Statistica Sinica Preprint No: SS-2020-0334	
Title	Quantification of model bias underlying the phenomenon of Einstein from Noise
Manuscript ID	SS-2020-0334
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0334
Complete List of Authors	Shao-Hsuan Wang, Yi-Ching Yao, Wei-Hau Chang and I-Ping Tu
Corresponding Author	I-Ping Tu
E-mail	iping@stat.sinica.edu.tw
Notice: Accepted version subject to English editing.	

Statistica Sinica

Quantification of model bias underlying the phenomenon of "Einstein from noise"

Shao-Hsuan Wang, Yi-Ching Yao, Wei-Hau Chang and I-Ping Tu

Academia Sinica

Abstract: Arising from cryogenic electron microscopy image analysis, "Einstein from noise" is a phenomenon of significant statistical interest because spurious patterns could easily emerge by averaging a large number of white-noise images aligned to a reference image through rotation and translation. While this phenomenon is often attributed to model bias, quantitative studies on such a bias are lacking. Here, we introduce a simple framework under which an image of p pixels is treated as a vector of dimension p and a white-noise image is a random vector uniformly sampled from the $(p-1)$-dimensional unit sphere. Moreover, we adopt the cross correlation of two images which is a similarity measure based on the dot product of image pixels. This framework geometrically explains how the bias results from averaging a properly chosen set of white-noise images that are most highly cross-correlated with the reference image. We quantify the bias in terms of three parameters: the number of white-noise images (n), the image dimension (p), and the size of the selection set (m). Under the conditions that n, p and m are all large and $(\ln n)^2/p$ and m/n are both small, we show that the bias is approximately $\sqrt{\frac{2\gamma}{1+2\gamma}}$ where $\gamma = \frac{m}{p} \ln\left(\frac{n}{m}\right)$.

Key words and phrases: Cryogenic electron microscopy; cross correlation; extreme value distribution; model bias; high dimensional data analysis; white-noise image

1. Introduction

The terminology of the phenomenon of "Einstein from noise" comes from the literature of cryogenic electron microscopy (cryo-EM). It refers to an artefact of model bias that arises from averaging a large number of cryo-EM images aligned to a reference (model) image. This artefact of model bias is strongly associated with the noisy nature of cryo-EM images.

Developed for imaging biological macromolecules preserved at a frozen-hydrated state, cryo-EM has become a major tool for high-resolution structure determination of molecules because of its recent breakthroughs in resolution. In contrast to X-ray crystallography, cryo-EM does not need crystals, and thereby is amenable to structure determination of proteins that are refractory to crystallization, including, in particular, membrane proteins (Liao et al., 2013) and molecular complexes that exhibit dynamic conformation behaviors (Yan et al., 2015). To recognize its great success with far-reaching applications, the Nobel Prize in Chemistry in 2017 was awarded to J. Dubochet, J. Frank and R. Henderson for their pioneering contributions to the development of cryo-EM.

A technical difficulty encountered by the cryo-EM technique is that during the experiment of imaging molecules, the orientation of each molecule is not recorded which needs to be estimated at the post-imaging stage. However, to mitigate radiation damages, only a minimal dose of electron can be used for acquiring the projection images of individual molecules (called 2D particle images). The resulting cryo-EM images are extremely noisy with the signal-to-noise ratio less than 0.1. A typical cryo-EM experiment tends to collect a large number of particle images in hope of compensating the noise contamination by averaging, where the dimension of a particle image is extremely high (larger than one hundred by one hundred). Hence, the data characters of cryo-EM images, including strong noise contamination, huge dimension and large sample size, make its processing and statistical analysis very challenging. Henderson (2013) further pointed out how spurious patterns could easily emerge by averaging a large number of white-noise images aligned to a reference image through rotation and translation. Specifically, he referred to the work of Stewart and Grigorieff (2004) in which an experiment was conducted by generating 1000 white-noise images and aligning each of them to Einstein's facial image through rotation and translation. A blurred Einstein's face emerged from averaging the 1000 aligned images, which Henderson (2013) dubbed "Einstein from noise" and used it to give unwary cryo-EM users a warning that an incorrect 3D density map could be constructed if data are blindly fitted to a reference model.

In a recent review paper, Lai et al. (2020) discussed the "Einstein from noise" phenomenon from a statistical perspective. To avoid the technical issue of how rotating an image may destroy the pixel format, they considered a simple mathematical framework under which an image of p pixels is treated as a vector of dimension p and a white-noise image is a random vector uniformly distributed on the $(p-1)$-dimensional unit sphere. The cross correlation of two images is adopted which is a similarity measure based on the dot product of image pixels and is widely used in image processing. Under this framework, we present in Section 2 a simulation study with $n = 2 \times 10^6$ white-noise images with the pixel number $p = 120 \times 120$. Among the 2×10^6 white-noise images, the largest cross correlation value with Einstein's facial image (the reference) is merely 0.039, while the cross correlation increases dramatically to 0.650 after averaging the $m = 800$ images that have the largest cross correlation values with Einstein's facial image. This illustrates the essence of the "Einstein from noise" phenomenon. The objective of the present paper is to provide a thorough study of the "Einstein from noise" phenomenon based on the statistical perspective laid out in Lai et al. (2020). A main task is to approximate the distribution of the cross correlation between the (normalized) average of the m selected images and the reference, which is referred to the (image selection) bias. While the bias depends on the three parameters n, p, and m in a convoluted manner, under the conditions that n, p and m are all large and $(\ln n)^2/p$ and m/n are both small, we show that the bias is approximately $\sqrt{\frac{2\gamma}{1+2\gamma}}$ where $\gamma = \frac{m}{p} \ln\left(\frac{n}{m}\right)$.

The rest of this paper is organized as follows. Section 2 introduces notation, terminology and the statistical model as well as demonstrates the phenomenon of "Einstein from noise". Section 3 consists of two parts: (i) presenting an extreme value theory for the distribution of the largest cross correlation value as n and p both tend to infinity and (ii) stating asymptotic results on the bias as n, p, and m all tend to infinity. The theoretical results in part (ii) are validated via simulation as presented in Section 4. Section 5 contains concluding remarks. Proofs of the asymptotic results in Section 3 are relegated to the Appendix. The online supplementary material contains the proofs of auxiliary lemmas.

2. Statistical Model

2.1 Notation, terminology, and model

Let $\boldsymbol{R}$ be the reference matrix (the digital version of the reference image) of dimension $d_1 \times d_2$. We assume that $\|\boldsymbol{R}\| = 1$ where $\| \cdot \|$ denotes the Frobenius norm of a matrix or Euclidean norm of a vector. We generate n independent and identically distributed (iid) white-noise images as follows. Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ be iid $d_1 \times d_2$ random matrices such that the $d_1 d_2$ components of each $\boldsymbol{Z}_i$ are iid standard normal. We refer to $\boldsymbol{Z}_i / \|\boldsymbol{Z}_i\|$, $i = 1, \ldots, n$ (the normalized version of $\boldsymbol{Z}_i$) as n iid white-noise images.

Let $\boldsymbol{r} = \text{vec}(\boldsymbol{R})$, the p-dimensional column vector which is the vectorized version of $\boldsymbol{R}$, where $p = d_1 d_2$. The fact that $\|\boldsymbol{r}\| = 1$ implies $\boldsymbol{r} \in \mathcal{S}^{p-1}$ (the $(p-1)$-dimensional unit sphere). Let $\boldsymbol{X}_i = \text{vec}(\boldsymbol{Z}_i)/\|\boldsymbol{Z}_i\|$. Thus, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are iid uniformly distributed on $\mathcal{S}^{p-1}$. We refer to both $\boldsymbol{Z}_i/\|\boldsymbol{Z}_i\|$ and $\boldsymbol{X}_i$ as the i-th white-noise image. With $\boldsymbol{r}^\top$ denoting the transpose of $\boldsymbol{r}$, the cross correlation of $\boldsymbol{X}_i$ and $\boldsymbol{r}$ (or equivalently $\boldsymbol{Z}_i/\|\boldsymbol{Z}_i\|$ and $\boldsymbol{R}$) is defined as $\boldsymbol{r}^\top \boldsymbol{X}_i$, the inner product (dot product) of $\boldsymbol{X}_i$ and $\boldsymbol{r}$, which is a similarity measure of two images. Note that $\boldsymbol{r}^\top \boldsymbol{X}_i = \cos \Theta_i$, where Θ_i is the angle between $\boldsymbol{r}$ and $\boldsymbol{X}_i$.

The n white-noise images are ordered (and denoted by $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}$)

according to their cross correlation values with $\boldsymbol{r}$. In other words, $(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)})$ is a permutation of $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ such that $\boldsymbol{r}^\top \boldsymbol{X}^{(1)} \geq \boldsymbol{r}^\top \boldsymbol{X}^{(2)} \geq \cdots \geq \boldsymbol{r}^\top \boldsymbol{X}^{(n)}$. Let $\Theta_{1:n} \leq \Theta_{2:n} \leq \cdots \leq \Theta_{n:n}$ be the order statistics of the angles $(\Theta_1, \cdots, \Theta_n)$, so that $\cos\Theta_{i:n} = \boldsymbol{r}^\top \boldsymbol{X}^{(i)}$, $i = 1, \ldots, n$. Let $\overline{\boldsymbol{X}}_m = m^{-1}\sum_{i=1}^m \boldsymbol{X}^{(i)}$. Then $\overline{\boldsymbol{X}}_m / \|\overline{\boldsymbol{X}}_m\| \in \mathcal{S}^{p-1}$ is the normalized average of the m white-noise images that are most highly cross-correlated with the reference image. Our goal is to find a good approximation of the distribution of $\rho_{n,p,m} = \boldsymbol{r}^\top \overline{\boldsymbol{X}}_m / \|\overline{\boldsymbol{X}}_m\|$ when n, p, and m are large. Note that for $m = 1$, $\rho_{n,p,1} = \boldsymbol{r}^\top \boldsymbol{X}^{(1)} = \cos\Theta_{1:n}$, is the largest cross correlation value. Note also that the distribution of $\rho_{n,p,m}$ does not depend on $\boldsymbol{r}$ which is due to the fact that if $\boldsymbol{X}$ is uniformly distributed on $\mathcal{S}^{p-1}$, then the distribution of $\boldsymbol{r}^\top \boldsymbol{X}$ is independent of $\boldsymbol{r}$.

2.2 Demonstration of the "Einstein from noise" phenomenon

We now present two figures summarizing the simulation study described in Section 1, where $n = 2 \times 10^6$, $p = d_1 \times d_2 = 120 \times 120 = 14400$, and $m = 1, 200, 400, 800$. In Figure 1, the leftmost (reference) image is Einstein's face, and the other 4 images correspond to $\overline{\boldsymbol{X}}_m / \|\overline{\boldsymbol{X}}_m\|$ for $m = 1, 200, 400, 800$. The second image from the left corresponds to $\boldsymbol{X}^{(1)}$, whose cross correlation (CC) value with Einstein's facial image is 0.039 (which is the largest among the 2×10^6 white-noise images generated in the simulation). While this image is rather noisy, Einstein's face emerges in the other 3 images with different degrees of blurring, corresponding to CC values 0.426, 0.536, and 0.650.

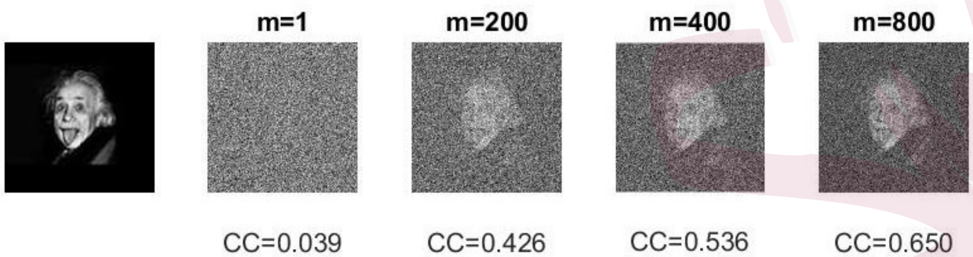

Figure 1: Example with Einstein's face as the reference image.

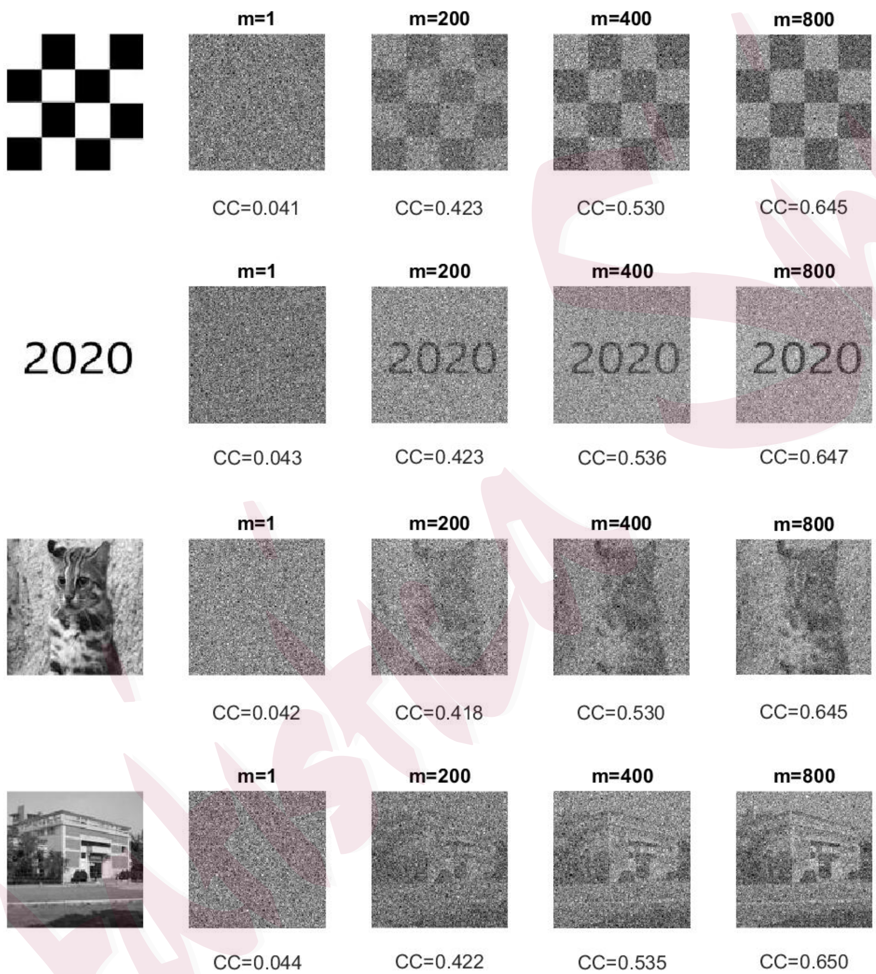

Figure 2: The phenomenon of "Einstein from noise" is shown across various reference images.

Figure 2 shows similar results with four different reference images of a simple chessboard, digits of 2020, a leopard cat and Statistics Building of Academia Sinica, indicating that the phenomenon of "Einstein from noise" is robust across various reference images. The cross correlation values in Figure 2 are about the same across different reference images, which can be explained by the previously mentioned fact that if $\boldsymbol{X}$ is uniformly distributed on $\mathcal{S}^{p-1}$, then the distribution of $\boldsymbol{r}^\top \boldsymbol{X}$ is independent of $\boldsymbol{r}$.

3. Asymptotic theory

3.1 Extreme value theory for the largest cross correlation

Recall that $\cos\Theta_{1:n}$ is the largest cross correlation. The following theorem provides an approximation to the distribution of $\cos\Theta_{1:n}$ when n and p are large.

Theorem 1. *Let*

$$K_{n,p} = -\ln n + \frac{1}{2}\ln\ln n - \frac{1}{2}\ln\left(\frac{2^{\frac{\ln n}{p}}}{1-\exp\left(-2^{\frac{\ln n}{p}}\right)}\right) + \frac{1}{2}\ln(4\pi). \tag{1}$$

We have

$$(p-1)\ln(\sin\Theta_{1:n}) - K_{n,p} \xrightarrow{d} G \quad \textit{uniformly as} \quad n\wedge p\to\infty, \tag{2}$$

where $n \wedge p = \min\{n, p\}$, $\xrightarrow{d}$ denotes convergence in distribution, and the cumulative distribution function of G is given by $G(t) = 1 - e^{-e^t}$, $t \in \mathbb{R}$, which is known as the extreme value distribution of Gumbel type.

Based on (2), for $0 < \alpha < 1$, the approximate 100α-th quantile of the distribution of $\cos\Theta_{1:n}$ is

$$M_{n,p}(\alpha) = \sqrt{1 - \exp\{2(K_{n,p} + \ln\ln\, \alpha^{-1})/(p-1)\}}.$$

Recall that $\cos\Theta_{1:n} = 0.039$ in the simulation study summarized in Figure 1, where $n = 2 \times 10^6$ and $p = 120 \times 120$. This observed value is compatible with the approximate 10th quantile $M_{n,p}(0.1) = 0.039$.

Figure 3 plots $M_{n,p}(\alpha)$ versus $\log_{10} n$ for $n \le 10^{100}$ with $p = 120 \times 120$ and $\alpha = .05, .5, .95$. Note that the three quantile curves are very close to each other, indicating that $\cos\Theta_{1:n}$ has a small standard deviation. Figure 3 suggests that for $\mathrm{P}(\cos\Theta_{1:n} \ge 0.1)$ to be at least 0.05, n is required to be greater than 10^{30}, and for $\mathrm{P}(\cos\Theta_{1:n} \ge 0.15)$ to be at least 0.05, n is required to be greater than 10^{70}. In other words, it is unlikely for any of n iid white-noise images of dimension 120×120 to have a cross correlation value with Einstein's face greater than 0.15 unless n is astronomically large.

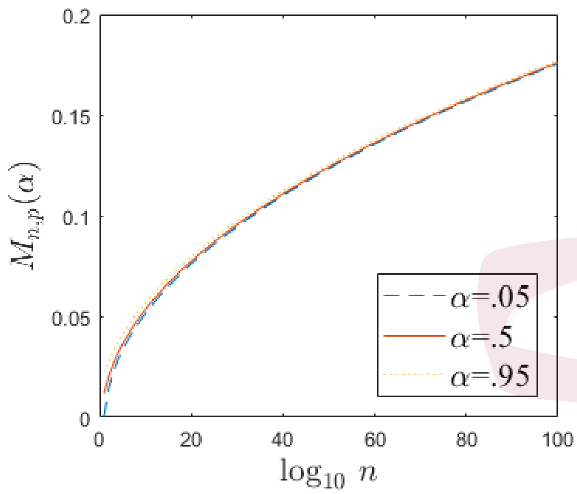

Figure 3: The approximate 100α-th quantile of the distribution of $\cos\Theta_{1:n}$ ($M_{n,p}(\alpha)$) versus $\log_{10} n$ with $p = 120 \times 120$, $\alpha = .05,\ .5,\ .95$.

3.2 Asymptotic results on $\rho_{n,p,m}$

When $p = p_n$ and $m = m_n$ both grow with n, asymptotic expansions for the distribution of $\rho_{n,p,m}$ are more involved. Our analysis requires the condition $(\ln n)^2/p = o(1)$ (which is stronger than $(\ln n)/p = o(1)$), so that terms such as $(\ln n)(\ln\ln n)/p$ become negligible. Let

$$\beta_{n,p,m} = \frac{m}{p}\left\{2\ln\frac{n}{m} - \ln\ln\frac{n}{m} - \ln(4\pi) + 2\right\},$$

which is a model bias index. While the quantity $\rho_{n,p,m}$ plays an important role in our asymptotic results below, we are unaware of any heuristic interpretation of this quantity.

Theorem 2. *Let $p = p_n \to \infty$ satisfy $(\ln n)^2/p = o(1)$ and $m = m_n \to \infty$ satisfy $m/n = o(1)$. Then*

$$\rho^2_{n,p,m} = \frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}\left(1 + o_p(1)\right).$$

Consequently, $\rho^2_{n,p,m} - \dfrac{\beta_{n,p,m}}{1+\beta_{n,p,m}} \to 0$ in probability.

Theorem 3. *Let $p = p_n \to \infty$ satisfy $(\ln n)^2/p = o(1)$ and $m = m_n \to \infty$ satisfy $m(\ln\ln n)^4/(\ln n)^2 = o(1)$. Then*

$$\alpha_{n,p,m}\left(\rho^2_{n,p,m} - \frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}\right) \xrightarrow{d} N(0,1),$$

where $\alpha_{n,p,m} = p\left(8m + 2p\,\beta^2_{n,p,m}\right)^{-1/2}\left(1+\beta_{n,p,m}\right)^2$ and $N(0,1)$ denotes the standard normal distribution.

Corollary 1. *Let $p = p_n \to \infty$ and $m = m_n \to \infty$.*

(i) If $(\ln n)^2/p = o(1)$ and $m/n = o(1)$, then

$$\frac{\rho_{n,p,m}}{\sqrt{\beta_{n,p,m}/(1+\beta_{n,p,m})}} = 1 + o_p(1).$$

Consequently,

$$\rho_{n,p,m} = \sqrt{\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}} + o_p(1) \quad \text{and} \quad \mathrm{E}(\rho_{n,p,m}) = \sqrt{\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}} + o(1).$$

(ii) In additional to the conditions specified in (i), if $m\,(\ln\ln\, n)^4/(\ln\, n)^2 = o(1)$, then

$$\tilde{\alpha}_{n,p,m}\left(\rho_{n,p,m} - \sqrt{\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}}\right) \xrightarrow{d} N(0,1),$$

where $\tilde{\alpha}_{n,p,m} = 2\alpha_{n,p,m}\sqrt{\beta_{n,p,m}/(1+\beta_{n,p,m})}$.

Remark 1. On top of the condition $(\ln n)^2/p = o(1)$, Theorem 2 only requires the mild condition $m/n = o(1)$. Let $\gamma_{n,p,m} = \frac{m}{p}\ln\frac{n}{m}$. Since $\beta_{n,p,m} = 2\gamma_{n,p,m}(1+o(1))$ (i.e. $2\gamma_{n,p,m}$ is the leading term of $\beta_{n,p,m}$), Theorem 2 implies

$$\rho^2_{n,p,m} = \frac{2\gamma_{n,p,m}}{1+2\gamma_{n,p,m}} + o_p(1).$$

Consequently,

$$\rho_{n,p,m} = \sqrt{\tfrac{2\gamma_{n,p,m}}{1+2\gamma_{n,p,m}}} + o_p(1) \quad \text{and} \quad \mathrm{E}(\rho_{n,p,m}) = \sqrt{\tfrac{2\gamma_{n,p,m}}{1+2\gamma_{n,p,m}}} + o(1). \qquad (3)$$

Remark 2. To establish asymptotic normality of $\rho^2_{n,p,m}$ (and $\rho_{n,p,m}$), Theorem 3 (and Corollary 1) requires the stringent condition $m(\ln\ln n)^4/(\ln n)^2 = o(1)$. It is unclear whether asymptotic normality still holds when m grows at a rate faster than $(\ln n)^2/(\ln\ln n)^4$. It should also be remarked that under the conditions as in Theorem 3, it is not true that $\alpha_{n,p,m}\left(\rho^2_{n,p,m} - \frac{2\gamma_{n,p,m}}{1+2\gamma_{n,p,m}}\right) \xrightarrow{d} N(0,1)$. This shows that while $2\gamma_{n,p,m}$ is the leading term of $\beta_{n,p,m}$, the remaining terms also play a non-negligible role in the proof of asymptotic normality.

Remark 3. Fan et al. (2018) developed an asymptotic theory to approximate the distribution of the maximum spurious correlation of a response variable Y with the best m linear combinations of p covariates $\boldsymbol{X}$ based on an iid sample of size n when $\boldsymbol{X}$ and Y are independent. See also Fan et al. (2012) for related results. In our setting, the quantity $\rho_{n,p,m}$ may be referred to as the spurious cross correlation of the reference with the normalized average of the m white-noise images that are most highly cross-correlated with the reference. Indeed, with the roles of n and p reversed, $\rho_{n,p,m}$ corresponds to another spurious correlation of the response variable Y with the the average of the m (standardized) covariates in $\boldsymbol{X}$ that are most highly correlated with Y when the p covariates in $\boldsymbol{X}$ and Y are all mutually independent.

4. Simulation Results on $\rho_{n,p,m}$

By Corollary 1(i), if m is small compared to n and $(\ln\, n)^2$ is small compared to p, then $\mathrm{E}(\rho_{n,p,m})$ is expected to be close to $\sqrt{\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}}$ while the standard deviation (s.d.) of $\rho_{n,p,m}$ is expected to be small. We conducted a simulation study of the distribution of $\rho_{n,p,m}$ for various combinations of (n, p, m) with $n = 10^4, 10^5$, $p = 10^4, 4 \times 10^4$, and $m = 100, 200, 400, 600$. The results are reported in Tables 1 and 2 where $\mathrm{E}(\rho_{n,p,m})$ and s.d.$(\rho_{n,p,m})$ were estimated based on 1000 replications for each case. While $\sqrt{\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}}$ approximates $\mathrm{E}(\rho_{n,p,m})$ well, it slightly overestimates $\mathrm{E}(\rho_{n,p,m})$, more notably for $n = 10^4$. Clearly, $\mathrm{E}(\rho_{n,p,m})$ increases as n or m increases or p decreases. On the other hand, s.d.$(\rho_{n,p,m})$ is small ($< .005$) in all cases. Besides, s.d.$(\rho_{n,p,m})$ decreases as n or p increases, and is about the same as m varies from 100 to 600. Also included in Tables 1 and 2 are $\tilde{\alpha}_{n,p,m}^{-1}$ and the empirical probability (denoted as Prob.) that

$$\left|\rho_{n,p,m} - \sqrt{\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}}\right| < 1.96\,\tilde{\alpha}_{n,p,m}^{-1}.$$

It is clear from the tables that $\tilde{\alpha}_{n,p,m}^{-1}$ approximates s.d.$(\rho_{n,p,m})$ reasonably well in all cases. By Corollary 1(ii), the Prob. value is expected to be close to .95 if the normal approximation is accurate. By Theorem 3 and Corollary 1, $\alpha_{n,p,m}\left(\rho_{n,p,m}^2 - \frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}\right)$ and $\tilde{\alpha}_{n,p,m}\left(\rho_{n,p,m} - \sqrt{\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}}\right)$ are approximately standard normal under somewhat stringent conditions on the growth rates of m and p as $n \to \infty$. While none of the combinations of (n,p,m) with $n = 10^4, 10^5$, $p = 10^4, 4\times 10^4$ and $m = 100, 200, 400, 600$ seems to satisfy the condition that $m\,(\ln\ln n)^4/(\ln n)^2$ be small, the normal approximation appears to be acceptable for $n = 10^5$ but less satisfactory for $n = 10^4$.

Table 1: $p = 10^4$.

	$n = 10^4$				$n = 10^5$			
m	100	200	400	600	100	200	400	600
$\mathrm{E}(\rho_{n,p,m})$	0.257	0.323	0.395	0.437	0.318	0.408	0.509	0.570
$\sqrt{\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}}$	0.258	0.325	0.399	0.442	0.319	0.409	0.510	0.571
$\mathrm{s.d.}(\rho_{n,p,m})$	0.0043	0.0045	0.0046	0.0048	0.0039	0.0039	0.0040	0.0037
$\tilde{\alpha}_{n,p,m}^{-1}$	0.0051	0.0053	0.0055	0.0057	0.0041	0.0042	0.0040	0.0039
Prob.	0.974	0.967	0.942	0.870	0.967	0.959	0.947	0.953

Table 2: $p = 4 \times 10^4$.

	$n = 10^4$				$n = 10^5$			
m	100	200	400	600	100	200	400	600
$\mathrm{E}(\rho_{n,p,m})$	0.132	0.168	0.210	0.236	0.165	0.218	0.283	0.327
$\sqrt{\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}}$	0.132	0.169	0.212	0.239	0.166	0.219	0.284	0.328
$\mathrm{s.d.}(\rho_{n,p,m})$	0.0022	0.0024	0.0026	0.0027	0.0019	0.0020	0.0021	0.0022
$\tilde{\alpha}_{n,p,m}^{-1}$	0.0026	0.0028	0.0031	0.0033	0.0021	0.0022	0.0023	0.0023
Prob.	0.977	0.978	0.946	0.871	0.968	0.967	0.955	0.953

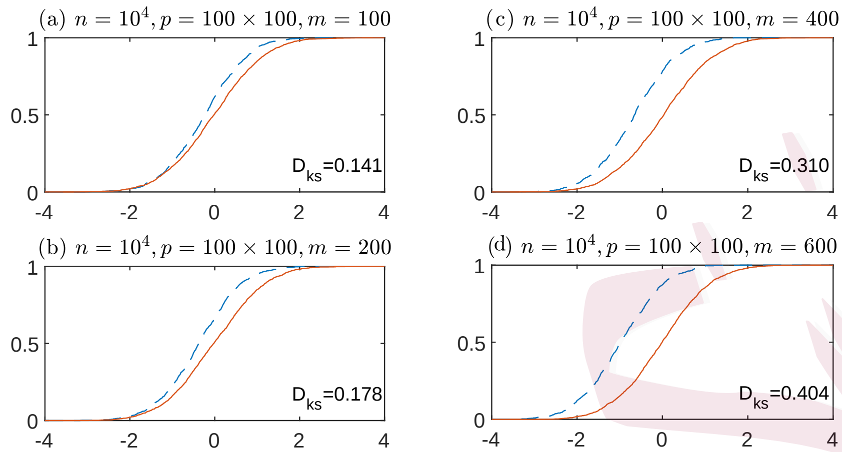

Figure 4: Empirical cdf of $\tilde{\alpha}_{n,p,m}(\rho_{n,p,m} - \sqrt{\beta_{n,p,m}/(1+\beta_{n,p,m})})$ (dashed curves) and standard normal cdf (solid curves): $n = 10^4$, $p = 10^4$.

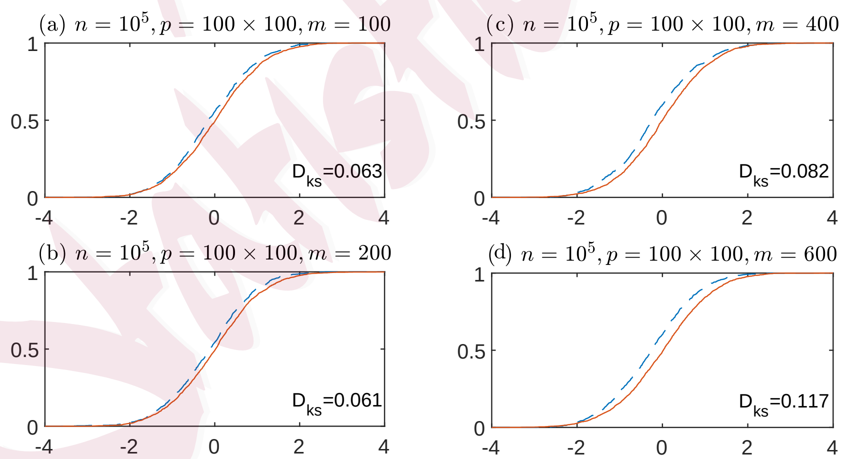

Figure 5: Empirical cdf of $\tilde{\alpha}_{n,p,m}(\rho_{n,p,m} - \sqrt{\beta_{n,p,m}/(1+\beta_{n,p,m})})$ (dashed curves) and standard normal cdf (solid curves): $n = 10^5$, $p = 10^4$.

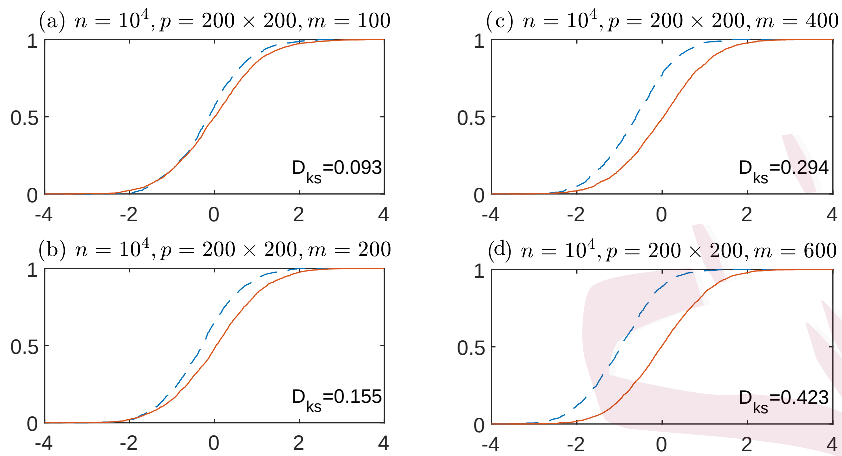

Figure 6: Empirical cdf of $\tilde{\alpha}_{n,p,m}(\rho_{n,p,m} - \sqrt{\beta_{n,p,m}/(1+\beta_{n,p,m})})$ (dashed curves) and standard normal cdf (solid curves): $n = 10^4$, $p = 4 \times 10^4$.

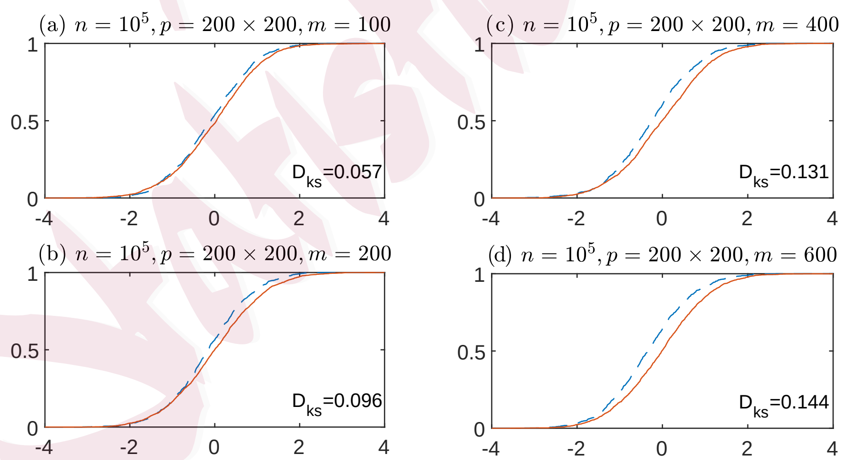

Figure 7: Empirical cdf of $\tilde{\alpha}_{n,p,m}(\rho_{n,p,m} - \sqrt{\beta_{n,p,m}/(1+\beta_{n,p,m})})$ (dashed curves) and standard normal cdf (solid curves): $n = 10^5$, $p = 4 \times 10^4$.

To get a more complete picture of the quality of the normal approximation in Corollary 1(ii), in Figures 4-7, we plot the empirical cumulative distribution function (cdf) of $\tilde{\alpha}_{n,p,m}\left(\rho_{n,p,m} - \sqrt{\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}}\right)$ (based on 1000 replications), along with the standard normal cdf for each combination of (n, p, m). (The value of D_{ks} is the Kolmogorov-Smirnov distance between the two cdfs.) Figures 4-7 are the cdf under four different scenarios, depending on the values of $n = 10^4, 10^5$ and $p = 10^4, 4 \times 10^4$. In each figure, it includes four plots, depending on the values of $m = 100, 200, 400, 600$. The empirical cdf is shifted to the left of the standard normal cdf (more notably for $n = 10^4$ in Figures 4 and 6), indicating that the mean of $\rho_{n,p,m} - \sqrt{\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}}$ is negative. This is consistent with the results in Tables 1 and 2 where $\sqrt{\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}}$ (slightly) overestimates $\mathrm{E}(\rho_{n,p,m})$ (more notably for $n = 10^4$).

5. Concluding Remarks

This paper studied a simple statistical model in order to quantitatively examine the phenomenon of "Einstein from noise". Specifically, for a given reference image of dimension p and a set S_n of n iid white-noise images (with the common uniform distribution on $\mathcal{S}^{p-1}$), we derived the asymptotic behavior of the cross correlation $\rho_{n,p,m}$ between the reference and the normalized average of the m "most biased" members in S_n in the sense that they have the largest cross correlation values with the reference. Our theoretical results indicate that for $m = 1$ and $p = 120 \times 120$, unless n is far beyond the practical range ($> 10^{70}$), $\rho_{n,p,1}$ is small (< 0.15) with high probability, implying that none of n white-noise images even remotely resembles the reference. On the other hand, for m moderately large (≥ 400), $\rho_{n,p,m}$ exceeds 0.5 with high probability if $n = 2 \times 10^6$, in which case a blurred version of the reference emerges from the normalized average of the m most biased members in S_n.

Given a set S_n of n iid white-noise images, Cai et al. (2013) derived the asymptotic distribution of the maximum of all pairwise cross correlations in S_n. See also Cai and Jiang (2011, 2012) and references therein. In the absence of a reference image, their results may be applied to test the null hypothesis that S_n consists of n iid white-noise images. On the other hand, given a reference image, our results can be used to test such a null hypothesis against the alternative that some of the n images in S_n are biased towards the reference by checking whether $\rho_{n,p,m}$ exceeds a threshold (which is determined by the null distribution of $\rho_{n,p,m}$).

Our approach can be directly generalized to tackle a special case of multiple references. Let $\boldsymbol{r}^{(1)}, \ldots, \boldsymbol{r}^{(k)}$ be k given references of dimension

p. Given a set S_n of n iid white-noise images, for $i = 1, \ldots, k$, let $\rho^{(i)}_{n,p,m}$ $(i = 1, \ldots, k)$ denote the cross correlation between $\boldsymbol{r}^{(i)}$ and the normalized average of those m members in S_n having the largest cross correlation values with $\boldsymbol{r}^{(i)}$. It would be of interest to derive the asymptotic distribution of $\max\{\rho^{(i)}_{n,p,m} : i = 1, \ldots, k\}$. If $\boldsymbol{r}^{(1)}, \ldots, \boldsymbol{r}^{(k)}$ are orthogonal (i.e. the pairwise cross correlations are all equal to 0), then it can be argued that $\rho^{(1)}_{n,p,m}, \ldots, \rho^{(k)}_{n,p,m}$ are asymptotically independent, so that the asymptotic distribution of $\max\{\rho^{(i)}_{n,p,m} : i = 1, \ldots, k\}$ can be readily derived by Corollary 1. However, it seems difficult to find the asymptotic distribution of $\max\{\rho^{(i)}_{n,p,m} : i = 1, \ldots, k\}$ when $\boldsymbol{r}^{(1)}, \ldots, \boldsymbol{r}^{(k)}$ are not orthogonal.

The phenomenon of "Einstein from noise" originally arose in the context of cryo-EM image analysis where a key component is image alignment (including rotation and translation). While to address this more complicated problem is beyond the scope of the present paper, it is worth noting that the geometric shape of the reference is likely to play a significant role in the asymptotic theory yet to be developed. As an example, consider a rotationally invariant reference, e.g. an image of a centered wheel. Because of rotational symmetry of the reference, a data image cannot fit the reference any better by rotation. We leave this challenging problem for future work.

Supplementary Material

The online Supplementary Material contains the proofs of Lemmas A6-A8 stated in the Appendix.

Acknowledgements

The authors gratefully acknowledge support by Academia Sinica grant AS-GCS-108-08 and Taiwan's Ministry of Science and Technology grant 106-2118-M-001-001-MY2.

A. Appendix

The Appendix consists of three sections. Section A.1 states some auxiliary lemmas, Section A.2 contains the proof of Theorem 1, and Section A.3 provides the proofs of Theorems 2 and 3 and Corollary 1. For easy reference, a complete list of notations is given in Supplementary Material. Note that if $\boldsymbol{X}$ is uniformly distributed on $\mathcal{S}^{p-1}$, then the distribution of $\boldsymbol{r}^\top \boldsymbol{X}$ is the same for all $\boldsymbol{r} \in \mathcal{S}^{p-1}$. Without loss of generality, we assume $\boldsymbol{r} = (1, 0, \ldots, 0)^\top \in \mathcal{S}^{p-1}$ in what follows.

A.1. Auxiliary lemmas

Lemma A1. *(Lemma 6.2 of Cai and Jiang (2012)) For $t \in (0,1)$, we have*

$$\left(1+\frac{1}{pt^2}\right)^{-1}\frac{1}{(p+2)t}(1-t^2)^{(p+2)/2} \le \int_t^1 (1-u^2)^{p/2}du \le \frac{1}{(p+2)t}(1-t^2)^{(p+2)/2}.$$

Since $\boldsymbol{X}_i$, $i=1,\ldots,n$ are iid uniformly distributed on $\mathcal{S}^{p-1}$ and Θ_i denotes the angle between $\boldsymbol{X}_i$ and $\boldsymbol{r}=(1,0,\ldots,0)^\top$, we have (cf. Eq (5) of Cai et al. (2013)) that Θ_i, $i=1,\ldots,n$ are iid with the common cdf

$$\begin{aligned} F_p(\theta) &= \int_0^\theta \frac{1}{\sqrt{\pi}}\frac{\Gamma(p/2)}{\Gamma((p-1)/2)}(\sin x)^{p-2}dx \\ &= \int_{\cos\theta}^1 \frac{1}{\sqrt{\pi}}\frac{\Gamma(p/2)}{\Gamma((p-1)/2)}(1-u^2)^{\frac{p-3}{2}}du, \ \theta\in[0,\pi]. \qquad (A.1)\end{aligned}$$

Let

$$\overline{F}_p(\theta)=\frac{1}{\sqrt{\pi}}\frac{\Gamma(p/2)}{\Gamma((p-1)/2)}\frac{\sin^{p-1}\theta}{(p-1)|\cos\theta|}. \qquad (A.2)$$

The following lemma is a consequence of Lemma A1.

Lemma A2. *For $\theta\in(0,\pi/2)$ and $p>3$, we have*

$$\left(1+\frac{1}{(p-3)\cos^2\theta}\right)^{-1}\overline{F}_p(\theta)\le F_p(\theta)\le \overline{F}_p(\theta).$$

Let $U_1, U_2, \ldots$ be iid uniform (0,1) random variables and let $U_{1:n} \le \cdots \le U_{n,n}$ denote the order statistics of $U_1,\ldots,U_n$. Let $S_0=0$, and $S_i=\xi_1+\cdots+\xi_i$, $i=1,2,\ldots,$ where $\xi_1,\xi_2,\ldots$ are iid exponential random variables with mean 1. The next lemma is well known; see e.g. Karlin and Taylor (1975). We write $\boldsymbol{X} \overset{d}{=} \boldsymbol{Y}$ if random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are equal in distribution.

Lemma A3. *(i)* $(U_{1:n}, \ldots, U_{n:n}) \overset{d}{=} (S_1, \ldots, S_n)/S_{n+1}$. *(ii)* $(S_1, \ldots, S_n)/S_{n+1}$ *is independent of* S_{n+1}.

Recall that $(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)})$ is a permutation of $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ such that $X_1^{(1)} \le \cdots \le X_1^{(n)}$, where $X_1^{(i)} = \boldsymbol{r}^\top \boldsymbol{X}^{(i)}$ (the first component of $\boldsymbol{X}^{(i)}$). Let $\boldsymbol{V}_i$ and $\boldsymbol{V}^{(i)}$ be defined by $\boldsymbol{X}_i = (X_{i1}, (1 - X_{i1}^2)^{1/2} \boldsymbol{V}_i^\top)^\top$ and $\boldsymbol{X}^{(i)} = (X_1^{(i)}, \nu_i \boldsymbol{V}^{(i)\top})^\top$, where $\nu_i = (1 - X_1^{(i)2})^{1/2}$. In other words, $\boldsymbol{V}_i$ ($\boldsymbol{V}^{(i)}$, respectively) $\in \mathcal{S}^{p-2}$ is the normalized subvector of $\boldsymbol{X}_i$ ($\boldsymbol{X}^{(i)}$, respectively) with the first component deleted.

Lemma A4.

(i) X_{i1} *and* $\boldsymbol{V}_i, i = 1, \ldots, n$ *are all independent.*

(ii) $X_{i1}, i = 1, \ldots, n$ *are iid.*

(iii) $\boldsymbol{V}_i, i = 1, \ldots, n$ *are iid with the uniform distribution on* $\mathcal{S}^{p-2}$.

(iv) $(\boldsymbol{V}^{(1)}, \ldots, \boldsymbol{V}^{(n)})$ *is independent of* $(X_{11}, \ldots, X_{n1})$ *and hence independent of* $(X_1^{(1)}, \ldots, X_1^{(n)})$.

(v) $\boldsymbol{V}^{(i)}, i = 1, \ldots, n$ *are iid with the uniform distribution on* $\mathcal{S}^{p-2}$.

To show Lemma A4, let Z_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$, be iid standard normal, and let

$$\boldsymbol{X}_i^* = (Z_{i1}, \dots, Z_{ip})^\top / \sqrt{\sum_{j=1}^{p} Z_{ij}^2} = (Z_{i1}/\sqrt{\sum_{j=1}^{p} Z_{ij}^2}, \nu_i^* \boldsymbol{V}_i^*)^\top, \; i = 1, \dots, n,$$

where $\nu_i^* = \sqrt{\sum_{j=2}^{p} Z_{ij}^2} / \sqrt{\sum_{j=1}^{p} Z_{ij}^2}$ and $\boldsymbol{V}_i^* = (Z_{i2}, \dots, Z_{ip})^\top / \sqrt{\sum_{j=2}^{p} Z_{ij}^2}$.

It is readily seen that $\boldsymbol{X}_i^*$ is uniformly distributed on $\mathcal{S}^{p-1}$ and independent of $\sum_{j=1}^{p} Z_{ij}^2$, and that $\boldsymbol{V}_i^*$ is uniformly distributed on $\mathcal{S}^{p-2}$ and independent of Z_{i1} and $\sum_{j=2}^{p} Z_{ij}^2$ (hence independent of $Z_{i1}/\sqrt{\sum_{j=1}^{p} Z_{ij}^2}$). Since $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) \stackrel{d}{=} (\boldsymbol{X}_1^*, \dots, \boldsymbol{X}_n^*)$ and $(\boldsymbol{V}_1, \dots, \boldsymbol{V}_n) \stackrel{d}{=} (\boldsymbol{V}_1^*, \dots, \boldsymbol{V}_n^*)$, Lemma A4 follows.

Recall that

$$\overline{\boldsymbol{X}}_m = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{X}^{(i)} = (m^{-1} \sum_{i=1}^{m} X_1^{(i)}, m^{-1} \sum_{i=1}^{m} \nu_i \boldsymbol{V}^{(i)\top})^\top$$

and that

$$\rho_{n,p,m}^2 = \left(\boldsymbol{r}^\top \frac{\overline{\boldsymbol{X}}_m}{\|\overline{\boldsymbol{X}}_m\|} \right)^2 = \frac{\left(\frac{1}{m} \sum_{i=1}^{m} X_1^{(i)} \right)^2}{\left(\frac{1}{m} \sum_{i=1}^{m} X_1^{(i)} \right)^2 + \left\| \frac{1}{m} \sum_{i=1}^{m} \nu_i \boldsymbol{V}^{(i)} \right\|^2}.$$

Let $\boldsymbol{V}_i'$, $i = 1, \dots, n$ be iid uniformly distributed on $\mathcal{S}^{p-2}$ and independent of $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$. Then the following lemma is a consequence of Lemma A4.

Lemma A5.

$$\rho^2_{n,p,m} \stackrel{d}{=} \frac{\left(m^{-1}\sum_{i=1}^m X_1^{(i)}\right)^2}{\left(m^{-1}\sum_{i=1}^m X_1^{(i)}\right)^2 + \|m^{-1}\sum_{i=1}^m \nu_i \boldsymbol{V}'_i\|^2}$$
$$= \frac{A_{n,p,m}}{A_{n,p,m}+V_{n,p,m}}, \tag{A.3}$$

where

$$A_{n,p,m} = \left(\frac{1}{m}\sum_{i=1}^m X_1^{(i)}\right)^2 \text{ and } V_{n,p,m} = \left\|\frac{1}{m}\sum_{i=1}^m \nu_i \boldsymbol{V}'_i\right\|^2. \tag{A.4}$$

The long proofs of Lemmas A6-A8 below are given in Supplementary Material.

Lemma A6. *Let* $m = m_n \to \infty$ *satisfy* $m/n = o(1)$ *and* $p = p_n \to \infty$ *satisfy* $(\ln n)^2/p = O(1)$. *Then*

(i)

$$\max_{1\le i\le m}\left|p\ln(\sin\Theta_{i:n}) + \ln\frac{n}{i} - \frac{1}{2}\ln\ln\frac{n}{i}\right| = O_p(1),$$

(ii)

$$\max_{1\le i\le m}\left|-\frac{p}{2}\cos^2\Theta_{i:n} + \ln\frac{n}{i} - \frac{1}{2}\ln\ln\frac{n}{i}\right| = O_p(1),$$

where $\Theta_{1:n}\le\Theta_{2:n}\le\cdots\le\Theta_{n:n}$ *are the order statistics of* $\Theta_1,\ldots,\Theta_n$.

Lemma A7. *Suppose that* $p = p_n \to \infty$ *satisfies* $(\ln n)^2/p = O(1)$.

(i) *If* $m = m_n \to \infty$ *satisfies* $m/n \to 0$, *then*

$$-pA_{n,p,m} + 2\ln\frac{n}{m} - \ln\ln\frac{n}{m} = O_p(1).$$

(ii) *If $m = m_n \to \infty$ satisfies $(\ln m)^3/(\ln n)^2 \to 0$, then*

$$-pA_{n,p,m} + 2\ln\frac{n}{m} - \ln\ln\frac{n}{m} - \ln(4\pi) + 2 - \frac{2}{p}\left(\ln\frac{n}{m}\right)^2 = o_p(1).$$

(iii) *If $m = m_n \to \infty$ satisfies $m(\ln\ln n)^4/(\ln n)^2 \to 0$, then*

$$\left(\frac{m}{8}\right)^{1/2}\left\{-pA_{n,p,m} + 2\ln\frac{n}{m} - \ln\ln\frac{n}{m} - \ln(4\pi) + 2 - \frac{2}{p}\left(\ln\frac{n}{m}\right)^2\right\} \xrightarrow{d} N(0,1).$$

Lemma A8. *Let $\boldsymbol{W}_1,\ldots,\boldsymbol{W}_n$ be iid uniformly distributed on $\mathcal{S}^{p-1}$. Then*

$$\sqrt{\frac{p}{2n^2}}\sum_{1\le i\ne\ell\le n}\langle \boldsymbol{W}_i, \boldsymbol{W}_\ell\rangle \xrightarrow{d} N(0,1) \text{ uniformly as } n\wedge p\to\infty,$$

where $\langle \boldsymbol{W}_i, \boldsymbol{W}_\ell\rangle$ denotes the inner product of $\boldsymbol{W}_i$ and $\boldsymbol{W}_\ell$.

A.2. Proof of Theorem 1

Theorem 1 is a special case of Theorem A1 below for $m = 1$.

Theorem A1. *Let*

$$T_{n,p} = (p-1)\ln(\sin\Theta_{m:n}) - K_{n,p},$$

where $K_{n,p}$ is defined as in (1). Let $G_m^*(t) = G_m(e^t), t \in \mathrm{R}$, where G_m denotes the gamma distribution with shape parameter m and scale parameter 1. Then for fixed $m = 1, 2, \ldots$, $T_{n,p} \xrightarrow{d} G_m^*$ uniformly as $n \wedge p \to \infty$.

Proof. We claim that

$$T_{n_\ell,p_\ell} \xrightarrow{d} G_m^* \tag{A.5}$$

for any increasing sequences $\{n_\ell\}$ and $\{p_\ell\}$ satisfying $n_\ell \to \infty, p_\ell \to \infty$ and $(\ln n_\ell)/p_\ell \to \alpha \in [0, \infty]$ as $\ell \to \infty$. Assume for now that the claim (A.5) holds. To show that $T_{n,p} \xrightarrow{d} G_m^*$ uniformly as $n \wedge p \to \infty$, suppose to the contrary that $\limsup_{n\wedge p\to\infty} \sup_{t\in\mathrm{R}} |\mathrm{P}(T_{n,p} \le t) - G_m^*(t)| > 0$. Then there exist an $\varepsilon > 0$ and a sequence $\{(n_\ell, p_\ell) : \ell = 1, 2, \ldots\}$ such that $\lim_{\ell\to\infty} n_\ell \wedge p_\ell = \infty$ and

$$\sup_{t\in\mathrm{R}} |\mathrm{P}(T_{n_\ell,p_\ell} \le t) - G_m^*(t)| > \varepsilon \text{ for } \ell = 1, 2, \ldots. \tag{A.6}$$

There exists a subsequence $\{(n_{\ell_k}, p_{\ell_k}) : k = 1, 2, \ldots\}$ such that $(\ln n_{\ell_k})/p_{\ell_k}$ converges to some value $\alpha \in [0, \infty]$. Then (A.6) contradicts (A.5), implying that $T_{n,p} \xrightarrow{d} G_m^*$ uniformly as $n \wedge p \to \infty$.

We now prove (A.5). For notational simplicity, we will deal only with the special case where $n_\ell = \ell$, $\ell = 1, 2, \ldots$. The general case can be treated similarly. Specifically, we show that if $p = p_n \to \infty$ satisfies $(\ln n)/p \to \alpha \in [0, \infty]$, then $T_{n,p} = T_{n,p_n} \xrightarrow{d} G_m^*$.

Suppose $p = p_n \to \infty$ satisfies $\lim_{n\to\infty}(\ln n)/p = \alpha \in [0, \infty]$. For fixed m, since $F_p(\Theta_{m:n}) \stackrel{d}{=} U_{m:n}$, we have by Lemma A3

$$\begin{aligned}\mathrm{P}(nF_p(\Theta_{m:n}) \le e^t) = \mathrm{P}\left(nU_{m:n} \le e^t\right) = \mathrm{P}\left(n\frac{S_m}{S_{n+1}} \le e^t\right)& \\ \longrightarrow \mathrm{P}(S_m \le e^t) = G_m\left(e^t\right) = G_m^*(t).&\end{aligned} \tag{A.7}$$

For fixed $t > 0$, let $t_n \in [0, 1)$ be such that

$$\frac{p-1}{2}\ln(1-t_n^2) = \min\{K_{n,p}+t, 0\}.$$

Noting that

$$K_{n,p} = K_{n,p_n} = -(\ln n)(1+o(1)) \text{ as } n \to \infty, \tag{A.8}$$

we have for large n

$$\frac{p-1}{2}\ln(1-t_n^2) = K_{n,p}+t < 0. \tag{A.9}$$

By Lemma A2,

$$\left(1+\frac{1}{(p-3)t_n^2}\right)^{-1}\overline{F}_p(\cos^{-1}t_n) \le F_p(\cos^{-1}t_n) \le \overline{F}_p(\cos^{-1}t_n),$$

implying that

$$\begin{aligned}\mathrm{P}(nF_p(\Theta_{m:n}) \le n\overline{F}_p(\cos^{-1}t_n)) &\ge \mathrm{P}(nF_p(\Theta_{m:n}) \le nF_p(\cos^{-1}t_n)) \\ &\ge \mathrm{P}\left(nF_p(\Theta_{m:n}) \le \left(1+\frac{1}{(p-3)t_n^2}\right)^{-1} n\overline{F}_p(\cos^{-1}\, t_n)\right).\end{aligned} \tag{A.10}$$

Recalling $\alpha = \lim_{n\to\infty}(\ln n)/p$, we claim that for every $\alpha \in [0, \infty]$, as $n \to \infty$

$$n\,\overline{F}_p(\cos^{-1}\, t_n) = e^t + o(1), \tag{A.11}$$

$$p\,t_n^2 \to \infty, \tag{A.12}$$

$$\mathrm{P}(\cos\,\Theta_{m:n} \le -t_n) \to 0. \tag{A.13}$$

By (A.7), (A.10), (A.11) and (A.12),

$$\mathrm{P}(\cos\,\Theta_{m:n} \geq t_n) = \mathrm{P}\left(nF_p(\Theta_{m:n}) \leq nF_p(\cos^{-1} t_n)\right) \to G^*_m(t). \qquad \text{(A.14)}$$

Furthermore,

$$\begin{aligned}
\mathrm{P}(T_{n,p} \leq t) &= \mathrm{P}\left(\frac{p-1}{2}\ln(1-\cos^2\Theta_{m:n}) - K_{n,p} \leq t\right) \\
&= \mathrm{P}(\cos^2\,\Theta_{m:n} \geq t_n^2) \;\; \text{(by (A.9))} \\
&= \mathrm{P}(\cos\,\Theta_{m:n} \geq t_n) + \mathrm{P}(\cos\,\Theta_{m:n} \leq -t_n) \\
&\to G^*_m(t) \text{ (by (A.13) and (A.14))}.
\end{aligned}$$

It remains to establish (A.11)-(A.13). Note that by Sterling's formula (see e.g. Tricomi and Erdélyi (1951)),

$$\frac{\Gamma(p/2)}{\Gamma((p-1)/2)} = \sqrt{\frac{p}{2}}\left(1 + O\left(\frac{1}{p}\right)\right) \text{ as } p \to \infty. \qquad \text{(A.15)}$$

We have

$$\begin{aligned}
\ln\left(n\overline{F}_p(\cos^{-1} t_n)\right) &= \ln\left\{\frac{n}{\sqrt{\pi}}\frac{\Gamma(p/2)}{\Gamma((p-1)/2)}\left(\frac{(1-t_n^2)^{p-1}}{(p-1)^2 t_n^2}\right)^{1/2}\right\} \;\; \text{(by (A.2))} \\
&= \ln\left\{n\left(\frac{(1-t_n^2)^{p-1}}{2\pi p t_n^2}\right)^{1/2}\right\} + O\left(\frac{1}{p}\right) \;\; \text{(by (A.15))} \\
&= \frac{p-1}{2}\ln(1-t_n^2) + \ln n - \frac{1}{2}\ln(pt_n^2) - \frac{1}{2}\ln(2\pi) + O\left(\frac{1}{p}\right) \\
&= K_{n,p} + t + \ln n - \frac{1}{2}\ln(pt_n^2) - \frac{1}{2}\ln(2\pi) + O\left(\frac{1}{p}\right) \;\; \text{(by (A.9))}.
\end{aligned} \qquad \text{(A.16)}$$

By (A.8) and (A.9),

$$\ln(1-t_n^2) = -\frac{2\ln n}{p}(1+o(1)), \tag{A.17}$$

implying that

$$t_n \to \left(1-e^{-2\alpha}\right)^{1/2}, \tag{A.18}$$

350 where $\lim_{n\to\infty}(\ln n)/p = \alpha \in [0,\infty]$ and $e^{-\infty} := 0$.

If $\alpha = 0$, we have $t_n \to 0^+$, so that by (A.17)

$$t_n^2 = \frac{2\ln n}{p}(1+o(1)), \tag{A.19}$$

from which it follows that $\ln(pt_n^2) = \ln(2\ln n) + o(1)$. By the definition of $K_{n,p}$, we have $K_{n,p} = -\ln n + (\ln\ln n)/2 + \ln(4\pi)/2 + o(1)$, so that $K_{n,p} + \ln n - \ln(pt_n^2)/2 - \ln(2\pi)/2 = o(1)$, which together with (A.16) establishes (A.11) for $\alpha = 0$. If $0 < \alpha < \infty$, we have $t_n^2 = 1 - e^{-2\alpha} + o(1)$ (by (A.18)) and $\ln(pt_n^2) = \ln\ln n - \ln\alpha + \ln(1-e^{-2\alpha}) + o(1)$, so that $K_{n,p} + \ln n - \ln(pt_n^2)/2 - \ln(2\pi)/2 = o(1)$, which together with (A.16) establishes (A.11) for $0 < \alpha < \infty$. If $\alpha = \infty$, we have $t_n \to 1^-$, so that by the definition of $K_{n,p}$,

$$\begin{aligned}
&K_{n,p} + \ln n - \frac{1}{2}\ln(pt_n^2) - \frac{1}{2}\ln(2\pi) \\
&= -\ln n + \frac{1}{2}\ln\ln n - \frac{1}{2}\ln\left(\frac{2\ln n}{p}\right) + \frac{1}{2}\ln(4\pi) + \ln n - \frac{1}{2}\ln p - \frac{1}{2}\ln(2\pi) + o(1) \\
&= o(1),
\end{aligned}$$

which together with (A.16) establishes (A.11) for $\alpha = \infty$.

Next, (A.19) holds for $\alpha = 0$, which implies (A.12). For $0 < \alpha \leq \infty$, it follows from (A.18) that $t_n \to (1 - e^{-2\alpha})^{1/2} > 0$, which implies (A.12).

Finally, to prove (A.13), note that

$$\mathrm{P}(\cos \Theta_{m:n} \leq -t_n) \leq \mathrm{P}(\Theta_{m:n} \geq \pi/2) = \mathrm{P}(B(n, 1/2) < m) \to 0,$$

where $B(n, 1/2)$ denotes a binomial random variable with parameters n and $1/2$ (success probability). This establishes (A.13) and completes the proof of Theorem A1. $\square$

A.3. Proofs of Theorems 2-3 and Corollary 1

We first show that if $m = m_n \to \infty$ satisfies $m/n \to 0$ and $p = p_n \to \infty$ satisfies $(\ln n)^2/p \to 0$, then

$$m\sqrt{\frac{p}{2}}\left(V_{n,p,m} - \frac{1}{m}\right) \xrightarrow{d} N(0,1), \tag{A.20}$$

where $V_{n,p,m} = \|\frac{1}{m}\sum_{i=1}^m \nu_i \boldsymbol{V}'_i\|^2$ with $\nu_i^2 = 1 - \cos^2 \Theta_{i:n}$, and $\boldsymbol{V}'_1, \ldots, \boldsymbol{V}'_m$ are iid uniformly distributed on $\mathcal{S}^{p-2}$, and $(\boldsymbol{V}'_1, \ldots, \boldsymbol{V}'_m)$ is independent of $(\nu_1, \ldots, \nu_m)$.

We have

$$
\begin{aligned}
V_{n,p,m} &= \frac{1}{m^2}\sum_{i=1}^{m}\nu_i^2\|\boldsymbol{V}_i'\|^2 + \frac{1}{m^2}\sum_{1\le i\ne\ell\le m}\nu_i\nu_\ell\langle\boldsymbol{V}_i',\boldsymbol{V}_\ell'\rangle \\
&= \frac{1}{m} + \frac{1}{m^2}\sum_{i=1}^{m}(\nu_i^2-1) + \frac{1}{m^2}\sum_{1\le i\ne\ell\le m}\{1+(\nu_i\nu_\ell-1)\}\langle\boldsymbol{V}_i',\boldsymbol{V}_\ell'\rangle \\
&= \frac{1}{m} + V_{1,n}' + V_{2,n}' + V_{3,n}', \qquad \text{(A.21)}
\end{aligned}
$$

where

$$
\begin{aligned}
V_{1,n}' &= \frac{1}{m^2}\sum_{i=1}^{m}(\nu_i^2-1) = -\frac{1}{m^2}\sum_{i=1}^{m}\cos^2\Theta_{i:n}, \\
V_{2,n}' &= \frac{1}{m^2}\sum_{1\le i\ne\ell\le m}\langle\boldsymbol{V}_i',\boldsymbol{V}_\ell'\rangle, \\
V_{3,n}' &= \frac{1}{m^2}\sum_{1\le i\ne\ell\le m}(\nu_i\nu_\ell-1)\langle\boldsymbol{V}_i',\boldsymbol{V}_\ell'\rangle.
\end{aligned}
$$

By Lemma A8, we have

$$
m\sqrt{\frac{p}{2}}V_{2,n}' \xrightarrow{d} N(0,1). \qquad \text{(A.22)}
$$

366 It remains to prove

$$
mp^{1/2}V_{i,n}' = o_p(1),\ i=1,3. \qquad \text{(A.23)}
$$

By Lemma A6(ii),

$$
\max_{1\le i\le m}\cos^2\Theta_{i:n} = O_p\left(\frac{\ln\, n}{p}\right),
$$

368 implying that $mp^{1/2}V_{1,n}' = O_p\left(\frac{\ln\, n}{p^{1/2}}\right) = o_p(1)$. To show $mp^{1/2}V_{3,n}' = o_p(1)$, note that $(\nu_1,\ldots,\nu_m)$ is independent of $(\boldsymbol{V}_1',\ldots,\boldsymbol{V}_m')$ and $\mathrm{E}[\langle\boldsymbol{V}_i',\boldsymbol{V}_\ell'\rangle\langle\boldsymbol{V}_{i'}',\boldsymbol{V}_{\ell'}'\rangle] =$

0 if $i \neq \ell$, $i' \neq \ell'$ and $\{i, \ell\} \neq \{i', \ell'\}$. Also, for $i \neq \ell$, $\mathrm{E}\langle \boldsymbol{V}'_i, \boldsymbol{V}'_\ell \rangle^2 = \int_0^\pi \cos^2(\theta) dF_{p-1}(\theta) = \frac{1}{p-1}$, where F_p is defined as in (A.1). We have

$$\begin{aligned}
\mathrm{E}V'^2_{3,n} &= \frac{2}{m^4} \sum_{1 \leq i \neq \ell \leq m} \mathrm{E}[(\nu_i \nu_\ell - 1)^2] \mathrm{E}\langle \boldsymbol{V}'_i, \boldsymbol{V}'_\ell \rangle^2 \\
&= \frac{2}{m^4} \sum_{1 \leq i \neq \ell \leq m} \mathrm{E}[(\nu_i \nu_\ell - 1)^2] \frac{1}{p-1} \\
&= o\left(\frac{1}{m^2 p}\right), \qquad \text{(A.24)}
\end{aligned}$$

since $|\nu_i| \leq 1$ and $\nu_i \nu_\ell - 1 \to 0$ in probability uniformly in $1 \leq i \neq \ell \leq m$. It follows from (A.24) that $mp^{1/2} V'_{3,n} = o_p(1)$. This proves (A.23) and completes the proof of (A.20).

Proof of Theorem 2. Since by (A.3) $\rho^2_{n,p,m} \overset{d}{=} \frac{A_{n,p,m}}{A_{n,p,m} + V_{n,p,m}}$, we have

$$\rho^2_{n,p,m} - \frac{\beta_{n,p,m}}{1 + \beta_{n,p,m}} \overset{d}{=} \frac{A_{n,p,m} - \beta_{n,p,m}/m}{(A_{n,p,m} + V_{n,p,m})(1 + \beta_{n,p,m})} + \frac{(1/m - V_{n,p,m})\beta_{n,p,m}}{(A_{n,p,m} + V_{n,p,m})(1 + \beta_{n,p,m})}. \qquad \text{(A.25)}$$

Since $\beta_{n,p,m} = \frac{m}{p}\left\{2 \ln \frac{n}{m} - \ln \ln \frac{n}{m} - \ln(4\pi) + 2\right\}$, it follows from Lemma A7(i) and (A.20) that

$$p(A_{n,p,m} - \frac{1}{m}\beta_{n,p,m}) = O_p(1), \quad mV_{n,p,m} = 1 + o_p(1), \quad p\beta_{n,p,m} V_{n,p,m} = (2 + o_p(1)) \ln\left(\frac{n}{m}\right).$$

Thus,

$$\begin{aligned}
\frac{A_{n,p,m} - \beta_{n,p,m}/m}{(A_{n,p,m} + V_{n,p,m})(1 + \beta_{n,p,m})} &= \frac{p(A_{n,p,m} - \beta_{n,p,m}/m)}{(p\beta_{n,p,m} A_{n,p,m} + p\beta_{n,p,m} V_{n,p,m})} \frac{\beta_{n,p,m}}{(1 + \beta_{n,p,m})} = o_p(1) \frac{\beta_{n,p,m}}{(1 + \beta_{n,p,m})}, \\
\frac{(1/m - V_{n,p,m})\beta_{n,p,m}}{(A_{n,p,m} + V_{n,p,m})(1 + \beta_{n,p,m})} &= \frac{(1 - mV_{n,p,m})}{(mA_{n,p,m} + mV_{n,p,m})} \frac{\beta_{n,p,m}}{(1 + \beta_{n,p,m})} = o_p(1) \frac{\beta_{n,p,m}}{(1 + \beta_{n,p,m})}.
\end{aligned}$$

We have by (A.25),

$$\rho^2_{n,p,m} = \frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}(1+o_p(1)).$$

The proof is complete. $\square$

Proof of Theorem 3. By (A.21)-(A.23),

$$\begin{aligned} m\sqrt{\frac{p}{2}}\left(V_{n,p,m}-\frac{1}{m}\right) &= m\sqrt{\frac{p}{2}}\left(V'_{1,n}+V'_{2,n}+V'_{3,n}\right) \\ &= m\sqrt{\frac{p}{2}}V'_{2,n}+o_p(1). \end{aligned} \tag{A.26}$$

Let

$$\begin{aligned} Z_{1,n} &= p\sqrt{\frac{m}{8}}\left(A_{n,p,m}-\frac{1}{m}\beta_{n,p,m}+\frac{2}{p^2}\left(\ln\frac{n}{m}\right)^2\right), \\ Z_{2,n} &= m\sqrt{\frac{p}{2}}V'_{2,n}, \\ \gamma_n &= (A_{n,p,m}+V_{n,p,m})(1+\beta_{n,p,m}). \end{aligned}$$

We have by (A.25) and (A.26)

$$\begin{aligned} &\rho^2_{n,p,m}-\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}} \\ \stackrel{d}{=}\ & \gamma_n^{-1}\left\{\frac{1}{p\sqrt{m/8}}Z_{1,n}-\frac{\beta_{n,p,m}}{m\sqrt{p/2}}m\sqrt{\frac{p}{2}}\left(V_{n,p,m}-\frac{1}{m}\right)\right\}-\gamma_n^{-1}\frac{2}{p^2}\left(\ln\frac{n}{m}\right)^2 \\ =\ & \gamma_n^{-1}\left\{\sqrt{\frac{8}{mp^2}}Z_{1,n}-\sqrt{\frac{2}{m^2p}}\beta_{n,p,m}(Z_{2,n}+o_p(1))\right\}-\gamma_n^{-1}\frac{2}{p^2}\left(\ln\frac{n}{m}\right)^2 \\ =\ & \gamma_n^{-1}\left(\frac{8}{mp^2}+\frac{2}{m^2p}\beta^2_{n,p,m}\right)^{1/2}\{c_{1,n}Z_{1,n}+c_{2,n}(Z_{2,n}+o_p(1))\}-\gamma_n^{-1}\frac{2}{p^2}\left(\ln\frac{n}{m}\right)^2, \end{aligned} \tag{A.27}$$

where

$$
\begin{aligned}
c_{1,n} &= \sqrt{\frac{8}{mp^2}}\left(\frac{8}{mp^2}+\frac{2}{m^2p}\beta^2_{n,p,m}\right)^{-1/2},\\
c_{2,n} &= -\sqrt{\frac{2}{m^2p}}\beta_{n,p,m}\left(\frac{8}{mp^2}+\frac{2}{m^2p}\beta^2_{n,p,m}\right)^{-1/2}.
\end{aligned}
$$

Since $\rho^2_{n,p,m} \overset{d}{=} A_{n,p,m}/(A_{n,p,m}+V_{n,p,m})$, we have by Theorem 2

$$\frac{A_{n,p,m}}{A_{n,p,m}+V_{n,p,m}} = \frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}(1+o_p(1)). \tag{A.28}$$

It follows from Lemma A7(i) and $(p/m)\beta_{n,p,m} = 2\ln\frac{n}{m}(1+o(1))$ that

$$\frac{mA_{n,p,m}}{\beta_{n,p,m}} = \frac{pA_{n,p,m}}{(p/m)\beta_{n,p,m}} = 1+o_p(1). \tag{A.29}$$

So we have

$$
\begin{aligned}
\gamma_n\left(\frac{8}{mp^2}+\frac{2}{m^2p}\beta^2_{n,p,m}\right)^{-1/2} &= \frac{pm}{\sqrt{8m+2p\beta^2_{n,p,m}}}(A_{n,p,m}+V_{n,p,m})(1+\beta_{n,p,m})\\
&= \frac{pmA_{n,p,m}}{\sqrt{8m+2p\beta^2_{n,p,m}}}\frac{A_{n,p,m}+V_{n,p,m}}{A_{n,p,m}}(1+\beta_{n,p,m})\\
&= \frac{pmA_{n,p,m}/\beta_{n,p,m}}{\sqrt{8m+2p\beta^2_{n,p,m}}}(1+\beta_{n,p,m})^2(1+o_p(1)) \quad \text{(by(A.28))}\\
&= \frac{p}{\sqrt{8m+2p\beta^2_{n,p,m}}}(1+\beta_{n,p,m})^2(1+o_p(1)) \quad \text{(by (A.29))}\\
&= \alpha_{n,p,m}(1+o_p(1)), \qquad \text{(A.30)}
\end{aligned}
$$

where $\alpha_{n,p,m} = p\left(8m+2p\,\beta^2_{n,p,m}\right)^{-1/2}(1+\beta_{n,p,m})^2$.

Also,

$$
\begin{aligned}
0 &< \frac{2}{p^2}\left(\ln\frac{n}{m}\right)^2\left(\frac{8}{mp^2}+\frac{2}{m^2p}\beta^2_{n,p,m}\right)^{-1/2}\\
&\le \frac{2}{p^2}\left(\ln\frac{n}{m}\right)^2\left(\frac{2}{m^2p}\beta^2_{n,p,m}\right)^{-1/2}\\
&= \frac{2}{p^2}\left(\ln\frac{n}{m}\right)^2\left\{\frac{2}{p^3}\left(\frac{p}{m}\beta_{n,p,m}\right)^2\right\}^{-1/2}\\
&= \sqrt{\frac{2}{p}}\left(\ln\frac{n}{m}\right)^2\left(2\ln\frac{n}{m}(1+o(1))\right)^{-1}\\
&= \frac{1}{\sqrt{2p}}\ln\frac{n}{m}(1+o(1)) = o(1),
\end{aligned}
$$

390 which together with (A.30) implies that

$$
\begin{aligned}
&\frac{\alpha_{n,p,m}}{\gamma_n}\frac{2}{p^2}\left(\ln\frac{n}{m}\right)^2\\
=&\left\{\frac{\alpha_{n,p,m}}{\gamma_n}\left(\frac{8}{mp^2}+\frac{2}{m^2p}\beta^2_{n,p,m}\right)^{1/2}\right\}\left\{\frac{2}{p^2}\left(\ln\frac{n}{m}\right)^2\left(\frac{8}{mp^2}+\frac{2}{m^2p}\beta^2_{n,p,m}\right)^{-1/2}\right\}\\
=&\ (1+o_p(1))o(1) = o_p(1). \qquad \text{(A.31)}
\end{aligned}
$$

It follows from (A.27), (A.30), and (A.31) that

$$
\begin{aligned}
&\alpha_{n,p,m}\left(\rho^2_{n,p,m}-\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}\right)\\
\overset{d}{=}&\ \frac{\alpha_{n,p,m}}{\gamma_n}\left(\frac{8}{mp^2}+\frac{2}{m^2p}\beta^2_{n,p,m}\right)^{1/2}\{c_{1,n}Z_{1,n}+c_{2,n}Z_{2,n}(1+o_p(1))\}-\frac{\alpha_{n,p,m}}{\gamma_n}\frac{2}{p^2}\left(\ln\frac{n}{m}\right)^2\\
=&\ (1+o_p(1))\{c_{1,n}Z_{1,n}+c_{2,n}Z_{2,n}(1+o_p(1))\}+o_p(1). \qquad \text{(A.32)}
\end{aligned}
$$

392 Note that $c_{1,n}$ and $c_{2,n}$ are constants (depending on n, p_n, m_n), which satisfy $c^2_{1,n}+c^2_{2,n}=1$. By Lemma A7(iii),

$$
-Z_{1,n} = \sqrt{\frac{m}{8}}\left\{-p\,A_{n,p,m}+2\ln\frac{n}{m}-\ln\ln\frac{n}{m}-\ln(4\pi)+2-\frac{2}{p}\left(\ln\frac{n}{m}\right)^2\right\} \xrightarrow{d} N(0,1).
$$

By (A.22), $Z_{2,n} \xrightarrow{d} N(0,1)$. Note that $Z_{1,n}$ and $Z_{2,n}$ are independent (since $A_{n,p,m}$ and $V'_{2,n}$ are independent). We have

$$c_{1,n}Z_{1,n} + c_{2,n}Z_{2,n} \xrightarrow{d} N(0,1),$$

which together with (A.32) implies that

$$\alpha_{n,p,m}\left(\rho^2_{n,p,m} - \frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}\right) \xrightarrow{d} N(0,1).$$

The proof is complete.

Proof of Corollary 1. Part (i) follows immediately from Theorem 2. To prove part (ii), we have by part (i) and Theorem 3 that

$$\begin{aligned} & 2\alpha_{n,p,m}\sqrt{\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}}\left(\rho_{n,p,m} - \sqrt{\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}}\right) \\ = & \frac{2\sqrt{\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}}}{\rho_{n,p,m} + \sqrt{\frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}}}\alpha_{n,p,m}\left(\rho^2_{n,p,m} - \frac{\beta_{n,p,m}}{1+\beta_{n,p,m}}\right) \xrightarrow{d} N(0,1), \end{aligned}$$

completing the proof. □

References

Cai, T. T., Fan, J., and Jiang, T. (2013). Distributions of angles in random packing on spheres. *Journal of Machine Learning Research* **14,** 1837–1864.

Cai, T. T. and Jiang, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Annals of Statistics* **39,** 1496–1525.

Cai, T. T. and Jiang, T. (2012). Phase transition in limiting distributions of coherence of high-dimensional random matrices. *Journal of Multivariate Analysis* **107,** 24–39.

Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B* **74,** 37–65.

Fan, J., Shao, Q. M., and Zhou, W. X. (2018). Are discoveries spurious? Distributions of maximum spurious correlations and their applications. *Annals of Statistics* **46,** 989–1017.

Henderson, R. (2013). Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proceedings of the National Academy of Sciences U.S.A.* **110,** 18037–18041.

Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes.* Academic Press.

Lai, T. L., Wang, S.-H., Yao, Y.-C., Chung, S.-C., Chang, W.-H., and Tu, I-P. (2020). *Cryo-EM: Breakthroughs in chemistry, structural biology, and statistical underpinnings.* Technical Report, Center for Innovative Study Design, Stanford University.

Liao, M., Cao, E., Julius, D., and Cheng, Y. (2013). Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504,** 107–112.

Stewart, A. and Grigorieff, N. (2004). Noise bias in the refinement of structures derived from single particles. *Ultramicroscopy* **102,** 67–84.

Tricomi, F. G. and Erdélyi, A. (1951). The asymptotic expansion of a ratio of gamma functions. *Pacific Journal of Mathematics* **1,** 133–142.

Yan, C., Hang, J., Wan, R., Huang, M., Wong, C. C., and Shi, Y. (2015). Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science* **349,** 1182–1191.