Marenko-Pastur law for Kendall’s tau

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1214/17-ECP59">http://dx.doi.org/10.1214/17-ECP59</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Institute of Mathematical Statistics</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Sun Feb 03 03:35:35 EST 2019</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/115158">http://hdl.handle.net/1721.1/115158</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Creative Commons Attribution 4.0 International License</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td><a href="http://creativecommons.org/licenses/by/4.0/">http://creativecommons.org/licenses/by/4.0/</a></td>
</tr>
</tbody>
</table>
DOI: 10.1214/17-ECP59
ISSN: 1083-589X

ELECTRONIC COMMUNICATIONS in PROBABILITY

Marčenko-Pastur law for Kendall’s tau

Afonso S. Bandeira*  Asad Lodhia †  Philippe Rigollet ‡

Abstract

We prove that Kendall’s Rank correlation matrix converges to the Marčenko Pastur law, under the assumption that observations are i.i.d random vectors $X_1, \ldots, X_n$ with components that are independent and absolutely continuous with respect to the Lebesgue measure. This is the first result on the empirical spectral distribution of a multivariate $U$-statistic.

Keywords: statistics; random matrix theory.
AMS MSC 2010: 60B20; 62H20.


1 Introduction

Estimating the association between two random variables $Y, Z \in \mathbb{R}$ is a central statistical problem. As such, many methods have been proposed, most notably Pearson’s correlation coefficient. While this measure of association is well suited to the case where $(Y, Z)$ is Gaussian, it may be inaccurate in other cases. This observation has led statisticians to consider other measures of association such as Spearman’s $\rho$ and Kendall’s $\tau$ that are proven to be more robust to heavy-tailed distributions (see, e.g., [11]). In a multivariate setting, covariance and correlation matrices are preponderant tools to understand the interaction between variables. They are also used as building blocks for more sophisticated statistical methods arising in principal component analysis or graphical models for example.

The past decade has witnessed a fertile interaction between random matrix theory and high-dimensional statistics (see [13] for a recent survey). Indeed, in high-dimensional settings, traditional asymptotics where the sample size tends to infinity but the dimension of the model is held fixed fail to capture a delicate interaction between sample size and dimension. Random matrix theory has allowed statisticians and practitioners alike to gain valuable insight on a variety of multivariate problems.

The terminology “Wishart matrices” is often, though sometimes abusively, used to refer to $p \times p$ random matrices of the form $X^\top X/n$, where $X$ is an $n \times p$ random matrix with independent rows (throughout this paper we restrict our attention to real random matrices). The simplest example arises where $X$ has i.i.d standard Gaussian entries but the main characteristics are shared by a much wider class of random matrices. This

*NYU, Courant Institute of Mathematical Sciences, USA. E-mail: bandeira@cims.nyu.edu
†MIT, Department of Mathematics, USA. E-mail: lodhia@math.mit.edu
‡MIT, Department of Mathematics, USA. E-mail: rigollet@math.mit.edu
Marčenko-Pastur law for Kendall’s tau

universality phenomenon manifests itself in various aspects of the limit distribution, and in particular in the limiting behavior of the empirical spectral distribution of the matrix. Let \( W = X^\top X/n \) be a \( p \times p \) Wishart matrix and let \( \lambda_1, \ldots, \lambda_p \) be its eigenvalues; then the empirical spectral distribution \( \hat{\mu}_p^W \) of \( W \) is the distribution on \( \mathbb{R} \) defined by the following mixture of Dirac point masses at the \( \lambda_j \)'s:

\[
\hat{\mu}_p^W = \frac{1}{p} \sum_{k=1}^{p} \delta_{\lambda_k}.
\]

Assuming that the entries of \( X \) are independent, centered and of unit variance, it can be shown that \( \hat{\mu}_p^W \) converges weakly to the Marčenko-Pastur distribution under weak moment conditions (see for example [1, Theorem 3.10]).

While this development alone has led to important statistical advances, it fails to capture more refined notions of correlation, notably more robust ones involving dependent observations such as rank-based statistics. A first step in this direction was made in [14], where the matrix \( X \) is assumed to have independent rows with isotropic distribution. More recently, this result was extended in [2, 12] and covers the case of Spearman’s \( \rho \) matrix which is based on ranks and is also a Wishart matrix of the form \( X^\top X/n \).

The main contribution of this paper is to derive the limiting distribution of Kendall’s \( \tau \) matrix, a cousin of Spearman’s \( \rho \) matrix which is not of the Wishart type but rather is a matrix whose entries are \( U \)-statistics. The Kendall \( \tau \) matrix is a popular surrogate for correlation matrices but an understanding of the fluctuations of its eigenvalues is still missing. Interestingly, the Marčenko-Pastur law has been used as a heuristic for Kendall’s \( \tau \) without justification in certain financial applications [3].

As it turns out, the limit of the empirical spectral distribution of \( \tau \), denoted by \( \hat{\mu}_p^\tau \), is not exactly the Marčenko-Pastur law, but rather an affine transformation of it. Our main theorem below gives the precise form of this transformation.

**Theorem 1.1.** Let \( X_1, \ldots, X_n \) be \( n \) independent random vectors in \( \mathbb{R}^p \) whose components \( X_i(k) \) are independent random variables that have a density with respect to the Lebesgue measure on \( \mathbb{R} \). Then as \( n \to \infty \) and \( \frac{p}{n} \to \gamma > 0 \) the empirical spectral distribution of \( \tau \) converges in probability to

\[
\frac{2}{3} Y + \frac{1}{3},
\]

where \( Y \) is distributed according to the standard Marčenko-Pastur law with parameter \( \gamma \) (see Theorem A.1 for the appropriate definition).

The proof of Theorem 1.1 combines standard tools from random matrix theory and asymptotic theory of \( U \)-statistics. Specifically, we apply a multivariate version of Hoeffding’s decomposition, also known as the ANOVA decomposition, to the matrix \( \tau \). This allows us to represent \( \tau \) as a sum of three terms. The first term is the deterministic identity matrix coming from the diagonal of \( \tau \). The identity matrix only causes an additive shift of the spectrum. The second term is a sum of independent identically distributed rank-one matrices whose diagonals are set to 0. These terms cause the appearance of the Marčenko-Pastur law by an application of Theorem A.1 once the missing diagonal is accounted for. The last residual terms vanish in Frobenius norm, so by Lemma A.2 they make no asymptotic contribution to the spectrum. Figure 1 illustrates numerically the result of Theorem 1.1.

Notation: For any integer \( k \geq 1 \) we write \([k] = \{1, \ldots, k\}\). We denote by \( I_p \) the identity matrix of \( \mathbb{R}^p \). For a vector \( x \in \mathbb{R}^p \), we denote by \( x(j) \) its \( j \)th coordinate. For any \( p \times p \) matrix \( M \), we denote by \( \text{diag}(M) \) the \( p \times p \) diagonal matrix with the same diagonal
elements as $M$ and we define $D_0(M) = M - \text{diag}(M)$. In other words, the operator $D_0$ replaces each diagonal element of a matrix by zero. We denote $\text{sign}(x)$ the sign of $x \in \mathbb{R}$ with convention that $\text{sign}(0) = 1$. The Frobenius norm of a $p \times p$ matrix $H$ is denoted by $\|H\|_F$ and we recall that $\|H\|_F^2 := \text{Tr}(H^\top H)$. Finally, we define $\text{Unif}([a, b])$ to be the uniform distribution on the interval $[a, b]$.

## 2 Kendall’s tau

The (univariate) Kendall $\tau$ statistic [5, 9, 10, 7] is defined as follows. Let $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ be $n$ independent samples of a pair $(Y, Z) \in \mathbb{R} \times \mathbb{R}$ of real-valued random variables. Then the (empirical) Kendall $\tau$ between $Y$ and $Z$ is defined as

$$\tau(Y, Z) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \text{sign}(Y_i - Y_j) \cdot \text{sign}(Z_i - Z_j).$$

The statistic $\tau$ takes values in $[-1, 1]$ and it can be expressed as

$$\tau = \frac{1}{\binom{n}{2}} \left( \# \{ \text{concordant pairs} \} - \# \{ \text{discordant pairs} \} \right),$$

where a pair $(i, j)$ is said to be concordant if $Y_i - Y_j$ and $Z_i - Z_j$ have the same sign and discordant otherwise.

It is known that the Kendall $\tau$ statistic is asymptotically Gaussian [7]. Specifically, if $Y$ and $Z$ are independent, then as $n \to \infty$,

$$\sqrt{n}\tau(Y, Z) \rightsquigarrow N\left(0, \frac{4}{9}\right).$$

(2.1)

This property has been central to construct independence tests between two random variables $X$ and $Y$ (see, e.g., [8]).

Kendall’s $\tau$ statistic can be extended to the multivariate case. Let $X_1, \ldots, X_n$, be $n$ independent copies of a random vector $X \in \mathbb{R}^p$, with coordinates $X(1), \ldots, X(p)$. The
(empirical) Kendall $\tau$ matrix of $X$ is defined to be the $p \times p$ matrix whose entries $\tau_{kl}$ are given by

$$\tau_{kl} := \tau(X(k), X(l)) = \frac{1}{n} \sum_{1 \leq i < j \leq n} \text{sign} \left( X_i(k) - X_j(k) \right) \cdot \text{sign} \left( X_i(l) - X_j(l) \right) \quad 1 \leq k, l \leq p. \tag{2.2}$$

Note that $\tau$ can be written as the sum of $\binom{n}{2}$ rank-one random matrices:

$$\tau = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \text{sign} \left( X_i - X_j \right) \otimes \text{sign} \left( X_i - X_j \right), \tag{2.3}$$

where the sign function is taken entrywise. Our goal is to describe aspects of the asymptotic behavior of $\tau$ when the coordinates of $X$ are independent. This case is often referred to as the null case in statistics to emphasize the absence of a signal.

It is easy to see that $\tau_{ii} = 1$ for all $i$. Together with (2.1), it implies that the matrix

$$\tau = \frac{3}{2} \tau - \frac{1}{2} I_p$$

is such that $\mathbb{E}[\sqrt{n}(\tau - I_p)] \to 0$ and $\text{Var}[\sqrt{n}(\tau - I_p)] \to 1$, if $i \neq j$ as $n \to \infty$. This suggests that if the empirical spectral distribution of $\tau$ converges to a Marčenko-Pastur distribution, it should be a standard Marčenko-Pastur distribution. This heuristic argument supports the affine transformation arising in Theorem 1.1. However, the matrix $\tau$ is not Wishart and the Marčenko-Pastur limit distribution does not follow from standard arguments. Nevertheless, Kendall’s $\tau$ is a $U$-statistic which are known to satisfy the weakest form of universality, namely a Central Limit Theorem under general conditions [6, 4]. In this paper, we show that in the case of the Kendall $\tau$ matrix, this universality phenomenon extends to the empirical spectral distribution.

3 Proof of Theorem 1.1

For any pair $(i, j)$ such that $1 \leq i, j \leq n$ and $i \neq j$, let $A_{(i,j)} \in \mathbb{R}^p$ be the vector

$$A_{(i,j)} := \text{sign} \left( X_i - X_j \right),$$

and recall from (2.3) that

$$\tau = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} A_{(i,j)} \otimes A_{(i,j)}. \tag{3.1}$$

Akin to most asymptotic results on $U$-statistics, we utilize a variant of Hoeffding’s (a.k.a. Efron-Stein, a.k.a ANOVA) decomposition [6]:

$$A_{(i,j)} = \tilde{A}_{(i,j)} + \hat{A}_{(i,j)} + \check{A}_{(i,j)}$$

where

$$A_{(i,:)} := \mathbb{E}[A_{(i,j)} | X_j], \quad A_{(:,j)} := \mathbb{E}[A_{(i,j)} | X_i], \quad \text{and} \quad A_{(i,:)} := A_{(i,j)} - A_{(:,j)} - A_{(i,:)}. \nonumber$$

It is easy to check that each of the random vectors in the right-hand side of (3.1) are centered and are orthogonal to each other with respect to the inner product $\mathbb{E}[V^TW]$ where $V, W \in \mathbb{R}^p$. These random vectors can be expressed conveniently thanks to the following Lemma.

**Lemma 3.1.** For $k \in [p]$, let $F_k$ denote the cumulative distribution function of $X(k)$. Fix $i \in [n]$ and let $U_i \in \mathbb{R}^p$ be a random vector with $k$th coordinate given by $U_i(k) = 2F_k(X_i(k)) - 1 \sim \text{Unif}([-1, 1])$. Then

$$\tilde{A}_{(i,:)} = -\hat{A}_{(:,i)} = U_i.$$
Proof. For any \( i \in [n] \), observe that since the components of \( X \) have a density, then \( \mathbb{P}(X_i(k) = X_j(k)|X_i) = 0 \) so that

\[
\mathbb{E}[\text{sign}(X_i(k) - X_j(k))|X_i] = \mathbb{P}(X_i(k) > X_j(k)|X_i) - \mathbb{P}(X_i(k) < X_j(k)|X_i) = 2F_k(X_i(k)) - 1.
\]

The observation that \( \hat{A}_{(i,\cdot)} = -\hat{A}_{(\cdot,i)} \) follows from the fact that \( A_{(i,j)} = -A_{(j,i)} \). \( \square \)

Using (3.1) we obtain the representation:

\[
A_{(i,j)} \otimes A_{(i,j)} = M^{(1)}_{(i,j)} + M^{(2)}_{(i,j)} + \left( M^{(2)}_{(i,j)} \right)^T + M^{(3)}_{(i,j)},
\]

where

\[
M^{(1)}_{(i,j)} := I_p + D_0[A_{(i,\cdot)} + A_{(\cdot,j)}] \otimes \{ A_{(i,\cdot)} + A_{(\cdot,j)} \},
\]

\[
M^{(2)}_{(i,j)} := D_0[A_{(i,j)} \otimes \{ A_{(i,\cdot)} + A_{(\cdot,j)} \}],
\]

\[
M^{(3)}_{(i,j)} := D_0[A_{(i,j)} \otimes A_{(i,j)}].
\]

By the relation \( \hat{A}_{(i,\cdot)} = -\hat{A}_{(\cdot,i)} \) from Lemma 3.1 we have

\[
\sum_{1 \leq i<j \leq n} \{ A_{(i,\cdot)} + A_{(\cdot,j)} \} \otimes \{ A_{(i,\cdot)} + A_{(\cdot,j)} \} = (n-1) \sum_{i=1}^n A_{(i,\cdot)} \otimes A_{(i,\cdot)} - \sum_{(i,j) \in [n]^2: i \neq j} A_{(i,\cdot)} \otimes A_{(j,\cdot)}.
\]

Using Lemma 3.1 yields:

\[
\frac{1}{n} \sum_{1 \leq i<j \leq n} M^{(1)}_{(i,j)} = I_p + \frac{2}{n} \sum_{i=1}^n D_0[U_i \otimes U_i] - \frac{1}{(2)} D_0 \left[ \sum_{(i,j) \in [n]^2: i \neq j} U_i \otimes U_j \right]. \quad (3.3)
\]

Next, note that, the coordinates of each \( U_i \), \( i = 1, \ldots, n \) are mutually independent and \( \mathbb{E}[U_i] = 0 \) so that

\[
\mathbb{E}[U_i \otimes U_i] = \mathbb{E}[T^2] I_p = \frac{1}{3} I_p, \quad (3.4)
\]

where \( T \sim \text{Unif}([-1,1]) \). Theorem A.1 implies that as \( n \to \infty \) and \( \frac{2}{n} \to \gamma > 0 \), the empirical spectral distribution of

\[
\frac{2}{n} \sum_{i=1}^n U_i \otimes U_i
\]

converges in probability to \( (2/3)Y \), where \( Y \) is distributed according to the standard Marčenko-Pastur law with parameter \( \gamma \). Moreover,

\[
\frac{1}{p} \mathbb{E} \left\| \frac{2}{n} \sum_{i=1}^n \text{diag}(U_i \otimes U_i) - \frac{2}{3} I_p \right\|_F^2 = \frac{4}{pn^2} \sum_{k=1}^p \mathbb{E} \left( \sum_{i=1}^n \{ U_i(k)^2 - \mathbb{E}[U_i(k)^2] \} \right)^2 \leq C/n \to 0,
\]

\[
\frac{1}{p} \mathbb{E} \left\| D_0 \left[ \sum_{(i,j) \in [n]^2: i \neq j} U_i \otimes U_j \right] \right\|_F^2 = \frac{1}{p \binom{n}{2}^2} \sum_{(k,l) \in [p]^2: k \neq l} \mathbb{E} \left( \sum_{(i,j) \in [n]^2: i \neq j} U_i(k)U_j(l) \right)^2 \leq \frac{Cp}{n^2} \to 0,
\]

for some constant \( C > 0 \) independent of \( n \). By Lemma A.2, together with (3.3) and the triangle inequality, we get the following result.
Theorem A.1

To see this, expand \(E_{\mathbb{C}^2} \mathbf{a} \mathcal{I} \mathbf{E} \mathbf{b} \mathcal{I} \), and satisfy \(X\) copies of a random vector of (3.5) is bounded by:

\[
\frac{1}{n} \sum_{1 \leq i < j \leq n} M_{(i,j)}^{(1)}
\]

and we also have \(I E\) some \(C > 0\). Lemma 3.1. Note that when \((3.2)\) and the triangle inequality:

To show that (3.5) vanishes asymptotically, notice that the collection of matrices \(\{M_{(i,j)}^{(3)}\}_{1 \leq i < j \leq n}\) satisfies

\[
\mathbb{E} \mathbf{Tr}\left\{\left(\mathbf{M}_{(i,j)}^{(3)}\right)^{\top} \mathbf{M}_{(i',j')}^{(3)}\right\} = \begin{cases} 
\mathbb{E}\left\|\mathbf{M}_{(i,j)}^{(3)}\right\|_F^2 & \text{for } (i,j) = (i',j') \\
0 & \text{otherwise}
\end{cases}
\]

(3.6)

To see this, expand

\[
\mathbb{E} \mathbf{Tr}\left\{\left(\mathbf{M}_{(i,j)}^{(3)}\right)^{\top} \mathbf{M}_{(i',j')}^{(3)}\right\} = \sum_{(k,l) \in [p]^2 : k \neq l} \left(\mathbb{E}\left[\left(\mathbf{A}_{(i,j)}(k) - U_i(k) + U_j(k)\right)\left(\mathbf{A}_{(i',j')}^{(3)}(k) - U_i'(k) + U_j'(k)\right)\right]\right) \\
\times \mathbb{E}\left[\left(\mathbf{A}_{(i,j)}(l) - U_i(l) + U_j(l)\right)\left(\mathbf{A}_{(i',j')}^{(3)}(l) - U_i'(l) + U_j'(l)\right)\right]
\]

and notice that each expectation is zero unless \((i,j) = (i',j')\) by Tower property and Lemma 3.1. Note that when \((i,j) = (i',j')\), the expression (3.7) is bounded by \(Cp^2\) for some \(C > 0\). The equation (3.6) also holds for the collection of matrices \(\{M_{(i,j)}^{(2)}\}_{1 \leq i < j \leq n}\) and we also have \(\mathbb{E}\|M_{(i,j)}^{(2)}\|^2 \leq Cp^2\) by a similar argument. Therefore the right side of (3.5) is bounded by:

\[
\frac{Cp}{n^2} \times \text{card}\{(i,j,i',j') \in [n]^4 : (i,j) = (i',j')\} \leq \frac{Cp}{n^2},
\]

for some constant \(C > 0\), which vanishes as \(n \to \infty\). This concludes the proof of Theorem 1.1.

A  The standard Marčenko-Pastur law

We include here the definition of the standard Marčenko-Pastur law and a bound on the distance between empirical spectral distributions of two matrices.

Theorem A.1 (Marčenko-Pastur law [1, Theorem 3.6]). Let \(X_1, \ldots, X_n\) be independent copies of a random vector \(X \in \mathbb{R}^p\) whose components \(X(k), k = 1, \ldots, p\) are i.i.d and satisfy \(\mathbb{E}[X(k) = 0], \mathbb{V}[X(k)] = 1\). Suppose that \(n \to \infty, \frac{p}{n} \to \gamma > 0\) and define \(a = (1 - \sqrt{\gamma})^2, b = (1 + \sqrt{\gamma})^2\). Then the empirical spectral distribution of the matrix

\[
\frac{1}{n} \sum_{i=1}^{n} X_i \otimes X_i,
\]
Marčenko-Pastur law for Kendall’s tau

converges almost surely to the standard Marčenko-Pastur law which has density:

\[ p_\gamma(x) = \begin{cases} 
\frac{1}{2\pi x} \sqrt{(b-x)(x-a)}, & \text{if } a \leq x \leq b, \\
0, & \text{otherwise,}
\end{cases} \]

and has a point mass \( 1 - \frac{1}{\gamma} \) at the origin if \( \gamma > 1 \).

**Lemma A.2** (\cite[Corollary A.41]{BaiSilverstein2010}). Let \( A \) and \( B \) be two \( p \times p \) normal matrices, with empirical spectral distributions \( \hat{\mu}^A \) and \( \hat{\mu}^B \). Then

\[ L(\hat{\mu}^A, \hat{\mu}^B) \leq \frac{1}{p} \| A - B \|_F^2, \]

where \( L(\hat{\mu}^A, \hat{\mu}^B) \) is the Lévy distance between the distribution functions \( \hat{\mu}^A \) and \( \hat{\mu}^B \).

**References**


**Acknowledgments.** The authors would like to thank Alice Guionnet for helpful comments and discussions. A. S. B. was supported by Grant NSF-DMS-1541100. Part of this work was done while A. S. B. was with the Department of Mathematics at MIT. A. L. was supported by Grant NSF-MS-1307704. P. R was supported by Grants NSF CAREER DMS-1541099, NSF DMS-1541100, DARPA W911NF-16-1-0551, ONR N00014-17-1-2147 and a grant from the MIT NEC Corporation. The authors would also like to thank the anonymous reviewer for their helpful comments.
Advantages of publishing in EJP-ECP

- Very high standards
- Free for authors, free for readers
- Quick publication (no backlog)
- Secure publication (LOCKSS\(^1\))
- Easy interface (EJMS\(^2\))

Economical model of EJP-ECP

- Non profit, sponsored by IMS\(^3\), BS\(^4\), ProjectEuclid\(^5\)
- Purely electronic

Help keep the journal free and vigorous

- Donate to the IMS open access fund\(^6\) (click here to donate!)
- Submit your best articles to EJP-ECP
- Choose EJP-ECP over for-profit journals

\(^{1}\)LOCKSS: Lots of Copies Keep Stuff Safe http://www.lockss.org/
\(^{3}\)IMS: Institute of Mathematical Statistics http://www.imstat.org/
\(^{4}\)BS: Bernoulli Society http://www.bernoulli-society.org/
\(^{5}\)Project Euclid: https://projecteuclid.org/
\(^{6}\)IMS Open Access Fund: http://www.imstat.org/publications/open.htm