Theory of the GMM Kernel

Ping Li
Department of Statistics and Biostatistics
Department of Computer Science
Rutgers University
Piscataway, NJ 08854, USA
pingli@stat.rutgers.edu

Cun-Hui Zhang
Department of Statistics and Biostatistics
Rutgers University
Piscataway, NJ 08854, USA
cunhui@stat.rutgers.edu

Abstract

We[1] develop some theoretical results for a robust similarity measure named "generalized min-max" (GMM). This similarity has direct applications in machine learning as a positive definite kernel and can be efficiently computed via probabilistic hashing. Owing to the discrete nature, the hashed values can also be used for efficient near neighbor search. We prove the theoretical limit of GMM and the consistency result, assuming that the data follow an elliptical distribution, which is a very general family of distributions and includes the multivariate t-distribution as a special case. The consistency result holds as long as the data have bounded first moment (an assumption which essentially holds for datasets commonly encountered in practice). Furthermore, we establish the asymptotic normality of GMM. Compared to the "cosine" similarity which is routinely adopted in current practice in statistics and machine learning, the consistency of GMM requires much weaker conditions. Interestingly, when the data follow the t-distribution with ν degrees of freedom, GMM typically provides a better measure of similarity than "cosine" roughly when $\nu < 8$ ($\nu = 8$ means the distribution is already very close to a normal). These theoretical results will help explain the recent success of the use of the GMM kernel [11, 12, 13] in machine learning tasks.

1 Introduction

In statistics and machine learning, it is often crucial to choose, either explicitly or implicitly, some measure of data similarity. The most commonly adopted measure might be the "cosine" similarity:

$$Cos(x, y) = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}} \tag{1}$$

where x and y are n-dimensional data vectors. This measure implicitly assumes that the data have bounded second moment otherwise it will not converge to a fixed limit as the sample size increases. The data encountered in the real-world, however, are virtually always heavy-tailed [9, 4, 5]. [15] argued that the many natural datasets follow the power law with exponent (denote by ν) varying between 1 and 2. For example, $\nu = 1.2$ for the frequency of use of words, $\nu = 2.04$ for the number of citations to papers, $\nu = 1.4$ for the number of hits on the web sites, etc. Basically, $\nu > 2$ means that data have bounded second moment. The cosine similarity (1) will not converge (as $n \to \infty$) to a fixed constant if the data do not have bounded second moment.

[1]Presented at ICSA Conference on Data Science and Stanford Statistics Seminar.

In this study, we analyze the "generalized min-max" (GMM) similarity. First, we define

$$x_{i+} = \begin{cases} x_i & \text{if } x_i \geq 0 \\ 0 & \text{otherwise} \end{cases}, \quad x_{i-} = \begin{cases} -x_i & \text{if } x_i < 0, \\ 0 & \text{otherwise} \end{cases}, \quad x_i = x_{i+} - x_{i-}$$

Then we compute GMM as follows:

$$GMM(x,y) = \frac{\sum_{i=1}^n \left[\min(x_{i+}, y_{i+}) + \min(x_{i-}, y_{i-})\right]}{\sum_{i=1}^n \left[\max(x_{i+}, y_{i+}) + \max(x_{i-}, y_{i-})\right]} \triangleq g_n(x,y) \tag{2}$$

Note that for nonnative data, GMM becomes the original "min-max" kernel, which has been studied in the literature [8, 3, 14, 7, 10]. This paper focuses on analyzing theoretical properties of GMM. In particular, we are interested in the limit of $g_n(x,y)$ as $n \to \infty$ and how fast g_n converges to the limit. The convergency and speed of convergence are important. For example, the cosine similarity (1) is popular largely because, as long as the data have bounded second moments, $Cos(x,y)$ converges to a fixed limit which is believed to be a good characterization of the similarity between x and y.

To proceed with the analysis, we will have to make assumptions on the data. In this paper, we adopt the "elliptical distribution" model [1] which is very broad and includes many common distributions (such as Gaussian and Cauchy) as special cases. We first provide a simulation study.

2 Simulations Based on t-Distribution

The bivariate t-distribution has an explicit density and is a special case of the elliptical distribution. Denote by $t_{\Sigma,\nu}$ the bivariate t-distribution with covariance matrix Σ and ν degrees of freedom. Basically, if two independent variables $Z \sim N(0,\Sigma)$ and $u \sim \chi^2_\nu$, then we have $Z\sqrt{\nu/u} \sim t_{\Sigma,\nu}$. Here, we let $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, where $-1 \leq \rho \leq 1$. We consider n iid samples $(x_i, y_i) \sim t_{\Sigma,\nu}$ and compute $GMM(x,y) = g_n(x,y)$ according to (2). We are interested in the mean an standard deviation of GMM for $n \in \{1, 10, 100, 1000, 10000\}$ and $\nu \in \{3, 2, 1, 0.5\}$, as shown in Figure 1.

The panels in the first (top) row present the mean of GMM (g_n). The curves of GMM lie between two fixed curves, f_1 and f_∞, which we will calculate to be the following expressions:

$$f_1 = \rho + \frac{1}{\pi}\left[\sqrt{1-\rho^2}\log(2-2\rho) - 2\rho\sin^{-1}\left(\sqrt{(1-\rho)/2}\right)\right], \quad f_\infty = \frac{1-\sqrt{(1-\rho)/2}}{1+\sqrt{(1-\rho)/2}} \tag{3}$$

For better clarity, the panels in the second (middle) row plot the magnified portion. In each panel, the top (dashed and green if color is available) curve represent f_1 and the bottom (dashed and red) curve represent f_∞. We can see that for $\nu = 3$ and $\nu = 2$, g_n converges to f_∞ fast. For $\nu = 1$, g_n also converges to f_∞ but much slower. With $\nu = 0.5$, g_n does not converge to f_∞.

The panels in the third (bottom) row plot the standard deviation (std). For $\nu \geq 1$, the std curves converge to 0, although at $\nu = 1$ the convergence is much slower. When $\nu = 0.5$, the standard deviation does not converge to 0.

Basically, the simulations suggest that g_n converges to f_∞ as long as the data have bounded first moment (i.e., $\nu > 1$) and the convergence still holds for the boundary case (i.e., $\nu = 1$). We will provide thorough theoretical analysis on g_n for the general elliptical distribution.

Because ρ measures data similarity, the fact that $g_n \to f_\infty$ as long as $\nu \geq 1$ is important because it means we have a robust measure of ρ as long as the data are "reasonably" distributed. As shown by [15], most natural datasets have the equivalent $\nu > 1$.

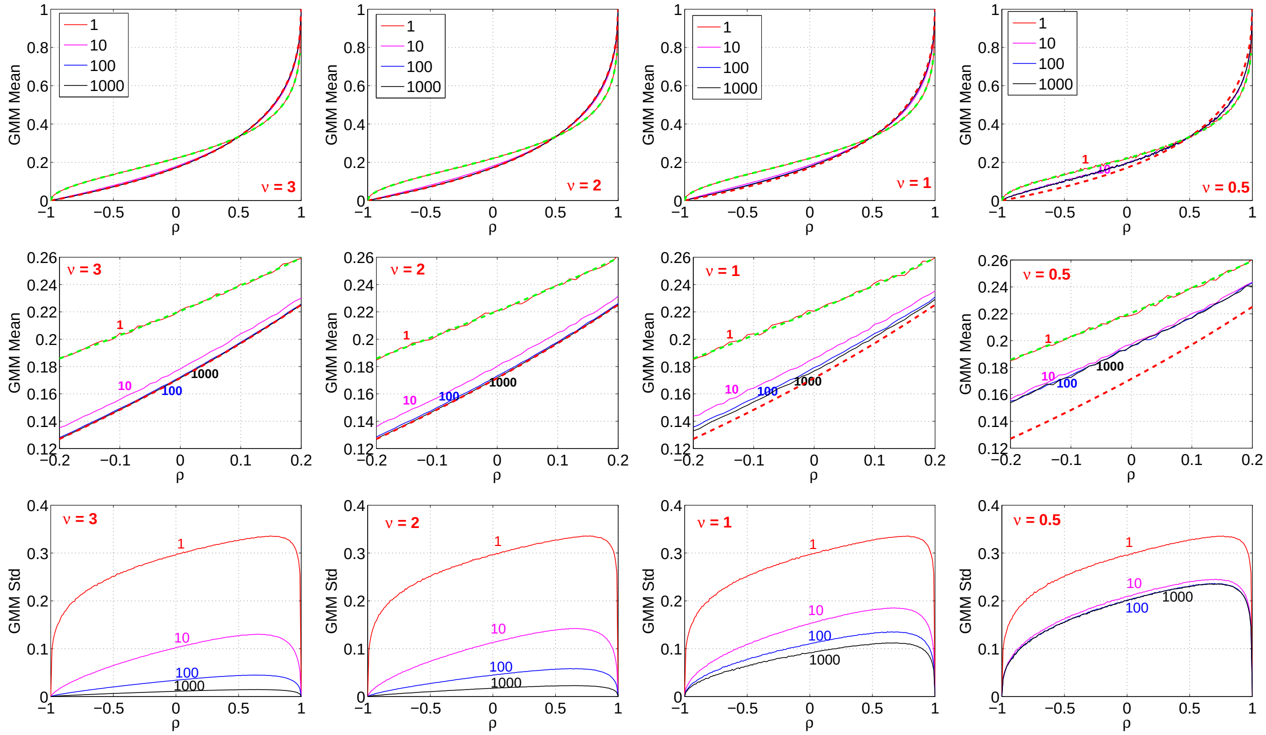

Figure 1: We simulate $GMM = g_n$ defined in (2) from the bivariate t-distribution with $\nu = 0.5, 1, 2, 3$ degrees of freedom, and $n = 1, 10, 100, 1000$, for 10000 repetitions. In the panels of the first two rows, we plot the mean curves together with two fixed (dashed) curves f_1 and f_∞ defined in (3). The panels in the second row are the zoomed-in version of the panels in the first row. The bottom panels plot the empirical standard deviation of g_n.

3 Analysis Based on Elliptical Distributions

We consider (x_i, y_i), $i = 1$ to n, are iid copy of (X, Y). Our goal is to analyze the statistical behavior of GMM, especially as $n \to \infty$,

$$GMM(x, y) = g_n(x, y) = \frac{\sum_{i=1}^n [\min(x_{i+}, y_{i+}) + \min(x_{i-}, y_{i-})]}{\sum_{i=1}^n [\max(x_{i+}, y_{i+}) + \max(x_{i-}, y_{i-})]}$$

To proceed with the theoretical analysis, we make a very general distributional assumption on the data. We say the vector (X, Y) has an elliptical distribution if

$$(X, Y)^T = AUT = \begin{pmatrix} a_1^T UT \\ a_2^T UT \end{pmatrix} \tag{4}$$

where $A = (a_1, a_2)^T$ is a deterministic 2×2 matrix, U is a vector uniformly distribution in the unit circle and T is a positive random variable independent of U. See [1] for an introduction.

In the family of elliptical distributions, there are two important special cases:

1. **Gaussian distribution**: In this case, we have $T^2 \sim \chi_2^2$ and

$$(X, Y)^T \sim N(0, \Sigma) \sim AU\sqrt{\chi_2^2}, \qquad \text{where } \Sigma = AA^T = \begin{pmatrix} 1 & \sigma\rho \\ \sigma\rho & \sigma^2 \end{pmatrix}. \tag{5}$$

Note that for analyzing g_n, it suffices to set $\mathrm{Var}(X) = 1$, due to cancelation in GMM.

2. t-**distribution**: In this case, we have $T \sim \sqrt{\chi_2^2 \nu / \chi_\nu^2}$ and

$$(X, Y)^T \sim N(0, \Sigma)\sqrt{\nu/\chi_\nu^2}. \tag{6}$$

Note that in Σ we consider $\sigma \neq 1$ to allow the situation that two vectors have different scales. For the convenience of presenting our theoretical results, we summarize the notations:

- $\Sigma = \begin{pmatrix} 1 & \sigma\rho \\ \sigma\rho & \sigma^2 \end{pmatrix}$, where $\rho \in [-1, 1]$ and $\sigma > 0$.
- $\alpha = \sin^{-1}\left(\sqrt{1/2 - \rho/2}\right) \in [0, \pi/2]$.
- $\tau \in [-\pi/2 + 2\alpha, \pi/2]$ is the solution of $\cos(\tau - 2\alpha)/\cos\tau = \sigma$, i.e., $\tau = \arctan(\sigma/\sin(2\alpha) - \cot(2\alpha))$. Note that $\tau = \alpha$ if $\sigma = 1$.

In addition, we need the following definitions of $f_1(\rho, \sigma)$ and $f_\infty(\rho, \sigma)$, for general σ as well as $\sigma = 1$:

$$\begin{aligned} f_1(\rho,\sigma) =& \frac{1}{\sigma\pi}\Big((\tau + \pi/2 - 2\alpha)\cos(2\alpha) + \sin(2\alpha)\log\frac{\cos(2\alpha - \pi/2)}{\cos\tau}\Big) \\ &+ \frac{\sigma}{\pi}\Big((\pi/2 - \tau)\cos(2\alpha) + \sin(2\alpha)\log\frac{\cos(2\alpha - \pi/2)}{\cos(2\alpha - \tau)}\Big), \qquad (7) \\ \overset{\sigma=1}{=}& \rho + \frac{1}{\pi}\left[\sqrt{1-\rho^2}\log(2 - 2\rho) - 2\rho\sin^{-1}\left(\sqrt{(1-\rho)/2}\right)\right] \qquad (8) \end{aligned}$$

$$\begin{aligned} f_\infty(\rho,\sigma) =& \frac{1 - \sin(2\alpha - \tau) + \sigma(1 - \sin\tau)}{\sigma(1 + \sin\tau) + 1 + \sin(2\alpha - \tau)} \qquad (9) \\ \overset{\sigma=1}{=}& \frac{1 - \sqrt{(1-\rho)/2}}{1 + \sqrt{(1-\rho)/2}} \qquad (10) \end{aligned}$$

Theorem 1 presents the results for consistency.

Theorem 1. ***(Consistency)*** *Assume (X, Y) has an elliptical distribution with $(X, Y)^T = AUT$ and $\Sigma = AA^T = \begin{pmatrix} 1 & \sigma\rho \\ \sigma\rho & \sigma^2 \end{pmatrix}$. Let (x_i, y_i), $i = 1$ to n, be iid copies of (X, Y), and $GMM(x, y) = g_n(x, y)$ as defined in (2). Then the following statements hold:*

- $g_1 = f_1(\rho, \sigma)$
- *If $\mathbb{E}T < \infty$, then $g_n \to f_\infty(\rho, \sigma)$, almost surely.*
- *If we have*

$$\lim_{t\to\infty} \frac{t\,\mathbb{P}(T > t)}{\mathbb{E}\min(T, t)} = 0, \tag{11}$$

  then $g_n \to f_\infty(\rho, \sigma)$, in probability.
- *If (X, Y) has a t-distribution with ν degrees of freedom, then $g_n \to f_\infty(\rho, \sigma)$ almost surely if $\nu > 1$ and $g_n \to f_\infty(\rho, \sigma)$ in probability if $\nu = 1$.*

Theorem 2 presents the results for asymptotic normality.

Theorem 2. ***(Asymptotic Normality)***
With the same notation and definitions as in Theorem 1, the following statements hold:

- *If* $\mathbb{E}T^2 < \infty$, *then*

$$n^{1/2}\left(g_n(x,y) - f_\infty(\rho,\sigma)\right) \xrightarrow{D} N\left(0, \frac{V}{H^4}\frac{\mathbb{E}T^2}{\mathbb{E}^2T}\right) \tag{12}$$

where

$$\begin{aligned} &V && (13)\\ =&\frac{1}{4\pi^3}\left\{2\tau+\pi-4\alpha+\sin(2\tau-4\alpha)+\sigma^2\left(\pi-2\tau-\sin(2\tau)\right)\right\}\left\{\sigma(1+\sin\tau)+1+\sin(2\alpha-\tau)\right\}^2 \\ +&\frac{1}{4\pi^3}\left\{\sigma^2\left(2\tau+\sin(2\tau)+\pi\right)+\left(\pi+4\alpha-2\tau-\sin(2\tau-4\alpha)\right)+4\sigma\left(\sin 2\alpha-2\alpha\cos 2\alpha\right)\right\} \\ &\quad\times\left\{1-\sin(2\alpha-\tau)+\sigma(1-\sin\tau)\right\}^2 \\ -&\frac{\sigma}{\pi^3}\left((\pi-2\alpha)\cos 2\alpha+\sin 2\alpha\right)\left\{1-\sin(2\alpha-\tau)+\sigma(1-\sin\tau)\right\}\left\{\sigma(1+\sin\tau)+1+\sin(2\alpha-\tau)\right\} \\ \overset{\sigma=1}{=}&\frac{4}{\pi^3}\sin^2\alpha\left(3\pi-8\cos\alpha+2\sin 2\alpha+\pi\cos 2\alpha-8\alpha\sin\alpha-4\alpha\cos 2\alpha\right) && (14)\end{aligned}$$

and

$$H = \frac{1}{\pi}\left\{\sigma(1+\sin\tau)+1+\sin(2\alpha-\tau)\right\} \overset{\sigma=1}{=} \frac{2}{\pi}(1+\sin\alpha) \tag{15}$$

- *If* (X,Y) *has a* t*-distribution with* ν *degrees of freedom and* $\nu > 2$, *then*

$$n^{1/2}\left(g_n(x,y) - f_\infty(\rho,\sigma)\right) \xrightarrow{\mathrm{D}} N\left(0, \frac{V}{H^4}\frac{\mathbb{E}T^2}{\mathbb{E}^2T}\right). \tag{16}$$

where $\mathbb{E}T^2 = \frac{2\nu}{\nu-2}$ *and* $\mathbb{E}T = \frac{\sqrt{\nu}\,\Gamma(\nu/2-1/2)\Gamma(1/2)}{2\,\Gamma(\nu/2)}$

- *If* (X,Y) *has a* t*-distribution with* $\nu = 2$ *degrees of freedom, then*

$$\left(\frac{n}{\log n}\right)^{1/2}\left(g_n(x,y) - f_\infty(\rho,\sigma)\right) \xrightarrow{\mathrm{D}} N\left(0, \frac{V}{H^4}\frac{4}{\pi^2}\right). \tag{17}$$

Figure 2 presents a simulation study to verify the asymptotic normality, in particular, the asymptotic variance formula

$$Var\left(g_n\right) = \frac{1}{n}\frac{V}{H^4}\frac{\mathbb{E}T^2}{\mathbb{E}^2T} + O\left(\frac{1}{n^2}\right) \tag{18}$$

by considering that the data follow a t-distribution with ν degrees of freedom and $\nu = 2.5,\ 3,\ 4,\ 5$. The simulation results confirm the asymptotic variance formula at large enough sample size n. When n is not too large, the asymptotic variance formula (18) can be conservative.

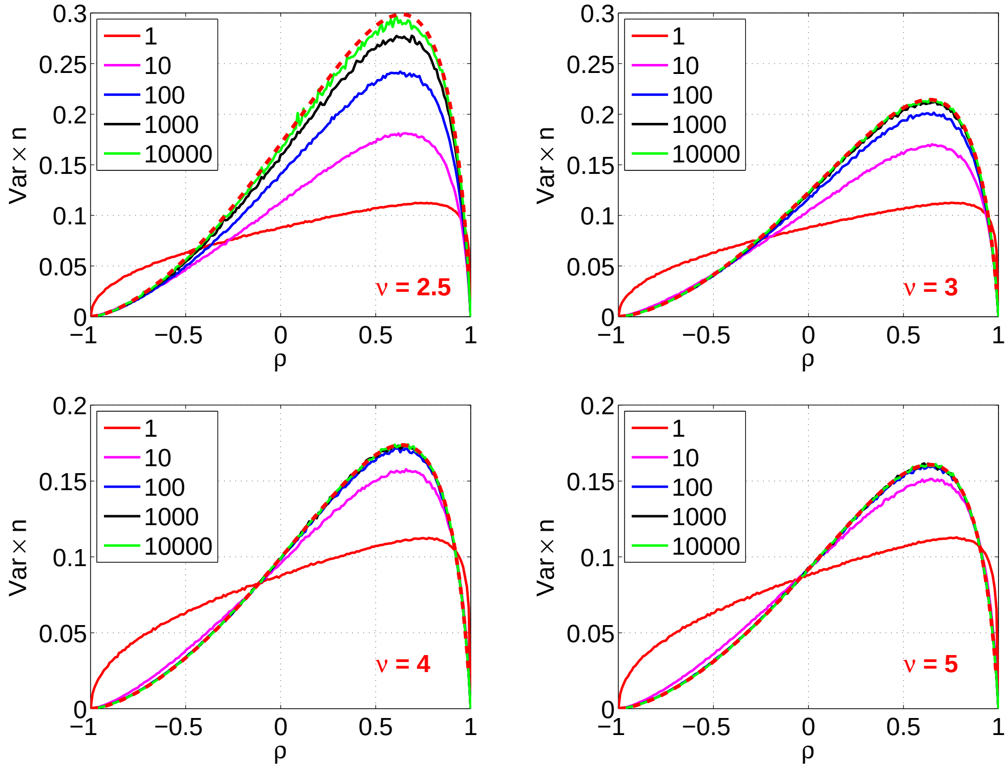

Figure 2: Simulations for verifying the asymptotic variance formula (18) based on t-distribution with ν degrees of freedom where $\nu \in \{2.5, 3, 4, 5\}$ and $\rho \in [-1, 1]$ spaced at 0.01. For each case, we repeat the simulation 10000 times. We report the empirical $Var(g_n) \times n$ with the theoretical asymptotic value $\frac{V}{H^4}\frac{\mathbb{E}T^2}{\mathbb{E}^2T}$ plotted as dashed curves. For n large enough, the asymptotic variance formula (18) becomes accurate. For small n values, however, the formula can be quite conservative.

4 Estimation of ρ

The fact that $g_n(x, y) \to f_\infty(\rho, \sigma)$ also provides a robust and convenient way to estimate the similarity between data vectors. Here, for convenience we consider $\sigma = 1$. For this case, we have $f_\infty = \frac{1-\sqrt{(1-\rho)/2}}{1+\sqrt{(1-\rho)/2}}$. This suggests an estimator of ρ:

$$\hat{\rho}_g = 1 - 2\left(\frac{1-g_n}{1+g_n}\right)^2 \tag{19}$$

As $n \to \infty$, $g_n \to f_\infty$ and $\hat{\rho}_g \to \rho$. In other words, the estimator $\hat{\rho}_g$ is asymptotically unbiased. The asymptotic variance of $\hat{\rho}_g$ can be computed using "delta method":

$$\begin{aligned} Var\left(\hat{\rho}_g\right) =& \left[8\frac{1-f_\infty}{(1+f_\infty)^3}\right]^2 Var\left(g_n\right) + O\left(\frac{1}{n^2}\right) \\ =& \frac{1}{n} 2\left(1-\rho\right)\left(1+\sqrt{(1-\rho)/2}\right)^4 \frac{V}{H^4}\frac{\mathbb{E}T^2}{\mathbb{E}^2T} + O\left(\frac{1}{n^2}\right) \end{aligned} \tag{20}$$

See (18) and Theorem 2 for more details. Again, we emphasize that this estimator $\hat{\rho}_g$ is meaningful as long as $\mathbb{E}T < \infty$ and $Var\left(\hat{\rho}_g\right) < \infty$ as long as $\mathbb{E}T^2 < \infty$.

It is interesting to compare this estimator with the commonly used estimator based on the "cosine" similarity:

$$Cos(x,y) = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}} \triangleq c_n(x,y)$$

When the data are bivariate normal, it is a known result [1] that $c_n(x,y)$, when appropriately normalized, converges in distribution to a normal

$$n^{1/2}\left(c_n - \rho\right) \xrightarrow{D} N\left(0, (1-\rho^2)^2\right) \tag{21}$$

This asymptotic normality (with difference in the variance term) holds as long as the data have bounded fourth moment. Here, we present the generalization as a theorem.

Theorem 3. *If $\mathbb{E}T^4 < \infty$, then*

$$n^{1/2}\left(c_n - \rho\right) \xrightarrow{D} N\left(0, \frac{\mathbb{E}T^4}{2\mathbb{E}^2T^2}(1-\rho^2)^2\right) \tag{22}$$

Based on Theorem 3, a natural estimator of ρ and its asymptotic variance would be

$$\hat{\rho}_c = c_n, \qquad Var\left(\hat{\rho}_c\right) = \frac{1}{n}\frac{\mathbb{E}T^4}{2\mathbb{E}^2T^2}\left(1-\rho^2\right)^2 + O\left(\frac{1}{n^2}\right) \tag{23}$$

When the data follow a t-distribution with ν degrees of freedom, we have

$$\mathbb{E}T^2 = \frac{2\nu}{\nu-2}, \qquad \mathbb{E}T^4 = \frac{4\nu^3}{(\nu-2)^2(\nu-4)} + \frac{4\nu^2}{(\nu-2)^2}$$

Figure 3 and Figure 4 provide a simulation study for comparing two estimators $\hat{\rho}_g$ and $\hat{\rho}_c$. We assume t-distribution with ν degrees of freedom, where $\nu \in \{2.5, 3, 4, 4.5, 5, 6, 8, 10\}$ as well $\nu = \infty$ (i.e., normal distribution). In each panel, we plot the empirical mean square errors (MSEs): $MSE(\hat{\rho}_g)$ and $MSE(\hat{\rho}_c)$ (computed from 10000 repetitions), along with the (asymptotic) theoretical variance of $\hat{\rho}_g$: $\frac{1}{n}2\left(1-\rho\right)\left(1+\sqrt{(1-\rho)/2}\right)^4 \frac{V}{H^4}\frac{\mathbb{E}T^2}{\mathbb{E}^2T}$. For clarity, we did not plot the theoretical variance of $\hat{\rho}_c$, which is fairly simple and more straightforward to be verified.

The results in Figure 3 and Figure 4 confirm that $\hat{\rho}_g$, the estimator based on GMM, is substantially more accurate than $\hat{\rho}_c$, the commonly used estimator based on cosine. Roughly speaking, when $\nu < 8$, $\hat{\rho}_g$ is more preferable. Even when the data are perfectly Gaussian (the bottom row in Figure 4), the use of $\hat{\rho}_g$ does not result in much loss of accuracy compared to $\hat{\rho}_c$.

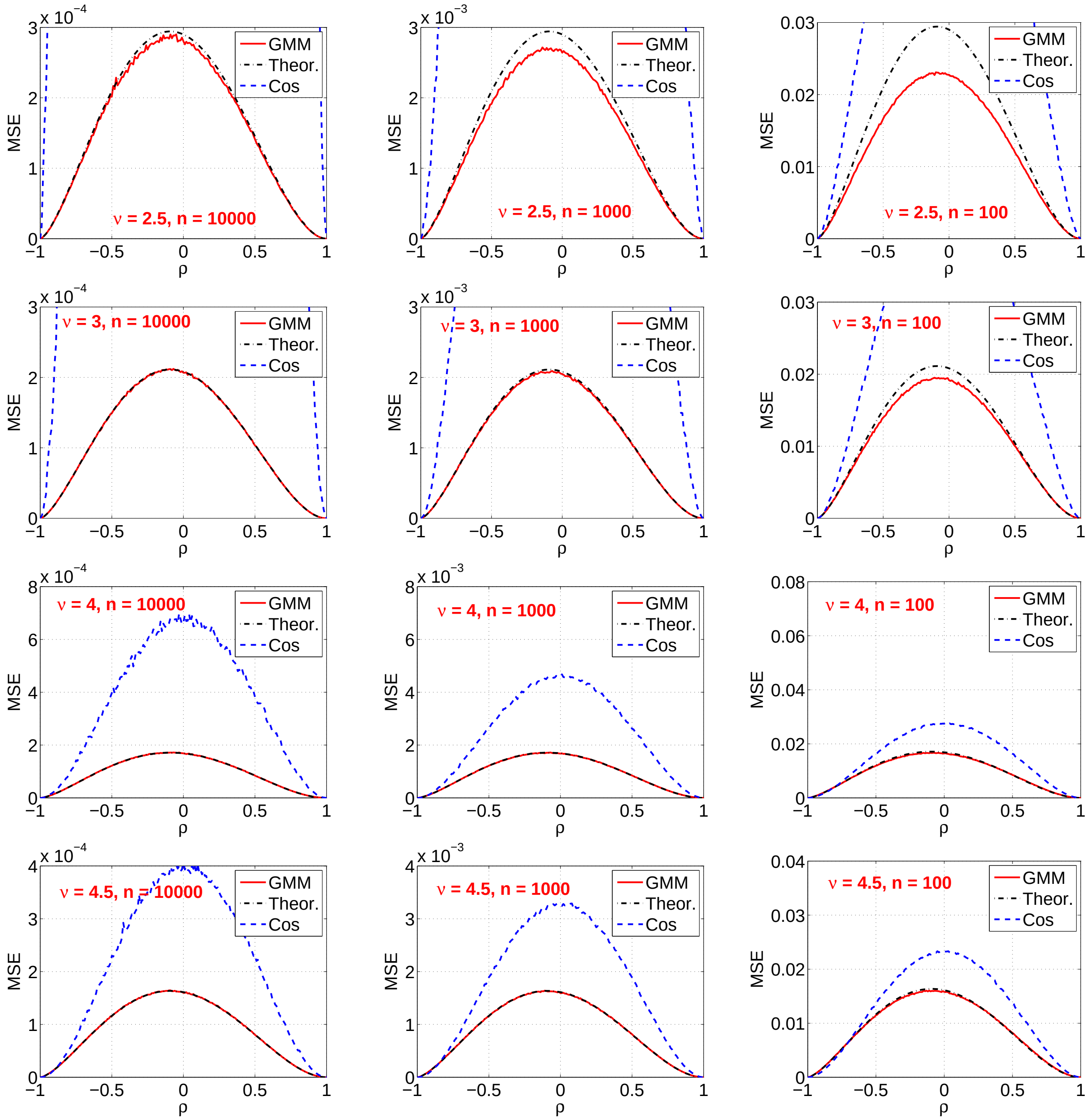

Figure 3: Simulations for comparing two estimators of data similarity ρ: 1) $\hat{\rho}_g$, the estimator based on GMM, and 2) $\hat{\rho}_c$, the estimator based on cosine. We assume the data follow a t distribution with ν degrees of freedom. In each panel (for each ν), we plot the empirical MSE($\hat{\rho}_g$) and MSE($\hat{\rho}_c$) as well as the theoretical asymptotic variance of $\hat{\rho}_g$: $\frac{1}{n}2\left(1-\rho\right)\left(1+\sqrt{(1-\rho)/2}\right)^4\frac{V}{H^4}\frac{\mathbb{E}T^2}{\mathbb{E}^2T}$. It is clear from the results that $\hat{\rho}_g$ is substantially more accurate than $\hat{\rho}_c$. The theoretical asymptotic variance formula, despite the complexity of its expression, is accurate when ν is not too close to 2.

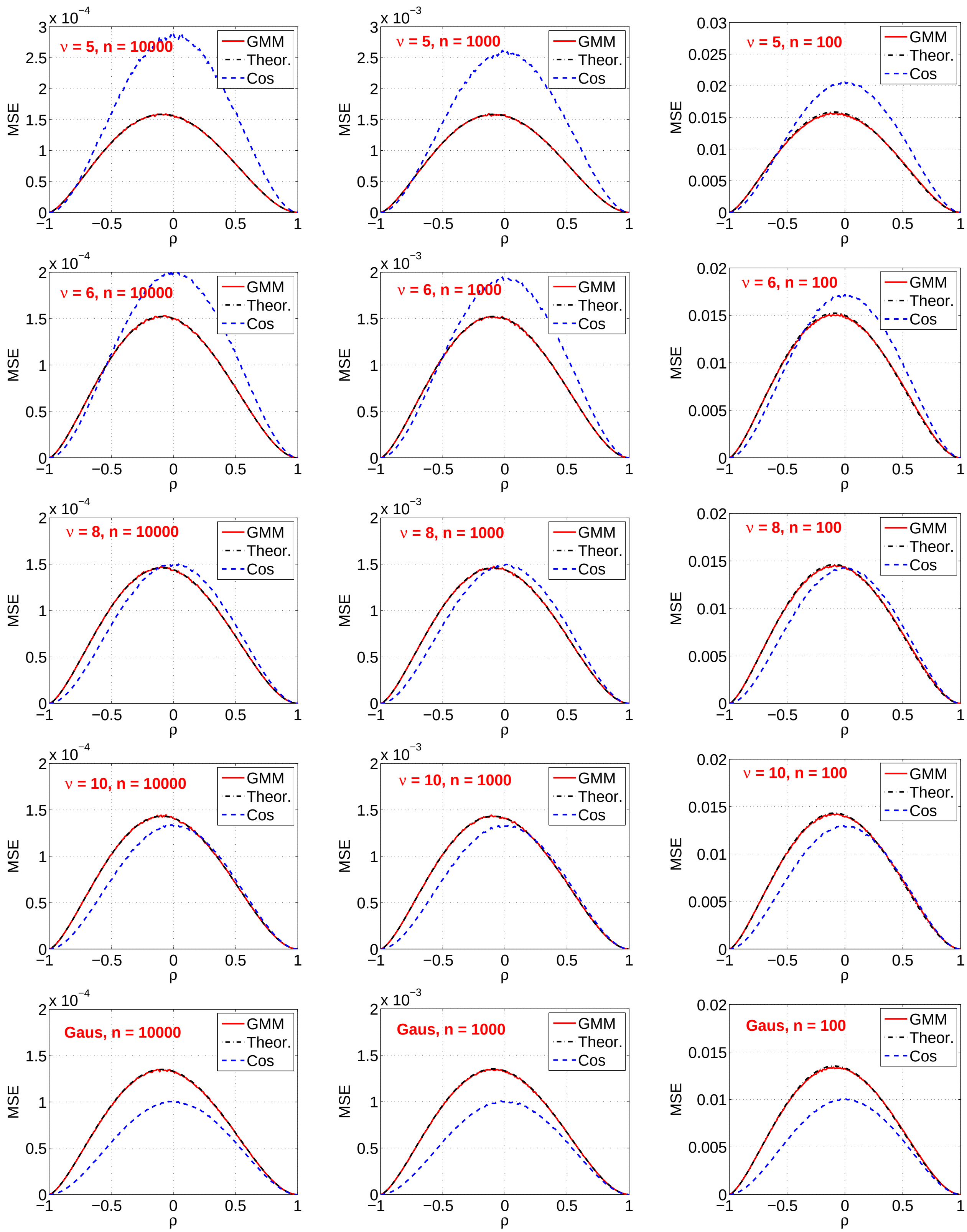

Figure 4: Continued from Figure 3. We present results for larger ν (5, 6, 8, 10) and $\nu = \infty$ (i.e., Gaussian data, the bottom row). Roughly speaking, when $\nu < 8$, it is preferable to use $\hat{\rho}_g$, the estimator based on GMM. In fact, even when data are perfectly Gaussian, using $\hat{\rho}_g$ does not result in too much loss of accuracy.

5 Concluding Remarks

The "cosine" similarity commonly used in practice essentially assumes that the data are normally (or equivalently) distributed. The data in reality, however, are typically heavy-tailed and sparse. A concurrent line of work [11, **?**] has shown that the new measure named "generalized min-max" (GMM) is particularly effective as a positive definite kernel and there is an efficient computational procedure to convert this nonlinear kernel into linear kernel. Extensive experiments on more than 50 datasets [11, 12] have demonstrated the promising performance in machine learning tasks. This motivates us to develop the theoretical results for analyzing GMM.

We show that, under mild conditions, GMM converges to a limit as long as the data have bounded first moment. In contrast, the cosine similarity requires that data to have bounded second moment. We derive the explicit expression for the limit and establish the asymptotic normality of GMM with explicit (and sophisticated) variance expressions. Those theoretical results will be useful for further analyzing of GMM in statistics, machine learning, and other applications.

References

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, Hoboken, New Jersey, third edition, 2003.

[2] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*. Cambridge University Press, 1987. Cambridge Books Online.

[3] M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *STOC*, pages 380–388, Montreal, Quebec, Canada, 2002.

[4] M. E. Crovella and A. Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. *IEEE/ACM Trans. Networking*, 5(6):835–846, 1997.

[5] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *SIGMOD*, pages 251–262, Cambridge,MA, 1999.

[6] B. V. Gnedenko and A. N. Kolmogorov. *Limit Distributions for Sum of Independent Random Variables*. Addison Wesley, Reading, MA, 1954.

[7] S. Ioffe. Improved consistent sampling, weighted minhash and L1 sketching. In *ICDM*, pages 246–255, Sydney, AU, 2010.

[8] J. Kleinberg and E. Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and Markov random fields. In *FOCS*, pages 14–23, New York, 1999.

[9] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic. *IEEE/ACM Trans. Networking*, 2(1):1–15, 1994.

[10] P. Li. 0-bit consistent weighted sampling. In *KDD*, Sydney, Australia, 2015.

[11] P. Li. Linearized GMM kernels and normalized random fourier features. Technical report, arXiv:1605.05721, 2016.

[12] P. Li. Nystrom method for approximating the gmm kernel. Technical report, arXiv:1605.05721, 2016.

[13] P. Li. Tunable GMM kernels. Technical report, arXiv:1701.02046, 2017.

[14] M. Manasse, F. McSherry, and K. Talwar. Consistent weighted sampling. Technical Report MSR-TR-2010-73, Microsoft Research, 2010.

[15] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):232–351, 2005.

A Proof of Theorem 1

For a random vector (X, Y), we are interested in quantities

$$\mu_1 = \mathbb{E}\,\frac{X_+ \wedge Y_+ + X_- \wedge Y_-}{X_+ \vee Y_+ + X_- \vee Y_-}, \quad \mu_\infty = \frac{\mathbb{E}(X_+ \wedge Y_+ + X_- \wedge Y_-)}{\mathbb{E}(X_+ \vee Y_+ + X_- \vee Y_-)}.$$

Without any assumption, we have

$$\mu_1 = \mathbb{E}\,\frac{X_+ \wedge Y_+}{X_+ \vee Y_+} + \mathbb{E}\,\frac{X_- \wedge Y_-}{X_- \vee Y_-} = \mathbb{E}\,\frac{|X| \wedge |Y|}{|X| \vee |Y|} I\{XY > 0\} = \mathbb{E}\,\frac{|X/Y| \wedge 1}{|X/Y| \vee 1} I\{X/Y > 0\}.$$

When $\mathbb{E}(|X| \wedge |Y|) < \infty$,

$$\mu_\infty = \frac{\mathbb{E}(|X| \wedge |Y|) I\{XY > 0\}}{\mathbb{E}\big[(|X| + |Y|) I\{XY \le 0\} + (|X| \vee |Y|) I\{XY > 0\}\big]}.$$

If (X, Y) is symmetric in the sense of $(X, Y) \sim (-X, -Y)$, then

$$\mu_1 = 2\,\mathbb{E}\,\frac{X_+ \wedge Y_+}{X_+ \vee Y_+}$$

and

$$\mu_\infty = \frac{\mathbb{E}(X_+ \wedge Y_+)}{\mathbb{E}(X_+ \vee Y_+)}.$$

The vector (X, Y) has an elliptical distribution if

$$(X, Y)^T = AUT = \begin{pmatrix} a_1^T UT \\ a_2^T UT \end{pmatrix}$$

where $A = (a_1, a_2)^T$ is a deterministic 2×2 matrix, U is a vector uniformly distribution in the unit circle and T is a positive random variable independent of U. In this case, $U \sim -U$, so that (X, Y) is symmetric. If T has a finite expectation, then T can be can cancelled in the calculation of μ_1 and μ_∞, so that

$$\mu_1 = 2\,\mathbb{E}\,\frac{(a_1^T U)_+ \wedge (a_2^T U)_+}{(a_1^T U)_+ \vee (a_2^T U)_+}$$

and

$$\mu_\infty = \frac{\mathbb{E}\{(a_1^T U)_+ \wedge (a_2^T U)_+\}}{\mathbb{E}\{(a_1^T U)_+ \vee (a_2^T U)_+\}}.$$

Since a bivariate Gaussian distribution is elliptical with $T^2 \sim \chi_2^2$, the elliptical case with finite $\mathbb{E}T$ is equivalent to the bivariate Gaussian case

$$(X,Y) \sim N(0,\Sigma) \ \text{ with } \ \Sigma = AA^T = \begin{pmatrix} 1 & \sigma\rho \\ \sigma\rho & \sigma^2 \end{pmatrix}.$$

Note that we set $\mathrm{Var}(X) = 1$ due to scale invariance of μ_1 and μ_∞.

For $\sigma > 0$ and $\rho \in [-1,1]$, let $\alpha = \sin^{-1}\left(\sqrt{1/2 - \rho/2}\right) \in [0,\pi/2]$, and $\tau \in [-\pi/2 + 2\alpha, \pi/2]$ be the solution of $\cos(\tau - 2\alpha)/\cos\tau = \sigma$. We have $\tau = \arctan(\sigma/\sin(2\alpha) - \cot(2\alpha))$. Define

$$\begin{aligned} f_1(\rho,\sigma) &= \frac{1}{\sigma\pi}\Big((\tau + \pi/2 - 2\alpha)\cos(2\alpha) + \sin(2\alpha)\log\frac{\cos(2\alpha - \pi/2)}{\cos\tau}\Big) \\ &\quad + \frac{\sigma}{\pi}\Big((\pi/2 - \tau)\cos(2\alpha) + \sin(2\alpha)\log\frac{\cos(2\alpha - \pi/2)}{\cos(2\alpha - \tau)}\Big), \end{aligned}$$

and

$$f_\infty(\rho,\sigma) = \frac{1 - \sin(2\alpha - \tau) + \sigma(1 - \sin\tau)}{\sigma(1 + \sin\tau) + 1 + \sin(2\alpha - \tau)}.$$

We note that $\sin(2\alpha) = 2\sin\alpha\cos\alpha = 2\sqrt{1/2 - \rho/2}\sqrt{1/2 + \rho/2} = \sqrt{1 - \rho^2}$, $\cos(2\alpha) = \rho$, and $\cos(2\alpha - \pi/2) = \sin(2\alpha) = \sqrt{1 - \rho^2}$. Moreover, $\tan\tau = \{\sigma - \cos(2\alpha)\}/\sin(2\alpha) = (\sigma - \rho)/\sqrt{1 - \rho^2}$, so that $1/\cos^2\tau = 1 + \tan^2\tau = 1 + (\sigma - \rho)^2/(1 - \rho^2) = (1 - \rho^2 + \sigma^2 - 2\sigma\rho + \rho^2)/(1 - \rho^2) = (1 + \sigma^2 - 2\sigma\rho)/(1 - \rho^2)$. Thus,

$$\frac{\cos^2(2\alpha - \pi/2)}{\cos^2\tau} = 1 + \sigma^2 - 2\sigma\rho, \ \ \frac{\cos^2(2\alpha - \pi/2)}{\cos^2(2\alpha - \tau)} = \frac{\cos^2(2\alpha - \pi/2)}{\sigma^2\cos^2\tau} = 1 + \sigma^{-2} - 2\rho/\sigma.$$

Consider the Gaussian case

$$(X,Y) \sim N(0,\Sigma) \ \text{ with } \ \Sigma = AA^T = \begin{pmatrix} 1 & \sigma\rho \\ \sigma\rho & \sigma^2 \end{pmatrix}.$$

Let

$$A = \begin{pmatrix} \cos\alpha & \sin\alpha \\ \sigma\cos\alpha & -\sigma\sin\alpha \end{pmatrix}.$$

We have

$$AA^T = \begin{pmatrix} 1 & \sigma(\cos^2\alpha - \sin^2\alpha) \\ \sigma(\cos^2\alpha - \sin^2\alpha) & \sigma^2 \end{pmatrix} = \begin{pmatrix} 1 & \sigma\rho \\ \sigma\rho & \sigma^2 \end{pmatrix}.$$

Let θ be a uniform variable in $(-\pi,\pi)$. Since $U \sim (\cos\theta, \sin\theta)^T$,

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \begin{pmatrix} \cos\alpha\cos\theta + \sin\alpha\sin\theta \\ \sigma(\cos\alpha\cos\theta - \sin\alpha\sin\theta) \end{pmatrix} T = \begin{pmatrix} \cos(\theta - \alpha) \\ \sigma\cos(\theta + \alpha) \end{pmatrix} T \sim \begin{pmatrix} \cos(\theta - 2\alpha) \\ \sigma\cos\theta \end{pmatrix} T$$

As $\alpha \in (0,\pi/2)$, it follows that

$$\begin{aligned} \mu_1 &= \frac{2}{2\pi}\int_{-\pi}^{\pi} \frac{(\cos(\theta - 2\alpha))_+ \wedge (\sigma\cos\theta)_+}{(\cos(\theta - 2\alpha))_+ \vee (\sigma\cos\theta)_+} d\theta \\ &= \frac{1}{\pi}\int_{-\pi/2+2\alpha}^{\pi/2} \frac{(\cos(\theta - 2\alpha)/\cos\theta) \wedge \sigma}{(\cos(\theta - 2\alpha)/\cos\theta) \vee \sigma} d\theta. \end{aligned}$$

As $\cos(\theta-2\alpha)/\cos\theta = \cos(2\alpha)+\tan\theta\sin(2\alpha)$, for $\cos\theta>0$ $\cos(\theta-2\alpha)/\cos\theta=\sigma$ iff $\theta=\tau$ and $\tau\in[-\pi/2+2\alpha,\pi/2]$. Thus, with $t=2\alpha-\theta$,

$$\begin{aligned}
\mu_1 &= \frac{1}{\pi}\int_{-\pi/2+2\alpha}^{\tau}\frac{\cos(\theta-2\alpha)}{\sigma\cos\theta}d\theta+\frac{1}{\pi}\int_{\tau}^{\pi/2}\frac{\sigma\cos\theta}{\cos(\theta-2\alpha)}d\theta \\
&= \frac{1}{\sigma\pi}\int_{-\pi/2+2\alpha}^{\tau}\{\cos(2\alpha)+\tan\theta\sin(2\alpha)\}d\theta+\frac{\sigma}{\pi}\int_{-\pi/2+2\alpha}^{2\alpha-\tau}\frac{\cos(t-2\alpha)}{\cos t}dt \\
&= f_1(\rho,\sigma).
\end{aligned}$$

We note that $\tau=\alpha$ when $\sigma=1$. Similarly,

$$\begin{aligned}
\mu_\infty &= \frac{(2\pi)^{-1}\int_{-\pi}^{\pi}(\cos(\theta-2\alpha))_+\wedge(\sigma\cos\theta)_+d\theta}{(2\pi)^{-1}\int_{-\pi}^{\pi}(\cos(\theta-2\alpha))_+\vee(\sigma\cos\theta)_+d\theta} \\
&= \frac{\int_{-\pi/2+2\alpha}^{\tau}\cos(\theta-2\alpha)d\theta+\sigma\int_{\tau}^{\pi/2}\cos\theta d\theta}{\sigma\int_{-\pi/2}^{\tau}\cos\theta d\theta+\int_{\tau}^{\pi/2+2\alpha}\cos(\theta-2\alpha)d\theta} \\
&= \frac{1-\sin(2\alpha-\tau)+\sigma(1-\sin\tau)}{\sigma(1+\sin\tau)+1+\sin(2\alpha-\tau)}. \\
&= f_\infty(\rho,\sigma).
\end{aligned}$$

It is well known [6, 2] that

$$\frac{\max_{i\le n}T_i}{T_1+\cdots+T_n}=o_{\mathbb{P}}(1)$$

if and only if

$$\lim_{t\to\infty}\frac{t\,\mathbb{P}(T>t)}{\mathbb{E}\min(T,t)}=0. \tag{24}$$

Suppose (24) holds. Let (X_i,Y_i) be a sequence of iid variables from (X,Y). Then,

$$\frac{\sum_{i=1}^n\{(X_i)_+\wedge(Y_i)_+ +(X_i)_-\wedge(Y_i)_-\}}{\sum_{i=1}^n\{(X_i)_+\vee(Y_i)_+ +(X_i)_-\vee(Y_i)_-\}}=f_\infty(\rho,\sigma)+o_{\mathbb{P}}(1).$$

This can be seen as follows. Write

$$(X_i,Y_i)^T=AU_iT_i,\quad\text{with}\quad A=\begin{pmatrix}\cos\alpha & \sin\alpha\\ \sigma\cos\alpha & -\sigma\sin\alpha\end{pmatrix}.$$

We have

$$\mathrm{Var}\Big(\frac{\sum_{i=1}^n(X_i)_+\wedge(Y_i)_+}{\sum_{i=1}^n T_i}\Big|T_1,\ldots,T_n\Big)\le\frac{C_0\sum_{i=1}^nT_i^2}{(\sum_{i=1}^nT_i)^2}\le\frac{C_0\max_{i\le n}T_i}{T_1+\cdots+T_n}=o_{\mathbb{P}}(1).$$

After applying this argument to $(X_i)_-\wedge(Y_i)_-$, $(X_i)_+\vee(Y_i)_+$ and $(X_i)_-\vee(Y_i)_-$, the conclusion follows from

$$\frac{\mathbb{E}\big[\sum_{i=1}^n\big((X_i)_+\wedge(Y_i)_+ +(X_i)_-\wedge(Y_i)_-\big)\big|T_1,\ldots,T_n\big]}{\mathbb{E}\big[\sum_{i=1}^n\big((X_i)_+\vee(Y_i)_+ +(X_i)_-\vee(Y_i)_-\big)\big|T_1,\ldots,T_n\big]}=f_\infty(\rho,\sigma).$$

Now consider the bivariate t-distribution as an example:

$$(X,Y)^T\sim N(0,\Sigma)\sqrt{\nu/\chi^2_\nu}.$$

where χ^2_ν is independent of $N(0,\Sigma)$. Since $N(0,\Sigma)$ can be written as $AU\sqrt{\chi^2_2}$, the bivariate t-distribution can be written as

$$(X,Y)^T \sim AUT \quad \text{with } T \sim \sqrt{\chi^2_2\nu/\chi^2_\nu} \sim \sqrt{2F_{2,\nu}}$$

with two independent chi-square variables, where $F_{2,\nu}$ denotes the F distribution. It can be shown that

$$\mathbb{E}T = \frac{\sqrt{\nu}\,\Gamma(\nu/2-1/2)\Gamma(1/2)}{2\,\Gamma(\nu/2)}, \quad \nu > 1.$$

For example, $\mathbb{E}T = \pi/\sqrt{2}$ for $\nu = 2$. For $\nu = 1$, we still have (25), as (24) follows from

$$\frac{t\,\mathbb{P}(T > t)}{\mathbb{E}\min(T,t)} = \frac{t(1+t^2)^{-1/2}}{\int_0^t (1+x^2)^{-1/2}dx} = \frac{1+o(1)}{\log t} \to 0.$$

B Proof of Theorem 2

Let T be independent of (ξ,ζ) and (T_i,ξ_i,ζ_i) be iid copies of (T,ξ,ζ). Assume that $\mathbb{E}T^2 + \mathbb{E}(\xi\mathbb{E}\zeta - \zeta\mathbb{E}\xi)^2 < \infty$. Then,

$$n^{1/2}\left(\frac{\sum_{i=1}^n T_i\xi_i}{\sum_{i=1}^n T_i\zeta_i} - \frac{\mathbb{E}\xi}{\mathbb{E}\zeta}\right) = n^{1/2}\frac{\sum_{i=1}^n T_i(\xi_i\mathbb{E}\zeta - \zeta_i\mathbb{E}\xi)}{\mathbb{E}\zeta\sum_{i=1}^n T_i\zeta_i} \xrightarrow{\mathrm{D}} N\left(0, \frac{V\mathbb{E}T^2}{(\mathbb{E}T)^2(\mathbb{E}\zeta)^4}\right).$$

with

$$V = \mathbb{E}(\xi\mathbb{E}\zeta - \zeta\mathbb{E}\xi)^2.$$

Alternatively, if the condition $\mathbb{E}T^2 < \infty$ is replaced by

$$\lim_{t\to\infty}\frac{t\,\mathbb{P}(T^2 > t)}{\mathbb{E}\min(T^2,t)} = 0, \tag{25}$$

then,

$$\begin{aligned}
&\frac{\sum_{i=1}^n T_i}{(\sum_{i=1}^n T_i^2)^{1/2}}\left(\frac{\sum_{i=1}^n T_i\xi_i}{\sum_{i=1}^n T_i\zeta_i} - \frac{\mathbb{E}\xi}{\mathbb{E}\zeta}\right) \\
=\;& \left(\frac{\sum_{i=1}^n T_i}{\mathbb{E}\zeta\sum_{i=1}^n T_i\zeta_i}\right)\frac{\sum_{i=1}^n T_i(\xi_i\mathbb{E}\zeta - \zeta_i\mathbb{E}\xi)}{(\sum_{i=1}^n T_i^2)^{1/2}} \\
=\;& (1+o(1))\frac{\sum_{i=1}^n T_i(\xi_i\mathbb{E}\zeta - \zeta_i\mathbb{E}\xi)}{(\mathbb{E}\zeta)^2(\sum_{i=1}^n T_i^2)^{1/2}} \\
\xrightarrow{\mathrm{D}}\;& N\left(0, \frac{V}{(\mathbb{E}\zeta)^4}\right).
\end{aligned}$$

Suppose (X,Y) is elliptical and

$$\xi = \{(X)_+ \wedge (Y)_+ + (X)_- \wedge (Y)_-\}/T, \quad \zeta = \{(X)_+ \vee (Y)_+ + (X)_- \vee (Y)_-\}/T.$$

As in the computation of f_∞, we have

$$\mathbb{E}\xi \;=\; \frac{2}{2\pi}\int_{-\pi}^{\pi}(\cos(\theta-2\alpha))_+ \wedge (\sigma\cos\theta)_+ d\theta$$

$$
\begin{aligned}
&= \frac{1}{\pi}\left\{\int_{-\pi/2+2\alpha}^{\tau}\cos(\theta-2\alpha)d\theta+\sigma\int_{\tau}^{\pi/2}\cos\theta d\theta\right\}\\
&= \frac{1}{\pi}\left\{1-\sin(2\alpha-\tau)+\sigma(1-\sin\tau)\right\}\\
&\overset{\sigma=1}{=} \frac{2}{\pi}\left\{1-\sin\alpha\right\}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}\zeta &= \frac{2}{2\pi}\int_{-\pi}^{\pi}(\cos(\theta-2\alpha))_+\vee(\sigma\cos\theta)_+d\theta\\
&= \frac{1}{\pi}\left\{\sigma\int_{-\pi/2}^{\tau}\cos\theta d\theta+\int_{\tau}^{\pi/2+2\alpha}\cos(\theta-2\alpha)d\theta\right\}\\
&= \frac{1}{\pi}\left\{\sigma(1+\sin\tau)+1+\sin(2\alpha-\tau)\right\}\\
&\overset{\sigma=1}{=} \frac{2}{\pi}\left\{1+\sin\alpha\right\}
\end{aligned}
$$

Moreover

$$
\begin{aligned}
\mathbb{E}\xi^2 &= \mathbb{E}\left[\{(X)_+\wedge(Y)_++(X)_-\wedge(Y)_-\}/T\right]^2\\
&= \frac{2}{2\pi}\int_{-\pi}^{\pi}\left\{(\cos(\theta-2\alpha))_+\wedge(\sigma\cos\theta)_+\right\}^2d\theta\\
&= \frac{1}{\pi}\left\{\int_{-\pi/2+2\alpha}^{\tau}\cos^2(\theta-2\alpha)d\theta+\sigma^2\int_{\tau}^{\pi/2}\cos^2\theta d\theta\right\}\\
&= \frac{1}{2\pi}\left\{\left(\theta+\frac{1}{2}\sin(2\theta-4\alpha)\right)\Big|_{-\pi/2+2\alpha}^{\tau}+\sigma^2\left(\theta+\frac{1}{2}\sin(2\theta)\right)\Big|_{\tau}^{\pi/2}\right\}\\
&= \frac{1}{2\pi}\left\{\left(\tau+\frac{1}{2}\sin(2\tau-4\alpha)\right)-(-\pi/2+2\alpha)+\sigma^2\left(\pi/2-\tau-\frac{1}{2}\sin(2\tau)\right)\right\}\\
&= \frac{1}{2\pi}\left\{\tau+\pi/2-2\alpha+\frac{1}{2}\sin(2\tau-4\alpha)+\sigma^2\left(\pi/2-\tau-\frac{1}{2}\sin(2\tau)\right)\right\}\\
&\overset{\sigma=1}{=} \frac{1}{2\pi}\left(\pi-2\alpha-\sin 2\alpha\right)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\zeta^2 &= \mathbb{E}\left[\{(X)_+\vee(Y)_++(X)_-\vee(Y)_-\}/T\right]^2\\
&= \frac{2}{2\pi}\int_{-\pi}^{\pi}\left\{(\cos(\theta-2\alpha))_+\vee(\sigma\cos\theta)_+\right\}^2d\theta+\frac{4}{2\pi}\int_{-\pi}^{\pi}\left\{(\cos(\theta-2\alpha))_+\times(\sigma\cos\theta)_-\right\}d\theta\\
&= \frac{1}{\pi}\left\{\sigma^2\int_{-\pi/2}^{\tau}\cos^2\theta d\theta+\int_{\tau}^{\pi/2+2\alpha}\cos^2(\theta-2\alpha)d\theta\right\}\\
&\quad -\frac{\sigma}{\pi}\int_{\pi/2}^{\pi/2+2\alpha}\left\{\cos(2\alpha)+\cos(2\theta-2\alpha)\right\}d\theta\\
&= \frac{1}{\pi}\left\{\sigma^2\left(\frac{\theta}{2}+\frac{1}{4}\sin(2\theta)\right)\Big|_{-\pi/2}^{\tau}+\left(\frac{\theta}{2}+\frac{1}{4}\sin(2\theta-4\alpha)\right)\Big|_{\tau}^{\pi/2+2\alpha}\right\}\\
&\quad -\frac{\sigma}{\pi}\left(\cos(2\alpha)\theta+\frac{1}{2}\sin(2\theta-2\alpha)\right)\Big|_{\pi/2}^{\pi/2+2\alpha}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\pi}\left\{\sigma^2\left(\frac{\tau}{2}+\frac{1}{4}\sin(2\tau)+\frac{\pi}{4}\right)+\left(\frac{\pi}{4}+\alpha-\frac{\tau}{2}-\frac{1}{4}\sin(2\tau-4\alpha)\right)\right\}+\frac{\sigma}{\pi}\left(\sin 2\alpha-2\alpha\cos 2\alpha\right)\\
&\overset{\sigma=1}{=} \frac{1}{\pi}\left(\frac{\pi}{2}+\alpha+\frac{3}{2}\sin 2\alpha-2\alpha\cos 2\alpha\right)
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}(\xi\zeta) &= \mathbb{E}\left[\{(X)_+\wedge(Y)_+ + (X)_-\wedge(Y)_-\}\{(X)_+\vee(Y)_+ + (X)_-\vee(Y)_-\}/T^2\right]\\
&= \frac{2}{2\pi}\int_{-\pi}^{\pi}\{(\cos(\theta-2\alpha))_+\times(\sigma\cos\theta)_+\}\,d\theta\\
&= \frac{\sigma}{2\pi}\int_{-\pi/2+2\alpha}^{\pi/2}\{\cos(2\alpha)+\cos(2\theta-2\alpha)\}\,d\theta\\
&= \frac{\sigma}{2\pi}\left\{\theta\cos(2\alpha)+2^{-1}\sin(2\theta)\right\}\Big|_{-\pi/2+\alpha}^{\pi/2-\alpha}\\
&= \frac{\sigma}{2\pi}\left((\pi-2\alpha)\cos 2\alpha+\sin 2\alpha\right)\\
&\overset{\sigma=1}{=} \frac{1}{2\pi}\left((\pi-2\alpha)\cos 2\alpha+\sin 2\alpha\right)
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
V =&\mathbb{E}\xi^2(\mathbb{E}\zeta)^2+\mathbb{E}\zeta^2(\mathbb{E}\xi)^2-2\mathbb{E}\xi\mathbb{E}\zeta\mathbb{E}(\xi\zeta)\\
=&\frac{1}{4\pi^3}\left\{2\tau+\pi-4\alpha+\sin(2\tau-4\alpha)+\sigma^2\left(\pi-2\tau-\sin(2\tau)\right)\right\}\{\sigma(1+\sin\tau)+1+\sin(2\alpha-\tau)\}^2\\
&+\frac{1}{4\pi^3}\left\{\sigma^2\left(2\tau+\sin(2\tau)+\pi\right)+\left(\pi+4\alpha-2\tau-\sin(2\tau-4\alpha)\right)+4\sigma\left(\sin 2\alpha-2\alpha\cos 2\alpha\right)\right\}\\
&\quad\times\{1-\sin(2\alpha-\tau)+\sigma(1-\sin\tau)\}^2\\
&-\frac{\sigma}{\pi^3}\left((\pi-2\alpha)\cos 2\alpha+\sin 2\alpha\right)\{1-\sin(2\alpha-\tau)+\sigma(1-\sin\tau)\}\{\sigma(1+\sin\tau)+1+\sin(2\alpha-\tau)\}
\end{aligned}
$$

The expression can be simplified when $\sigma=1$:

$$
\begin{aligned}
\mathbb{E}\xi &\overset{\sigma=1}{=} \frac{2}{\pi}\{1-\sin\alpha\}\\
\mathbb{E}\zeta &\overset{\sigma=1}{=} \frac{2}{\pi}\{1+\sin\alpha\}\\
\mathbb{E}\xi^2 &\overset{\sigma=1}{=} \frac{1}{2\pi}\left(\pi-2\alpha-\sin 2\alpha\right)\\
\mathbb{E}\zeta^2 &\overset{\sigma=1}{=} \frac{1}{\pi}\left(\frac{\pi}{2}+\alpha+\frac{3}{2}\sin 2\alpha-2\alpha\cos 2\alpha\right)\\
\mathbb{E}(\xi\zeta) &\overset{\sigma=1}{=} \frac{1}{2\pi}\left((\pi-2\alpha)\cos 2\alpha+\sin 2\alpha\right)
\end{aligned}
$$

Thus, when $\sigma=1$, we have

$$
\begin{aligned}
V =&\mathbb{E}\xi^2(\mathbb{E}\zeta)^2+\mathbb{E}\zeta^2(\mathbb{E}\xi)^2-2\mathbb{E}\xi\mathbb{E}\zeta\mathbb{E}(\xi\zeta)\\
=&\frac{1}{2\pi}\left(\pi-2\alpha-\sin 2\alpha\right)\left[\frac{2}{\pi}\left(1+\sin\alpha\right)\right]^2+\frac{1}{\pi}\left(\frac{\pi}{2}+\alpha+\frac{3}{2}\sin 2\alpha-2\alpha\cos 2\alpha\right)\left[\frac{2}{\pi}\left(1-\sin\alpha\right)\right]^2\\
&-2\frac{1}{2\pi}\left((\pi-2\alpha)\cos 2\alpha+\sin 2\alpha\right)\frac{2}{\pi}\left(1-\sin\alpha\right)\frac{2}{\pi}\left(1+\sin\alpha\right)\\
=&\frac{4}{\pi^3}\sin^2\alpha\left(3\pi-8\cos\alpha+2\sin 2\alpha+\pi\cos 2\alpha-8\alpha\sin\alpha-4\alpha\cos 2\alpha\right)
\end{aligned}
$$

For a bivariate t-distribution with ν degrees of freedom, we have $T \sim \sqrt{\chi_2^2 \nu / \chi_\nu^2} \sim \sqrt{2F_{2,\nu}}$, and

$$\mathbb{E}T^2 = 2\mathbb{E}\{F_{2,\nu}\} = \frac{2\nu}{\nu-2}, \qquad \mathbb{E}T = \frac{\sqrt{\pi}}{2}\frac{\Gamma(\nu/2-1/2)\sqrt{\nu}}{\Gamma(\nu/2)}$$

Thus, when $\mathbb{E}T^2 < \infty$, we have the asymptotic normality

$$n^{1/2}\left(\frac{\sum_{i=1}^n T_i\xi_i}{\sum_{i=1}^n T_i\zeta_i} - f_\infty(\rho,\sigma)\right) \xrightarrow{\mathrm{D}} N\left(0, \frac{V\mathbb{E}T^2}{(\mathbb{E}T)^2(\mathbb{E}\zeta)^4}\right).$$

For t-distribution with $\nu = 2$, condition (25) holds as

$$\frac{t\,\mathbb{P}(T^2 > t)}{\mathbb{E}\min(T^2,t)} = \frac{t(1+t/2)^{-1}}{\int_0^t (1+x/2)^{-1}dx} \asymp \frac{1}{\log t} \to 0.$$

Moreover, $\mathbb{P}(\max_{i\le n} T_i^2 > n/\epsilon) = O(\epsilon)$, $\mathbb{E}(T^2\wedge(n/\epsilon)) \approx 2\log n$ and $\mathbb{E}(T^2\wedge(n/\epsilon))^2 = O(n)$, so that

$$\frac{\sum_{i=1}^n T_i^2}{2n\log n} = 1 + O_P(1/\log n).$$

Thus, for $\nu = 2$,

$$\left(\frac{n}{\log n}\right)^{1/2}\left(\frac{\sum_{i=1}^n T_i\xi_i}{\sum_{i=1}^n T_i\zeta_i} - f_\infty(\rho,\sigma)\right) \xrightarrow{\mathrm{D}} N\left(0, \frac{4V}{\pi^2(\mathbb{E}\zeta)^4}\right).$$

C Proof of Theorem 3

[1] provides the result for the normal case. We extend the results of [1] to the general elliptical family. Again, a vector (X,Y) has an elliptical distribution if

$$(X,Y) = T(\xi,\zeta),\ (\xi,\zeta)^T = AU = \begin{pmatrix} a_1^T U \\ a_2^T U \end{pmatrix}$$

where $A = (a_1, a_2)^T$ is a deterministic 2×2 matrix, U is a vector uniformly distribution in the unit circle and T is a positive random variable independent of U. We want to compute the asymptotic variance of the sample correlation

$$\widehat{\rho}_n = \frac{\sum_{i=1}^n X_iY_i}{\sqrt{\sum_{i=1}^n\sum_{j=1}^n X_i^2Y_j^2}}$$

Due to scale invariance, it suffices to consider the case of $\mathbb{E}X^2 = \mathbb{E}Y^2 = 1$.

$$\begin{aligned}
\widehat{\rho}_n - \rho &= \frac{\sum_{i=1}^n X_iY_i/n - \rho}{\sqrt{\sum_{i=1}^n\sum_{j=1}^n X_i^2Y_j^2/n^2}} + \rho\frac{1-\sqrt{\sum_{i=1}^n\sum_{j=1}^n X_i^2Y_j^2/n^2}}{\sqrt{\sum_{i=1}^n\sum_{j=1}^n X_i^2Y_j^2/n^2}} \\
&= \sum_{i=1}^n \frac{X_iY_i}{n} - \rho + \rho\frac{1-\sum_{i=1}^n\sum_{j=1}^n X_i^2Y_j^2/n^2}{2} + O_P(1/n) \\
&= \sum_{i=1}^n \frac{X_iY_i}{n} - \rho + \frac{\rho}{2}\Big(1-\sum_{i=1}^n\frac{X_i^2}{n}\Big) + \frac{\rho}{2}\Big(1-\sum_{i=1}^n\frac{Y_i^2}{n}\Big) + O_P(1/n).
\end{aligned}$$

Thus, the asymptotic variance of $\widehat{\rho}$ is

$$
\begin{aligned}
V &= \mathbb{E}\Big(XY - \rho - (\rho/2)(X^2 + Y^2 - 2)\Big)^2 \\
&= \mathbb{E}\Big(XY - (\rho/2)(X^2 + Y^2)\Big)^2 \\
&= \mathbb{E}\Big(T^2\{\xi\zeta - (\rho/2)(\xi^2 + \zeta^2)\}\Big)^2. \\
&= \mathbb{E}T^4\ \mathbb{E}\Big(\xi\zeta - (\rho/2)(\xi^2 + \zeta^2)\Big)^2.
\end{aligned}
$$

Let $\mathbb{E}_0$ be the expectation in the Gaussian case. We have $T^2 \sim \chi_2^2$ under $\mathbb{E}_0$, $\mathbb{E}_0 T^2 = 2$, $\mathbb{E}_0 T^4 = \mathrm{Var}_0(T^2) + (\mathbb{E}T_0^2)^2 = 4 + 4 = 8$, and $V_0 = (1-\rho^2)^2$. A comparison with the solution in the Gaussian case yields

$$
\begin{aligned}
V &= \frac{\mathbb{E}T^4(\mathbb{E}_0T^2)^2}{(\mathbb{E}T^2)^2\mathbb{E}_0T^4}\left\{\frac{\mathbb{E}_0T^4}{(\mathbb{E}_0T^2)^2}\mathbb{E}\Big(\xi\zeta - (\rho/2)(\xi^2 + \zeta^2)\Big)^2\right\} \\
&= \frac{4\mathbb{E}T^4}{8(\mathbb{E}T^2)^2}\left\{\frac{\mathbb{E}_0T^4}{(\mathbb{E}_0T^2)^2}\mathbb{E}_0\Big(\xi\zeta - (\rho/2)(\xi^2 + \zeta^2)\Big)^2\right\} \\
&= \frac{\mathbb{E}T^4}{2(\mathbb{E}T^2)^2}\mathbb{E}_0\Big(XY - \rho - (\rho/2)(X^2 + Y^2 - 2)\Big)^2\Bigg|_{\mathbb{E}_0X^2 = \mathbb{E}_0Y^2 = 1} \\
&= \frac{\mathbb{E}T^4}{2(\mathbb{E}T^2)^2}(1-\rho^2)^2.
\end{aligned}
$$