

https://dspace.org/

Permutation invariant Gaussian matrix models for financial correlation matrices

George Barnes; Sanjaye Ramgoolam; Michael Stephanou

ELSEVIER SCIENCE BV
https://hdl.handle.net/10539/42118
WOS

Downloaded from a DSpace repository. https://dspace.org/

Physica A 651 (2024) 130015



Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa





Permutation invariant Gaussian matrix models for financial correlation matrices

George Barnes [a,*], Sanjaye Ramgoolam [a,b], Michael Stephanou [c]

[a] Centre for Theoretical Physics, School of Physical and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK

[b] National Institute for Theoretical Physics, School of Physics and Centre for Theoretical Physics, University of the Witwatersrand, Wits, 2050, South Africa

[c] Rand Merchant Bank, 1 Merchant Place, Fredman Drive, Johannesburg, 2196, South Africa

ARTICLE INFO

Keywords:
Permutation invariant matrix models
Symmetric group representation theory
Statistical mechanics
Gaussianity
Financial correlations
High-frequency foreign exchange data

ABSTRACT

We construct an ensemble of correlation matrices from high-frequency foreign exchange market data, with one matrix for every day for 446 days. The matrices are symmetric and have vanishing diagonal elements after subtracting the identity matrix. For such ensembles, we construct the general permutation invariant Gaussian matrix model, which has 4 parameters characterised using the representation theory of symmetric groups. The permutation invariant polynomial functions of the symmetric, diagonally vanishing matrices have a basis labelled by undirected loop-less graphs. Using the expectation values of the general linear and quadratic permutation invariant functions of the matrices in the dataset, the 4 parameters of the matrix model are determined. The model then predicts the expectation values of the cubic and quartic polynomials. These predictions are compared to the data to give strong evidence for a good overall fit of the permutation invariant Gaussian matrix model. The linear, quadratic, cubic and quartic polynomial functions are then used to define low-dimensional feature vectors for the days associated to the matrices. These vectors, with choices informed by the refined structure of small non-Gaussianities, are found to be effective as a tool for anomaly detection in market states: statistically significant correlations are established between atypical days as defined using these feature vectors, and days with significant economic events as recognised in standard foreign exchange economic calendars. They are also shown to be useful as a tool for ranking pairs of days in terms of their similarity, yielding a strongly statistically significant correlation with a ranking based on a higher dimensional proxy for visual similarity.

1. Introduction

Permutation invariant Gaussian matrix models have recently been introduced [1], with motivations coming from the study of the statistics of ensembles of matrices arising in computational linguistics [2]. They have been applied to show the existence of approximate Gaussianity in these ensembles [3,4]. Permutation invariant polynomials of matrix variables, which are the key observables in the study of Gaussianity, have also been used as tools for lexical semantic tasks in computational linguistics [4].

In this paper, we construct and analyse financial correlation matrices for a sequence of days obtained by calculating correlations between price movements in high frequency foreign exchange market data. This is an interesting study in itself and adds to a rather sparse literature on high frequency forex correlation matrices ([5] is one other recent example). The correlation matrices are

* Corresponding author.
E-mail addresses: george.barnes.3108@gmail.com (G. Barnes), s.ramgoolam@qmul.ac.uk (S. Ramgoolam), michael.stephanou@gmail.com (M. Stephanou).

https://doi.org/10.1016/j.physa.2024.130015
Received 9 July 2023; Received in revised form 14 July 2024
Available online 8 August 2024
0378-4371/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

symmetric and have vanishing matrix elements along the diagonal (by subtracting the identity matrix). We also briefly note the positive-semidefinite property of these correlation matrices which we do not explicitly model. The permutation invariant matrix model constructed in [1] uses an integration over general matrices and has 13 parameters, 2 are coefficients in the action of linear invariants while 11 are coefficients of quadratic invariants. For the case of the restricted matrices here, there is a reduction of the 13-parameter model to a 4-parameter model. In this paper we explicitly construct the 4-parameter model and use it to demonstrate approximate permutation invariant Gaussianity in the ensemble of forex correlation matrices we construct.

The permutation invariant polynomial functions of generic matrices of size D and degree k have a basis labelled by directed graphs with k edges and any number of nodes, when $D \geq 2k$. This condition is satisfied for the cases of interest in this paper since we consider k up to 8 and $D = 19$. The nodes of the graph correspond to indices being summed and the edges of the graph correspond to the matrix variables. In the case of permutation invariant polynomial functions of symmetric, diagonally vanishing matrices of degree k, there exists a basis labelled by undirected graphs containing no loops (i.e. no edges starting and ending at the same node). See Appendix B.4 for an explanation of this connection between graphs and polynomial invariant functions, as well as examples of the loopless graphs and the corresponding polynomials.

In part the motivation from physics for applying the permutation invariant Gaussian matrix (PIGM) models is to establish universality properties of these permutation invariant Gaussianities in diverse matrix ensembles in data sciences. In the case of traditional random matrix theory (RMT) [6,7], which is also Gaussian but has a continuous, as opposed to a discrete symmetry, the areas of application extend across many-body quantum physics [8] and various areas of data science [9]. This diverse range of applications defines a universality class of random matrix statistics. There are already hints of such a universality in existing applications of PIGM models, where different matrix ensembles in computational linguistics constructed from different kinds of algorithms show approximate Gaussianity. The Gaussianity analysed in [4] was based on the construction of matrices by linear regression in [2] while [3] extended the analysis of [4] and also analysed the matrices constructed by neural network methods in [10].

Further motivation for investigating the broad theme of Gaussianity within a financial setting comes from the known statistical properties of correlation matrices. Existing results establish the asymptotic D^2-variate Gaussianity of the sampling distribution of $D \times D$ dimensional correlation matrices under fairly general conditions for example (related to finite fourth moments of the underlying observations from which the correlation matrices are constructed, see [11], Theorem 3.4.4 and subsequent comments). This differs from our non-asymptotic i.e. finite sample setting, but does imply that the distribution of ensembles of correlation matrices approaches a multivariate Gaussian in the large observation sample limit.

Other known statistical results include the fact that the non-asymptotic sampling distribution of correlation matrices is the Wishart distribution, under certain specific conditions [12]. In particular, the Wishart distribution only arises in the case where the observations from which the correlation matrix estimates are constructed are themselves multivariate Gaussian. We do not make this assumption. Thus our setting is one of finite observation samples which may not be Gaussian distributed. It has also been noted that the correlation information encoded in correlation matrices is not sensitive to the ordering of basis vectors (in [13] for financial correlation matrices for example). This motivates a permutation invariant model for correlation matrices. The Wishart distribution is not permutation invariant in general, in contrast to the PIGM model. To summarise then, in our setting we apply the PIGM model as a phenomenological model for ensembles of correlation matrices. The PIGM model is permutation invariant and is able to capture leading Gaussian structure which forms the basis for quantifying small non-Gaussian corrections.

We note a rich history of applying random matrix theory (RMT) to the study of financial correlation matrices centring on the eigenvalue distributions of these matrices [14–16]. In particular, random matrix models have been fit to the empirical eigenvalue distributions of such correlation matrices, predominantly the so-called Marchenko–Pastur (M–P) distribution [17] derived for identity correlation matrices, as well as more realistic models [18,19]. Evidence has also been presented that large eigenvalues are associated with overall market and sector correlation structure (see [16] for example). Practical applications of these findings have been developed such as cleaning/de-noising correlation matrices, amongst others [20–23].

The PIGM model provides a new approach to studying and describing financial correlation matrices that is distinct from existing approaches based on RMT. It focuses on low degree permutation invariant polynomial functions of matrices (which we refer to as observables) instead of eigenvalue distributions, which are the focus of traditional RMT. This perspective is based on the postulate that near-Gaussian permutation invariant sectors of real world matrix data contain useful information. The PIGM model furnishes a parsimonious specification of the probability density function of these matrices using only 4 free parameters for the symmetric, vanishing diagonal model. This is close to the one or two parameters of the simplest RMT and far smaller than a multi-variate Gaussian distribution for a $D \times D$ matrix, which has order D^2 parameters. It provides an analytical solution to the expectation values of permutation invariant products of matrix elements. The empirical higher order observables (cubic, quartic etc.) that agree closely with the model – which is only fit to linear and quadratic observables – reveal consistency with random matrices implied by the fitted Gaussian model. The empirical higher order observables that depart from theoretical expectations indicate informative structure beyond that encoded in the random model. A vector of observables therefore provides a signature for a particular correlation matrix, which may provide a useful, lower dimensional, permutation invariant representation of correlation matrices. The effectiveness of this representation is explored in anomaly detection and measurement of similarity as initial examples.

It should be noted that the PIGM model can be applied beyond symmetric correlation matrices to general cross-correlation matrices with only permutation invariance (e.g. constructed from returns at different times) through the full 13-parameter PIGM model developed in [1]. In this article we lay the foundations for these future studies.

In Section 2 we summarise the theoretical results of the paper on permutation invariant Gaussian models. We define general permutation invariant Gaussian matrix models and consider the restriction of these models appropriate to the financial data described

in Section 3, namely that the matrices must be symmetric and have vanishing diagonal elements. We also define the permutation invariant observables of the model and explain a useful bijection between these observables and loopless graphs, examples of which are given. The detailed construction and solution of the PIGM model is contained within Appendix B. It is achieved with the help of representation theory of the symmetric group and builds on the results of [1,24]. We find that these models are characterised by 1 linear and 3 quadratic couplings. Linear and quadratic expectation values of observables can be expressed simply in terms of these coupling parameters, see Eqs. (B.72) and (B.71) respectively. Higher order expectation values are simply constructed from these with the application of Wick's theorem.

Section 3 gives details of the high-frequency forex data used to construct the matrix ensemble studied in the remainder of the paper, as well as the method by which the members of this ensemble are constructed from the underlying data.

Section 4 contains a description of the empirical statistical properties of the observables. This includes measures of their Gaussianity and comparison of their properties with those predicted by the model presented in Section 2.[1]

In Section 5 we construct vectors of observables for each correlation matrix. These observable vectors are low dimensional representations of the correlation matrices. There are 31 cubic and quartic observables for general matrix size D (as long as $D \geq 8$, a condition which is generally satisfied in large D applications such as the one here).

In general, we find that the observable vectors provide a good representation of the original correlation matrices, performing well in anomaly detection and similarity ranking applications. The best performances are achieved by selecting subsets of the cubic and quartic observables, based on the ranking of their small non-Gaussianities, and on the postulate that the more non-Gaussian observables are most informative of economic factors driving atypicality of the days. We conclude in Section 6 and suggest interesting future research directions arising from this work. We collect some derivations, figures and robustness checks in supplementary material, which are referred to within the main text. The code and data required to reproduce the results in this article are available at https://github.com/pigm-finance/codedata, with certain aspects further described in the supplementary material.

2. Summary of results on the 4-parameter Gaussian matrix model

Here we summarise the main results and outline the key ideas behind the construction of the general PIGM model for an ensemble of symmetric matrices which have vanishing matrix elements along the diagonal. This section is intended to provide, for a reader with background in mathematical finance or statistics, an understanding of the key theoretical points of the paper, without getting into the full details of the construction presented in Appendix B which will be more easily accessible to a reader with knowledge of group representation theory at a level covered in mathematical texts such as in chapters 3, 5 and 7 of [25]. We will review the description of probability distributions using a Euclidean action which is Gaussian or near-Gaussian using the simple case of one-variable statistics and motivate the measure of non-Gaussianity we use later in the case of permutation invariant matrix distributions. We explain the structure of the 4-parameter permutation invariant Gaussian matrix model and the connection between the permutation invariant polynomial functions of the matrices with loopless graphs. Finally we present formulae for the one and two-point functions of matrix variables derived in detail in Appendix B.

It is useful to recall that a one-variable Gaussian distribution, with mean μ and standard deviation σ, for a random variable x, is described by a probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2.1}$$

The moments of the distribution are expectation values $\langle x^k \rangle$, defined as

$$\langle x^k \rangle = \int_{-\infty}^{\infty} dx \;\; f(x) \;\; x^k. \tag{2.2}$$

It is also useful to define, by analogy with statistical physics and quantum mechanical path integrals, the action $S = \frac{(x-\mu)^2}{2\sigma^2}$. The partition function is

$$Z = \int dx \;\; e^{-S}, \tag{2.3}$$

while the moments are

$$\langle x^k \rangle = \frac{1}{Z} \int dx \;\; e^{-S} \;\; x^k. \tag{2.4}$$

The action S is a quadratic function of x.

It is often the case that the action of a theory is approximately Gaussian — deviating from Gaussianity by some small higher order terms. The full action of the system S' is then written as a Gaussian piece S, plus an additional non-Gaussian piece δS

$$S' = S + \lambda \delta S. \tag{2.5}$$

[1] We include a further, practical comparison between empirical observations and theoretical results from the PIGM model in section 3 of the supplementary material.

The smallness of the higher order terms is governed by the interaction strength λ, whose smallness is required to ensure

$$\langle \mathcal{O}_\alpha \rangle_S - \langle \mathcal{O}_\alpha \rangle_{S'} < \sigma_{\langle \mathcal{O}_\alpha \rangle_{S'}}, \tag{2.6}$$

i.e. the expectation value of some observable $\mathcal{O}_\alpha$, is largely insensitive to the non-Gaussian contribution to the true action governing the theory, δS.

More concretely, take $\mu = 0$ in the simple pure Gaussian, one parameter toy model defined above. The partition function is

$$Z = \int \mathrm{d}x \;\; e^{-S} = \int \mathrm{d}x \;\; e^{-\frac{x^2}{2\sigma^2}}. \tag{2.7}$$

View this as an approximation to some true physical partition function which includes some small, non-Gaussian perturbation. We explain by way of a simple example which captures the mechanism, that the absolute differences between expectation values of low order polynomials in the random variables (observables) in the Gaussian model and the perturbed model are small compared to the standard deviation of the observable. The simple example consists of a Gaussian action perturbed by a small quartic correction, so that the perturbed partition function Z' is

$$Z' = \int \mathrm{d}x e^{-S'} = \int \mathrm{d}x e^{-\frac{x^2}{2\sigma^2} - \frac{\lambda}{4!}x^4} = Z\left(1 - \frac{\lambda}{4!}\langle x^4 \rangle + \cdots\right). \tag{2.8}$$

Using (2.4) we calculate the absolute difference between the fourth moment of each of the theories and obtain (see the supplementary material for the derivation),

$$\frac{\left|\langle x^4 \rangle - \langle x^4 \rangle_{S'}\right|}{\sigma_{\langle x^4 \rangle_{S'}}} \sim \lambda \sigma^4. \tag{2.9}$$

Therefore, as long as the physical theory is approximately Gaussian, i.e. λ is small, its normalised fourth moment is well approximated by that of the purely Gaussian theory.

We postulate that real market effects governing the interactions between currency rates – as encoded in the distribution of the correlation matrices we study – are modelled analogously by a Gaussian action (incorporating linear and quadratic permutation invariant polynomial functions of the matrices) plus some small non-Gaussian perturbation (cubic and higher order permutation invariant polynomial functions of the matrices). The smallness of the non-Gaussian terms allows us to approximate expectation values using a purely Gaussian theory. Evidence for this near-Gaussianity is provided primarily by the smallness of the measured observable deviations from those predicted by a purely Gaussian theory. These are listed in Table 4 of Section 4.

Permutation invariant Gaussian matrix models for generic real matrices of size D are described by a partition function

$$\mathcal{Z} = \int \prod_{i,j=1}^{D} d\,M_{ij} \;\; e^{-S^{\mathrm{PIGMM}}}. \tag{2.10}$$

The integration measure is the standard Euclidean measure on $\mathbb{R}^{D^2}$. The general Gaussian action is defined by a general quadratic permutation invariant function of the matrix elements M_{ij}, depending on 13 parameters, which is compatible with convergence of the integral [1]. The set of all permutations of D objects forms the symmetric group S_D. The permutations $\gamma \in S_D$ act as

$$\gamma : M_{ij} \to M_{\gamma(i)\gamma(j)}. \tag{2.11}$$

A precise description of the general parameter space for such a Gaussian action is found using the representation theory of the symmetric group. A key outcome of the representation theoretic treatment is that there exists a convenient set of variables $S_a^{V_A;\tau_A}$, which are linear combinations of the D^2 variables M_{ij}:

$$S_a^{V_A;\tau_A} = \sum_{i,j=1}^{D} C_{a\,;\,ij}^{V_A,\tau_A} M_{ij}. \tag{2.12}$$

The variables V_A are vector spaces $\{V_{[D]} = V_0, V_{[D-1,1]} = V_H, V_{[D-2,2]} = V_2, V_{[D-2,1,1]} = V_3\}$. The meaning of this list is explained further in Appendix B. For now the only important point is that this is a list of 4 elements. The index a ranges over a set of basis elements for each vector space V_A, numbering respectively $\{1, (D-1), D(D-3)/2, (D-1)(D-2)/2\}$, which are the dimensions $\mathrm{Dim}V_A$ of the vector spaces. The index τ_A runs over $\{2, 3, 1, 1\}$ values respectively. The key result is that the action for the general permutation invariant Gaussian model takes the form

$$S^{\mathrm{PIGMM}} = -\mu_1 S^{V_0;1} - \mu_2 S^{V_0;2} + \sum_A \sum_{a=1}^{\mathrm{Dim}V_A} g^{(A)}_{\tau_A,\tau'_A} \sum_{\tau,\tau'} S_a^{V_A;\tau_A} S_a^{V_A;\tau'_A}, \tag{2.13}$$

where μ_1 and μ_2 are linear coupling parameters. The parameters $g^{(A)}_{\tau_A,\tau'_A}$ are symmetric matrix parameters of matrix size $2, 3, 1, 1$. Thus they define parameter spaces of dimension $\{3, 6, 1, 1\}$. Convergence is guaranteed by the condition that these matrices have positive eigenvalues. Thus the general Gaussian model has 2 parameters for the linear invariants and 11 parameters for the quadratic invariants.

We show in Appendix B that the general permutation invariant Gaussian matrix model for symmetric matrices is a 9-parameter model, and for symmetric matrices with diagonally vanishing matrix elements the permutation invariant Gaussian model is a 4-parameter model. The 9-dimensional parameter space for the symmetric matrices and the 4-dimensional parameter space for the

symmetric diagonally vanishing matrices are subspaces of the 13-dimensional parameter space for generic matrices. The embedding of the 4-dimensional parameter space in the 13-parameter space is described in Appendix B.

The diagonal action of S_D simultaneously permutes the rows and columns of a matrix as in (2.11). We wish to find an action that is invariant under this group action. We refer to the space of symmetric matrices with vanishing diagonal as the physical subspace V^{phys} of general $D \times D$ matrices and label matrices in the space with a superscript i.e. $M^{\text{phys}} \in V^{\text{phys}}$. They obey the conditions

$$M_{ij} = M_{ji}, \qquad M_{ii} = 0, \qquad 1 \leq i, j \leq D. \tag{2.14}$$

Since these conditions are S_D equivariant with respect to the action defined in (2.11) the physical subspace is invariant under S_D

$$M_{ij} \in V^{\text{phys}} \Rightarrow M_{\gamma(i)\gamma(j)} \in V^{\text{phys}}, \qquad \forall \gamma \in S_D. \tag{2.15}$$

By physical we mean only to restrict to the non-trivial data of interest. In the correlation matrices described in Section 3 all the data is contained within symmetric matrices with vanishing diagonal.

An important ingredient in understanding permutation invariant random matrix models is the structure of the permutation invariant polynomial functions of matrices, which are closely related to graphs. The key point behind this relation is that permutation invariant polynomial functions can be constructed by summing over the matrix indices. For general matrices there are 2 linear and 11 quadratic functions, these polynomials and their graphs are described in Appendix B.4. For symmetric matrices with vanishing diagonal elements there is 1 linear invariant function and 3 quadratic invariant functions, which can be represented by the following graphs



(2.16)

Each index summation is represented by a node of the graph and matrices span these nodes as edges. No loops are permitted due to the matrices having vanishing diagonal and the edges are undirected due to the symmetry of the matrices.

For a fixed degree, the permutation invariant polynomial functions form a vector space. As long as the matrix dimension is larger than twice the degree of the polynomial the graphs are in one-to-one correspondence with basis elements of this vector space. In our present financial application this degree condition is always satisfied. A detailed discussion of this condition and the effects of going beyond it are given in [24].

The action of the reduced PIGM model is given by the most general combination of permutation invariant linear and quadratic terms

$$S = -\mu \sum_{i,j=1}^{D} M_{ij} + \tau_1 \sum_{i,j=1}^{D} M_{ij} M_{ij} + \tau_2 \sum_{i,j,k=1}^{D} M_{ij} M_{jk} + \tau_3 \sum_{i,j,k,l=1}^{D} M_{ij} M_{kl}\,, \tag{2.17}$$

where μ is the linear coupling strength and τ_1, τ_2 and τ_3 are the quadratic couplings. Label observables of this theory $\mathcal{O}_\alpha(M_{ij})$, where α indexes the particular observable, they are permutation invariant polynomials of the random matrix variables of symmetric matrices with vanishing diagonal. Expectation values of these variables are defined as

$$\langle \mathcal{O}_\alpha(M_{ij}) \rangle = \frac{1}{\mathcal{Z}} \int \mathrm{d}M \mathcal{O}_\alpha e^{-S}, \tag{2.18}$$

where the measure is defined as

$$\mathrm{d}M \equiv \prod_{i<j} \mathrm{d}M_{ij}. \tag{2.19}$$

The quadratic, cubic and quartic observables along with their corresponding graphs are listed in Appendix B.4. In order to compute (2.18) for any choice of $\mathcal{O}_\alpha$ we must find a change of basis that factorises the RHS. This is possible with the application of appropriate projectors. Given these the action for the 4-parameter model can be written

$$S = \sum_{i,j,k,l=1}^{D} \frac{1}{2} \Big(\tau_{V_0} M_{ij} P^{\text{phys};V_0}_{ij,kl} M_{kl} + \tau_{V_H} M_{ij} P^{\text{phys};V_H}_{ij,kl} M_{kl} + \tau_{V_2} M_{ij} P^{\text{phys};V_2}_{ij,kl} M_{kl} \Big) - \sum_{i,j=1}^{D} \mu_{V_0} C^{\text{phys};V_0}_{ij} M_{ij}. \tag{2.20}$$

Where $\mu_{V_0}, \tau_{V_0}, \tau_{V_H}$ and τ_{V_2} are the couplings in the transformed basis. Performing the projections in (2.20) allows for the solution of the linear and quadratic expectation values via standard techniques of Gaussian integration. These can then be transformed into the original M_{ij} basis to give expressions for the matrix observables. Defining

$$\widetilde{\mu}_{V_0} \equiv \tau_{V_0}^{-1} \mu_{V_0}, \tag{2.21}$$

this procedure produces the one-point function

$$\begin{aligned} \langle M^{\text{phys}}_{ij} \rangle &= C^{\text{phys};V_0}_{ij} \langle S^{\text{phys};V_0} \rangle \\ &= \Big(\sqrt{\frac{D-1}{D^3}} - \frac{1}{\sqrt{D(D-1)}} F_{i,j} \Big) \widetilde{\mu}_{V_0}, \end{aligned} \tag{2.22}$$

and the two-point function

$$\begin{aligned}\langle M_{ij}^{\text{phys}} M_{kl}^{\text{phys}}\rangle =&\Big(\sqrt{\frac{D-1}{D^3}}-\frac{1}{\sqrt{D(D-1)}}F_{i,j}\Big)\Big(\sqrt{\frac{D-1}{D^3}}-\frac{1}{\sqrt{D(D-1)}}F_{k,l}\Big)\widetilde{\mu}_{V_0}^2\frac{1}{D}\Big(\frac{1}{D-1}F_{i,j}F_{k,l}-\frac{1}{D}\big(F_{i,j}+F_{k,l}\big)+\frac{D-1}{D^2}\Big)\tau_{V_0}^{-1}\\&+\frac{1}{2(D-2)}\big(1-\delta_{ij}\big)\big(1-\delta_{kl}\big)\big(F_{i,k}+F_{j,k}+F_{i,l}+F_{j,l}\big)\tau_{V_H}^{-1}\\&+\left(\frac{1}{2}F_{i,k}F_{j,l}+\frac{1}{2}F_{i,l}F_{j,k}-\frac{D}{D-2}\sum_{p,q=1}^{D}F_{i,p}F_{j,p}F_{k,q}F_{l,q}F_{p,q}-\frac{1}{D-1}F_{i,j}F_{k,l}\right)\tau_{V_2}^{-1}.\end{aligned}\tag{2.23}$$

Identifying and then summing over indices in (2.22) and (2.23) produces analytic expressions in D for the expectation values of the linear and quadratic observables

$$\sum_{i,j}\langle M_{ij}^{\text{phys}}\rangle=\sqrt{D(D-1)}\widetilde{\mu}_{V_0},\tag{2.24}$$

and

$$\sum_{i,j}\langle M_{ij}^{\text{phys}}M_{ij}^{\text{phys}}\rangle=\widetilde{\mu}_{V_0}^2+\tau_{V_0}^{-1}+(D-1)\tau_{V_H}^{-1}+\frac{D(D-3)}{2}\tau_{V_2}^{-1},\tag{2.25}$$

$$\sum_{i,j,k}\langle M_{ij}^{\text{phys}}M_{jk}^{\text{phys}}\rangle_{\text{conn}}=(D-1)\widetilde{\mu}_{V_0}^2+(D-1)\tau_{V_0}^{-1}+\frac{(D-1)(D-2)}{2}\tau_{V_H}^{-1},\tag{2.26}$$

$$\sum_{i,j,k,l}\langle M_{ij}^{\text{phys}}M_{kl}^{\text{phys}}\rangle_{\text{conn}}=D(D-1)\widetilde{\mu}_{V_0}^2+D(D-1)\tau_{V_0}^{-1}\tag{2.27}$$

respectively.

Cubic, quartic and higher expectation values can be calculated from these with the application of Wick's theorem, which allows them to be expressed as sums of products of linear and quadratic expectation values. Details of Wick's theorem as applied to cubic and quartic expectation values are given in Appendix B.4.2.

The equations for the 13-parameters appearing in the action (2.13) in terms of the 4-parameters of the physical action (2.20) are given in (B.80) and (B.81). As expected both models give consistent results for the expectation values of physical observables.

To determine how well this PIGM model predicts the statistics of the forex correlation data described in Section 3 we first use the correlation matrices to define the Gaussian model i.e. to fix the linear and quadratic couplings of the model. We then calculate the theoretical expectation values $\langle\mathcal{O}_\alpha(M)\rangle_{\text{T}}$ defined by (2.18) using this action with the application of Wick's theorem. These can then be compared to the experimental expectation values $\langle\mathcal{O}_\alpha(M)\rangle_{\text{E}}$ calculated from the financial correlation matrices themselves using the following similarity measure

$$\Delta_\alpha=\frac{\big|\langle\mathcal{O}_\alpha(M)\rangle_{\text{T}}-\langle\mathcal{O}_\alpha(M)\rangle_{\text{E}}\big|}{\sigma_{E,\alpha}(M)},\tag{2.28}$$

where $\sigma_{E,\alpha}$ is the standard deviation of the expectation value with respect to the ensemble of correlation matrices. This measure of similarity is used to identify the observables which deviate most significantly from Gaussianity. In Sections 5.1 and 5.2 these least Gaussian observables are shown to be the optimal candidates for data reduction in a variety of anomaly detection and similarity tests.

3. Daily correlation matrices from high-frequency forex data

3.1. Forex data description

The high-frequency forex data that we analyse pertain to 19 of the most liquidly traded currency pairs and cover the date range from 1 April 2020 to 31 January 2022. The data is sourced from TrueFX [26] and is comprised of all updates/ticks of the best price quotes at which any market participant is willing to buy (top-of-book bid quotes) or sell (top-of-book offer quotes). Market participants providing such price quotes include banks, brokers and asset managers on the Integral OCX platform.[2] We exclude United States currency settlement holidays (days where no settlements of prior transactions are made) due to the central importance of the US Dollar to FX trading. We also exclude the 24th, 25th, 26th, 31st of December and the 1st, 2nd of January due to reduced liquidity. The rationale behind this is to remove days which would likely have a different data generating process a priori. In total, around one billion pricing updates were analysed.

For each currency pair, the mid-price series, $p_j^{(I)}$, is calculated from the bid and offer quotes as follows:

$$p_j^{(I)}=(b_j^{(I)}+a_j^{(I)})/2\quad I\in\{1,\dots,19\},\ j\in\{1,\dots,n^I\},\tag{3.1}$$

where $b^{(I)}$ and $a^{(I)}$ are contemporaneous bid and offer quotes respectively, $I\in\{1,\dots,19\}$ indexes the currency pairs, j indexes the quotes and n^I corresponds to the number of quotes for the currency pair I per day. See Table 8 for a mapping of these indices to

[2] This platform is an ECN i.e. an Electronic Communication Network.

actual currency pair names. These mid-prices are then sampled on a regular time grid using the last-tick methodology, where the most recent quotes in each currency pair are used to calculate the mid-price for that time interval. The regularly sampled mid-prices are then:

$$p^{(I)}_{t_{(1)}}, \dots, p^{(I)}_{t_{(n)}}, \quad t_{(i+1)} - t_{(i)} = 5 \text{ min}, \quad i \in \{1, \dots, n\}, \tag{3.2}$$

where $t_{(i)}$, $i \in \{1, \dots, n\}$ are the time stamps on a regularly sampled grid and n is the number of 5 min intervals per day. If we denote the time stamp of each quote as τ_j, $j \in \{1, \dots, n^I\}$, then the quote used for each 5 min interval can be described as,

$$p^{(I)}_{t_{(i)}} = p^{(I)}_{\max\{1 \le j \le n^I \,|\, \tau_j \le t_{(i)}\}}. \tag{3.3}$$

We note that the choice of 5 min as a time interval is common in high frequency financial correlation analyses. We obtain the (log) mid-price returns via:

$$r^{(I)}_{(i)} = \log \frac{p^{(I)}_{t_{(i+1)}}}{p^{(I)}_{t_{(i)}}}, \quad i \in \{1, \dots, n-1\}, \tag{3.4}$$

where the log function as used here and throughout the article is the natural logarithm. Note that the first time interval of each day, for all currency pairs, begins at 00:00:00.000 UTC/GMT and ends at 00:04:59:59.999 UTC/GMT. The last time interval begins at 23:55:00.000 UTC/GMT and ends at 23:59:59.999 UTC/GMT. The advantage of determining calendar date based on UTC/GMT is that the major FX trading sessions are all captured on the same calendar date, namely Asia, then Europe, then North America. There are $n = 288$ five minute intervals per day. The time intervals are not only regular, but also aligned across all the currency pairs. Indeed, it is important to note that this procedure downsamples the rapidly updating and irregularly spaced tick data to a regular grid on a much larger timescale.

See Table 1 for the statistics on the number of quotes per time interval for each currency pair, aggregated across all days. See Table 2 for the descriptive statistics of the regularly sampled (log) returns per currency pair, again aggregated across all days in the data set. It is readily apparent from the descriptive statistics in Table 2 that the means of the (log) return distributions are very close to zero and that the standard deviations vary between currency pairs. In addition, the returns have high kurtosis consistent with the expected behaviour of price movements of financial instruments with a calendar time clock. The only currency pair that has a markedly asymmetric distribution is USD/TRY as evidenced by a large negative skewness (i.e. a left-skewed distribution). The large volatility, kurtosis and negative skewness of the USD/TRY distribution can be related to the sharp depreciation of the Turkish Lira during the Turkish currency and debt crisis which occurred during the period of analysis.

A few further comments are in order. The data we analyse are mid-quote data, to be contrasted with transaction price data. Bid and offer quote data represent the prices at which market participants are willing to trade at any given moment (and mid-quotes are the average of these prices), whereas transaction prices are those prices where trades have actually occurred. Quote data typically present a more continuous view of price formation whereas transaction data represent the most definitive, but less frequent measurements of the price. The FX market is a highly fragmented, over-the-counter (OTC) market. Trading on any given currency pair is distributed across a number of primary and secondary venues along with bilateral transactions between trading counterparties. For the majority of such venues, market participants do not have access to detailed transaction data other than their own trades and thus quote data is usually the main source of price formation information. Our analyses rely exclusively on quote data for price information.

We take the view that our results may well generalise beyond quote data however, based on the thesis that "transaction prices and mid-quote prices are both noisy measures of the latent "efficient prices", polluted by market microstructure effects." As per [27]. In the context of realised volatility, [27] go further and propose that a good estimator is one that yields approximately the same estimates with transaction data and mid-quote data. While [27] present this view with reference to realised volatility, in follow on work [28], empirical evidence is gathered that correlation estimates based on mid-quotes and transaction prices are similar for many correlation estimators (for equities data at least). This includes the realised covariance measure estimated on a regular time grid with the last-tick sampling scheme (which we utilise in this article). While these results are obtained for equities data, it is plausible that they generalise to forex data.

We make a final note that the data we analyse is obtained from a single venue i.e. the Integral OCX platform. This should also not preclude generalisability given that quotes are not expected to drastically differ between venues due to arbitrage.

3.2. Correlation matrix methodology

In statistics, various measures of association between two random variables have been defined. In our context, we apply certain measures of correlation to ascertain the degree to which currency (log) returns are concordant or discordant. Intuitively, this should capture an important aspect of the relationship of one currency pair with another.

Calculating correlations on high frequency financial data is complicated by two main effects. The first is the fact that observations occur irregularly in time and moreover, asynchronously across instruments. The second is the presence of microstructure noise due to various factors such as bid–ask bounce (relevant mainly for transaction based data), minimum tick intervals, latency effects etc. There are several covariance/correlation estimators that have been proposed in the literature to address these issues (i.e. asynchronous updates and market microstructure noise). Given that these estimators are not the focus of this article, and that the literature is extensive, we point the reader to the following (non-exhaustive) list of Refs. [28–34] for more detail on these two complicating

Table 1
Descriptive statistics of number of quote updates per 5 min time interval.

Currency pair	Mean	Std Dev.	Q1	Med.	Q3
AUD/JPY	419	374.5	187	319	531
AUD/NZD	304	279.3	135	231	383
AUD/USD	404	396.9	165	296	511
CAD/JPY	272	257.0	113	197	348
CHF/JPY	287	271.7	118	212	368
EUR/CHF	269	295.5	90	173	338
EUR/GBP	308	316.7	103	206	407
EUR/JPY	506	435.8	202	387	685
EUR/PLN	277	487.4	47	125	296
EUR/USD	512	506.4	181	376	683
GBP/JPY	553	471.5	240	433	724
GBP/USD	492	472.6	174	361	668
NZD/USD	269	269.3	114	199	333
USD/CAD	373	371.7	152	271	470
USD/CHF	237	255.4	82	161	305
USD/JPY	322	319.9	135	234	399
USD/MXN	447	434.1	143	324	610
USD/TRY	205	580.4	9	36	138
USD/ZAR	445	498.5	139	325	588

Table 2
Summary statistics of regular 5 min (log) mid-price returns.

Currency pair	Mean (x 10^{-4})	Std Dev. (x 10^{-4})	Skewness	Kurtosis
AUD/JPY	0.0	3.8	0.0	24.4
AUD/NZD	0.0	2.1	0.2	36.6
AUD/USD	0.0	3.8	0.0	15.3
CAD/JPY	0.0	3.1	0.1	16.2
CHF/JPY	0.0	2.4	0.0	12.5
EUR/CHF	0.0	1.8	−0.6	80.5
EUR/GBP	0.0	2.6	−0.2	37.5
EUR/JPY	0.0	2.4	0.1	16.8
EUR/PLN	0.0	2.7	0.0	31.1
EUR/USD	0.0	2.4	−0.2	32.9
GBP/JPY	0.0	3.1	0.0	20.0
GBP/USD	0.0	3.1	0.0	14.9
NZD/USD	0.0	3.8	0.1	19.8
USD/CAD	0.0	2.7	0.0	18.8
USD/CHF	0.0	2.5	−0.2	19.8
USD/JPY	0.0	2.1	0.2	17.3
USD/MXN	0.0	5.5	0.0	18.8
USD/TRY	0.1	11.7	−7.3	795.7
USD/ZAR	0.0	6.4	−0.1	17.2

issues and various approaches to jointly handle them. In this article we utilise the correlation estimator (3.5) on (log) mid-price returns. This estimator is referred to as the realised correlation estimator in the finance literature and is defined as,

$$\hat{\rho}^{IJ} = \frac{\sum_{i=1}^{n-1} r_{(i)}^{(I)} r_{(i)}^{(J)}}{\sqrt{\sum_{i=1}^{n-1} (r_{(i)}^{(I)})^2 \sum_{i=1}^{n-1} (r_{(i)}^{(J)})^2}}, \quad I, J \in \{1, \dots, 19\}, \tag{3.5}$$

where I, J are currency pair indices. This estimator captures the normalised, aggregated co-movement (i.e. covariance) of two series of returns over a given time period (one day in our case). The resultant $\hat{\rho}^{IJ}$ correlation matrix is a symmetric, real matrix with $19(19-1)/2 = 171$ independent entries. This figure accounts for the fact that the diagonal elements are fixed and equal and do not contribute to the degrees of freedom (we subtract the identity to get a correlation matrix with vanishing diagonal elements). As mentioned previously, we are concerned with the ensemble statistics of such matrices. There are several ways to construct such an ensemble. We choose to calculate the correlation matrix for each trading day (aligned with UTC/GMT boundaries), and study the sampling distribution of the matrices. In particular, we study, $\hat{\rho}_A^{IJ}$, where $A \in \{1, \dots, N_D\}$ indexes trading dates. In our data, there are 446 unique trading days, i.e. $N_D = 446$. We plot examples of the evolution of two elements of these correlation matrices over time in Fig. 1.

It is well established that the realised correlation estimator is, in general, sensitive to the issues described above. However, it is widely acknowledged in the literature that the impact of these issues can be mitigated by sampling regularly at a lower frequency i.e. 5 - 15 min intervals. We utilise 5 min time intervals in particular, as is common in analysing high frequency financial data. We have also verified empirically that the correlation results are not very sensitive to the choice of the time interval length — see Table 3 where we demonstrate that the correlation values are fairly similar (in aggregate) at different timescales (1, 5, 10 and

Table 3
Summary statistics for the values of correlation obtained from the realised correlation estimator with different sampling intervals. Here the mean (Mean) is the average correlation across all currency pair combinations. The standard deviation (Std Dev.) is the average standard deviation of the correlation elements across all currency pair combinations. Finally cor(., 5 mins) gives the average correlation of the individual correlation elements with those obtained for a 5 min sampling period across all currency pair combinations.

Time scale (mins)	Mean	Std Dev.	Cor(., 5 mins)
1	0.05	0.10	0.77
5	0.04	0.12	1.00
10	0.04	0.15	0.86
15	0.04	0.16	0.77

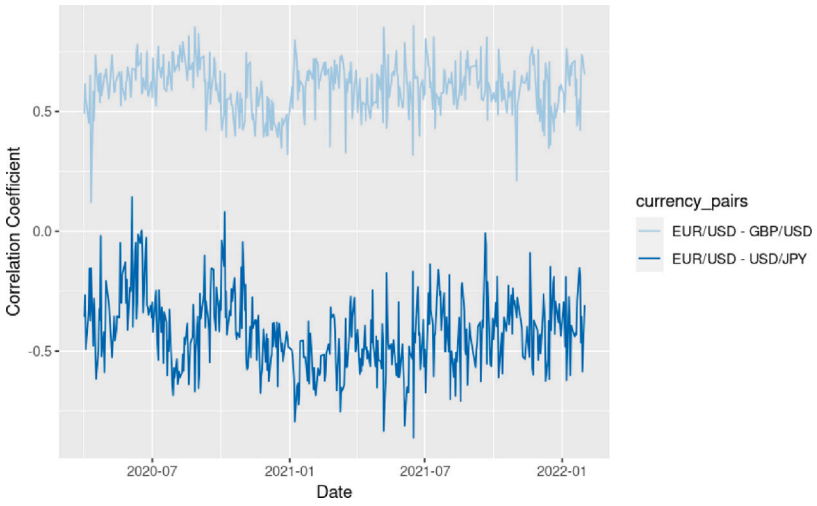


Fig. 1. Examples of realised daily correlation estimates over time.

15 min) and moreover, vary in a concordant manner. We have also confirmed that the model fit results are similar for 1, 5, 10 and 15 min suggesting some robustness to this choice (see Table 4). Finally, we have validated that our anomaly detection results presented in Section 5.1 are qualitatively similar for all timescales beyond a certain value (the effectiveness of the anomaly detection procedure for 5, 10 and 15 min is similar, with 1 min results being slightly worse), see section 5 of the supplementary material.

We do acknowledge however that the procedure we have applied is not likely to be the most efficient nor the most robust to microstructure noise (see the aforementioned references for approaches that could be more efficient and robust). However, the simplicity and ubiquity of the realised correlation estimator is appealing and it allows us to make contact with asymptotic Gaussian sampling properties as discussed in the introduction. The main focus of the present article is to explore the phenomenological modelling of ensembles of correlation matrices and not particular correlation estimators. Indeed, the PIGM model can be just as readily applied to correlation matrices arising from other correlation estimators, some of which naturally handle the asynchronous nature of the data, and do not require a choice of timescale e.g. [29,35]. Our results may well generalise to a broader class of correlation estimators. Investigating the degree to which the results generalise to other correlation estimators is an area for future research.

4. Matrix theory and matrix data: near-Gaussianity

In this section we apply the Gaussian model of Section 2 to the ensemble of correlation matrices defined in Section 3. We show that the vast majority of cubic and quartic observables closely match the predictions of the Gaussian theory.

We form a compact representation of the original correlation matrix data by defining a vector of observables for each correlation matrix. The observable vectors are shown to perform well in a variety of anomaly detection and similarity ranking tasks in Section 5. Optimal performance is achieved in these tasks by constructing observable vectors from the least Gaussian observables.

4.1. Theory/experiment deviations normalised by standard deviations of the observables

We begin by elucidating some empirical statistical properties of the observables — the permutation invariant polynomials of the correlation matrix elements. The empirical distributions of these observables typically exhibit a right/positive skew along with heavier tails than the normal distribution. It is also noteworthy that all estimated Pearson product-moment correlations between

the observables are positive and that most correlations are very strongly positive (see the supplementary material for a plot of these correlations). The strength of the correlations is particularly relevant in our choice of statistical distance measure in the anomaly detection analysis presented in Section 5.1. Indeed, this motivated utilising the Mahalanobis distance which typically performs well even in the presence of such correlations. In addition, the fact that the empirical distributions of the observables are mostly unimodal, suggests this distance measure is appropriate. Equipped with the Gaussian model and its solution, given in Appendix B, and the financial data described in Section 3 we now perform a variety of tests to assess how well this model describes the statistics of the data. Firstly, we calculate the normalised absolute error

$$\Delta_\alpha = \frac{|\langle \mathcal{O}_\alpha \rangle_{\mathrm{T}} - \langle \mathcal{O}_\alpha \rangle_{\mathrm{E}}|}{\sigma_{\mathrm{E},\alpha}}, \tag{4.1}$$

between the experimental cubic and quartic observable average values

$$\langle \mathcal{O}_\alpha \rangle_{\mathrm{E}} = \frac{1}{n_A} \sum_{A=1}^{N_D} \mathcal{O}_\alpha(\rho_A^{ij}), \tag{4.2}$$

and the Gaussian model's prediction of those expectation values $\langle \mathcal{O}_\alpha \rangle_{\mathrm{T}}$. In both equations α labels the observable and in (4.1) we have normalised by the standard deviation of the experimental observable values.

As we argued in Section 2 we expect these normalised errors to be small where the underlying data is approximately Gaussian. The normalised absolute errors for each observable with a sampling timescale of 1, 5, 10, and 15 min are listed in the third, fourth, fifth, and sixth columns of Table 4 respectively. In general these normalised absolute errors are in very good agreement: at each timescale 4 or fewer observables differ from the theoretical prediction by more than 1 standard deviation, and the average normalised absolute error of the cubic and quartic observables is at most 0.43 standard deviations. This is evidence for Gaussianity in the permutation invariant sector of the FX-rate correlation matrix data. The fact that the absolute errors are small across the range of sampling timescales considered, from 1 min up to 15 min, suggests that this is a robust feature of the data and not merely an artefact of the choice of a particular sampling timescale. We have chosen to use the 5 min data for the ratio of the experimental and theoretical standard deviation as well as the subsequent statistical analyses.

The model defined in Appendix B can be used to predict the standard deviation of cubic and quartic observables. We call this the theoretical standard deviation, and define it for each observable α as

$$\sigma_{\mathrm{T},\alpha} \equiv \sqrt{|\langle (\mathcal{O}_\alpha)^2 \rangle_{\mathrm{T}} - \langle \mathcal{O}_\alpha \rangle_{\mathrm{T}}^2|}\,. \tag{4.3}$$

In itself this is an interesting quantity to compare to the experimental observable standard deviations $\sigma_{\mathrm{E},\alpha}$. The ratio of the two standard deviations is shown for each observable in the seventh column of Table 4 for data sampled at 5 min intervals. The values of $\sigma_{\mathrm{T},3}$ and $\sigma_{\mathrm{T},22}$ provided by the model are much smaller than the observed values. This is consistent with the absolute errors in Table 4, where we see the expectation values of these observables deviating the most from the model. It is these large deviations from Gaussianity that lend these observables their power in the construction of lower-dimensional representations of the correlation matrices (see Section 5.1.3).

We briefly note an alternative approach to estimating the theoretical standard deviation, also employed in [4] to give good theoretical predictions of the experimental standard deviations of observables. This estimate is obtained by calculating the absolute difference between $\langle \mathcal{O}_\alpha \rangle_{\mathrm{T}}$ and $\langle \mathcal{O}_\alpha \rangle_{\mathrm{T}'}$, where $\langle \mathcal{O}_\alpha \rangle_{\mathrm{T}'}$ is the expectation value evaluated with the quadratic couplings that parametrise the model shifted by one standard deviation. Taking the average of this difference over all 8 possible permutations of sign for the shifts of the 3 parameters gives us our estimate of the standard deviation. This method was used to estimate the standard deviations of $\mathcal{O}_{12}$ and $\mathcal{O}_{19}$ due to the prohibitive computational demands of calculating the octic expectation values $\langle \mathcal{O}_{12}^2 \rangle$ and $\langle \mathcal{O}_{19}^2 \rangle$.

4.2. Absolute errors relative to standard deviations and standard errors of observables

Thus far, we have analysed the differences between theoretical observable mean values and experimental observable mean values, normalised by the experimental standard deviation of the observable values. We have found that 27 out of 31 observables deviate less than 1 experimental standard deviation.[3] These normalised differences have a physical interpretation. Small normalised differences in this case are suggestive of small coupling constants for higher order corrections in the action (see Section 2) which is evidence for near-Gaussianity of the experimental data generating process. Furthermore, we have found that "physical" tests of the theoretical model versus experiment such as calculating the proportion of days captured and balanced accuracy of a typicality classifier provide additional evidence for the pure Gaussian model being a good approximation (see section 3 of the supplementary material).

Another possible choice for normalising the differences between the theoretical observable mean values and the experimental observable mean values is the experimental standard error. The standard error in this case is the standard deviation of the experimental observable mean values. We denote the standard error, $\sigma_{\bar{x}}$. Given the definition of the mean estimator, i.e. $\bar{x} = 1/n \sum_{i=1}^{n} x_i$, the standard error is equal to the standard deviation of the permutation invariant polynomial values for each matrix, σ_{E}, divided by $\sqrt{n}$ i.e. $\sigma_{\bar{x}} = \sigma_{\mathrm{E}}/\sqrt{n}$. The standard error is useful in determining whether the differences between the theoretical observable mean values and experimental observable mean values are plausibly due to sampling variation. The larger the sample size, the smaller the departures between theory and experiment that can be distinguished from sampling variation i.e. "experimental

[3] The four observables that differ by more than one standard deviation are $\mathcal{O}_3, \mathcal{O}_{17}, \mathcal{O}_{19}$ and $\mathcal{O}_{22}$.

Table 4
For each observable in the first two columns the third, fourth, fifth and sixth columns list the absolute difference between the experimental value and theoretical prediction normalised by the experimental standard deviation for data sampled at 1, 5, 10 and 15 min intervals respectively. The seventh column lists the ratio of the experimental and theoretical standard deviations for data sampled at 5 min intervals.

Label	Observable	$\Delta_\alpha^{(1)}$	$\Delta_\alpha^{(5)}$	$\Delta_\alpha^{(10)}$	$\Delta_\alpha^{(15)}$	σ_E/σ_T
$\mathcal{O}_1$	$\sum_{i,j}\hat{\rho}_{ij}^3$	0.09	0.02	0.07	0.10	0.90
$\mathcal{O}_2$	$\sum_{i,j,k}\hat{\rho}_{ij}^2\hat{\rho}_{jk}$	0.43	0.33	0.28	0.23	1.49
$\mathcal{O}_3$	$\sum_{i,j,k}\hat{\rho}_{ij}\hat{\rho}_{jk}\hat{\rho}_{ki}$	2.02	2.04	1.98	1.83	9.39
$\mathcal{O}_4$	$\sum_{i,j,k,l}\hat{\rho}_{ij}^2\hat{\rho}_{kl}$	0.05	0.01	0.00	0.01	1.06
$\mathcal{O}_5$	$\sum_{i,j,k,l}\hat{\rho}_{ij}\hat{\rho}_{jk}\hat{\rho}_{kl}$	0.96	0.97	0.96	0.90	3.36
$\mathcal{O}_6$	$\sum_{i,j,k,l}\hat{\rho}_{ij}\hat{\rho}_{ik}\hat{\rho}_{il}$	0.36	0.33	0.33	0.28	0.73
$\mathcal{O}_7$	$\sum_{i,j,k,l,m}\hat{\rho}_{ij}\hat{\rho}_{jk}\hat{\rho}_{lm}$	0.11	0.12	0.12	0.13	1.05
$\mathcal{O}_8$	$\sum_{i,j,k,l,m,n}\hat{\rho}_{ij}\hat{\rho}_{kl}\hat{\rho}_{mn}$	0.01	0.01	0.02	0.02	0.65
$\mathcal{O}_9$	$\sum_{i,j}\hat{\rho}_{ij}^4$	0.16	0.54	0.67	0.69	1.10
$\mathcal{O}_{10}$	$\sum_{i,j,k}\hat{\rho}_{ij}^2\hat{\rho}_{jk}^2$	0.49	0.42	0.38	0.33	2.42
$\mathcal{O}_{11}$	$\sum_{i,j,k}\hat{\rho}_{ij}\hat{\rho}_{jk}^3$	0.09	0.05	0.12	0.16	1.30
$\mathcal{O}_{12}$	$\sum_{i,j,k}\hat{\rho}_{ij}\hat{\rho}_{ik}\hat{\rho}_{jk}^2$	0.90	0.88	0.85	0.79	3.27[a]
$\mathcal{O}_{13}$	$\sum_{i,j,k,l}\hat{\rho}_{ij}\hat{\rho}_{kj}\hat{\rho}_{ij}^2$	0.23	0.19	0.17	0.15	1.67
$\mathcal{O}_{14}$	$\sum_{i,j,k,l}\hat{\rho}_{ij}\hat{\rho}_{kl}^3$	0.12	0.04	0.00	0.03	0.64
$\mathcal{O}_{15}$	$\sum_{i,j,k,l}\hat{\rho}_{ij}\hat{\rho}_{jk}^2\hat{\rho}_{kl}$	0.25	0.21	0.17	0.14	1.07
$\mathcal{O}_{16}$	$\sum_{i,j,k,l}\hat{\rho}_{ij}\hat{\rho}_{jk}\hat{\rho}_{kl}^2$	0.53	0.49	0.45	0.41	2.45
$\mathcal{O}_{17}$	$\sum_{i,j,k,l}\hat{\rho}_{ij}\hat{\rho}_{jk}\hat{\rho}_{ik}\hat{\rho}_{kl}$	1.01	1.00	0.96	0.88	7.44
$\mathcal{O}_{18}$	$\sum_{i,j,k,l}\hat{\rho}_{ij}^2\hat{\rho}_{kl}^2$	0.08	0.07	0.08	0.08	2.06
$\mathcal{O}_{19}$	$\sum_{i,j,k,l}\hat{\rho}_{ij}\hat{\rho}_{jk}\hat{\rho}_{kl}\hat{\rho}_{li}$	1.21	1.24	1.20	1.08	8.57[a]
$\mathcal{O}_{20}$	$\sum_{i,j,k,l,m}\hat{\rho}_{ik}\hat{\rho}_{jk}\hat{\rho}_{lk}\hat{\rho}_{mk}$	0.37	0.36	0.36	0.29	0.79
$\mathcal{O}_{21}$	$\sum_{i,j,k,l,m}\hat{\rho}_{ij}\hat{\rho}_{jk}\hat{\rho}_{lk}\hat{\rho}_{mk}$	0.40	0.39	0.37	0.33	1.67
$\mathcal{O}_{22}$	$\sum_{i,j,k,l,m}\hat{\rho}_{ij}\hat{\rho}_{kl}\hat{\rho}_{lm}\hat{\rho}_{mk}$	1.25	1.31	1.27	1.15	10.1
$\mathcal{O}_{23}$	$\sum_{i,j,k,l,m}\hat{\rho}_{ij}^2\hat{\rho}_{kl}\hat{\rho}_{lm}$	0.07	0.05	0.05	0.05	1.63
$\mathcal{O}_{24}$	$\sum_{i,j,k,l,m}\hat{\rho}_{ij}\hat{\rho}_{kl}\hat{\rho}_{lm}^2$	0.31	0.24	0.20	0.16	0.81
$\mathcal{O}_{25}$	$\sum_{i,j,k,l,m}\hat{\rho}_{ij}\hat{\rho}_{jk}\hat{\rho}_{kl}\hat{\rho}_{lm}$	0.79	0.79	0.78	0.71	5.98
$\mathcal{O}_{26}$	$\sum_{i,j,k,l,m,n}\hat{\rho}_{ij}\hat{\rho}_{kl}\hat{\rho}_{km}\hat{\rho}_{kn}$	0.14	0.12	0.12	0.08	0.54
$\mathcal{O}_{27}$	$\sum_{i,j,k,l,m,n}\hat{\rho}_{ij}\hat{\rho}_{jk}\hat{\rho}_{lm}\hat{\rho}_{ln}$	0.10	0.10	0.11	0.12	1.47
$\mathcal{O}_{28}$	$\sum_{i,j,k,l,m,n}\hat{\rho}_{ij}^2\hat{\rho}_{kl}\hat{\rho}_{mn}$	0.06	0.02	0.00	0.00	0.58
$\mathcal{O}_{29}$	$\sum_{i,j,k,l,m,n}\hat{\rho}_{ij}\hat{\rho}_{kl}\hat{\rho}_{lm}\hat{\rho}_{mn}$	0.65	0.65	0.64	0.57	2.10
$\mathcal{O}_{30}$	$\sum_{i,j,k,l,m,n,o}\hat{\rho}_{ij}\hat{\rho}_{jk}\hat{\rho}_{lm}\hat{\rho}_{no}$	0.13	0.13	0.13	0.13	0.67
$\mathcal{O}_{31}$	$\sum_{i,j,k,l,m,n,o,p}\hat{\rho}_{ij}\hat{\rho}_{kl}\hat{\rho}_{mn}\hat{\rho}_{op}$	0.02	0.03	0.03	0.04	0.38
	Average	0.43	0.42	0.41	0.38	2.50

[a] Values were obtained using an estimate of σ_{T} described at the end of Section 4.1.

error''. In particular, genuine departures are sizeable in comparison to standard errors e.g. bigger than 3 standard errors. Such departures can be interpreted as highly statistically significant differences.

In our data set, which has a fairly large sample size of 446, we have observed that 13 out of 31 observables have a difference between the theoretical observable mean and experimental observable mean value of fewer than 3 standard errors (i.e. 18 out of 31 observables exhibit a departure of more than 3 standard errors). To further explore the statistical significance of differences in the theoretical versus experimental observable mean values we have also calculated the percentile bootstrap confidence intervals of the experimental observable mean values. The bootstrap procedure involved re-sampling from the original set of observable values, uniformly with replacement, to construct 1000 bootstrap samples of the same size as the original sample (i.e. 446). The mean of each such bootstrap sample was then calculated. Given the asymptotic normality of the mean estimator, it is expected that 99.7th percentile bootstrap confidence intervals will approximately correspond to 3 standard errors on either side of the original experimental mean estimate. This is reflected in our results, which reveal 12 of 31 observables with theoretical observable mean values within this confidence interval and 19 of 31 observables with theoretical mean values outside the interval. This is in close agreement with the aforementioned basic standard error results where we had 13 of 31 observables with theoretical observable mean values lying within 3 standard errors of the experimental observable mean values.

It is also worth noting that statistically significant differences may nonetheless be small in terms of relative error, which we recall is defined as

$$\frac{|\langle\mathcal{O}_\alpha\rangle_{\mathrm{T}} - \langle\mathcal{O}_\alpha\rangle_{\mathrm{E}}|}{\langle\mathcal{O}_\alpha\rangle_{\mathrm{E}}}\,. \tag{4.4}$$

Indeed we have observed that 20 out of 31 observables have a relative error of less than 30% when comparing theoretical versus experimental mean observable values.

The key point from this section is that when we consider the absolute error of observables in comparison to the standard deviation, a measure motivated by consideration of perturbative corrections to the toy Gaussian model, we have 27 of 31 observables which are within 1 standard deviation (all are within 3 standard deviations). On the other hand 6 of 31 are within 1 standard error (13 are within 3 standard errors). This suggests that developing computations of expectation values in theoretical models which

contain small cubic and quartic terms, as guided by the data, is likely to give statistically significant improvement, given our current sample sizes, of the agreement between theoretical and experimental expectation values of observables. This is technically more intricate than computing in the Gaussian model and is left for future investigation.

5. Applications of matrix theory to matrix data: anomaly detection and similarity measures

In establishing approximate Gaussianity in the matrix data of correlation matrices, observable vectors, formed using lists of permutation invariant polynomial functions labelled by graphs, provide the key bridge between permutation invariant Gaussian matrix theory and the matrix data. The observable vectors associated with the correlation matrices can be regarded as particular, lower-dimensional representations of the correlation matrices. The observable vectors themselves are random vectors, for which the statistics entailed by the PIGM model are a good approximation in general. A natural question to ask is whether the observable vectors provide a more compact representation of correlation matrices which accentuates statistical "signal" in the data as opposed to noise. Such a representation would be closely linked to an accurate characterisation of the market state and applications could include classification/regression models, clustering analysis, anomaly/outlier detection etc. In this section, we will consider two tasks: anomaly detection and similarity measurement, which we describe in more detail in Sections 5.1 and 5.2 respectively.

5.1. Anomaly detection and economically significant dates

In this section we will demonstrate that the observable vectors do indeed constitute a promising representation for anomaly/outlier detection. The task of anomaly/outlier detection pertains to identifying observations that differ significantly from the majority of the data set. In our context, we seek to identify unusual and noteworthy observable vectors, each of which is associated with the correlation matrix of a particular date. To verify that a meaningful result has been obtained, we need a notion of unusual and noteworthy dates in the forex market as a reference. The natural approach we take is to consider special dates in the forex trading calendar corresponding to the highest impact economic news announcements. These announcements often lead to a flurry of trading activity along with associated price movements, market volatility and changes in the relationships between currency pairs. Note that the main purpose of our analysis is to establish whether the observable space has physically/economically meaningful structure (thus constituting a good representation) rather than to introduce a state-of-the-art anomaly detection algorithm which would require significant tuning and comparison to several existing alternatives. As such, we combine the observable vectors with a straightforward, parameter-free, distance-based approach for anomaly detection. We only compare the observable vector representation to $D(D-1)/2$-dimensional vectors formed by stacking the daily correlation matrix elements, which we refer to as the raw correlation representation.

5.1.1. Anomaly detection algorithm

A common approach to detecting anomalous/outlier observations, is to utilise a statistical distance measure to determine the distance of each random vector from the mean vector. One can equivalently think of this as determining the length of the random vectors as measured from the origin for centred data. We will utilise the Mahalanobis distance measure to assess these distances (see [36] for example). These distances directly correspond to anomaly scores, allowing the vectors to be ordered by anomalousness. The Mahalanobis distance better treats the case of highly correlated elements in the random vectors, as is the case for the observable vectors (as noted in Section 4.1). Concretely, given a multivariate probability distribution F on $\mathbb{R}^N$ (i.e. generating random vectors $\vec{y} \in \mathbb{R}^N$), with mean vector $\vec{\mu}$ and covariance matrix Σ, the Mahalanobis distance of a point $\vec{x} \in \mathbb{R}^N$ from the mean $\vec{\mu}$ is defined as,

$$d(\vec{x}, \vec{\mu}) = \sqrt{(\vec{x} - \vec{\mu})\Sigma^{-1}(\vec{x} - \vec{\mu})}, \tag{5.1}$$

where $\mu_i = E(y_i)$ and $\Sigma_{ij} = E\left[(y_i - \mu_i)(y_j - \mu_j)\right]$. In the context of anomaly detection, the distribution F may be regarded as the distribution of the non-anomalous points.

One potential issue in practice is that the covariance matrix estimate used in the calculation above could be ill-conditioned, possibly leading to numerical issues. Another potential issue is that all points being evaluated contribute to the mean and covariance estimates in the Mahalanobis distance calculation. In particular, both ordinary points and those points that are, in fact, anomalous, affect the mean and covariance estimates. We have attempted to improve the situation by excluding each point in turn from the estimate of the mean and covariance before evaluating the distance of this point from the mean. We have found that this simple and easy-to-interpret adjustment gives better anomaly detection results. While more sophisticated approaches exist in the literature based on the robust estimation of the mean and covariance matrix, these are more complicated and require further choices to be made. It is also noteworthy that we have found that the procedure of excluding each point in turn from the mean and covariance estimates yields essentially the same result as using the average distance to all other points as an anomaly score (i.e. the average k-nearest neighbours approach with k set to the number of other points in the data set). The average k-NN approach using all other points in the data set is known to be effective in detecting anomalous points that are on the boundaries of the data set [36], whilst not being able to detect isolated points in the interior of the data set, which we regard as a more subtle type of anomaly for future study.

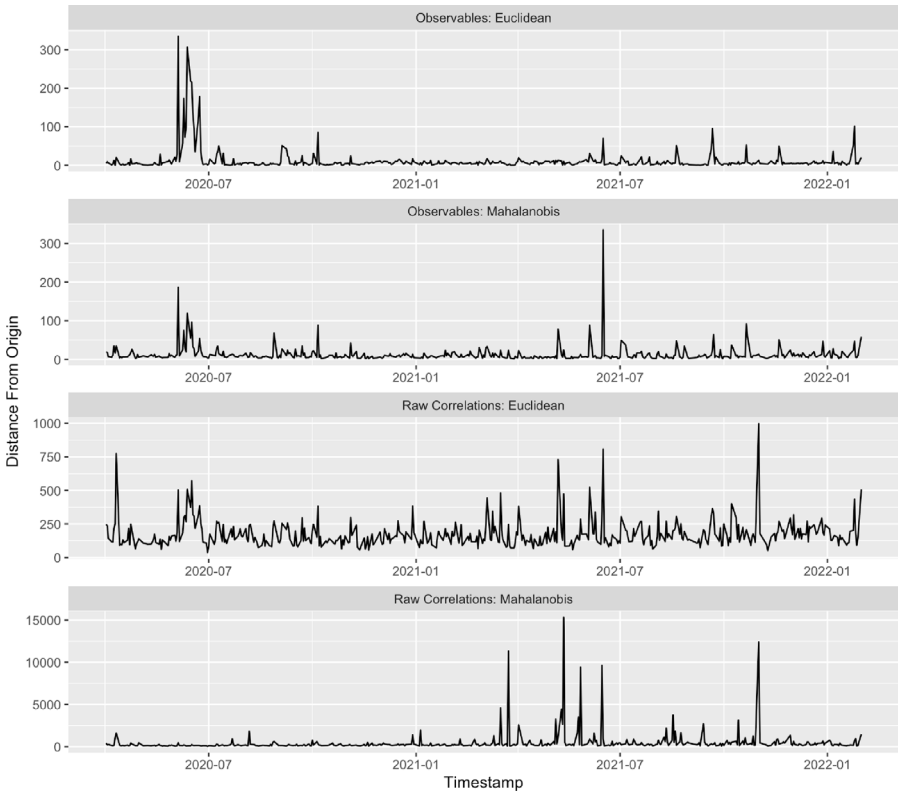


Fig. 2. Distances of observable vectors and raw correlation vectors from the origin using the standardised Euclidean and Mahalanobis metrics.

5.1.2. Economically significant dates

It is well known amongst forex trading practitioners that there are certain currencies and certain types of economic announcements that typically have the highest impact on the forex markets (see [37] for example). These currencies are associated with the countries or blocs with the largest economies, namely the US Dollar, Chinese Renminbi, Japanese Yen, European Union Euro and Great British Pound. Four of the most important classes of economic announcements are described below.

- Central bank meetings and announcements relating to interest rate decisions etc. These include the FOMC (Federal Open Market Committee), ECB (European Central Bank), BoE (Bank of England), PBoC (People's Bank of China) and BoJ (Bank of Japan) meetings associated with the United States, European Union, Great Britain, China and Japan respectively.
- Unemployment data releases. One of the most important examples of this is the US Non-Farm Payrolls release.
- Consumer price index releases. The most important release in this category is the US consumer price index release.
- Unplanned forex news including special central bank meetings and speeches, political speeches etc.

In our subsequent analyses we utilise the economic calendar sourced from [38] of high impact events and filter for only those events pertaining to the aforementioned currencies (and associated economies) and economic announcements specifically. The exact strings used for filtering the events based on event name are: "ECB Press Conference", "BoE MPC", "FOMC Press Conference", "BoJ Press Conference", "PBoC Interest Rate Decision", "Nonfarm Payrolls", "Consumer Price Index ex Food & Energy", "European Council Meeting", "EU Leaders Special Summit" and "ECB Special Strategy Meeting". During the period 2020-04-01 to 2022-01-31 one or more of these high impact events occurred on approximately 27% of business days.

5.1.3. Dimensionality reduction

The construction of lower-dimensional representations of the correlation matrices – namely the observable vectors – is effectively a dimensionality reduction procedure. As is common with such procedures (e.g. Principal Component Analysis, PCA), there is a trade-off between reducing dimensionality and preserving information content. Balancing these trade-offs through a good choice of the number of components often leads to better results in various applications. In PCA, the cumulative variance of the first principal components is typically used as an organising quantity to select how many such components to include. In our case, we take the normalised magnitude of the difference between the empirical expectation values of the cubic and quartic observables and the theoretical predictions of the PIGM model (2.28), as an organising quantity for determining which observables to retain. The thesis is that the empirical higher order observables that depart from theoretical expectations indicate additional information

Table 5
The 12 cubic and quartic observables that have the largest normalised difference from the PIGM predictions.

Observable label	Observable Def.	Observable order
$\mathcal{O}_3$	$\sum_{i,j,k} \hat{\rho}_{ij}\hat{\rho}_{jk}\hat{\rho}_{ki}$	Cubic
$\mathcal{O}_5$	$\sum_{i,j,k,l} \hat{\rho}_{ij}\hat{\rho}_{jk}\hat{\rho}_{kl}$	Cubic
$\mathcal{O}_9$	$\sum_{i,j} \hat{\rho}^4_{ij}$	Quartic
$\mathcal{O}_{10}$	$\sum_{i,j,k} \hat{\rho}^2_{ij}\hat{\rho}^2_{jk}$	Quartic
$\mathcal{O}_{12}$	$\sum_{i,j,k} \hat{\rho}_{ij}\hat{\rho}_{ik}\hat{\rho}^2_{jk}$	Quartic
$\mathcal{O}_{16}$	$\sum_{i,j,k,l} \hat{\rho}_{ij}\hat{\rho}_{jk}\hat{\rho}^2_{kl}$	Quartic
$\mathcal{O}_{17}$	$\sum_{i,j,k,l} \hat{\rho}_{ij}\hat{\rho}_{jk}\hat{\rho}_{ik}\hat{\rho}_{kl}$	Quartic
$\mathcal{O}_{19}$	$\sum_{i,j,k,l} \hat{\rho}_{ij}\hat{\rho}_{jk}\hat{\rho}_{kl}\hat{\rho}_{li}$	Quartic
$\mathcal{O}_{21}$	$\sum_{i,j,k,l,m} \hat{\rho}_{il}\hat{\rho}_{jk}\hat{\rho}_{lk}\hat{\rho}_{mk}$	Quartic
$\mathcal{O}_{22}$	$\sum_{i,j,k,l,m} \hat{\rho}_{ij}\hat{\rho}_{kl}\hat{\rho}_{lm}\hat{\rho}_{mk}$	Quartic
$\mathcal{O}_{25}$	$\sum_{i,j,k,l,m} \hat{\rho}_{ij}\hat{\rho}_{jk}\hat{\rho}_{kl}\hat{\rho}_{lm}$	Quartic
$\mathcal{O}_{29}$	$\sum_{i,j,k,l,m,n} \hat{\rho}_{ij}\hat{\rho}_{kl}\hat{\rho}_{lm}\hat{\rho}_{mn}$	Quartic

beyond the linear and quadratic structure encoded into the PIGM model. We have empirically determined that the 12 "least Gaussian" observables yield optimal anomaly detection results (i.e. median odds-ratio). Notably, the results broadly improve as more observables are added starting from a small number of observables, reach a peak and then decline somewhat as more observables are added. These "least Gaussian" observables are listed in Table 5. We have also observed that useful information remains in the other observables however. The cubic and quartic observables that are best predicted by the PIGM model still have reasonable effectiveness in anomaly detection for example, as do random subsets of observables and the complete set of observables. This aligns with our overarching findings that the PIGM model is a good fit overall and captures meaningful statistical structure. There does appear to be additional information captured in the least well fit observables however as supported by the results below.

5.1.4. Longest observable vectors and economically significant dates

We assess how strongly the lengths of the observable vectors constructed from the observables in Table 5 are associated with the presence or absence of economically significant events. To investigate the utility of the observable representation, we compare to the results obtained for the original correlation matrices, applying both standardised Euclidean and Mahalanobis distance measures. An interesting aside is that the Mahalanobis distance on the original correlation matrices can be regarded as a soft-threshold version of anomaly detection based on evaluating the PCA reconstruction error [36], thus constituting an interesting comparison involving PCA. The methodology is as follows.

1. Calculate the standardised Euclidean and Mahalanobis vector lengths for the observable vector associated with each date in the dataset (representing distance from the mean observable vector or equivalently the origin in this case). The maximal dimension of the vector space considered here is $N = 31$ while the optimal number of least Gaussian observables, as stated earlier, is 12.
2. Calculate the standardised Euclidean and Mahalanobis vector lengths for the correlation feature vector associated with each date in the dataset. The correlation feature vector for each date is comprised of the 171 pairwise correlations calculated between all 19 currency pairs. We term these features raw correlations.
3. Rank the dates in the dataset by Euclidean and Mahalanobis vector length, in descending order for the observable vectors and raw correlation feature vectors.
4. Assess whether the top 25, 50 and 100 dates have a statistically significantly higher number of economically significant events than the bottom 25, 50 and 100 dates ordered by distance in a descending manner. In addition, we calculate the ratio between the odds of observing a economically significant news event in the top/most anomalous 25, 50, 100 dates and the odds of such an event occurring in the bottom/most typical 25, 50, 100 dates. This odds-ratio (OR) is defined as,

$$\text{OR} = \frac{P_T/(1-P_T)}{P_B/(1-P_B)}, \tag{5.2}$$

where P_B corresponds to the proportion of the 25, 50, 100 closest dates to the origin that are associated with economically significant events. Similarly, P_T corresponds to the proportion of the 25, 50, 100 furthest dates from the origin that are associated with economically significant events.

The distances for the respective metrics and features are presented in Fig. 2. Notably, the observable feature vectors appear to have more distinct outlier days and less noise. The results of comparing the top 25, 50, 100 dates by distance from the origin with the bottom 25, 50, 100 dates respectively are collected in Table 6. The best contrast of the number of economic events appearing in the furthest days from the origin compared to the closest days from the origin respectively is given by the Mahalanobis distance evaluated on observable vectors. Indeed, for the Mahalanobis distance on observable vectors, there is a higher degree of statistical significance (i.e. smaller p-values) and higher odds-ratios than the other combinations of metric and features in almost all cases. The high odds-ratios imply that the odds of an economically significant event occurring in the anomalous groups (most distant)

Table 6
In-sample anomaly detection results for 5 min sampled data. In the table below, the proportions, P_B, P_T and the odds-ratio, OR, are as defined in Eq. (5.2). The p-value is obtained using Fisher's exact one-sided test. The * symbol following a p-value indicates significance at the 0.05 level, ** indicates significance at the 0.01 level and *** indicates significance at the 0.001 level.

Metric	Features	Subset size	P_T	P_B	p-value	OR
Euclidean	Observables	25	0.40	0.20	1.1×10^{-1}	2.67
Euclidean	Observables	50	0.38	0.24	1.0×10^{-1}	1.94
Euclidean	Observables	100	0.38	0.28	8.8×10^{-2}	1.58
Mahalanobis	Observables	25	0.44	0.16	3.1×10^{-2}*	4.12
Mahalanobis	Observables	50	0.38	0.16	1.2×10^{-2}*	3.22
Mahalanobis	Observables	100	0.39	0.13	2.1×10^{-5}***	4.28
Euclidean	Raw Correlations	25	0.52	0.32	1.3×10^{-1}	2.30
Euclidean	Raw Correlations	50	0.44	0.28	7.2×10^{-2}	2.02
Euclidean	Raw Correlations	100	0.39	0.22	6.8×10^{-3}**	2.27
Mahalanobis	Raw Correlations	25	0.24	0.16	3.6×10^{-1}	1.66
Mahalanobis	Raw Correlations	50	0.32	0.14	2.8×10^{-2}*	2.89
Mahalanobis	Raw Correlations	100	0.28	0.20	1.2×10^{-1}	1.56

Table 7
Out-of-sample anomaly detection results for 5 min sampled data. In the table below, the proportions, P_B, P_T and the odds-ratio, OR, are as defined in Eq. (5.2). The p-value is obtained using Fisher's exact one-sided test. The * symbol following a p-value indicates significance at the 0.05 level, ** indicates significance at the 0.01 level and *** indicates significance at the 0.001 level.

Metric	Features	Subset Size	P_T	P_B	p-value	OR
Euclidean	Observables	25	0.44	0.24	1.2×10^{-1}	2.49
Euclidean	Observables	50	0.44	0.24	2.8×10^{-2}*	2.49
Euclidean	Observables	100	0.29	0.23	2.1×10^{-1}	1.37
Mahalanobis	Observables	25	0.56	0.16	3.6×10^{-3}**	6.68
Mahalanobis	Observables	50	0.50	0.10	1.0×10^{-5}***	9.00
Mahalanobis	Observables	100	0.37	0.12	3.0×10^{-5}***	4.31
Euclidean	Raw Correlations	25	0.48	0.20	3.6×10^{-2}*	3.69
Euclidean	Raw Correlations	50	0.48	0.16	5.6×10^{-4}***	4.85
Euclidean	Raw Correlations	100	0.42	0.14	7.9×10^{-6}***	4.45
Mahalanobis	Raw Correlations	25	0.28	0.16	2.5×10^{-1}	2.04
Mahalanobis	Raw Correlations	50	0.36	0.20	5.9×10^{-2}	2.25
Mahalanobis	Raw Correlations	100	0.31	0.25	2.2×10^{-1}	1.35

are much higher than the odds in the typical groups (least distant). This provides evidence that the observable vectors are a good, low-dimensional characterisation of the market state and accentuate meaningful financial "signal".

An additional note relates to the correlation matrices and associated observable vectors for February 2022. During this period, there were several extremely anomalous dates (with extreme vector lengths), coinciding with the beginning of the war in Ukraine. These had the effect of masking the anomalous nature of earlier events and reducing the sensitivity of the detection algorithm. This is a well known consequence of applying the Mahalanobis distance to anomaly detection, known as the masking effect [39]. We therefore excluded February 2022 from all our analyses.

The analysis conducted thus far can be regarded as pertaining to anomaly detection with an in-sample selection of the observables to include. We also conducted an out-of-sample analysis using the same observables as the in-sample analysis, now for the date range 2022-03-01 to 2023-03-31. The results are collected in Table 7 and reveal that the in-sample anomaly detection results generalise well, and the Mahalanobis distance calculated on observable vectors continues to out-perform the other alternatives. Robustness checks that were conducted include rerunning the analysis with correlation matrices constructed with 1, 10 and 15 min sampling intervals for the log returns (as opposed to 5 min). The results of these analyses were qualitatively similar to those already presented, with only the 1 min results being somewhat worse than longer timescales, see section 5 of the supplementary material.

5.2. Visual similarity

Equipped with the observable vector representation of correlation matrices and the Mahalanobis distance metric, another interesting investigation is to assess how visually similar the correlation matrices are that correspond to the pairs of dates with the closest observable vectors and contrast this with the furthest pairs. The procedure involves calculating the Mahalanobis distance between each pair of dates, of which there are $446(445)/2 = 99{,}235$ such pairs, and then sorting by distance in ascending order. The closest 10 pairs and most distant 10 pairs are visualised in figures provided in the supplementary material. It is readily apparent that the proximity of the dates in the space of observables aligns quite closely with the visual similarity or dissimilarity of the correlation matrices.

Finally, we quantitatively assess whether there is a concordance between the closeness of the visual representations of correlation matrices with the closeness of the observable vectors. In particular, we determine the Spearman correlation between the distance ranking of pairs of dates as determined in two ways,

1. Using the Euclidean distance to measure distance between pairs of correlation matrices as a proxy for visual similarity (this is essentially the widely used mean squared error metric for comparing image similarity, known to be closely related to perceived visual similarity).
2. Using the Mahalanobis distance between pairs of observable vectors.

The result is a Spearman correlation of approximately 0.7 across 99,235 pairs of dates. Spearman correlation coefficients of between 0.7 and 1.0 are regarded as very strongly positive, with 1.0 representing a perfect monotonic relationship. Using the exact Spearman correlation test, the p-value is smaller than 2.2×10^{-16} and the hypothesis of the Spearman correlation being zero can be strongly rejected in favour of the alternate hypothesis that the Spearman correlation is positive. This implies that despite the much lower dimensionality of the observable vectors, they do indeed capture the essence of what we would regard as visually similar correlation matrices.

6. Discussion and conclusion

We have developed the most general 4-parameter permutation invariant Gaussian matrix model appropriate for ensembles of matrices which are symmetric and diagonally vanishing, as suitable for financial correlation matrices. We have used the models to find near-Gaussianity in ensembles of matrices, one matrix for every day over a period, constructed from high-frequency foreign exchange price quotes. The near-Gaussianity was found to be robust against changes in how the ensemble was constructed : we varied the sampling intervals used to construct the daily estimates as well as the number of days used in our ensemble. The near-Gaussianity is used to motivate a data-reduction technique based on the use of low degree permutation invariant functions of matrices (observables) as characteristics of the entities represented by the matrices in the ensemble, in this case the days in the period under consideration. The small non-Gaussianities of each observable were used to rank the observables in order of decreasing non-Gaussianity and to find an optimal number of least Gaussian observables for data analysis. The degree of non-Gaussianity is thus being used as an analog of the magnitude of singular values in principal component analysis (PCA). The sets of observables considered, either the full set of observables up to quartic degree or the subsets with optimal number of least Gaussian observables, are much smaller than the number of matrix elements in the matrices. We found successful results in anomaly detection based on the observables to find the most atypical and the most typical days in the ensemble. We demonstrated statistically significant matching between these typicality/atypicality results extracted from the data of financial correlation matrices and corresponding results based on human economic judgement of significant events affecting foreign exchange markets. We propose that the success of the use of a set of least Gaussian observables in anomaly detection should be interpreted as indicating that these ensembles of daily foreign exchange matrices capture an economic reality best described by the Gaussian model perturbed by specific small cubic and quartic couplings in the action. The non-Gaussianities capture system-specific non-universalities while the overall approximate Gaussianity is a universal characteristic which holds across diverse systems, as indeed already evidenced in ensembles of words [3,4]. Related discussions of Gaussianity, universality and non-universality appear in random matrix theory literature [40–42].

We have thus explored two hypotheses, which are, in principle, independent. The first hypothesis is that the daily correlation matrices estimated from the forex data are well described by the PIGM model. This is supported by Table 4 for our data set. The second hypothesis is that small departures from Gaussianity, as revealed by the PIGM model, are physically/economically meaningful characteristics of the day. We have gathered evidence in favour of this hypothesis on our particular data set in Section 5.1, see Tables 6 and 7 in particular.

Future directions include evaluating how the parameters of the PIGM model and our other key results change as the details of the correlation calculation methodology as well as the particular FX data set used for the analysis change. This should help definitively answer the question of whether our results capture intrinsic properties of the FX market or are more specific to a particular data set and correlation estimation procedure. Indeed, this includes investigating how non-stationarity in the FX market data affects the fit of the PIGM model and whether the aforementioned hypotheses survive different market regimes. Another interesting avenue of exploration related to statistical methodology is the investigation of the impact of using maximum likelihood estimation in place of the method of moments approach that was utilised in fitting the PIGM model in this paper. While we expect the approaches to yield qualitatively similar results, there could be subtle differences.

A distinct research direction involves studying the effect of different choices of time-scale for constructing the ensemble of correlation matrices, e.g. one hour, twelve hours, two days etc., on the linear and quadratic parameters. In addition, it should be illuminating to investigate whether the agreement of empirical observables with PIGM predictions (fit on linear and quadratic observables only) changes with correlation time-scale. Indeed, the evolution of the magnitude of cubic and quartic corrections with time-scale should assist in answering the question of whether the non-Gaussianity increases or decreases with time-scale.

There are also several practical directions to pursue with particular relevance to finance. The first involves clustering the correlation matrices in observable vector representation and elucidating how the cluster structure changes with correlation time-scale. This should provide insight into the states of the market at different time-scales and how they relate to each other (along the lines of [43,44] for example). Modelling transitions between market states is a related application. A further area of interest is to study the effectiveness of so-called nowcasting of market state, where the market state is identified as it is formed. These applications could have great utility in risk management for example. Assessing the effectiveness of the PIGM observables as features in machine learning algorithms for predicting quantities such as future price returns would inform viability in trading strategy applications.

While we have studied forex data in this article, another future direction is to apply the PIGM model to stock market data. We expect that some results will be similar and differences would be particularly illuminating.

Finally, generating realistic, random samples of financial correlation matrices has been recognised as having value in a number of contexts including enhancing trading strategies and stress testing portfolios of assets [13]. We are presently investigating drawing samples from the PIGM model. In this case we do not expect sampled matrices to be positive-semidefinite in general, where positive-semidefiniteness is a property of all valid correlation matrices. However this is readily addressed using approaches such as those in [45]. We are also contrasting to existing approaches [13]. It will be fascinating to determine whether there is a trade-off between the exceptional parsimony of the PIGM model (i.e. 4 parameters) and the ability to capture the known stylised facts of financial correlation matrices - particularly in contrast to the approach in [13] which utilises a generative adversarial neural network with several orders of magnitude more parameters than the PIGM model.

CRediT authorship contribution statement

George Barnes: Writing – review & editing, Writing – original draft, Conceptualization. **Sanjaye Ramgoolam:** Writing – review & editing, Writing – original draft, Conceptualization. **Michael Stephanou:** Writing – review & editing, Writing – original draft, Conceptualization.

Data availability

The data and code has been shared via a link given in the article.

Statement and acknowledgements

The views expressed in this article are those of the authors and do not necessarily reflect the views of Rand Merchant Bank. Rand Merchant Bank does not make any representations or give any warranties as to the correctness, accuracy or completeness of the information presented; nor does Rand Merchant Bank assume liability for any losses arising from errors or omissions in the information in this article. We would like to thank Steve Abel, Manuel Accettuli Huber, Stephon Alexander, Adrian Bevan, Graham Brown, Adrian Padellaro, Mehrnoosh Sadrzadeh, Alex Stapleton and Steve Thomas for useful discussions on this project. We would also like to thank the reviewers for their very helpful suggestions. SR is supported by the STFC consolidated grant ST/P000754/1 "String Theory, Gauge Theory and Duality". S.R. acknowledges the support of the Institut Henri Poincaré (UAR 839 CNRS-Sorbonne Université), and LabEx CARMIN (ANR-10-LABX-59- 01). SR also acknowledges the support of the Perimeter Institute for Theoretical Physics during a visit in the final stages of completion of the paper. Research at Perimeter Institute is supported by the Government of Canada through Industry Canada and by the Province of Ontario through the Ministry of Economic Development and Innovation. SR also acknowledges the hospitality of the high energy physics group of Brown University and a Visiting Professorship at the Dublin Institute for Advanced Studies during the completion of this project.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Currency pair mapping

See Table 8 for the mapping between currency pair indices and currency pair names.

Appendix B. 4-parameter Gaussian model: detailed construction

In this appendix we give a detailed account of the construction of the 4-parameter matrix model which was outlined in Section 2. The aim is to model the statistics of the particular ensemble of correlation matrices introduced in Section 3 as well as any matrix ensemble in the same universality class. This class is composed of symmetric square matrices with zeros along the diagonal, for which physical quantities are invariant under simultaneous permutations of the rows and columns. Given the universality of the model we label our matrices M and index them with lowercase indices $1 \leq i, j \leq D$, to distinguish them from the financial correlation matrices specifically, which we label $\hat{\rho}$. We will refer to matrices with zeros on the diagonal as "diagonally vanishing" throughout.

We begin by defining the action of permutations on the matrix variables and establishing their irreducible decomposition under this group action. We then define the most general action of a permutation invariant Gaussian matrix model, containing symmetric matrices with vanishing diagonal. It is parameterised by one linear and three quadratic couplings. Finding projectors that project from the original matrix basis M_{ij} to a basis transforming according to this irreducible decomposition allows us to rewrite this action in a diagonalised form. In turn, this diagonalisation permits the application of standard multi-dimensional Gaussian integration techniques which, along with the application of Wick's theorem, produce analytic formulae for the expectation values of observables as a function of D — the dimension of the matrices.

Table 8
Currency pair mapping.

Index (I)	Currency pair
1	AUD/JPY
2	AUD/NZD
3	AUD/USD
4	CAD/JPY
5	CHF/JPY
6	EUR/CHF
7	EUR/GBP
8	EUR/JPY
9	EUR/PLN
10	EUR/USD
11	GBP/JPY
12	GBP/USD
13	NZD/USD
14	USD/CAD
15	USD/CHF
16	USD/JPY
17	USD/MXN
18	USD/TRY
19	USD/ZAR

B.1. Symmetric group representation theory and matrix variables

The diagonal action of S_D simultaneously permutes the rows and columns of a matrix

$$M_{ij} \to M_{\gamma(i)\gamma(j)}, \qquad \forall \gamma \in S_D. \tag{B.1}$$

We wish to find an action S (which defines the measure) that is invariant under the symmetric group action in Eq. (B.1) on the matrix variables. We wish to find an action that is invariant under this group action, and again refer to the space of symmetric matrices with vanishing diagonal as the physical subspace V^{phys} of general $D \times D$ matrices and label matrices in the space with a superscript i.e. $M^{\text{phys}} \in V^{\text{phys}}$.

The natural representation of the symmetric group V_D, is a D-dimensional representation with basis $\{e_1, e_2, \dots, e_D\}$. For each permutation $\gamma \in S_D$ the linear operator acting on this space is defined by

$$\rho_{V_D}(\gamma) e_i = e_{\gamma^{-1}(i)}. \tag{B.2}$$

The natural representation is a reducible representation and decomposes as

$$V_D \cong V_0 \oplus V_H, \tag{B.3}$$

where V_0 is the trivial representation and V_H is the Hook representation [46], both of which are irreducible. This decomposition is given explicitly by forming the following linear combinations of natural basis elements

$$\begin{aligned}
E_0 &= \frac{1}{\sqrt{D}}(e_1 + e_2 + \cdots + e_D), \\
E_1 &= \frac{1}{\sqrt{2}}(e_1 - e_2), \\
E_2 &= \frac{1}{\sqrt{6}}(e_1 + e_2 - 2e_3), \\
&\vdots \\
E_a &= \frac{1}{\sqrt{a(a+1)}}(e_1 + e_2 + \cdots + e_a - a e_{a+1}), \\
&\vdots \\
E_{D-1} &= \frac{1}{\sqrt{D(D+1)}}(e_1 + e_2 + \cdots + e_{D-1} - (D-1)e_D).
\end{aligned} \tag{B.4}$$

E_0 is the trivial representation, and the remaining $D-1$ vectors $\{E_1, E_2 \dots, E_{D-1}\}$ form an irreducible representation of S_D which is V_H.

A general matrix M_{ij} transforming under (B.1) forms the representation $V_D \otimes V_D$, the tensor product of two copies of the natural representation of S_D. A natural basis for this tensor product is given by

$$e_i \otimes e_j, \qquad 1 \le i, j \le D, \tag{B.5}$$

on which S_D acts as

$$\rho_{V_D^{\otimes 2}}(\gamma) e_i \otimes e_j = e_{\gamma^{-1}(i)} \otimes e_{\gamma^{-1}(j)}. \tag{B.6}$$

$\{e_1, e_2, \ldots, e_D\}$ is an orthonormal basis of V_D under the inner product

$$\left(e_i, e_j\right) = \delta_{ij}\,. \tag{B.7}$$

This inner product is extended to $V_D \otimes V_D$ as

$$\left(e_i \otimes e_j, e_k \otimes e_l\right) = \delta_{ik}\delta_{jl}. \tag{B.8}$$

Using (B.7) we define the change of basis coefficients

$$C_{0,i} \equiv (E_0, e_i) = \frac{1}{\sqrt{D}}, \tag{B.9}$$

$$C_{a,i} \equiv (E_a, e_i). \tag{B.10}$$

$V_D \otimes V_D$ is a reducible representation of S_D which has the following irreducible decomposition

$$V_D \otimes V_D \cong 2V_0 \oplus 3V_H \oplus V_2 \oplus V_3. \tag{B.11}$$

This decomposition can be deduced using (B.3) along with the tensor product rule described in section 7.13 of [25]. Other useful references for symmetric group representation theory include [46,47]. In terms of Young diagrams, described by listing the number of columns in each row in descending order, the irreducible representations appearing on the right hand side of this decomposition are

$$\begin{aligned} V_0 &= [D], \\ V_H &= [D-1, 1], \\ V_2 &= [D-2, 2], \\ V_3 &= [D-2, 1, 1]. \end{aligned} \tag{B.12}$$

The space of symmetric matrices with vanishing diagonal form a subspace of the representations in (B.11). Firstly, symmetric matrices transform as $\mathrm{Sym}^2(V_D)$ i.e. the symmetric part of the product space $V_D \otimes V_D$. This is a reducible representation with the following decomposition

$$\mathrm{Sym}^2(V_D) \cong 2V_0 \oplus 2V_H \oplus V_2. \tag{B.13}$$

The matrix elements along the diagonal $\{M_{ii} | 1 \leq i \leq D\}$ transform like the natural representation V_D. Removing a copy of $V_D \cong V_0 \oplus V_H$ from the symmetric product of V_D in (B.13) gives the following decomposition of the physical subspace

$$V^{\mathrm{phys}} \cong \mathrm{Sym}^2(V_D)/V_D \cong V_0 \oplus V_H \oplus V_2. \tag{B.14}$$

The decomposition (B.14) tells us that the enforcement of permutation invariance on the action of a Gaussian theory containing symmetric D dimensional matrices without diagonal permits a single independent linear term (i.e. the number of trivial representations appearing on the RHS). Quadratic products of physical matrices transform as

$$V^{\mathrm{phys}} \otimes V^{\mathrm{phys}} \cong \left(V_0 \oplus V_H \oplus V_2\right) \otimes \left(V_0 \oplus V_H \oplus V_2\right). \tag{B.15}$$

Using the orthogonality property of characters as well as the reality property of irreducible representations of the symmetric group, it can be shown that the tensor product of two irreducible representations contains the trivial representation if and only if the irreducible representations are identical — and in this case the decomposition contains exactly one copy of the trivial representation. This enables us to count the number of independent quadratic terms. We find three independent quadratic contributions to the action corresponding to the following three terms in (B.15)

$$V_0 \otimes V_0 \cong V_0 + \cdots\,, \tag{B.16}$$

$$V_H \otimes V_H \cong V_0 + \cdots\,, \tag{B.17}$$

$$V_2 \otimes V_2 \cong V_0 + \cdots\,. \tag{B.18}$$

Much of our task in calculating expectation values of observables of the physical model for symmetric diagonally vanishing matrices amounts to finding a change of basis for $V_D \otimes V_D$ from the original $e_i \otimes e_j$ to one which transforms in the same manner as the irreducible decomposition of V^{phys}, i.e. from the LHS of (B.14) to the RHS. Once found it diagonalises the physical action and consequently permits the calculation of expectation values of observables. The coefficients that define this change of basis are called Clebsch–Gordon coefficients. For each irreducible representation on the RHS of (B.14) define $C_{ij}^{\mathrm{phys};V_0}, C_{ij,a}^{\mathrm{phys};V_H}, C_{ij,a}^{\mathrm{phys};V_2}$ respectively, where a is a state index running over the dimension of the irreducible representation.

Note that if we only imposed the condition of symmetry $M_{ij} = M_{ji}$ on the matrices, we would have two linear couplings corresponding to the two copies of V_0 in (B.13). We would have 3 parameters of a 2×2 symmetric matrix of couplings for quadratic terms arising from the two copies of V_0 in (B.13), 3 parameters of a 2×2 symmetric matrix of couplings for quadratic terms arising from the two copies of V_H in (B.13), and finally one parameter for V_2. For symmetric matrices, therefore, there is a 9 parameter permutation invariant matrix model. We will focus, in the following, on the 4-parameter model which incorporates the symmetry condition $M_{ij} = M_{ji}$ as well as the condition of vanishing diagonal.

B.2. Projector for V^{phys}

The projectors to the trivial representations appearing in the quadratic products of M_{ij} are given by squaring the relevant Clebsch coefficients and summing over intermediate states

$$P^{\text{phys};V_0}_{ij,kl} = C^{\text{phys};V_0}_{ij} C^{\text{phys};V_0}_{kl}\,, \tag{B.19}$$

$$P^{\text{phys};V_H}_{ij,kl} = \sum_{a=1}^{D-1} C^{\text{phys};V_H}_{ij,a} C^{\text{phys};V_H}_{kl,a}\,, \tag{B.20}$$

$$P^{\text{phys};V_2}_{ij,kl} = \sum_{a=1}^{\dim V_2} C^{\text{phys};V_2}_{ij,a} C^{\text{phys};V_2}_{kl,a}\,. \tag{B.21}$$

Here we find explicit formulae for these projectors. The V_0 and V_H projectors are constructed by finding the Clebschs on the RHS of (B.19) and (B.20) explicitly. In the case of the V_2 projector things are not so simple as the Clebsch is not so easily calculable, nonetheless we are able to construct the projector using general properties of Clebsch coefficients and other known projectors, without precise knowledge of the V_2 Clebsch itself.

To find $C^{\text{phys};V_0}_{ij}$ and $C^{\text{phys};V_H}_{ij,a}$ we first write down a representation theory basis for $V_D \otimes V_D$ in terms of the change of basis coefficients given in (B.9) and (B.10) as was done in [1] i.e. a basis that transforms like the RHS of (B.11),

$$S^{V_0;1} \equiv \sum_{i,j=1}^{D} C_{0,i} C_{0,j} M_{ij} = \frac{1}{D} \sum_{i,j=1}^{D} e_i \otimes e_j\,, \tag{B.22}$$

$$S^{V_0;2} \equiv \frac{1}{\sqrt{D-1}} \sum_{a=1}^{D-1} \sum_{i,j=1}^{D} C_{a,i} C_{a,j} M_{ij} = \frac{1}{\sqrt{D-1}} \sum_{a=1}^{D-1} E_a \otimes E_a\,, \tag{B.23}$$

$$S_a^{V_H;1} \equiv \sum_{i,j=1}^{D} C_{0,i} C_{a,j} M_{ij} = \frac{1}{\sqrt{D}} \sum_{i=1}^{D} e_i \otimes E_a\,, \tag{B.24}$$

$$S_a^{V_H;2} \equiv \sum_{i,j=1}^{D} C_{a,i} C_{0,j} M_{ij} = \frac{1}{\sqrt{D}} \sum_{i=1}^{D} E_a \otimes e_i\,, \tag{B.25}$$

$$S_a^{V_H;3} \equiv \sqrt{\frac{D}{D-2}} \sum_{b,c=1}^{D-1} \sum_{i,j,k=1}^{D} C_{a,k} C_{b,k} C_{c,k} C_{b,i} C_{c,j} M_{ij} = \sqrt{\frac{D}{D-2}} \sum_{b,c=1}^{D-1} \sum_{i=1}^{D} C_{a,i} C_{b,i} C_{c,i} E_b \otimes E_c\,. \tag{B.26}$$

We also note the orthogonal decomposition

$$V_D \otimes V_D \cong \text{Sym}^2(V_D) \oplus \Lambda^2(V_D) \cong V^{\text{phys}} \oplus V^{\text{diag}} \oplus \Lambda^2(V_D), \tag{B.27}$$

in which V^{diag} is the subspace of diagonal matrix elements and $\Lambda^2(V_D)$ is the antisymmetric subspace of $V_D \otimes V_D$.

Define further representation variables $S^{\text{diag};V_0}$ and $S_a^{\text{diag};V_H}$ composed of the diagonal elements of M_{ij}, that transform according to the first and second terms on the RHS of the V^{diag} decomposition respectively

$$S^{\text{diag};V_0} \equiv \frac{1}{\sqrt{D}} \sum_{i=1}^{D} e_i \otimes e_i, \tag{B.28}$$

$$S_a^{\text{diag};V_H} \equiv E_a \otimes E_a. \tag{B.29}$$

Using the inner product on $V_D \otimes V_D$ in (B.8) We can express these in terms of the representation variables

$$\begin{aligned} S^{\text{diag};V_0} &= (S^{\text{diag};V_0}, S^{V_0;1}) S^{V_0;1} + (S^{\text{diag};V_0}, S^{V_0;2}) S^{V_0;2} \\ &= \frac{1}{\sqrt{D}} S^{V_0;1} + \sqrt{\frac{D-1}{D}} S^{V_0;2}\,, \end{aligned} \tag{B.30}$$

and

$$\begin{aligned} S_a^{\text{diag};V_H} &= \frac{1}{2} \sum_{b=1}^{D-1} \left((S_a^{\text{diag};V_H}, S_b^{V_H;1}) S_b^{V_H;1} + (S_a^{\text{diag};V_H}, S_b^{V_H;2}) S_b^{V_H;2} \right) + \sum_{b=1}^{D-1} (S_a^{\text{diag};V_H}, S_b^{V_H;3}) S_b^{V_H;3} \\ &= \frac{1}{2} \sum_{b=1}^{D-1} \left(\sqrt{\frac{2}{D}} \delta_{ab} S_b^{V_H;1} + \sqrt{\frac{2}{D}} \delta_{ab} S_b^{V_H;2} \right) + \sum_{b=1}^{D-1} \sqrt{\frac{D-2}{D}} \delta_{ab} S_b^{V_H;3} \\ &= \frac{1}{\sqrt{2D}} \left(S_a^{V_H;1} + S_a^{V_H;2} \right) + \sqrt{\frac{D-2}{D}} S_a^{V_H;3}\,. \end{aligned} \tag{B.31}$$

The physical variables $S^{\text{phys};V_0}$ and $S_a^{\text{phys};V_H}$ span the orthogonal complement of the diagonal variables in the V_0 and V_H subspaces (as given in (B.28)) of $\text{Sym}^2(V_D)$

$$\begin{aligned} S^{\text{phys};V_0} &= \sqrt{\frac{D-1}{D}} S^{V_0;1} - \frac{1}{\sqrt{D}} S^{V_0;2} \\ &= \sqrt{\frac{D-1}{D}} \sum_{i,j=1}^{D} C_{0,i} C_{0,j} M_{ij} - \frac{1}{\sqrt{D(D-1)}} \sum_{a=1}^{D-1} \sum_{i,j=1}^{D} C_{a,i} C_{a,j} M_{ij} \,, \end{aligned} \tag{B.32}$$

$$\begin{aligned} S_a^{\text{phys};V_H} &= \sqrt{\frac{D-2}{2D}} \left(S_a^{V_H;1} + S_a^{V_H;2} \right) - \sqrt{\frac{2}{D}} S_a^{V_H;3} \\ &= \sqrt{\frac{D-2}{2D^2}} \sum_{i,j=1}^{D} \left(C_{a,i} + C_{a,j} \right) M_{ij} - \sqrt{\frac{2}{D}} \sum_{i,j=1}^{D} \sum_{b,c=1}^{D-1} C_{b,i} C_{c,j} C_{b,c}^{HH\to H}{}_{a} M_{ij} \,. \end{aligned} \tag{B.33}$$

From (B.32) and (B.33) we can read off the Clebsch coefficients needed in the construction of the projectors (B.19) and (B.20):

$$S^{\text{phys};V_0} = \sum_{i,j} C_{ij}^{\text{phys};V_0} M_{ij} \;\Rightarrow\; C_{ij}^{\text{phys};V_0} = \sqrt{\frac{D-1}{D}} C_{0,i} C_{0,j} - \frac{1}{\sqrt{D(D-1)}} \sum_{a=1}^{D-1} C_{a,i} C_{a,j} \,, \tag{B.34}$$

$$S_a^{\text{phys};V_H} = \sum_{i,j} C_{ij,a}^{\text{phys};V_H} M_{ij} \;\Rightarrow\; C_{ij,a}^{\text{phys};V_H} = \sqrt{\frac{(D-2)}{2D^2}} \left(C_{a,i} + C_{a,j} \right) - \sqrt{\frac{2}{(D-2)}} \sum_{b,c=1}^{D-1} \sum_{k=1}^{D} C_{b,i} C_{c,j} C_{a,k} C_{b,k} C_{c,k} \,. \tag{B.35}$$

Although we do not know the V_2 Clebsch coefficients $C_{b,c}^{HH\to V_2}{}_{a}$ appearing in the remaining physical variables

$$S_a^{\text{phys};V_2} = S_a^{V_2} = \sum_{i,j=1}^{D} \sum_{b,c=1}^{D-1} C_{b,i} C_{c,j} C_{b,c}^{HH\to V_2}{}_{a} M_{ij}, \tag{B.36}$$

we do know that they possess the usual Clebsch orthogonality property

$$\sum_{b,c=1}^{D-1} C_{bc,\ a_1}^{HH\to V_2} C_{bc,\ a_2}^{HH\to V_2} = \delta_{a_1 a_2} \,. \tag{B.37}$$

Also, we are able to write an expression for the V_2 projector using the decomposition

$$\text{Sym}^2(V_H) = V_0 \oplus V_H \oplus V_2. \tag{B.38}$$

This allows us to express the V_2 projector in terms of the projectors from $V_H \otimes V_H$ to $\text{Sym}^2(V_H)$, V_0 and V_H

$$P^{\text{phys};V_2} = P^{V_H,V_H\to V_2} = \left(1 - P^{V_H,V_H\to V_0} - P^{V_H,V_H\to V_H} \right) P^{V_H,V_H\to \text{Sym}^2(V_H)}. \tag{B.39}$$

The projectors on the RHS of this expression are given by

$$P_{ab,cd}^{V_H,V_H\to \text{Sym}^2(V_H)} = \frac{1}{2} \left(\delta_{ac}\delta_{bd} - \delta_{ad}\delta_{bc} \right), \tag{B.40}$$

$$P_{ab,cd}^{V_H,V_H\to V_0} = \frac{1}{D-1} \delta_{ab}\delta_{cd} \,, \tag{B.41}$$

$$P_{ab,cd}^{V_H,V_H\to V_H} = \frac{D}{D-2} \sum_{e=1}^{D-1} \sum_{i,j=1}^{D} C_{a,i} C_{b,i} C_{e,i} C_{c,j} C_{d,j} C_{e,j} \,. \tag{B.42}$$

The derivations of these expressions are given in section 2.2 of [1]

The action of the physical V_0, V_H projectors, given by the square of the Clebsch coefficients in (B.34) and (B.35), along with the V_2 projector given in (B.39) can be found by acting on a generic state $e_i \otimes e_j$ in $V_D \otimes V_D$. Doing so leaves us with the following delta expressions

$$P_{ij,kl}^{\text{phys};V_0} = \frac{1}{D(D-1)} \left(\delta_{ij} - 1 \right) \left(\delta_{kl} - 1 \right), \tag{B.43}$$

$$P_{ij,kl}^{\text{phys};V_H} = \frac{1}{2(D-1)} \left(1 - \delta_{ij} \right) \left(1 - \delta_{kl} \right) \left(\delta_{ik} + \delta_{il} + \delta_{jk} + \delta_{jl} - \frac{4}{D} \right), \tag{B.44}$$

$$P_{ij,kl}^{\text{phys};V_2} = \frac{1}{D-2} \left(-D\delta_{ij}\delta_{jk}\delta_{kl} + \delta_{ij}\delta_{ik} + \delta_{ij}\delta_{il} + \delta_{ik}\delta_{il} + \delta_{jk}\delta_{jl} + \frac{1}{D-1} \left(\delta_{ij} - 1 \right) \left(\delta_{kl} - 1 \right) - \frac{1}{2} \left(\delta_{ik} + \delta_{il} + \delta_{jk} + \delta_{jl} \right) \right). \tag{B.45}$$

The projector from $V_D \otimes V_D$ to the entire physical subspace, projects general $D \times D$ matrices onto the space of symmetric matrices with vanishing diagonal. The action of this projector on a general state $e_k \otimes e_l$ can be written using the inner product (B.8),

$$\begin{aligned} P^{\text{phys}} e_k \otimes e_l &= \frac{1}{2} \sum_{i<j}^{D} e_i \otimes e_j (e_i \otimes e_j, e_k \otimes e_l) \\ &= \frac{1}{2}(e_k \otimes e_l + e_l \otimes e_k) - \delta_{kl}(e_k \otimes e_l). \end{aligned} \tag{B.46}$$

Such that it can be written as a delta expression that acts by unrestricted summation as

$$P^{\text{phys}}_{ij,kl} = \frac{1}{2}\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right) - \delta_{ij}\delta_{jk}\delta_{kl}. \tag{B.47}$$

As expected, given the orthogonality of the physical projectors (B.43)–(B.45)

$$P^{\text{phys}} = P^{\text{phys};V_0} + P^{\text{phys};V_H} + P^{\text{phys};V_2}. \tag{B.48}$$

B.3. The action and physical projectors

The action of the most general permutation invariant Gaussian matrix model of symmetric matrices with vanishing diagonal is given by

$$S = \tau_1 \sum_{i,j=1}^{D} M^{\text{phys}}_{ij} M^{\text{phys}}_{ij} + \tau_2 \sum_{i,j,k=1}^{D} M^{\text{phys}}_{ij} M^{\text{phys}}_{jk} + \tau_3 \sum_{i,j,k,l=1}^{D} M^{\text{phys}}_{ij} M^{\text{phys}}_{kl} - \mu \sum_{i,j=1}^{D} M^{\text{phys}}_{ij}, \tag{B.49}$$

where μ is the linear coupling strength and τ_1, τ_2, τ_3 are quadratic couplings. Define the partition function

$$\mathcal{Z} \equiv \int \mathrm{d}M e^{-S}. \tag{B.50}$$

We are interested in calculating permutation invariant expectation values of operators composed of the M_{ij} variables. These are defined as

$$\langle \mathcal{O}_\alpha(M) \rangle \equiv \frac{1}{\mathcal{Z}} \int \mathrm{d}M \mathcal{O}_\alpha(M) e^{-S}. \tag{B.51}$$

In order to evaluate (B.51) we must factorise the integrals that appear. The action as written in (B.49) contains $\frac{D(D-1)}{2}$ matrix variables mixed in a non-trivial way. The solution to this problem exploits the decomposition (B.14). The presence of one copy of V_0 means that there is one linear term with coefficient μ_{V_0} employing the Clebsch–Gordan coefficient for V_0 in V^{phys}. There are three quadratic terms in the action which employ the projectors for each of the three irreducible representations

$$S = \sum_{i,j,k,l} \frac{1}{2}\left(\tau_{V_0} M_{ij} P^{\text{phys};V_0}_{ij,kl} M_{kl} + \tau_{V_H} M_{ij} P^{\text{phys};V_H}_{ij,kl} M_{kl} + \tau_{V_2} M_{ij} P^{\text{phys};V_2}_{ij,kl} M_{kl}\right) - \sum_{i,j} \mu_{V_0} C^{\text{phys};V_0}_{ij} M_{ij}. \tag{B.52}$$

Applying these projectors amounts to rewriting the action in terms of the representation theory variables $S^{\text{phys};V_0}, S^{\text{phys};V_H}_a, S^{\text{phys};V_2}_a$ as given in Eqs. (B.34), (B.35) and (B.36) :

$$S = -\mu_{V_0} S^{\text{phys};V_0} + \frac{\tau_{V_0}}{2} S^{\text{phys};V_0} S^{\text{phys};V_0} + \frac{\tau_{V_H}}{2} \sum_{a=1}^{D-1} S^{\text{phys};V_H}_a S^{\text{phys};V_H}_a + \frac{\tau_{V_2}}{2} \sum_{a=1}^{\dim V_2} S^{\text{phys};V_2}_a S^{\text{phys};V_2}_a. \tag{B.53}$$

With this change of variables the partition function reads

$$\mathcal{Z} \equiv \int \mathrm{d}S e^{-S}, \tag{B.54}$$

with

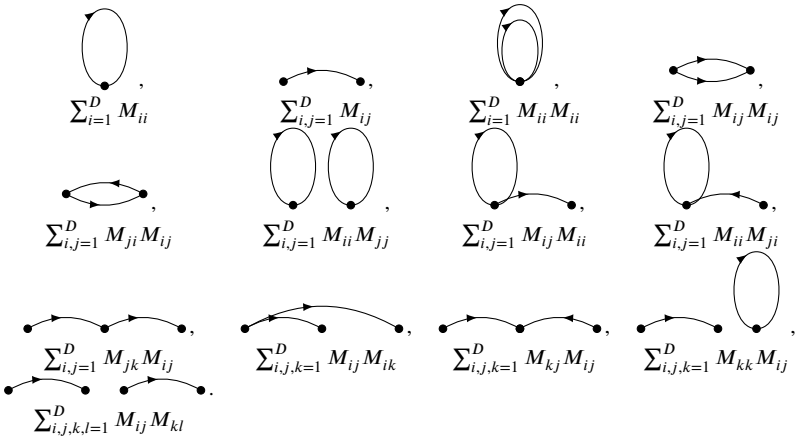
$$\mathrm{d}S = \mathrm{d}S^{\text{phys};V_0} \prod_{a_1=1}^{D-1} \mathrm{d}S^{\text{phys};V_H}_{a_1} \prod_{a_2=1}^{\dim V_2} \mathrm{d}S^{\text{phys};V_2}_{a_2}. \tag{B.55}$$

The factorised expression permits the application of standard techniques of Gaussian integration. Substituting the expressions for the projectors into (B.52) and performing the summations we can also write the action in terms of the full $V_D \otimes V_D$ representation variables

$$\begin{aligned} S = &- \mu_{V_0}\left(\sqrt{\frac{D-1}{D}} S^{V_0;1} - \frac{1}{\sqrt{D}} S^{V_0;2}\right) + \frac{\tau_{V_0}}{2}\left(\sqrt{\frac{D-1}{D}} S^{V_0;1} - \frac{1}{\sqrt{D}} S^{V_0;2}\right)^2 + \frac{\tau_{V_H}}{2} \sum_{a=1}^{D-1}\left(\sqrt{\frac{D-2}{2D}}(S^{V_H;1}_a + S^{V_H;2}_a) - \sqrt{\frac{2}{D}} S^{V_H;3}_a\right)^2 \\ &+ + \frac{\tau_{V_2}}{2} \sum_{a=1}^{\dim V_2} S^{V_2}_a S^{V_2}_a. \end{aligned} \tag{B.56}$$

B.4. Observables and correlators

For generic matrices there are 2 linear and 11 invariant quadratic functions, these are associated with the following graphs



(B.57)

The graphs are associated with these polynomials as follows: each vertex corresponds to a summed index and each directed edge to a matrix M_{ij} from the vertex corresponding to index i to the vertex corresponding to index j. This intuitive connection was used in [2] and it was recognised that it is valid for $D \geq 2k$, with D the size of the matrix and k the degree of the polynomial. A mathematical proof of the connection was given in section 2 of [24]. There is also a formulation of the counting in terms of partition algebras as developed in section 3.1 of [48] and there are related results in earlier papers [49][50].

The observables of our theory $\mathcal{O}_\alpha$ are similarly the permutation invariant functions of the matrix variables M_{ij}, obeying

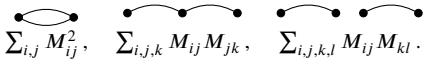
$$\mathcal{O}_\alpha(M_{ij}) = \mathcal{O}_\alpha(M_{\gamma(i)\gamma(j)}), \qquad \forall\, \gamma \in S_D. \tag{B.58}$$

The physical permutation invariant observables of order k are in one-to-one correspondence with undirected, loopless multigraphs with k edges. Requiring the matrices to be symmetric is equivalent to considering undirected edges. The further requirement that the matrices have vanishing diagonal entries is equivalent to restricting to loopless multigraphs.

Below we list the complete set of quadratic, cubic and quartic graphs of general matrices that survive the projection to the physical subspace.

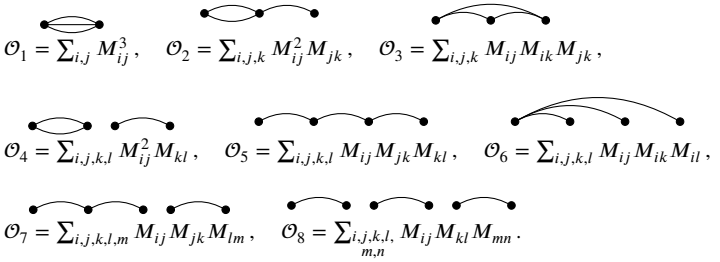
$$M^{\text{phys}}_{ij} = P^{\text{phys}}_{ij,kl} M_{kl}. \tag{B.59}$$

The counting of these graphs organised by number of edges is given by the OEIS sequence A050535 [51]. The three quadratic observables are



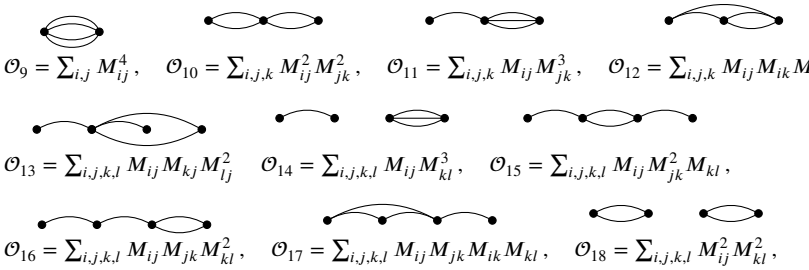
(B.60)

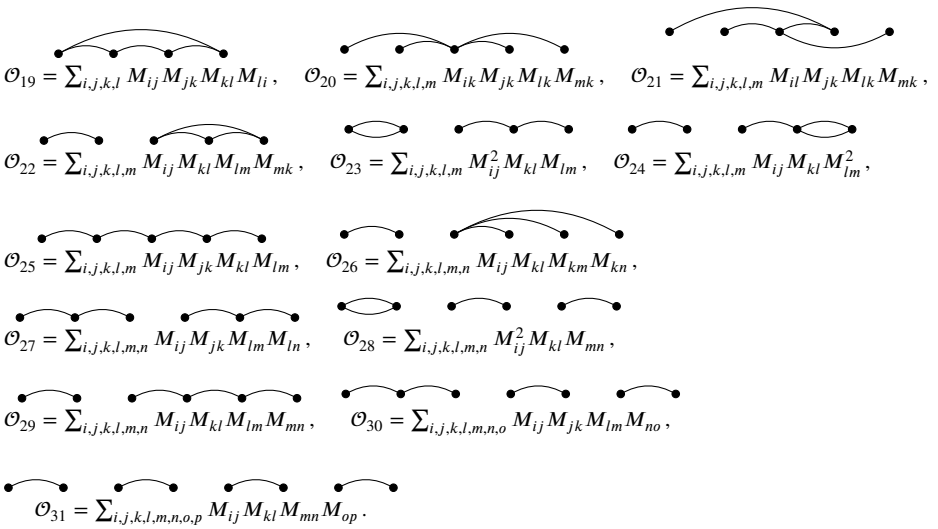
The eight cubic observables are



(B.61)

The 23 quartic observables are





$$\mathcal{O}_{19} = \sum_{i,j,k,l} M_{ij} M_{jk} M_{kl} M_{li}\,, \quad \mathcal{O}_{20} = \sum_{i,j,k,l,m} M_{ik} M_{jk} M_{lk} M_{mk}\,, \quad \mathcal{O}_{21} = \sum_{i,j,k,l,m} M_{il} M_{jk} M_{lk} M_{mk}\,,$$

$$\mathcal{O}_{22} = \sum_{i,j,k,l,m} M_{ij} M_{kl} M_{lm} M_{mk}\,, \quad \mathcal{O}_{23} = \sum_{i,j,k,l,m} M_{ij}^2 M_{kl} M_{lm}\,, \quad \mathcal{O}_{24} = \sum_{i,j,k,l,m} M_{ij} M_{kl} M_{lm}^2\,,$$

$$\mathcal{O}_{25} = \sum_{i,j,k,l,m} M_{ij} M_{jk} M_{kl} M_{lm}\,, \quad \mathcal{O}_{26} = \sum_{i,j,k,l,m,n} M_{ij} M_{kl} M_{km} M_{kn}\,,$$

$$\mathcal{O}_{27} = \sum_{i,j,k,l,m,n} M_{ij} M_{jk} M_{lm} M_{ln}\,, \quad \mathcal{O}_{28} = \sum_{i,j,k,l,m,n} M_{ij}^2 M_{kl} M_{mn}\,,$$

$$\mathcal{O}_{29} = \sum_{i,j,k,l,m,n} M_{ij} M_{kl} M_{lm} M_{mn}\,, \quad \mathcal{O}_{30} = \sum_{i,j,k,l,m,n,o} M_{ij} M_{jk} M_{lm} M_{no}\,,$$

$$\mathcal{O}_{31} = \sum_{i,j,k,l,m,n,o,p} M_{ij} M_{kl} M_{mn} M_{op}\,. \tag{B.62}$$

B.4.1. Expectation values

We can write the partition function in the following form

$$\begin{aligned}\mathcal{Z} &= \int \mathrm{d}S \exp\Big(\sum_{\Lambda,a} \mu_{\Lambda,a} S_a^{\mathrm{phys};\Lambda} - \frac{1}{2}\sum_{\Lambda,a} S_a^{\mathrm{phys};\Lambda} \tau_\Lambda S_a^{\mathrm{phys};\Lambda}\Big) \\ &= \frac{(2\pi)^{\frac{D(D-1)}{2}}}{(\det\tau)^{\frac{1}{2}}} \exp\Big(\frac{1}{2}\sum_{\Lambda}\sum_{a} \mu_{\Lambda,a}\tau_\Lambda^{-1}\mu_{\Lambda,a}\Big),\end{aligned} \tag{B.63}$$

in which we have included linear couplings for all S^{phys} variables. This is the usual trick employed to generate expectation values from the partition function, included so that we are able to source expectation values of all observables by taking derivatives with respect to the linear couplings. In order to recover the permutation invariant model all but the V_0 linear coupling are set to zero. The integral on the LHS is performed using standard techniques from multidimensional Gaussian integration.

Expectation values of the physical variables $\mathcal{O}(S)$ are defined by

$$\langle \mathcal{O}(S)\rangle \equiv \frac{\int \mathrm{d}S \mathcal{O}(S) e^{-S}}{\int \mathrm{d}S e^{-S}}. \tag{B.64}$$

Calculating these by taking derivatives of the RHS of (B.63), with respect to $\mu_{\Lambda,a}$ we find for example, the linear expectation values are given by

$$\begin{aligned}\langle S_a^{\mathrm{phys};\Lambda}\rangle &= \frac{1}{\mathcal{Z}}\int \mathrm{d}S S_a^{\mathrm{phys};\Lambda} e^{-S} = \frac{1}{\mathcal{Z}}\frac{\partial \mathcal{Z}}{\partial \mu_{\Lambda,a}}\Big|_{\mu_{\Lambda,a}\neq 0 \,\mathrm{iff}\, \Lambda = V_0} \\ &= \tau_\Lambda^{-1}\mu_\Lambda \delta(\Lambda, V_0).\end{aligned} \tag{B.65}$$

That is, the only non-zero linear expectation value is

$$\langle S^{\mathrm{phys};V_0}\rangle = \tau_{V_0}^{-1}\mu_{V_0}. \tag{B.66}$$

For later convenience we define

$$\widetilde{\mu}_{V_0} \equiv \tau_{V_0}^{-1}\mu_{V_0}. \tag{B.67}$$

Similarly, we can calculate the two-point function by taking two derivatives of the RHS of (B.63)

$$\begin{aligned}\langle S_a^{\mathrm{phys};\Lambda_1} S_b^{\mathrm{phys};\Lambda_2}\rangle &= \frac{1}{\mathcal{Z}}\int \mathrm{d}S S_a^{\mathrm{phys};\Lambda_1} S_b^{\mathrm{phys};\Lambda_2} e^{-S} = \frac{1}{\mathcal{Z}}\frac{\partial}{\partial \mu_{\Lambda_2,b}}\frac{\partial \mathcal{Z}}{\partial \mu_{\Lambda_1,a}}\Big|_{\mu_{\Lambda,a}\neq 0 \,\mathrm{iff}\, \Lambda = V_0} \\ &= \frac{1}{\mathcal{Z}}\frac{\partial}{\partial \mu_{\Lambda_2,b}}(\tau_{\Lambda_1})^{-1}_{cd}\mu_{\Lambda_1,c}\delta_{ac}\mathcal{Z}\Big|_{\mu_{\Lambda,a}\neq 0 \,\mathrm{iff}\, \Lambda = V_0} \\ &= \tau_{\Lambda_1}^{-1}\delta_{ab}\delta(\Lambda_1,\Lambda_2) + \langle S_a^{\mathrm{phys};\Lambda_1}\rangle\langle S_b^{\mathrm{phys};\Lambda_2}\rangle \\ &= \tau_{\Lambda_1}^{-1}\delta_{ab}\delta(\Lambda_1,\Lambda_2) + \widetilde{\mu}_{V_0}^2\delta(\Lambda_1,V_0)\delta(\Lambda_2,V_0)\,.\end{aligned} \tag{B.68}$$

Defining

$$F_{i,j} \equiv \sum_{a=1}^{D-1} C_{a,i} C_{a,j} = \left(\delta_{ij} - \frac{1}{D}\right), \tag{B.69}$$

we can find the one-point function of the M_{ij}^{phys} by writing it in terms of the physical S variables and applying (B.66)

$$\begin{aligned}\langle M_{ij}^{\text{phys}} \rangle &= C_{ij}^{\text{phys};V_0} \langle S^{\text{phys};V_0} \rangle \\ &= \left(\sqrt{\frac{D-1}{D^3}} - \frac{1}{\sqrt{D(D-1)}} F_{i,j} \right) \widetilde{\mu}_{V_0}.\end{aligned} \tag{B.70}$$

Similarly, we can find the two-point function of the M_{ij}^{phys} variables by writing each M^{phys} in terms of the physical S variables and then using (B.66) and (B.68) to evaluate each of the resulting expectation values

$$\begin{aligned}\langle M_{ij}^{\text{phys}} M_{kl}^{\text{phys}} \rangle =& \left(\sqrt{\frac{D-1}{D^3}} - \frac{1}{\sqrt{D(D-1)}} F_{i,j} \right) \left(\sqrt{\frac{D-1}{D^3}} - \frac{1}{\sqrt{D(D-1)}} F_{k,l} \right) \widetilde{\mu}_{V_0}^2 + \frac{1}{D} \left(\frac{1}{D-1} F_{i,j} F_{k,l} - \frac{1}{D} \left(F_{i,j} + F_{k,l} \right) + \frac{D-1}{D^2} \right) \tau_{V_0}^{-1} \\ &+ \frac{1}{2(D-2)} \left(1 - \delta_{ij}\right) \left(1 - \delta_{kl}\right) \left(F_{i,k} + F_{j,k} + F_{i,l} + F_{j,l} \right) \tau_{V_H}^{-1} \\ &+ \left(\frac{1}{2} F_{i,k} F_{j,l} + \frac{1}{2} F_{i,l} F_{j,k} - \frac{D}{D-2} \sum_{p,q=1}^{D} F_{i,p} F_{j,p} F_{k,q} F_{l,q} F_{p,q} - \frac{1}{D-1} F_{i,j} F_{k,l} \right) \tau_{V_2}^{-1}.\end{aligned} \tag{B.71}$$

Evaluating these expressions for the linear and quadratic permutation invariant observables gives

$$\sum_{i,j} \langle M_{ij}^{\text{phys}} \rangle = \sqrt{D(D-1)} \widetilde{\mu}_{V_0}, \tag{B.72}$$

and

$$\sum_{i,j} \langle M_{ij}^{\text{phys}} M_{ij}^{\text{phys}} \rangle = \widetilde{\mu}_{V_0}^2 + \tau_{V_0}^{-1} + (D-1) \tau_{V_H}^{-1} + \frac{D(D-3)}{2} \tau_{V_2}^{-1}, \tag{B.73}$$

$$\sum_{i,j,k} \langle M_{ij}^{\text{phys}} M_{jk}^{\text{phys}} \rangle_{\text{conn}} = (D-1) \widetilde{\mu}_{V_0}^2 + (D-1) \tau_{V_0}^{-1} + \frac{(D-1)(D-2)}{2} \tau_{V_H}^{-1}, \tag{B.74}$$

$$\sum_{i,j,k,l} \langle M_{ij}^{\text{phys}} M_{kl}^{\text{phys}} \rangle_{\text{conn}} = D(D-1) \widetilde{\mu}_{V_0}^2 + D(D-1) \tau_{V_0}^{-1}. \tag{B.75}$$

B.4.2. Wick's theorem

Since our theory is Gaussian, Wick's theorem allows us to calculate higher point expectation values from the linear and quadratic expectation values. The theorem states that a k-point expectation value is equal to the sum of all possible ways of breaking that expectation value into expectation values of order one and two. For instance, defining the connected piece of the quadratic expectation value as

$$\langle M_{ij}^{\text{phys}} M_{kl}^{\text{phys}} \rangle_{\text{conn}} \equiv \langle M_{ij}^{\text{phys}} M_{kl}^{\text{phys}} \rangle - \langle M_{ij}^{\text{phys}} \rangle \langle M_{kl}^{\text{phys}} \rangle, \tag{B.76}$$

cubic expectation values are given by

$$\begin{aligned}\langle M_{ij}^{\text{phys}} M_{kl}^{\text{phys}} M_{mn}^{\text{phys}} \rangle =& \langle M_{ij}^{\text{phys}} M_{kl}^{\text{phys}} \rangle_{\text{conn}} \langle M_{mn}^{\text{phys}} \rangle + \langle M_{ij}^{\text{phys}} M_{mn}^{\text{phys}} \rangle_{\text{conn}} \langle M_{kl}^{\text{phys}} \rangle + \langle M_{kl}^{\text{phys}} M_{mn}^{\text{phys}} \rangle_{\text{conn}} \langle M_{ij}^{\text{phys}} \rangle \\ &+ \langle M_{ij}^{\text{phys}} \rangle \langle M_{kl}^{\text{phys}} \rangle \langle M_{mn}^{\text{phys}} \rangle.\end{aligned} \tag{B.77}$$

B.5. Embedding within 13-parameter PIGM model

Previous work has solved the most general Gaussian matrix models for general $D \times D$ matrices [1]. These models are defined by 2 linear and 11 quadratic coupling parameters. We expect the $1+3$ parameter model considered in this paper to be embedded in the larger $2+11$ parameter model.

Indeed, comparing the action of the physical model (B.56) written in terms of S variables (B.22)–(B.26) (rewritten here for convenience)

$$\begin{aligned}S =& -\mu_{V_0} \left(\sqrt{\frac{D-1}{D}} S^{V_0;1} - \frac{1}{\sqrt{D}} S^{V_0;2} \right) + \frac{\tau_{V_0}}{2} \left(\sqrt{\frac{D-1}{D}} S^{V_0;1} - \frac{1}{\sqrt{D}} S^{V_0;2} \right)^2 + \frac{\tau_{V_H}}{2} \sum_{a=1}^{D-1} \left(\sqrt{\frac{D-2}{2D}} (S_a^{V_H;1} + S_a^{V_H;2}) - \sqrt{\frac{2}{D}} S_a^{V_H;3} \right)^2 \\ &+ \frac{\tau_{V_2}}{2} \sum_{a=1}^{\dim V_2} S_a^{V_2} S_a^{V_2}.\end{aligned} \tag{B.78}$$

to that of the $2+11$ parameter model written in terms of the same variables

$$S^{\text{PIGMM}} = -\sum_{\alpha=1}^{2} \mu_\alpha^{V_0} S^{V_0;\alpha} + \frac{1}{2} \sum_{\alpha,\beta=1}^{2} S^{V_0;\alpha} (\Lambda_{V_0})_{\alpha\beta} S^{V_0;\beta} + \frac{1}{2} \sum_{a=1}^{D-1} \sum_{\alpha,\beta=1}^{3} S_a^{H;\alpha} (\Lambda_H)_{\alpha\beta} S_a^{H;\beta} + \frac{1}{2} \Lambda_{V_2} \sum_{a=1}^{(D-1)(D-2)/2} S_a^{V_2} S_a^{V_2}$$

$$+\frac{1}{2}\Lambda_{V_3}\sum_{a=1}^{D(D-3)/2} S_a^{V_3} S_a^{V_3}\,. \tag{B.79}$$

we see the point at which the $2+11$ model reduces to the physical model is given by

$$\mu^{V_0} = \mu_{V_0}\left[\sqrt{\tfrac{D-1}{D}} \quad -\tfrac{1}{\sqrt{D}}\right], \tag{B.80}$$

and

$$\Lambda_{V_0} = \tau_{V_0}\left[\tfrac{D-1}{D} \quad -\tfrac{\sqrt{D-1}}{D} - \tfrac{\sqrt{D-1}}{D} \quad \tfrac{1}{D}\right], \quad \Lambda_{V_H} = \tau_{V_H}\begin{bmatrix} \frac{D-2}{2D} & \frac{D-2}{2D} & -2\frac{\sqrt{D-2}}{D} \\ \frac{D-2}{2D} & \frac{D-2}{2D} & -2\frac{\sqrt{D-2}}{D} \\ -2\frac{\sqrt{D-2}}{D} & -2\frac{\sqrt{D-2}}{D} & \frac{2}{D} \end{bmatrix}, \quad \Lambda_{V_2} = \tau_{V_2}, \quad \Lambda_{V_3} = 0. \tag{B.81}$$

In this appendix we have given a detailed construction of the most general permutation invariant matrix model containing diagonally vanishing symmetric matrices as well as formulae for the expectation values of permutation invariant observables. The model is defined by 4 parameters and is relevant to the study of the statistics of correlation matrices.

Appendix C. Inner product calculations

We present here some inner products which are useful in arriving at equations for the physical variables $S^{\text{phys};V_0}$ and $S_a^{\text{phys};V_H}$ obtained in (B.32) and (B.33) of Appendix B.

Using the inner product on the orthonormal basis of the natural representation

$$(e_i, e_j) = \delta_{ij}. \tag{C.1}$$

We calculate the following inner products used to determine the Clebsch coefficients from $V_D \otimes V_D$ to the physical V_0 and V_H irreducible representations

$$(S^{\text{diag};V_0}, S^{V_0;1}), \quad (S^{\text{diag};V_0}, S^{V_0;2}), \quad (S_{a_1}^{\text{diag};V_H}, S_{a_2}^{V_H;1,2}), \quad (S_{a_1}^{\text{diag};V_H}, S_{a_2}^{V_H;3}). \tag{C.2}$$

The normalised representation theory states are given by

$$S^{V_0;1} = \frac{1}{D}\sum_{i,j=1}^{D} e_i \otimes e_j, \tag{C.3}$$

$$S^{V_0;2} = \frac{1}{\sqrt{D-1}}\sum_{a=1}^{D-1} E_a \otimes E_a, \tag{C.4}$$

$$S^{\text{diag};V_0} = \frac{1}{\sqrt{D}}\sum_{i=1}^{D} e_i \otimes e_i, \tag{C.5}$$

$$S_a^{V_H;1,2} = \frac{1}{\sqrt{2D}}\sum_{i=1}^{D}\left(e_i \otimes E_a + E_a \otimes e_i\right), \tag{C.6}$$

$$S_a^{V_H;3} = \frac{1}{2}\sqrt{\frac{D}{D-2}}\sum_{b,c=1}^{D-1}\sum_{i=1}^{D} C_{a,i}C_{b,i}C_{c,i}\left(E_b \otimes E_c + E_c \otimes E_b\right), \tag{C.7}$$

$$S_a^{\text{diag};V_H} = E_a \otimes E_a. \tag{C.8}$$

The non-zero inner products are given by the following

$$\begin{aligned}(S^{\text{diag};V_0}, S^{V_0;1}) &= D^{-\frac{3}{2}}\sum_{i,j,k=1}^{D}(e_i \otimes e_i, e_j \otimes e_k) = D^{-\frac{3}{2}}\sum_{i,j,k=1}^{D}\delta_{ij}\delta_{ik} = \sum_{i=1}^{D} D^{-\frac{3}{2}} \\ &= \frac{1}{\sqrt{D}},\end{aligned} \tag{C.9}$$

$$\begin{aligned}(S^{\text{diag};V_0}, S^{V_0;2}) &= \frac{1}{\sqrt{D}\sqrt{D-1}}\sum_{i=1}^{D}\sum_{a=1}^{D-1}(e_i \otimes e_i, E_a \otimes E_a) \\ &= \frac{1}{\sqrt{D}\sqrt{D-1}}\sum_{i,j,k=1}^{D}\sum_{a=1}^{D-1} C_{a,j}C_{a,k}(e_i \otimes e_i, e_j \otimes e_k) \\ &= \frac{1}{\sqrt{D}\sqrt{D-1}}\sum_{i,j,k=1}^{D}\sum_{a=1}^{D-1} C_{a,j}C_{a,k}\delta_{ij}\delta_{ik}\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{D}\sqrt{D-1}} \sum_{j=1}^{D} \sum_{a=1}^{D-1} C_{a,j} C_{a,j} \\
&= \sum_{a=1}^{D-1} \frac{1}{\sqrt{D}\sqrt{D-1}} \\
&= \frac{\sqrt{D-1}}{\sqrt{D}},
\end{aligned} \tag{C.10}
$$

$$
\begin{aligned}
(S_{a_1}^{\mathrm{diag};V_H}, S_{a_2}^{V_H;1,2}) &= \frac{1}{\sqrt{2D}} \sum_{i,j,k=1}^{D} C_{a_1,j} C_{a_2,k} \left(e_j \otimes e_j, (e_i \otimes e_k + e_k \otimes e_i)\right) \\
&= \frac{1}{\sqrt{2D}} \sum_{i,j,k=1}^{D} C_{a_1,j} C_{a_2,k} \left(\delta_{ij}\delta_{jk} + \delta_{jk}\delta_{ij}\right) \\
&= \sqrt{\frac{2}{D}} \sum_{i,j=1}^{D} C_{a_1,j} C_{a_2,j} \delta_{ij} \\
&= \sqrt{\frac{2}{D}} \delta_{a_1 a_2},
\end{aligned} \tag{C.11}
$$

$$
\begin{aligned}
(S_{a_1}^{\mathrm{diag};V_H}, S_{a_2}^{V_H;3}) &= \frac{1}{2}\sqrt{\frac{D}{D-2}} \sum_{b,c=1}^{D-1} \sum_{i,j,k,l=1}^{D} C_{a_1,i} C_{b,i} C_{c,i} C_{a_2,j} C_{b,k} C_{c,l} \left(e_j \otimes e_j, (e_k \otimes e_l + e_l \otimes e_k)\right) \\
&= \sqrt{\frac{D}{D-2}} \sum_{b,c=1}^{D-1} \sum_{i,j,k,l=1}^{D} C_{a_1,i} C_{b,i} C_{c,i} C_{a_2,j} C_{b,k} C_{c,l} \delta_{jk} \delta_{jl} \\
&= \sqrt{\frac{D}{D-2}} \sum_{b,c=1}^{D-1} \sum_{i,j=1}^{D} C_{a_1,i} C_{b,i} C_{c,i} C_{a_2,j} C_{b,j} C_{c,j} \\
&= \sqrt{\frac{D}{D-2}} \sum_{i,j=1}^{D} C_{a_1,i} C_{a_2,j} \left(\delta_{ij} - \frac{1}{D}\right)\left(\delta_{ij} - \frac{1}{D}\right) \\
&= \sqrt{\frac{D}{D-2}} \sum_{i,j=1}^{D} C_{a_1,i} C_{a_2,j} \left(\delta_{ij}\left(1 - \frac{2}{D}\right) + \frac{1}{D^2}\right) \\
&= \sqrt{\frac{D}{D-2}} \left[\sum_{i=1}^{D} C_{a_1,i} C_{a_2,i} \left(1 - \frac{2}{D}\right) + \frac{1}{D^2} \sum_{i,j=1}^{D} C_{a_1,i} C_{a_2,j}\right] \\
&= \sqrt{\frac{D}{D-2}} \left(1 - \frac{2}{D}\right) \delta_{a_1,a_2}.
\end{aligned} \tag{C.12}
$$

Appendix D. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.physa.2024.130015.

References

[1] S. Ramgoolam, Permutation invariant Gaussian matrix models, Nuclear Phys. B 945 (2019) 114682.
[2] D. Kartsaklis, S. Ramgoolam, M. Sadrzadeh, Linguistic matrix theory, Ann. Inst. H. Poincare D Comb. Phys. Interact. 6 (3) (2019) 385–426.
[3] M.A. Huber, A. Correia, S. Ramgoolam, M. Sadrzadeh, Permutation invariant matrix statistics and computational language tasks, 2022.
[4] S. Ramgoolam, M. Sadrzadeh, L. Sword, Gaussianity and typicality in matrix distributional semantics, Ann. Inst. H. Poincare D Comb. Phys. Interact. 9 (1) (2022) 1–45.
[5] R. Gębarowski, P. Oświęcimka, M. Wątorek, S. Drożdż, Detecting correlations and triangular arbitrage opportunities in the forex by means of multifractal detrended cross-correlations analysis, Nonlinear Dynam. 98 (3) (2019) 2349–2364.
[6] E.P. Wigner, Characteristic vectors of bordered matrices with infinite dimensions, Ann. of Math. 62 (3) (1955) 548–564.
[7] F.J. Dyson, A Brownian-motion model for the eigenvalues of a random matrix, J. Math. Phys. 3 (6) (1962) 1191–1198.
[8] T. Guhr, A. Muller-Groeling, H.A. Weidenmuller, Random matrix theories in quantum physics: Common concepts, Phys. Rep. 299 (1998) 189–425.
[9] A. Edelman, Y. Wang, Random matrix theory and its innovative applications, Fields Inst. Commun. 66 (2013) 91–116.
[10] G. Wijnholds, M. Sadrzadeh, S. Clark, Representation learning for type-driven composition, in: Proceedings of the 24th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Online, 2020, pp. 313–324.
[11] T. Anderson, An Introduction to Multivariate Statistical Analysis, third ed., Wiley-Interscience, 2003.
[12] M. Potters, J.-P. Bouchaud, A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists, Cambridge University Press, 2020.
[13] G. Marti, Corrgan: Sampling realistic financial correlation matrices using generative adversarial networks, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2020, pp. 8459–8463.
[14] L. Laloux, P. Cizeau, J.-P. Bouchaud, M. Potters, Noise dressing of financial correlation matrices, Phys. Rev. Lett. 83 (1999) 1467–1470.
[15] L. Laloux, P. Cizeau, M. Potters, J.-P. Bouchaud, Random matrix theory and financial correlations, Int. J. Theor. Appl. Finance 03 (03) (2000) 391–397.

[16] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, T. Guhr, H.E. Stanley, Random matrix approach to cross correlations in financial data, Phys. Rev. E 65 (6) (2002) 066126.
[17] V.A. Marčenko, L.A. Pastur, Distribution of eigenvalues for some sets of random matrices, Math. USSR-Sbornik 1 (4) (1967) 457.
[18] Z. Burda, J. Jurkiewicz, Signal and noise in financial correlation matrices, Phys. A 344 (1) (2004) 67–72, Applications of Physics in Financial Analysis 4 (APFA4).
[19] G. Livan, S. Alfarano, E. Scalas, Fine structure of spectral properties for random correlation matrices: An application to financial markets, Phys. Rev. E 84 (2011) 016113.
[20] J.-P. Bouchaud, M. Potters, Financial applications of random matrix theory: A short review, in: The Oxford Handbook of Random Matrix Theory, 2009.
[21] J. Bun, J.-P. Bouchaud, M. Potters, Cleaning large correlation matrices: Tools from random matrix theory, Phys. Rep. 666 (2017) 1–109.
[22] M.M.L. de Prado, Machine Learning for Asset Managers, Cambridge University Press, 2020.
[23] M. Potters, J.-P. Bouchaud, L. Laloux, Financial applications of random matrix theory: Old laces and new pieces, Acta Phys. Pol. B 36 (9) (2005) 2767.
[24] G. Barnes, A. Padellaro, S. Ramgoolam, Permutation invariant Gaussian two-matrix models, J. Phys. A 55 (14) (2022) 145202.
[25] M. Hamermesh, Group Theory and Its Application to Physical Problems, Courier Corporation, 2012.
[26] Integral, TrueFX, 2021, https://www.truefx.com/. (Accessed: 15 October 2021).
[27] O.E. Barndorff-Nielsen, P.R. Hansen, A. Lunde, N. Shephard, Realized kernels in practice: Trades and quotes, Econom. J. 12 (3) (2009) C1–C32.
[28] O.E. Barndorff-Nielsen, P.R. Hansen, A. Lunde, N. Shephard, Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading, J. Econometrics 162 (2) (2011) 149–169.
[29] P. Malliavin, M.E. Mancino, A Fourier transform method for nonparametric estimation of multivariate volatility, Ann. Statist. 37 (4) (2009) 1983–2010.
[30] Y. Ait-Sahalia, J. Fan, D. Xiu, High-frequency covariance estimates with noisy and asynchronous financial data, J. Amer. Statist. Assoc. 105 (492) (2010) 1504–1517.
[31] M. Bibinger, Efficient covariance estimation for asynchronous noisy high-frequency data, Scand. J. Stat. 38 (1) (2011) 23–45.
[32] N. Hautsch, L.M. Kyj, R.C.A. Oomen, A blocking and regularization approach to high-dimensional realized covariance estimation, J. Appl. Econometrics 27 (4) (2012) 625–645.
[33] K. Christensen, M. Podolskij, M. Vetter, On covariation estimation for multivariate continuous Itô semimartingales with noise in non-synchronous observation schemes, J. Multivariate Anal. 120 (2013) 59–84.
[34] M. Bibinger, N. Hautsch, P. Malec, M. Reiß, Estimating the quadratic covariation matrix from noisy observations: Local method of moments and efficiency, Ann. Statist. 42 (4) (2014) 1312–1346.
[35] T. Hayashi, N. Yoshida, On covariance estimation of non-synchronously observed diffusion processes, Bernoulli 11 (2) (2005) 359–379.
[36] C. Aggarwal, Outlier Analysis, Springer International Publishing, 2016.
[37] FXSSI, FXSSI, 2022, https://fxssi.com/high-impact-forex-news. (Accessed: 08 December 2022).
[38] FXSTREET, FXSTREET, 2022, https://www.fxstreet.com/economic-calendar. (Accessed: 01 July 2022).
[39] H. Ghorbani, Mahalanobis distance and its application for detecting multivariate outliers, Facta Univ. Ser. Math. Inform. 34 (3) (2019) 583–595.
[40] A. Edelman, A. Guionnet, S. Pé ché, Beyond universality in random matrix theory, Ann. Appl. Probab. 26 (3) (2016).
[41] F. Luo, Y. Yang, J. Zhong, H. Gao, L. Khan, D.K. Thompson, J. Zhou, Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory, BMC Bioinform. 8 (1) (2007) 1–17.
[42] T. Tao, V. Vu, Random matrices: universality of local eigenvalue statistics, 2011.
[43] M. Marsili, Dissecting financial markets: Sectors and states, Quant. Finance 2 (4) (2002) 297.
[44] D. Hendricks, T. Gebbie, D. Wilcox, Detecting intraday financial market states using temporal clustering, Quant. Finance 16 (11) (2016) 1657–1678.
[45] N.J. Higham, Computing the nearest correlation matrix—A problem from finance, IMA J. Numer. Anal. 22 (3) (2002) 329–343.
[46] W. Fulton, J. Harris, Representation Theory: A First Course, vol. 129, Springer Science & Business Media, 2013.
[47] B.E. Sagan, The Symmetric Group: Representations, Combinatorial Algorithms, and Symmetric Functions, Springer Science & Business Media, 2013.
[48] G. Barnes, A. Padellaro, S. Ramgoolam, Hidden symmetries and large N factorisation for permutation invariant matrix observables, J. High Energy Phys. 2022 (8) (2022) 1–33.
[49] F. Gabriel, Combinatorial theory of permutation-invariant random matrices I: partitions, geometry and renormalization, 2015.
[50] F. Gabriel, A combinatorial theory of random matrices III: random walks on $\mathfrak{S}(N)$, ramified coverings and the $\mathfrak{S}(\infty)$ Yang-Mills measure, 2015.
[51] OEIS, OEIS sequence A050535, 2023, https://oeis.org/search?q=A050535. (Accessed: 19 December 2023).