, we can achieve

*Electronic address: riccardo.marcaccioli.16@ucl.ac.uk
†Electronic address: g.livan@ucl.ac.uk

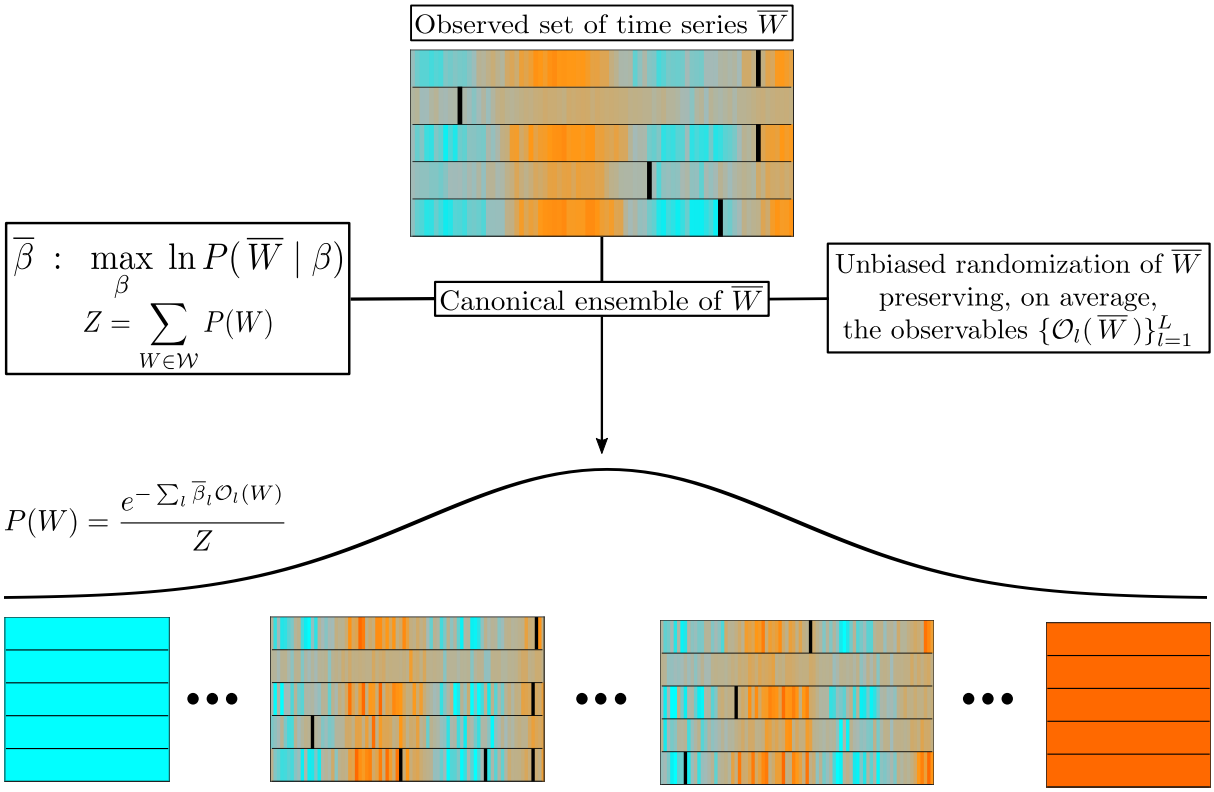

FIG. 1: Schematic representation of the model. Starting from an empirical set of time series $\overline{W}$, we construct its unbiased randomization by finding the probability measure $P(W)$ on the phase space $\mathcal{W}$ which maximises Gibbs' entropy while preserving the constraints $\{\mathcal{O}_l(\overline{W})\}_{l=1}^{L}$ as ensemble averages. The probability distribution $P(W)$ depends on L parameters that can be found by maximising the likelihood of drawing $\overline{W}$ from the ensemble. In the Figure, orange, turquoise and black are used to indicate positive, negative or empty values of the entries W_{it}, respectively, while brighter shades of each color are used to display higher absolute values. As it can be seen, the distribution $P(W)$ assigns higher probabilities to those sets of time series that are more consistent with the constraints and therefore more similar to $\overline{W}$. See [18] for a similar chart in the case of the canonical ensemble of complex networks.

the above by invoking the Maximum Entropy principle, i.e., by identifying $P(W)$ as the distribution that maximises the entropy functional $S(W) = -\sum_{W\in\mathcal{W}} P(W)\ln P(W)$, while satisfying the L constraints $\langle \mathcal{O}_\ell(W)\rangle = \sum_W \mathcal{O}_\ell(W)P(W) = \overline{O}_\ell$ and the normalisation condition $\sum_W P(W) = 1$. This reads (see Appendix A)

$$P(W) = \frac{e^{-H(W)}}{Z} \ , \tag{1}$$

where $H(W) = \sum_\ell \beta_\ell \ \mathcal{O}_\ell(W)$ is the Hamiltonian of the system, β_ℓ ($\ell = 1, \ldots, L$) are Lagrange multipliers introduced to enforce the constraints, and $Z = \sum_W e^{-H(W)}$ is the partition function of the ensemble, which verifies $\langle \mathcal{O}_\ell(W)\rangle = \partial \ln Z/\partial\beta_\ell \ , \forall\, \ell$.

Figure 1 provides a sketch representation of the ensemble theory just introduced. The rationale of enforcing the aforementioned constraints is that of shaping $P(W)$ in a way that assigns low probability to regions of the phase space $\mathcal{W}$ where the observables associated to the Lagrange multipliers β_ℓ take values that are exceedingly different to those measured in the empirical set $\overline{W}$, and high probability to regions where some degree of similarity with $\overline{W}$ is retained. This, in turn, allows to test whether other properties (not encoded in any of the constraints) of $\overline{W}$ are statistically significant by measuring how often they appear in instances drawn from the ensemble. The existence and uniqueness of the Lagrange multipliers ensuring the ensemble's ability to preserve the constraints $\overline{O}_\ell$ is a well known result, and they are equivalent to those that would be obtained from the maximization of the Likelihood of drawing the empirical matrix $\overline{W}$ from the ensemble [20].

Let us now put the above framework into practice by specifying an ensemble in detail. Let us consider an $N \times T$ empirical data matrix $\overline{W}$ whose rows have been rescaled to have zero mean, so that $\overline{W}_{it} > 0$ ($\overline{W}_{it} < 0$) will indicate that the time t value of the i-th variable is higher (lower) than its empirical mean. Also, without loss of generality, let us assume that $\overline{W}_{it} \in \mathbb{R}_{\neq 0}$, and that $\overline{W}_{it} = 0$ indicates missing data. For later convenience, let us define $A^{\pm} = \Theta(\pm W)$ and $w^{\pm} = \pm W\Theta(\pm W)$ (we shall denote the corresponding quantities measured on the empirical set as $\overline{A}^{\pm}$ and $\overline{w}^{\pm}$), and let us constrain

the ensemble to preserve the values of the following observables:

- The number of positive (above-average) and negative (below-average) values $\overline{N}_i^{\pm} = \sum_t \overline{A}_{it}^{\pm}$, and the number of missing values $\overline{N}_i^0 = T - \overline{N}_i^+ - \overline{N}_i^-$ recorded for each time series ($i = 1, \ldots, N$).
- The cumulative positive and negative values recorded for each time series: $\overline{S}_i^{\pm} = \sum_t \overline{w}_{it}^{\pm}$ ($i = 1, \ldots, N$).
- The number of positive, negative, and missing values recorded at each sampling time: $\overline{M}_t^{\pm} = \sum_i \overline{A}_{it}^{\pm}$, $\overline{M}_t^0 = N - \overline{M}_t^+ - \overline{M}_t^-$ ($t = 1, \ldots, T$).
- The cumulative positive and negative value recorded at each sampling time: $\overline{R}_t^{\pm} = \sum_i \overline{w}_{it}^{\pm}$ ($t = 1, \ldots, T$).

The above list amounts to $8(N+T)$ constraints, and the Hamiltonian H depends on the very same number of parameters:

$$H(W) = \sum_{i=1}^{N}\sum_{t=1}^{T}\Big[\left(\alpha_i^N + \alpha_t^T\right) A_{it}^+ + \left(\beta_i^N + \beta_t^T\right) A_{it}^- + \left(\gamma_i^N + \gamma_t^T\right) w_{it}^+ + \left(\sigma_i^N + \sigma_t^T\right) w_{it}^-\Big] , \tag{2}$$

where we have introduced the Lagrange multipliers associated to all constraints. As shown in Appendix A, the above expression leads to the following partition function:

$$\begin{aligned} Z &= \sum_{W\in\mathcal{W}} e^{-H(W)} = \prod_{i=1}^{N}\prod_{t=1}^{T} Z_{it} \\ &= \prod_{i=1}^{N}\prod_{t=1}^{T}\left[1 + \frac{e^{-(\alpha_i^N+\alpha_t^T)}}{\gamma_i^N+\gamma_t^T} + \frac{e^{-(\beta_i^N+\beta_t^T)}}{\sigma_i^N+\sigma_t^T}\right] \\ &= \prod_{i=1}^{N}\prod_{t=1}^{T}\left(1 + e^{\frac{\mu_{it}^1-\epsilon_{it}}{T_{it}}} + e^{\frac{\mu_{it}^2-\epsilon_{it}}{T_{it}}}\right) , \end{aligned} \tag{3}$$

where the quantities $\mu_{it}^{1,2}$, ϵ_{it}, and T_{it} are functions of the Lagrange multipliers (see Appendix A).

Some considerations about Eq. (3) are now in order. First of all, the partition function factorises into the product of independent factors Z_{it}, and therefore into a collection of $N \times T$ *statistically independent* sub-systems. However, it is crucial to notice that their parameters (i.e., the Lagrange multipliers) are coupled through the system of equations specifying the constraints ($\langle \mathcal{O}_\ell(W)\rangle = \partial \ln Z/\partial\beta_\ell \ , \forall \ell$). As we shall demonstrate later, this ensures that part of the original system's correlation structure is retained within the ensemble. Moreover, with the above positions, the system described by Eq. (3) can be interpreted as a system of $N \times T$ orbitals with energies ϵ_{it} and local temperatures T_{it} that can be populated by fermions belonging to two different species (corresponding to positive and negative values in the time series) characterised by local chemical potentials μ_{it}^1 and μ_{it}^2, respectively.

Returns		**Significance null hypothesis**			**median**
Stat	**Sample**	**0.01-0.99**	**0.05-0.95**	**0.1-0.9**	**rel. err.**
Var	stock	0.95	0.76	0.59	0.2
	day	0.88	0.78	0.69	0.14
Skew	stock	1	0.98	0.95	0.13
	day	0.78	0.58	0.49	0.46
Kurt	stock	0.78	0.61	0.51	0.60
	day	0.85	0.68	0.55	0.1

TABLE I: Fraction of empirical moments compatible with their corresponding ensemble distribution at different significance levels specified in terms of quantiles (e.g., 0.01-0.99 denotes that the 1st and 99th percentiles of the ensemble distribution are used as bounds to determine whether the null hypothesis of an empirical moment being compatible with the ensemble distribution can be rejected or not). Moments are calculated both for each stock and each trading day. In the last column, we also report, for each moment, the median relative error between the empirical value and its ensemble average.

Ratio empirical aggregated pdfs not rejected by a K-S test	**Aggregation level**	**K-S test significance**	
		0.01	**0.05**
	stocks	0.92	0.68
	days	0.82	0.75

TABLE II: Fraction of empirical return distributions (both for stocks and trading days) that are compatible with their ensemble counterparts based on Kolmogorov-Smirnov tests at different significance levels.

From the partition function in Eq. (3) we can finally calculate the probability distribution $P(W)$:

$$P(W) = \prod_{i=1}^{N}\prod_{t=1}^{T}\left[P_{it}^+\right]^{A_{it}^+}\left[P_{it}^-\right]^{A_{it}^-}\left[1 - P_{it}^+ - P_{it}^-\right]^{1-A_{it}^+-A_{it}^-} \left[Q_{it}^+(w_{it}^+)\right]^{A_{it}^+}\left[Q_{it}^-(w_{it}^-)\right]^{A_{it}^-} , \tag{4}$$

where $P_{it}^{\pm}$ and $Q_{it}^{\pm}(w_{it}^{\pm})$ are functions of the Lagrange multipliers (see Appendix A) and correspond, respectively, to the probability of drawing a positive (negative) value for the i-th variable at time t and to its probability distribution.

As an example application of the ensemble defined above, let us consider the daily returns of the $N = 100$ most capitalized NYSE stocks over $T = 560$ trading days (spanning October 2016 - November 2018). In this example, the aforementioned constraints force the ensemble to preserve, on average, the number of positive and negative returns and the overall positive and negative return for each time series and for each trading day, leading to $6(N \times T)$ constraints. When these constraints are enforced, an explicit expression for the marginal distribu-

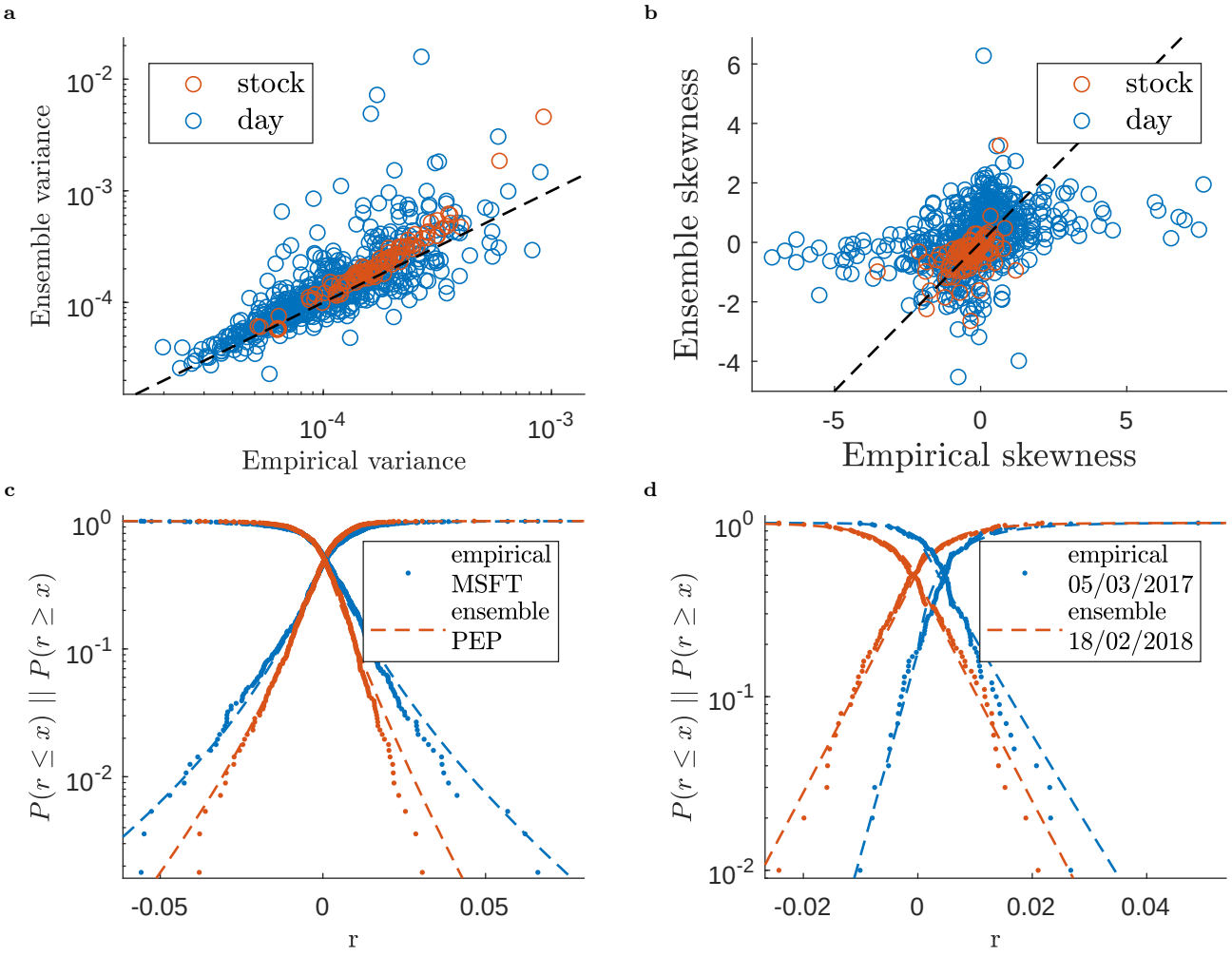

FIG. 2: Comparisons between empirical statistical properties and ensemble averages. In these plots we demonstrate the model's ability to partially reproduce non-trivial statistical properties of the original set of time series that are not explicitly encoded as ensemble constraints. **a**) Empirical vs ensemble average values of the variances of the returns calculated for each stock (red dots) and each day (blue dots). **b**) Same plot for the skewness of the returns. **c**) Comparison between the ensemble and empirical cumulative distributions (and associated survival functions) for the returns of two randomly selected stocks (Microsoft and Pepsi Company). Remarkably, a Kolmogorov-Smirnov test (0.01 significance) shows that 92% of the stocks returns empirical distributions are compatible with their ensemble counterparts. **d**) Same plot for the returns of all stocks on two randomly chosen days. In this case, 82% of daily returns empirical distributions are compatible with their ensemble counterparts (K-S test at 0.01 significance).

tions can be obtained (see Eq. (A8) in Appendix A):

$$P(W_{it} = x) = (1 - P_{it}^{+})\, \lambda_{it}^{-}\, e^{\lambda_{it}^{-} x}\, \Theta(-x) + \\ + P_{it}^{+}\, \lambda_{it}^{+}\, e^{-\lambda_{it}^{+} x}\, \Theta(x)\ , \tag{5}$$

where $\lambda_{it}^{\pm}$ are also functions of the Lagrange multipliers (see Appendix A). The above distribution allows both to efficiently sample the ensemble numerically and to obtain analytical results for several observables.

Figure 2 and Tables I and II illustrate how the above first-moment constraints translate into explanatory power of higher-order statistical properties. Indeed, in the large majority of cases, the empirical return distributions of individual stocks and trading days and their higher-order moments (variance, skewness, and kurtosis) are statistically compatible with the corresponding ensemble distributions. Notably, this is the case without constraints explicitly aimed at enforcing such level of agreement. This, in turn, further confirms that the ensemble can indeed be exploited to perform reliable hypothesis testing by sampling random scenarios that are however closely based on the empirically available data.

In that spirit, in Figure 3 we show an example of ex-post anomaly detection, where the original time series of a stock is plotted against the 95% confidence intervals obtained from the ensemble for *each* data point W_{it}. As it can be seen, the results are non-trivial, since the returns flagged as anomalous are not necessarily the largest

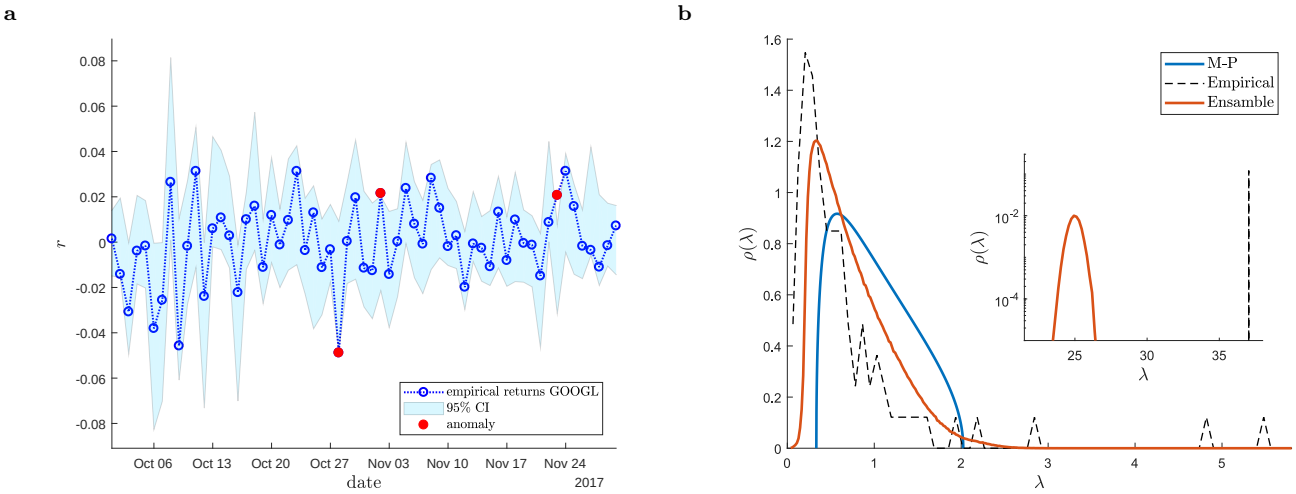

FIG. 3: Applications of the ensemble theory we propose to a system of stocks. a) Anomaly detection performed on each single trading day of a randomly selected stock (Google). A return measured on a specific day for a specific stock is marked as anomalous if it exceeds the associated 95% confidence interval on that specific return. **b)** Comparison between the empirical spectrum of the estimated correlation matrix (black dashed line), its ensemble counterpart (orange line) and the one prescribed by the Marchenko-Pastur law (blue line). The inset shows the empirical larges eigenvalue (dahsed line) against the ensemble distribution for it.

ones in absolute value. This is because the constraints imposed on the ensemble reflect the *collective* nature of financial market movements, thus resulting in the statistical validation of events that are anomalous with respect to the overall heterogeneity present in the market.

Following the above line of reasoning, in Figure 3 we show a comparison between the eigenvalue spectrum of the empirical correlation matrix of the data, and the average eigenvalue spectrum of the ensemble. As it is well known, the correlation matrix spectrum of most complex interacting systems normally features a large bulk of small eigenvalues which is often approximated by the Marchenko-Pastur (MP) distribution [21] of Random Matrix Theory (i.e., the average eigenvalue spectrum of the correlation matrix of a large system of uncorrelated variables with finite second moments) [22–24], plus a few large and isolated eigenvalues that contain information about the relevant correlation structure of the system (e.g., they can be associated to clusters of strongly correlated variables [25]). As it can be seen in the Figure, the ensemble's average eigenvalue spectrum captures the spectral bulk much better than the MP distribution, and the ensemble distribution for the largest eigenvalue is very close to the one empirically observed, demonstrating that the main source of correlation in the market is well captured by the ensemble. Conversely, the average distance between the empirically observed largest eigenvalue and its ensemble distribution can be interpreted as the portion of the market's collective movement which cannot be explained by the constraints imposed on the ensemble.

The above are just a few possible applications of one variant of the framework we have introduced. The framework itself is rather flexible, can be easily modified by adding or removing constraints from the Hamiltonian, and can be applied to analyze data from a large variety of systems. Indeed, in Appendix B we provide details of another example application based on weekly temperature time series recorded in North-American cities, where we show that our ensemble approach captures time periodicities very well.

Appendix A: Methods

We want to find a probability density function $P(W)$ on $\mathcal{W}$ such that the expectation values of a set of observables coincide with their empirical value, i.e $\langle \mathcal{O}_\ell(W) \rangle = \overline{O}_\ell$ $(\ell = 1, \ldots, L)$, where $\overline{W} \in \mathcal{W}$ is the empirical set of measurements. At first, this problem may appear almost impossible to solve, given that $P(W)$ may be determined by a number of degrees of freedom way larger than the number of constraints we are imposing. However, as introduced in the main text, this can be done by using the Maximum Entropy principle, or, in other words, by adding another (functional) constraint on the probability distribution, which requires that $P(W)$ should also maximise the Gibbs entropy:

$$S(W) = -\sum_{W \in \mathcal{W}} P(W) \ln P(W) \ , \qquad \text{(A1)}$$

while preserving the constraints:

$$\langle \mathcal{O}_\ell(W) \rangle = \sum_{W \in \mathcal{W}} \mathcal{O}_\ell(W) \, P(W) \ = \ \mathcal{O}_\ell(\overline{W}) = \overline{O}_\ell \ , \quad \text{(A2)}$$

and the normalization:

$$\sum_{W\in\mathcal{W}} P(W) \;=\; 1 \;. \tag{A3}$$

Eq. A1-A3 define a constrained optimization problem, whose solution is found by solving the following equation:

$$\frac{\partial}{\partial P}\left[S+\alpha\left(1-\sum_{W\in\mathcal{W}}P(W)\right)+ \right. \\ \left. +\sum_{\ell=1}^{L}\beta_\ell\left(O_\ell-\sum_{W\in\mathcal{W}}\mathcal{O}_\ell(W)\,P(W)\right)\right]=0\;, \tag{A4}$$

where, as usual in such scenarios, each constraint has been coupled with a Lagrange multiplier α, β_ℓ ($\ell = 1,\dots,L$). Defining $H(W)=\sum_\ell \beta_\ell \mathcal{O}_\ell(W)$ as the Hamiltonian of the ensemble and $Z=e^{\alpha+1}=\sum_W e^{-H(W)}$ its partition function, the solution of Eq. (A4) reads:

$$P(W)=\frac{e^{-H(W)}}{Z}\;. \tag{A5}$$

This is the general probability density function ruling the ensemble theory we are proposing. Of course, the sum $\sum_W$ on the phase space of the system used in the above equations still needs to be properly specified.

We are going to do so while considering the Hamiltonian in Eq. (2). As pointed out above, in order to find the partition function Z of the system, we just need to sum $e^{-H(W)}$ over all possible configurations, i.e., over the set of all the $N\times T$ real valued matrices $\mathcal{W}$. Recalling the notations introduced in the main text $A^\pm=\Theta(\pm W)$ and $w^\pm=\pm W\Theta(\pm W)$, we can write the sum over the phase space as follows:

$$\sum_{W\in\mathcal{W}} \equiv \prod_{i=1}^{N}\prod_{t=1}^{T}\sum_{\substack{(A^+_{it},A^-_{it})=(1,0)\\(0,0)}}^{(0,1)}\int_0^{+\infty}dw^+_{it}\int_0^{+\infty}dw^-_{it}\;. \tag{A6}$$

Looking at Eq. (A6), we first see that negative and positive events [26], i.e., values above and below the empirical mean of each variable, cannot coexist in an entry (an orbital in our analogy) W_{it}, which, once occupied, cannot hold any other event, even of the same species. We can now fully understand the origin of the two species of fermions appearing in the partition function (3). Moreover, in Eq. (A6) the role of w^+_{it} and w^-_{it} becomes clear: they effectively work as general coordinates for each of the two fermionic species. In principle, the integrals in Eq. (A6) could have as upper limits some quantities $U^\pm_{it}$ to incorporate any possible prior knowledge on the bounds of the variables of interest.

We can now calculate the partition function Z of the ensemble in Eq. (2):

$$\begin{aligned}
Z &= \sum_{W\in\mathcal{W}} e^{-H(W)} = \\
&= \prod_{i=1}^{N}\prod_{t=1}^{T}\sum_{\substack{(A^+_{it},A^-_{it})=(1,0)\\(0,0)}}^{(0,1)}\int_0^{\infty}dw^+_{it}\int_0^{\infty}dw^-_{it}\;\times \\
&\quad\times\; e^{\left[\left(\alpha_i^N+\alpha_t^T\right)A^+_{it}+\left(\beta_i^N+\beta_t^T\right)A^-_{it}+\left(\gamma_i^N+\gamma_t^T\right)w^+_{it}+\left(\sigma_i^N+\sigma_t^T\right)w^-_{it}\right]} = \\
&= \prod_{i=1}^{N}\prod_{t=1}^{T}\sum_{\substack{(A^+_{it},A^-_{it})=(1,0)\\(0,0)}}^{(0,1)}\left(\frac{e^{-\alpha_i^N-\alpha_t^T}}{\gamma_i^N+\gamma_t^T}\right)^{A^+_{it}}\left(\frac{e^{-\beta_i^N-\beta_t^T}}{\sigma_i^N+\sigma_t^T}\right)^{A^-_{it}} = \\
&= \prod_{i=1}^{N}\prod_{t=1}^{T}\left[1+\frac{e^{-(\alpha_i^N+\alpha_t^T)}}{\gamma_i^N+\gamma_t^T}+\frac{e^{-(\beta_i^N+\beta_t^T)}}{\sigma_i^N+\sigma_t^T}\right] = \\
&= \prod_{i=1}^{N}\prod_{t=1}^{T}\left(1+e^{\frac{\mu^1_{it}-\epsilon_{it}}{T_{it}}}+e^{\frac{\mu^2_{it}-\epsilon_{it}}{T_{it}}}\right)\;,
\end{aligned} \tag{A7}$$

where we have defined the following quantities in order to make apparent the analogy with the two species fermionic gas introduced in the main text:

$$\begin{aligned}
T_{ij} &= \frac{1}{\log\left(\sigma_i^T+\sigma_j^e\right)+\log\left(\gamma_i^T+\gamma_j^e\right)}\;, \\
\epsilon_{ij} &= \frac{1}{2}+\frac{T_{ij}}{2}\left(\alpha_i^T+\alpha_j^e+\beta_i^T+\beta_j^e\right)\;, \\
\mu^2_{ij} &= \frac{kT_{ij}}{2}\left(\alpha_i^T+\alpha_j^e-\beta_i^T-\beta_j^e-\log\frac{\sigma_i^T+\sigma_j^e}{\gamma_i^T+\gamma_j^e}\right)=-\mu^1_{ij}\;.
\end{aligned}$$

From the above partition function, via Eq. (A5) we can derive the probability density function in Eq. (4), which quantifies the probability of drawing a specific instance W from the ensemble. The quantities defining such probability distribution have a well defined physical meaning, and read as follows:

- $P^+_{it}=\frac{e^{-\left(\alpha_i^N+\alpha_t^T\right)}}{\left(\gamma_i^N+\gamma_t^T\right)Z_{it}}$ Probability of observing a positive value in the i-th time series at time t
- $P^-_{it}=\frac{e^{-\left(\beta_i^N+\beta_t^T\right)}}{\left(\sigma_i^N+\sigma_t^T\right)Z_{it}}$ Probability of observing a negative value in the i-th time series at time t
- $1-P^+_{it}-P^-_{it}$ Probability of observing a missing value in the i-th time series at time t
- $Q^+_{it}(w)=(\gamma_i^N+\gamma_t^T)e^{-(\gamma_i^N+\gamma_t^T)w}$ Probability distribution of a positive value w for the i-th time series at time t
- $Q^-_{it}(w)=(\sigma_i^N+\sigma_t^T)e^{-(\sigma_i^N+\sigma_t^T)w}$ Probability distribution of a negative value w for the i-th time series at time t

When no data are missing, i.e. $(A^+_{it}, A^-_{it}) \neq (0,0)$, the sum defined in Eq. (A6) changes and, as a result, the partition function (A7) becomes :

$$Z = \prod_{i,t=1}^{N,T} Z_{it} = \prod_{i,t=1}^{N,T} \left[\frac{e^{-(\alpha_i^N + \alpha_t^T)}}{\gamma_i^N + \gamma_t^T} + \frac{1}{\sigma_i^N + \sigma_t^T} \right] \ .$$

After noticing that $A^+_{it} = 0 \Rightarrow w^+_{it} = 0 \wedge w^-_{it} > 0$, the probability of drawing from the ensemble an instance W can be easily found:

$$P(W) = \prod_{i,t=1}^{N,T} \left[P^+_{it} \ \ Q^+_{it}(w^+_{it}) \right]^{A^+_{it}} \left[P^-_{it} \ \ Q^-_{it}(w^-_{it}) \right]^{1-A^+_{it}} \ , \tag{A8}$$

where the quantities in the above expression are defined as those above.

Looking at Eq. (A8), we can understand how we have obtained Eq. (5) in the main text. In order to simulate a drawing of a set of time series W from the ensemble, we first need to construct a "topology" of positive events by placing a positive event in entry W_{it} with probability P^+_{it} and a negative event otherwise. Then we need to place a weight $W_{it} = x$ using one of the two exponential distributions $Q^\pm_{it}$ defined above, depending on the type of event that was assigned to W_{it}. This procedure is encompassed by the hyperexponential distribution in Eq. (5), which can be obtained via the standard generating function approach, and whose parameters read $\lambda^+_{it} = (\gamma^N_i + \gamma^T_i)^{-1}$, and $\lambda^-_{it} = (\sigma^N_i + \sigma^T_i)^{-1}$.

Appendix B: Application to a set of temperature time series

We now apply the framework introduced in the main text to sets of time series featuring temperatures recorded at different frequencies (week/day/8 hours) in $N = 30$ different North American cities [27] (weekly data range from July 2013 to July 2018, daily data range from July 2016 to July 2018, 8 hour data range from January 2017 to July 2018). We do so in order to test the ability of our ensemble approach to capture the main features of time series whose most relevant statistical properties are markedly different from those of financial returns, which we instead studied in the main paper. In particular, our main focus will be on the ability of the ensemble to capture the periodicities that characterize temperature data at different time scales.

As done in the main text, we indicate as $\overline{W}$ the $N \times T$ data matrix (with $T = 264, 730, 2321$ in the case of temperatures recorded at the weekly, daily, and 8 hour frequency, respectively) with values rescaled to have zero mean, and we indicate as W any generic instance drawn from the corresponding ensemble. We also redefine here for convenience the matrices $A^\pm = \Theta(\pm W)$, $w^\pm = \pm W \Theta(\pm W)$. The ensemble we are going to use is fully specified by the $6(N+T)$ constraints enforced in the following Hamiltonian (there are no missing data, which leads to $2(N+T)$ fewer constraints with respect to the general formulation outlined in the main paper):

$$\begin{aligned} H(W) = \sum_{i=1}^{N} \sum_{t=1}^{T} \big[&\left(\alpha_i^N + \alpha_t^T\right) A^+_{it} \\ &+ \left(\gamma_i^N + \gamma_t^T\right) w^+_{it} + \left(\sigma_i^N + \sigma_t^T\right) w^-_{it} \big] \ , \end{aligned} \tag{B1}$$

leading to the partition function:

$$Z = \prod_{i=1}^{N} \prod_{t=1}^{T} Z_{it} = \prod_{i=1}^{N} \prod_{t=1}^{T} \left[\frac{e^{-(\alpha_i^N + \alpha_t^T)}}{\gamma_i^N + \gamma_t^T} + \frac{1}{\sigma_i^N + \sigma_t^T} \right] \ . \tag{B2}$$

In Figure 4 we show that, independently from the frequency at which temperatures are sampled, the average ensemble power spectral density captures remarkably well the relevant frequencies that characterize the empirical time series of each city. Indeed, as can be seen from panels **a** and **b**, the ensemble power spectra based on the data recorded at the weekly and daily frequency perfectly capture the six-months periodicity associated with the seasons' cycle. Panel **c** shows that the same frequency is also captured in the data recorded every 8 hours, and that when calibrating the ensemble on such data, the power spectrum also perfectly captures the daily frequency associated with the day-night cycle (see inset).

In Fig. 5 we expand the above analysis to the periodicities of moments. Panel **a** shows the empirical daily variance of temperatures recorded across the 30 cities mentioned above against the corresponding ensemble average. At first sight, the latter seems to be largely uncorrelated from the former. Yet, the corresponding power spectrum shown in panel **b** highlights that the relevant frequencies in the data (six months and one day) are captured very well, although the ensemble places additional power on such frequencies.

A somewhat similar phenomenon is shown in panels **c** and **d**, which show the daily skewness computed across all cities and its corresponding power spectra. Once again, the average ensemble spectrum places more power on the six-months and daily frequencies with respect to the empirical one. This results in a clearly discernible oscillating pattern, which significantly deviates from the empirical behavior. Nevertheless, these results are interesting. Indeed, as it can be seen in panel **c** positive (negative) skewness values take place during the summer (winter) months, as a reflection of higher (lower) average temperatures. Although this is a fairly trivial example, it highlights how the ensemble approach can reveal stylized trends that are genuinely informative about the dynamics of the system under study.

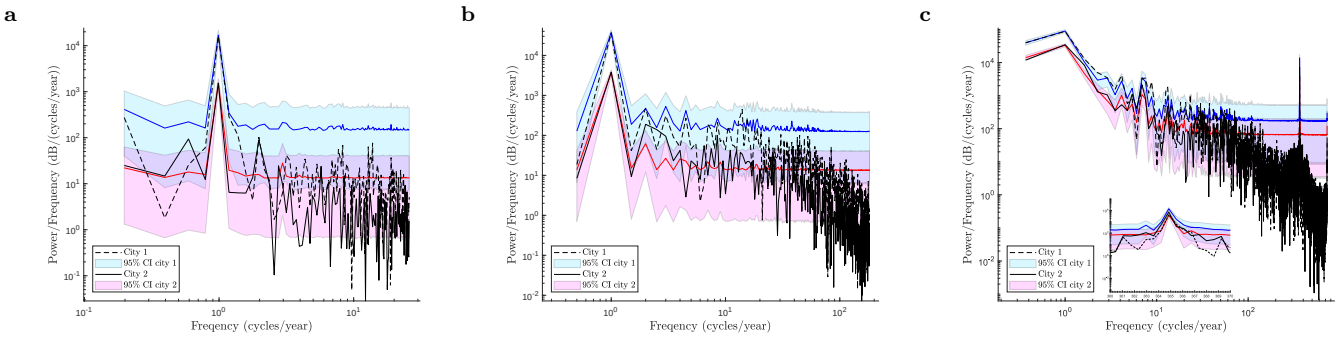

FIG. 4: Ability of the esemble to preserve periodicities in the data. a) Empirical power spectrum of weekly temperatures against the average ensemble spectrum for two different cities (city 1 is Boston and city 2 is Los Angeles). **b)** Same plot for daily temperatures. **c)** Same plot for 8 hours temperatures.

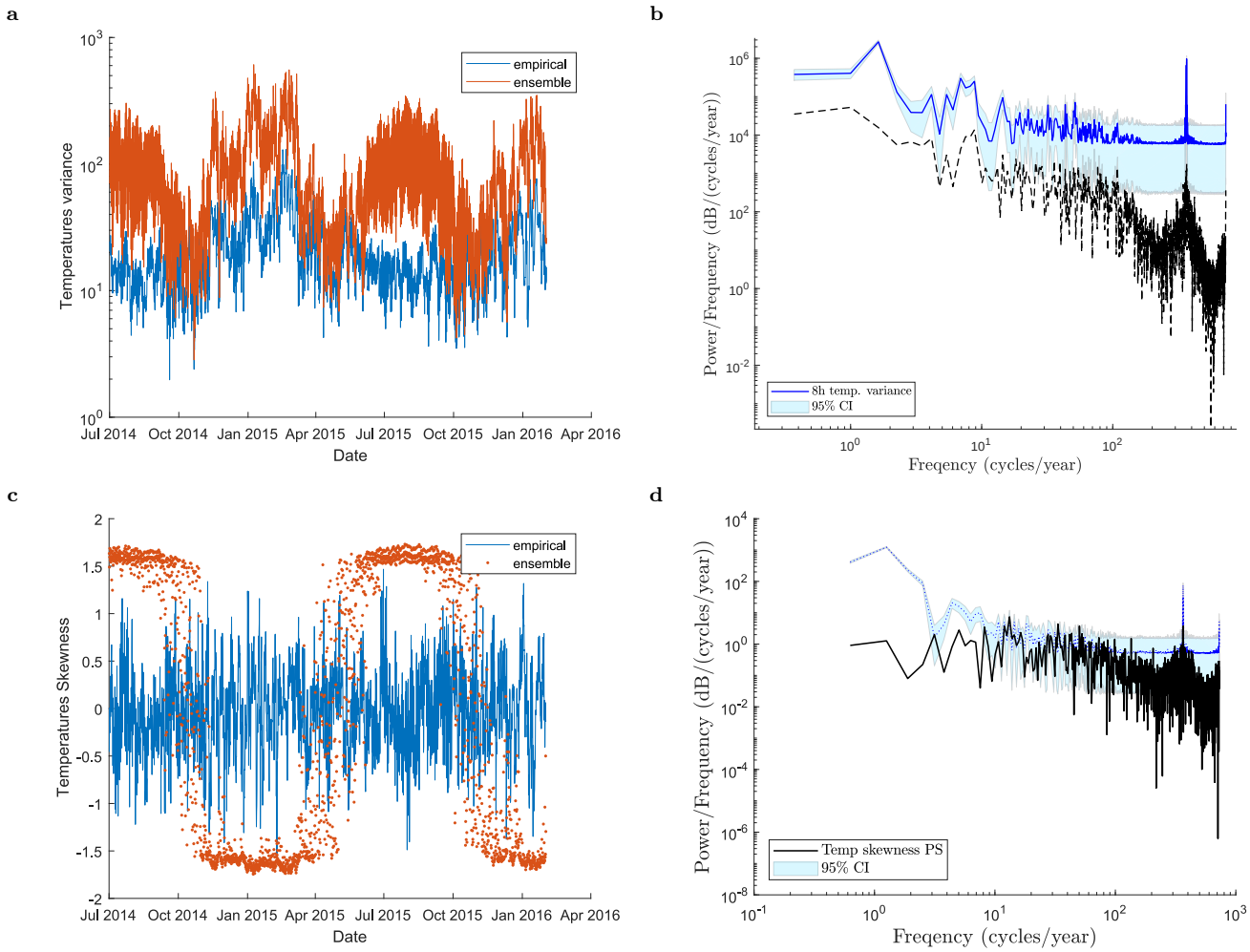

FIG. 5: Ability of the ensemble to preserve periodicities in the data. a) Variance of the temperatures recorded at 8 hour intervals across all 30 cities (the blue line denotes empirical values, the orange one denotes the ensemble average). **b)** Comparison between the empirical spectrum of the 8-hours temperature variance across cities (dashed line) and the ensemble spectrum (blue line). **c)** Skewness of the temperatures recorded at 8 hour intervals across all 30 cities (the blue line denotes empirical values, the orange one denotes the ensemble average). **d)** Comparison between the empirical spectrum of the 8-hours temperature skewness across cities (dashed line) and the ensemble spectrum (blue line).

Code availability

The MATLAB code used to implement the ensemble methodology described in this paper is available at mathworks.com/matlabcentral/fileexchange/72000-canonical-ensemble-for-time-series

Acknowledgments

G.L. acknowledges support from an EPSRC Early Career Fellowship in Digital Economy (Grant No. EP/N006062/1).

[1] C. Chatfield, *Introduction to multivariate analysis* (Routledge, 2018).
[2] H. E. Tinsley and S. D. Brown, *Handbook of applied multivariate statistics and mathematical modeling* (Academic press, 2000).
[3] J. Vaze, D. Post, F. Chiew, J.-M. Perraud, N. Viney, and J. Teng, Journal of Hydrology **394**, 447 (2010).
[4] G. Drótos, T. Bódai, and T. Tél, Phys. Rev. E **94**, 022214 (2016), URL `https://link.aps.org/doi/10.1103/PhysRevE.94.022214`.
[5] P. von Bünau, F. C. Meinecke, F. C. Király, and K.-R. Müller, Phys. Rev. Lett. **103**, 214101 (2009).
[6] C. Tsallis, C. Anteneodo, L. Borland, and R. Osorio, Physica A: Statistical Mechanics and its Applications **324**, 89 (2003).
[7] R. CONT, Quantitive Finance **1**, 223 (2001).
[8] G. Livan, J.-i. Inoue, and E. Scalas, Journal of Statistical Mechanics: Theory and Experiment **2012**, P07025 (2012).
[9] A. C. Davison and D. V. Hinkley, *Bootstrap methods and their application*, vol. 1 (Cambridge university press, 1997).
[10] D. Kuonen, WBL in Angewandter Statistik ETHZ 2017 **19**, 1 (2018).
[11] J. S. Haukoos and R. J. Lewis, Academic emergency medicine **12**, 360 (2005).
[12] H. Lütkepohl, *New introduction to multiple time series analysis* (Springer Science & Business Media, 2005).
[13] D. Qin, Journal of Economic Surveys **25**, 156 (2011).
[14] P. Whittle, Biometrika **39**, 309 (1952), ISSN 00063444, URL `http://www.jstor.org/stable/2334027`.
[15] T. Bollerslev, Journal of econometrics **31**, 307 (1986).
[16] J. Park and M. E. Newman, Physical Review E **70**, 066117 (2004).
[17] A. Gabrielli, R. Mastrandrea, G. Caldarelli, and G. Cimini, Phys. Rev. E **99**, 030301 (2019), URL `https://link.aps.org/doi/10.1103/PhysRevE.99.030301`.
[18] G. Cimini, T. Squartini, F. Saracco, D. Garlaschelli, A. Gabrielli, and G. Caldarelli, Nature Reviews Physics **1**, 58 (2019).
[19] T. Squartini and D. Garlaschelli, New Journal of Physics **13**, 083001 (2011).
[20] D. Garlaschelli and M. I. Loffredo, Physical Review E **78**, 015101 (2008).
[21] V. A. Marchenko and L. A. Pastur, Matematicheskii Sbornik **114**, 507 (1967).
[22] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, and H. E. Stanley, Physical review letters **83**, 1471 (1999).
[23] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, Physical review letters **83**, 1467 (1999).
[24] G. Livan, M. Novaes, and P. Vivo, *Introduction to random matrices: theory and practice* (Springer, 2018).
[25] G. Livan, S. Alfarano, and E. Scalas, Physical Review E **84**, 016113 (2011).
[26] This in general holds for any discretization of the distribution of the entries of W.
[27] Vancouver, Portland, San Francisco, Seattle, Los Angeles, San Diego, Las Vegas, Phoenix, Albuquerque, Denver, San Antonio, Dallas, Houston, Kansas City, Minneapolis, Saint Louis, Chicago, Nashville, Indianapolis, Atlanta, Detroit, Jacksonville, Charlotte, Miami, Pittsburgh, Toronto, Philadelphia, New York, Montreal, Boston