Exploratory Factor Analysis of Data on a Sphere
Fan Dai, Karin S. Dorman, Somak Dutta and and Ranjan Maitra

Data on high-dimensional spheres arise frequently in many disciplines either naturally or as a consequence of preliminary processing and can have intricate dependence structure that needs to be understood. We develop exploratory factor analysis of the projected normal distribution to explain the variability in such data using a few easily interpreted latent factors. Our methodology provides maximum likelihood estimates through a novel fast alternating expectation profile conditional maximization algorithm. Results on simulation experiments on a wide range of settings are uniformly excellent. Our methodology provides interpretable and insightful results when applied to tweets with the #MeToo hashtag in early December 2018, to time-course functional Magnetic Resonance Images of the average pre-teen brain at rest, to characterize handwritten digits, and to gene expression data from cancerous cells in the Cancer Genome Atlas.

Index Terms
AECM, l2-normalization, Lanczos algorithm, matrix-free computations, #MeToo, MNIST, projected normal distribution, resting state fMRI, term-document matrix, TCGA.

I. INTRODUCTION
Observations that lie on a unit sphere [1], [2], [3], [4] arise naturally in disciplines such as astronomy (e.g., the orientation of planetary and cometary orbits), biology (e.g., homing behavior or the direction of capillaries in tissue), geology (e.g., the orientation of paleomagnetism in rocks or the axes of sand dunes), meteorology (e.g., wind direction), oceanography (e.g., the direction of ice flows) or physics (e.g., the direction of particle diffusion). Many distributions [3], [4] have been proposed to model such observations, but the von Mises (or circular normal) distribution [5] is commonly used to model angular data, while observations on a three-dimensional sphere are often modeled with the Fisher distribution [6]. The von Mises and Fisher distributions are the 2- and 3-dimensional cases of the Langevin [1] distribution – also called the Fisher-von Mises [7] or von Mises-Fisher [4] distribution – that models data on a multi-dimensional sphere. However, this distribution is unable to adequately characterize relationships between variables, especially in the context of modern applications such as text processing, genetics or imaging, where observations end up on unit spheres after appropriate processing. We describe four motivating applications with high-dimensional data on unit spheres that can benefit from a distribution that can characterize dependence between variables.

A. Motivating Applications
We introduce four scenarios where preprocessing yields unit-sphered data. In each case, the goal is to characterize variability and relationships between variables, whether they are words in tweets containing the hashtag #MeToo, the time-points in temporal functional Magnetic Resonance Imaging (fMRI) data of an averaged typically developing child brain at rest, pixels in bitmap images, or cancerous gene expressions in the Cancer Genome Atlas (TCGA).

1) Summarizing #MeToo tweets: The #MeToo hashtag on Twitter provided an online forum for women to narrate their harassment or abuse at the hands of men in positions of authority. The term had been percolating in social media since 2006, but received a major fillip in 2017 with the outing of high-profile abusers in entertainment, politics, business and media. Our interest is in characterizing variability in tweets having the hashtag and to countenance its description by a few underlying factors. Each tweet is akin to a text document, so all the tweets can be processed into a weighted document-term frequency matrix [8], [9], [10], with the most common weighting scheme normalizing each tweet to lie on a high-dimensional unit sphere [11]. Modeling each tweet in terms of the Langevin distribution, as commonly done for text data [12], would require marginal exchangeability and essentially no correlation between words and a characterization totally at odds with the reality of tweet structure and characteristics. We return to this dataset in Section IV-A.

2) Characterizing the brain at rest: Characterizing changes in cerebral blood-flow at rest [13], or in the absence of any stimulus or task, is of interest to researchers [14]. A typical resting state experiment using fMRI [15], [16], [17], [18], [19] involves collecting a time series of blood-oxygen-level-dependent (BOLD) images of the brain at rest, and after preprocessing, fitting a linear model at each voxel with motion and other structured noise as covariates. The time series of the residuals, or the part of the BOLD response vector unexplained by the linear model, is the resting state component of the dataset [20], [21]. Functionally connected voxels may be detected from correlations between the residual time series sequences at each voxel, or equivalently by calculating the Euclidean distance between the centered and standardized time series that lie on the unit sphere [22]. However, it may also be of interest to characterize the distribution and variability of activity patterns in the resting state brain, for which the Langevin distribution is again inadequate. We revisit this application in Section IV-B.
3) Assessing variability in handwritten digits: The Modified National Institute of Standards and Technology (MNIST) database [23] has 70,000 normalized and anti-aliased $28 \times 28$ bitmap images of handwritten digits of 0 through 9, written by 500 writers. The standardization of each pixel value to the interval $[0, 255]$, as a consequence of the digitization of the image, means that the raw photon value of each image pixel is no longer available. The loss of a meaningful absolute scale for the image suggests use of correlation as a similarity measure. This similarity measure is akin to normalizing each image to lie on the unit sphere. Characterizing inter-pixel relationships can summarize variability in handwritten digits but again, requires a distribution richer than the Langevin for the same reasons as before. We investigate this application further in Section IV-C.

4) Identifying genetic pathways underlying cancer: The Cancer Genome Atlas (TCGA) dataset provides messenger ribonucleic acid (mRNA) sequencing data of tumor samples from more than 11,000 patients collected over a 12-year period. After initial processing, each tumor has gene compositional data, where each observation lies on a simplex, and a square root transformation on each feature maps the observation to lie on a unit sphere [24]. It is expected that a few groups of genes can explain variability between the tumor samples (and patients) and that identification of such gene groups could lead to future diagnostic tools and treatments [25]. Once again, however, such relationships cannot be modeled by the Langevin distribution with its lack of flexible relationships between coordinates. We study this application in detail in Section IV-D.

Our four showcase applications point to the inadequacy of the Langevin distribution in characterizing heterogeneity and relationships between the observed variables that lie on a unit sphere. A more general approach to modeling data on a sphere involves taking the inverse stereographic projection of a general multivariate normal distribution [26], [27]. However, this approach is not practical in high dimensions: indeed, in their application, [27] use a priori dimension reduction (into the first three principal components) on the raw data as a preprocessing step before modeling and analysis. Other distributions such as the complex Bingham [28], Fisher-Bingham [29] and the real/complex Watson distributions [30], [31] also suffer from intractability with dimensions greater than three.

A more flexible distributional family arises from the projected normal (PN) distribution that is obtained by radial projection of the multivariate normal (MVN) distribution onto the unit sphere [4]. In general, the PN density does not have a closed-form expression. [32] and [33] provided Bayesian inference for this model in two and higher dimensions, respectively, under the (mild) constraint that the first element of the dispersion matrix parameter of the projected normal density is unity. The structure of this dispersion matrix parameter is of interest in many applications, so we investigate if a small number of latent factors can explain the variability in the (observed) data on a sphere. Our approach posits parameter estimation from a projected normal distribution as a missing data problem, where the complete data is from the generative MVN distribution that we specify in terms of a factor model. Two Alternating Expectation Conditional Maximization (AECM) algorithms [34] are developed for parameter estimation within our framework. Because our interest stems from high-dimensional problems, we also develop accelerated implementations that incorporate squared iterative (SQUAREM) methods [35] to achieve superlinear convergence.

The remainder of this paper is organized as follows. Section II develops the factor model and analysis for data on a sphere (FADS) and proposes two novel AECM algorithms called FADS-D (Expectation-Maximization, i.e. EM, with a duet of latent variables) and FADS-P (EM with one latent variable and a profile maximization step) for model-fitting. We use the extended Bayesian Information Criterion (eBIC) of [36] to decide on the number of factors. Section III evaluates the performance of FADS-P and FADS-D in simulation experiments. We next apply our methodology to the motivating applications of Section I-A. Section V discusses possible extensions and future work. An online supplement, with sections, tables, theorems and figures referenced here with the prefix “S”, is available.

II. METHODOLOGY

A. Background and preliminary development

1) The projected normal (PN) distribution: A popular method of constructing probability distributions on the unit sphere is by radially projecting a multivariate distribution, typically Gaussian and has been well-studied in low-dimensional problems by [4], p. 178], [37], [38], and, from a Bayesian perspective, by [32] and [33]. The PN distribution $\mathcal{P}\mathcal{N}_p(\mu, \Sigma)$ on the $p$-unit sphere $S_{p-1}$ is defined as the distribution of $Y/\|Y\|$ where $Y$ has an MVN distribution with mean $\mu$ and covariance matrix $\Sigma$. Consequently, the $\mathcal{P}\mathcal{N}_p(\mu, \Sigma)$ density function is

$$f(x; \mu, \Sigma) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left\{ -\frac{(\mu^\top \Sigma^{-1} \mu)}{2} + \frac{m^2}{2v} \right\} \int_0^\infty R^{v-1} \exp \left\{ -\frac{1}{2v} (R - m)^2 \right\} dR,$$

where $m = x^\top \Sigma^{-1} \mu / x^\top \Sigma^{-1} x$, $v = 1/x^\top \Sigma^{-1} x$ (see Theorem S1). The integral in (1) arises from integrating out the unobserved length $R = \|Y\|$ and is proportional to the $p$th raw moment of a univariate normal distribution truncated above zero. However, this family of distributions is not identifiable [4], so there is need for further constraints, such as fixing a diagonal entry of $\Sigma$ or, more commonly, setting $\mu \in S_{p-1}$ that we adopt in this work. The matrix $\Sigma$ flexibly models the dependence among coordinates, allowing different shapes and orientations. For example, Fig. 1 shows the isodensity lines of a 3D PN distribution with the same $\mu$ but three different $\Sigma$s. With spherical $\Sigma$, the isodensity lines on the surface of the sphere are circular, but for the other two heteroscedastic cases, we obtain ellipsoids of different eccentricities and orientations. [37] provide method-of-moments estimators for the parameters of the PN distribution but this method is impractical for high
dimensions, especially when the sample size is relatively small. An alternative approach to specifying general distributions on the sphere is by stereographic projection of the MVN distribution [27]. Both approaches result in multimodal densities as the diagonal entries of $\Sigma$ increase. However, the stereographic projection approach is not easily analyzed in high dimensions, so [27] suggest reducing the dimensionality of the MVN through principal components analysis before applying stereographic projection. We provide computationally efficient EM algorithms for maximum likelihood (ML) estimators for the parameters of the PN distribution that can be used in high dimensions. Our approach also decomposes $\Sigma$ into a factor structure, allowing for interpretation and further understanding of the variability in a dataset. We describe our factor model setup next.

2) A latent factor model for the PN distribution: Let $X_1, X_2, \ldots, X_n$ be independent identically distributed (i.i.d.) realizations from (1). Our latent factor model assumes $X_i = Y_i / \|Y_i\|$, where the generative latent variables $Y_i$ have the decomposition

$$Y_i = \mu + \Lambda Z_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

and $\Lambda$ is a $p \times q$ matrix called the factor loading matrix with rank $q$. $Z_i$s are latent factors of dimension $q$ and $\epsilon_i$s are residual noise vectors. Typically, $q \ll \min(n, p)$ and so (2) explains the variability in the $Y_i$s in terms of a few ($q$) underlying latent factors. Usually, $Z_i$s are assumed independent $\mathcal{N}_q(0, I)$ and $\epsilon_i$s independent $\mathcal{N}_p(0, \Psi)$ random vectors, where $\Psi$ is a diagonal matrix with positive entries called the uniquenesses. Hence, the covariance matrix $\Sigma$ admits a lower-dimensional representation:

$$\Sigma = \Lambda \Lambda^\top + \Psi.$$ Parameter estimation in (2) has received attention [26], [39], [40], [41], [42], [43] over several decades, with recent work [44] providing near-instantaneous solutions even for very high dimensions.

The Gaussian factor model (2) is not invariant to orthogonal transformation [26], [42] unless all uniquenesses are equal (i.e. $\Psi = \sigma^2 I_p$ for some $\sigma^2 > 0$). Also, (2) is not identifiable because both $\Lambda$ and $\Lambda \Psi$ give rise to the same model for any orthogonal matrix $Q$ and so additional constraints are needed for identifiability. For computational simplicity, we assume that $\Lambda^\top \Psi^{-1} \Lambda$ is diagonal with decreasing diagonal entries. However, for better interpretability, and as is standard in traditional factor analysis, we present estimates of $\Lambda$ after applying some appropriate orthogonal (e.g. varimax, quartimax or oblimin) rotation [45].

B. Estimation of parameters

The PN factor model of Section II-A2 has two generative latent components – the lengths $R_i = \|Y_i\|$ and the factors $Z_i$, $i = 1, 2, \ldots, n$. Given $q$ and these inherent latencies, the EM algorithm provides a natural way to obtain the ML estimators of $\mu$, $\Lambda$ and $\Psi$. But vanilla EM is slow to converge, so we develop an AECM algorithm [34] for parameter estimation using the latent components $(R_i, Z_i)$s. However, [44] recently showed a matrix-free profile likelihood approach to be more computationally efficient than EM [43] for the Gaussian factor model. Building on that approach, we develop a AECM algorithm, where the expectation step (E-step) only integrates over the latent $R_i$s, while the conditional maximization step (CM-step) maximizes over $\mu$ with the latent factor model parameters estimated after profiling the expected complete log-likelihood. Our algorithms are further accelerated by SQUAREM [35] to achieve superlinear convergence.

1) An AECM algorithm with a duet of latent variables: The EM algorithm for ML estimation in a Gaussian factor model [46] can be naturally extended to the PN factor model by considering both $R = (R_1, R_2, \ldots, R_n)$ and $Z = [Z_1, Z_2, \ldots, Z_n]$ as missing data. Then the $Q$-function, that is, the conditional expectation of the complete data log-likelihood given the data $X$, is

$$Q_D(\mu, \Lambda, \Psi; \mu_1, \Lambda_1, \Psi_1) = c - \frac{1}{2}n \log \det \Psi + \text{Tr} \{ \Psi^{-1} \mathbb{E}_{Z,R|X}(S_{Z,R}) \},$$

where $c$ is a constant free of $\mu$, $\Lambda$ and $\Psi$, and $\mathbb{E}_{Z,R|X}(S_{Z,R})$ is the conditional expectation of $S_{Z,R} = n^{-1} \sum_{i=1}^n (R_i X_i - \mu - \Lambda Z_i)^\top (R_i X_i - \mu - \Lambda Z_i)$ given $X$, at the current parameter values $(\mu_1, \Lambda_1, \Psi_1)$. The E-step needs this conditional expectation, but it is straightforward to show (Section S1.5) that the conditional distribution of $Z_i$ given $R$ and $X$ at the current parameter values $(\mu_1, \Lambda_1, \Psi_1)$ is $\mathcal{N}_q(\Lambda_1^\top (\Lambda_1 \Lambda_1^\top + \Psi_1)^{-1}(R_i X_i - \mu_1), (I + \Lambda_1^\top \Psi_1^{-1} \Lambda_1)^{-1})$. Further, the conditional expectations of $R_i$ and $R_i^2$ given $X$ are $\mathbb{E}_{R_i|X}(R_i) = \mathcal{I}_p(m_i, v_i)/\mathcal{I}_p-1(m_i, v_i)$ and $\mathbb{E}_{R_i|X}(R_i^2) = \mathcal{I}_{p+1}(m_i, v_i)/\mathcal{I}_{p-1}(m_i, v_i)$.

![Fig. 1: Sample 3D PN densities (with higher values in darker shades) for three different $\Sigma$s.](image)
Proposition 1. Subject to the constraint $Z$ employing matrix-free profile likelihood maximization \[44\]. We propose to speed up our AECM algorithm by incorporating \[\text{E-step} \] has two sets (or a duet) of latent variables, so we use FADS-D to specify alternate an E-step with each CM-step that we now discuss individually. It is important to note that all CM-steps in our AECM algorithm have closed-form solutions. The E-step has two sets (or a duet) of latent variables, so we use FADS-D to specify.

Proof. The proposition is a corollary of Theorem S2 in Section S1.2 with $B = \Psi$.

For the CM-step for $\Lambda$, we maximize $Q_D(\mu_{t+1}, \Lambda_t, \Psi_t; \mu_t, \Lambda_t, \Psi_t)$ w.r.t. $\Lambda$ at the current values of the other parameters, that is at $\mu_{t+1}$ and $\Psi_t$ to get the update

$$\Lambda_{t+1} = \left[ \sum_{i=1}^{n} \mathbb{E}_{R_i|X_i} \left\{ (R_i X_i - \mu_{t+1} Z_i^\top) \right\} \right] \left[ \sum_{i=1}^{n} \mathbb{E}_{Z_i|X_i} \left\{ (Z_i Z_i^\top) \right\} \right]^{-1},$$

where the conditional expectations are at the current parameter values $\mu_{t+1}, \Lambda_t, \Psi_t$.

Finally, the CM-step for $\Psi$ maximizes $Q_D(\mu_{t+1}, \Lambda_{t+1}, \Psi; \mu_{t+1}, \Lambda_{t+1}, \Psi_t)$ w.r.t. $\Psi$ while keeping fixed the values of the other parameters at $\mu_{t+1}$ and $\Lambda_{t+1}$. The CM-step update is then

$$\Psi_{t+1} = \text{diag} \left[ \left( \sum_{i=1}^{n} \mathbb{E}_{R_i|Z_i} \left\{ (R_i X_i - \mu_{t+1} Z_{i+1}) (R_i X_i - \mu_{t+1} Z_i) \right\} \right) \right],$$

where the conditional expectation is at the current parameter values $\mu_{t+1}, \Lambda_{t+1}, \Psi_t$.

2 A faster AECM algorithm with profiling: The methodology of Section II-B1 has two sets of latent variables in the lengths $R_i$s and the factors $Z_i$s. The latter also arise in the case of EM methods for Gaussian factor models, but can be eliminated by employing matrix-free profile likelihood maximization \[44\]. We propose to speed up our AECM algorithm by incorporating these methods. We drop $Z_i$s from our erstwhile duet and only consider $R_i$s as our latent variables. Then the $Q$-function is

$$Q_P(\mu, \Lambda, \Psi; \mu_t, \Lambda_t, \Psi_t) = c - \frac{n}{2} \log \det(\Lambda \Lambda^\top + \Psi) + \text{Tr}(\Lambda \Lambda^\top + \Psi)^{-1} \mathbb{E}_{R_i|X_i}(S_R),$$

where $c$ is a constant free of $\mu$, $\Lambda$ and $\Psi$, and $\mathbb{E}_{R_i|X_i}(S_R)$ is the conditional expectation of $S_R = n^{-1} \sum_{i=1}^{n}(R_i X_i - \mu)(R_i X_i - \mu)^\top$ at the current parameter values $(\mu_t, \Lambda_t, \Psi_t)$. The calculation of the conditional expectation $E_{R_i|X_i}(S_R)$ involves the same conditional expectations of $R_i$ and $R_i^2$ given $X$ as in FADS-D and are computed similarly. We now discuss the CM-steps. The calculation of the parameter space into which we partition the parameter space naturally in terms of $\mu_t$ and $\Lambda_t$.

The CM-step for $\mu$ (at the $(t+1)$th cycle) involves maximizing $Q_P(\mu, \Lambda_t, \Psi_t; \mu_t, \Lambda_t, \Psi_t)$ as a function of $\mu$ subject to the constraint $||\mu|| = 1$. Analogous to Proposition 1, a unique closed-form solution for the CM-step for $\mu$ can be obtained following the through

Proposition 2. The global maximizer of $\mu \to Q_P(\mu, \Lambda_t, \Psi_t)$ is $\mu_{t+1} = (I + \Lambda \Sigma_t)^{-1} c_t$, where $c_t = n^{-1} \sum_{i=1}^{n} X_i \mathbb{E}(R_i|X_i, \mu_t, \Lambda_t, \Psi_t)$. $\Sigma_t = \Lambda_t \Lambda_t^\top + \Psi_t$ and $\lambda$ is a root of $g(u) = c_t^\top (I + u \Sigma_t)^{-2} c_t - 1$. Moreover, if $||c_t|| > 1$, then $\lambda$ is the unique root of $g(u)$ between 0 and $c_t^\top \Sigma_t^{-1} c_t/2$, otherwise it is the unique negative root between $(|u| c_t^\top c_t/2 - 1)/\sigma^2$ and 0, where $\sigma^2$ is the largest eigenvalue of $\Sigma_t$ with eigenvector $v_\star$.

Proof. The proposition follows as a corollary to Theorem S2 in Section S1.2 with $B = \Sigma$. \[45\]
The root \( \lambda \) in Proposition 2 can be computed by bisection while \( v_* \) can be computed using a restarted Lanczos algorithm, as seen shortly in Section II-B3. Also, the inverses of \( \Sigma_t \) or \( I + \lambda \Sigma_t \) are computed using the Sherman-Morrison-Woodbury identity [47].

Given \( \mu_{t+1} \) and \( X_i, \) \( i = 1, 2, \ldots, n, \) we obtain \( \mathbb{E}(R_t | X_i) \) and \( \mathbb{E}(R_{t+1}^2 | X_i) \) and let

\[
\tilde{S}_t = \mathbb{E}_{R|X}(S_R) \bigg|_{\mu = \mu_{t+1}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{R_i | X_i} \left( (R_t X_i - \mu_{t+1}) (R_t X_i - \mu_{t+1})^\top \right)
\]

at the current parameter values. With \( \mu \) held fixed at \( \mu_{t+1}, \) maximizing (8) w.r.t. \( \Lambda \) and \( \Psi \) is equivalent to maximizing the log-likelihood of a Gaussian factor model where the mean is profiled out and the “sample covariance” matrix is \( \tilde{S}_t. \) Therefore, we jointly update \( \Lambda \) and \( \Psi \) by following [44] and profiling out \( \Lambda, \) using the common (in Gaussian factor analysis) and computationally useful identifiability constraint on \( \Lambda \) that the signal matrix \( \Gamma = \Lambda \Psi^{-1} \Lambda^\top \) is diagonal with non-increasing diagonal entries. This scale-invariant constraint is determined up to sign in the columns of \( \Lambda. \) Under this constraint, \( \Lambda \) is profiled out from \( Q_p(\mu_{t+1}; \Lambda, \Psi; \mu_{t+1}, \Lambda, \Psi) \) for a given \( \Psi \) as per the following

**Proposition 3.** For a positive-definite diagonal matrix \( \Psi, \) let \( \theta_1 \geq \theta_2 \geq \cdots \geq \theta_q, \) be the \( q \) largest eigenvalues of \( W = \Psi^{-1/2} S_t \Psi^{-1/2}. \) Let the columns of \( V_q \) store the eigenvectors corresponding to these eigenvalues. Then the function \( Q_p(\mu_{t+1}; \Lambda, \Psi; \mu_{t+1}, \Lambda, \Psi) \) is maximized w.r.t. \( \Lambda \) at \( \Lambda = \Psi^{1/2} V_q \Delta, \) where \( \Delta \) is a \( q \times q \) diagonal matrix with \( i \)th diagonal entry as \( \max(\theta_i - 1, 0)^{1/2}. \) The profile Q-function is

\[
Q_p(\Psi) \doteq c' - \frac{n}{2} \{ \log \det \Psi + \text{Tr } \Psi^{-1} \tilde{S}_t + \sum_{i=1}^{q} (\log \theta_i - \theta_i + 1) \},
\]

where \( c' \) is a constant free of \( \Psi. \) Further, \( \nabla Q_p(\Psi) = -\frac{1}{n} \text{diag}(\Delta \Psi^{-1} + \Psi - \tilde{S}_t). \)

**Proof.** See Section S1.3.

Our algorithm here involves profiling in the second CM-step, so we call it FADS-P. We use the limited-memory Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm [48] with box-constraints (L-BFGS-B) to maximize (9). Instead of the exact Hessian matrix, this quasi-Newton algorithm uses values of \( Q_p(\Psi) \) and \( \nabla Q_p(\Psi) \) from the last few (typically five) iterations to find an effective approximation to the exact Newton-step, thereby reducing storage costs from \( O(p^2) \) to \( O(p). \) Having obtained \( \Psi_{t+1} = \arg \max Q_p(\Psi), \) we use Proposition 3 to compute the eigenvalues and eigenvectors of \( \Psi_{t+1} \) and update \( \Lambda_{t+1} = \Psi_{t+1}^{1/2} V_q \Delta. \)

We conclude with a few remarks on the CM-update of \( \Psi. \) The function \( Q_p(\Psi) \) in (9) depends on \( \tilde{S}_t \) only through the \( q \) largest eigenvalues of \( W, \) so for \( Q_p(\Psi) \) and \( \nabla Q_p(\Psi), \) we need to compute only the \( q \) largest eigenvalues and corresponding eigenvectors of \( W. \) For \( q \ll \min(n,p), \) the largest eigenvalues and eigenvectors are speedily computed using the restarted Lanczos algorithm, with constraints on \( \Psi \) (e.g., \( \Psi = \sigma^2 I_p, \sigma^2 > 0 \)) easily incorporated. Also, \( \nabla Q_p(\Psi) \) is available in closed-form, enabling a check for first-order optimality. Finally, \( Q_p(\Psi) \) is expressed in terms of \( \tilde{S}_t, \) but ML estimators are scale-equivariant, so each iteration of FADS-P estimates \( \Lambda \) and \( \Psi \) using the correlation matrix and scales the estimates back to that of \( \tilde{S}_t. \)

3) Implementation via fast statistical computations: We want our FADS algorithms to also be used for datasets with large \( n \) or \( p, \) as in our applications, so we now discuss and investigate practical ways to speed up computations.

a) Computation of log-likelihood and conditional expectations: Both the log-likelihood (1) and the conditional expectations require computing (4) for \( k \in \{p-1, p, p+1\}. \) To reduce clutter, we write \( \mathcal{I}_k \equiv \mathcal{I}_k(m, v). \) Integration by parts yields the recurrence relation [37]

\[
\mathcal{I}_{k+2} = m \mathcal{I}_{k+1} + (k+1)v \mathcal{I}_k.
\]

However, this recurrence relation is vulnerable to numerical overflow for large \( p, \) so we derive (Section S1.4) a recurrence relation for the ratio

\[
\frac{\mathcal{I}_{k+2}}{\mathcal{I}_{k+1}} = m + (k+1)v \frac{\mathcal{I}_k}{\mathcal{I}_{k+1}}.
\]

Together, (10) and (11) also yield \( \mathcal{I}_{k+2}/\mathcal{I}_k = (k+1)v + m \mathcal{I}_{k+1}/\mathcal{I}_k. \) When \( m > 0, \) the recurrence relation (11) is numerically stable and computed very quickly, even for large \( k. \) However, for \( m < 0, \) both (10) and (11) are unstable and often yield negative values for \( \mathcal{I}_k \) with \( k > 10. \) So, for \( m < 0, \) we use quadrature to compute \( \mathcal{I}_k, \) for \( k \in \{p-1, p, p+1\}, \) using the fact that the integrand in (4) is strictly log-concave and maximized at \( q = (m + \sqrt{m^2 + 4vk})/2. \) We write

\[
\log \mathcal{I}_k = k \log q - \frac{1}{2v} (q - m)^2 + \log \int_0^\infty \left( \frac{R^2}{q} \right)^k \exp \left\{ -\frac{1}{2v} (R - \bar{q})(R + \bar{q} - 2m) \right\} dR.
\]

The integrand is bounded above by unity so we can easily find (e.g., using bisection) two quantities \( 0 < \underline{q} < \bar{q} \) where the integrand is less than some small value (e.g., \( 10^{-15} \)). The integration in (12) is by non-adaptive Gauss-Kronrod quadrature over \([\underline{q}, \bar{q}]\) with 10, 21, 43 and 87 points [49]: in our experiments, 43 points gave high numerical accuracy.
b) Computation of partial eigenvalues and eigenvectors: We can obtain the $q$ largest eigenvalues and eigenvectors of $W$ via the implicitly restarted Lanczos algorithm [50]. Suppose that $m = \max\{2q + 1, 20\}$ and that $f_1 \in \mathbb{R}^p$ is any vector with $\|f_1\| = 1$ and initialize $F_1 = f_1$. We then employ the Lanczos iterations [51] as follows. For $k = 1, 2, \ldots, m$,

- Compute $u_k = Wf_k$ and $\alpha_k = f_k^T u_k$.
- Compute $r_k = u_k - \alpha_k f_k - \hat{\beta}_{k-1} f_{k-1}$ (assuming $\hat{\beta}_0 = 0$ and $f_0 = 0$).
- Let $\beta_k = \|r_k\|$ and if $k < q$ and $\beta_k \neq 0$, compute $f_{k+1} = \frac{r_k}{\beta_k}$ and set $F_{k+1} = [f_k, f_{k+1}]$.

Suppose that $T_m$ is the $m \times m$ symmetric tridiagonal matrix with diagonal entries $\alpha_1, \alpha_2, \ldots, \alpha_m$, and $j$th-off diagonal entries $\beta_j$ for $j = 1, \ldots, m - 1$. We compute the eigenvalues $e_1 > e_2 > \cdots > e_m$ of $T_m$ with eigenvectors $g_1, g_2, \ldots, g_m$ via a Sturm sequencing algorithm [52]. Also let $v_j = F_m g_j$, for $1 \leq j \leq m$. The $e_j$’s and $v_j$’s are called Ritz values and Ritz vectors of $W$. It can be shown that $\|Wv_j - v_j e_j\| = \beta_m |g_{j,m}|$, for $g_{j,m}$ the $m$th entry of vector $g_j$, for $j = 1, 2, \ldots, m$. The algorithm stops if

$$\beta_m \max_{1 \leq j \leq m} |g_{j,m}| < \delta$$

for some prespecified tolerance $\delta$ and $e_1, \ldots, e_q$ and $v_1, v_2, \ldots, v_q$ are accurate approximations of the $q$ largest eigenvalues and corresponding eigenvectors of $W$.

However, in practice, more iterations are needed for the Ritz vectors and Ritz values to converge to the eigenvalues and eigenvectors of $W$, so [50] suggests implicitly restarting the Lanczos algorithm and also shifting the spectrum of the symmetric tridiagonal matrices iteratively to force the new residuals $r_m$ to zero, thereby accelerating the convergence rate. So we compute the QR-decompositions: $T_m - e_j f_m^T = Q_j R_j$, for $j = q + 1, \ldots, m$, let $Q = Q_{q+1} Q_{q+2} \cdots Q_m$, and reset $F_m = F_m Q$ and $T_m = Q^T T_m Q$. Then

$$WF_q = F_q T_q + \beta^* f_{q+1} e_q^T$$

(14)

where $\beta^*$ is the $(q + 1, q)$th entry of $T_m$, $e_q$ is the $q$th canonical basis vector in $\mathbb{R}^q$, and $T_q$ is the $q \times q$ principal sub-matrix of $T_m$ [50]. Therefore, (14) is itself a $q$th-order Lanczos factorization of $W$. Next, we “restart” the Lanczos iterations from $k = q + 1, \ldots, m$ instead of 1 through $m$, terminating if (13) is satisfied, and restarting the algorithm otherwise.

The only way $W$ enters our algorithm is through matrix-vector products $Wf_k$ that can be computed without storing $W$ or $S_l$. Additionally, $W$ is symmetric, so we do not need separate computations for the left and right singular vectors in partial singular value decomposition using the Lanczos algorithm (as needed in [44]), yielding substantial savings in terms of both compute time and storage. Overall, our algorithm calculates the $q$ largest eigenvalues and eigenvectors with $O(nqp)$ computational cost and $O(qp)$ additional memory.

4) Initialization: As with most iterative algorithms, initial values can significantly impact performance of our EM algorithms. We devise methods for initializing ($\mu, \Lambda, \Psi$), borrowing ideas from [53], [54]. Our $\mu$ was simulated from $\mathcal{N}_p(0, I)$ and $l_2$-normalized, while the diagonal entries of $\Psi$ were each i.i.d. $\mathcal{U}(0.2, 0.8)$ draws and the elements of $\Lambda$ were i.i.d. $\mathcal{N}(0, 1)$ realizations. Starting with $M$ such initial values, the AECA algorithm was run for $J$ “short” iterations after which only the $L \leq M$ streams with the highest log-likelihood values were iterated all the way to convergence. The final estimates are those obtained from the run achieving the highest final log-likelihood values. In this paper, we used $(M, J, L) = (1000, 10, 10)$.

C. Choosing $q$

The factor model explains variability in a large number of variables through a small number ($q$) of latent factors, so the choice of $q$ is important. There is little theoretical work on the selection of $q$ for Gaussian factor models. The Bayesian information criterion (BIC) of [55] is one possibility, however it does not always perform well in high dimensions [36]. We use eBIC that minimizes $\text{eBIC}(q) = -2 \log \hat{\ell}_q + pq \{\log(n) + 2\gamma \log(p)\}$, with $\gamma = \max\{1 - 1/(2 \log_n(p)), 0\}$, which performs well in our experiments.

D. Estimating the factor scores

Once $q$ has been chosen and the ML estimators of the parameters obtained, we may compute factor scores for use in subsequent analyses. For example, these scores can be used to rank each observational unit on the factors, for post-hoc clustering [56], [57], [58], [59], in scRNA-seq analysis [60], [61], or they can be used as covariates in generalized linear regression models [62], for example, to account for population structure in genome-wide association studies [63], [64], [65]. We extend the approach of [66, p 226–231] to estimate the factor score for the $i$th observation by minimizing the expected weighted squared error loss $\mathbb{E}\{||\Psi^{-1/2} (Y_i - \mu - \Lambda Z_i)\|^2 | X_i\}$. The resulting solution is the conditional expectation of Bartlett score with complete data: $\tilde{Z}_i = (\Lambda^T \Psi^{-1} \Lambda)^{-1} \Lambda^T \Psi^{-1} \{\mathbb{E}_{R_i | X_i}(R_i)X_i - \mu\}$. We use the unobserved $Y_i$’s instead of $X_i$’s in the minimization problem because $\Psi$ and $\Lambda$ are estimated on the latent scale. As a result, the factor scores are also estimated on the latent scale. In practice, the ML estimates of the parameters are plugged in, and the value of the conditional expectation $\mathbb{E}_{R_i | X_i}(R_i)$ at the ML estimates is available as a byproduct of our FADS algorithms.
E. Standard errors of the ML estimates

The standard errors of the ML estimates can be obtained by using the missing information principle [67] when $n \geq (q + 2)p$. To that end, suppose that $\Theta = \{\mu, \Lambda, \Psi\}$. The observed information at the ML estimates is given by, $I_X(\Theta) = (1/n) \sum_{i=1}^{n} \nabla q_i \nabla q_i^\top$, where for $i = 1, 2, \ldots, n$, $\nabla q_i = \mathbb{E}_{R_i(\theta)}(\partial\ell(Y_i; \Theta)) / \partial \Theta$ is the complete data score statistic and

$$\frac{\partial \ell(Y_i; \Theta)}{\partial \mu} = (\Lambda\Lambda^\top + \Psi)^{-1}(R_i X_i - \mu)$$

$$\frac{\partial \ell(Y_i; \Theta)}{\partial \Lambda} = -\frac{1}{2} \left( (\Lambda\Lambda^\top + \Psi)^{-1} - (\Lambda\Lambda^\top + \Psi)^{-1}(R_i X_i - \mu)(R_i X_i - \mu)^\top (\Lambda\Lambda^\top + \Psi)^{-1} \right)$$

$$\frac{\partial \ell(Y_i; \Theta)}{\partial \Psi} = -\frac{1}{2} \mathrm{diag} \left\{ (\Lambda\Lambda^\top + \Psi)^{-1} - (\Lambda\Lambda^\top + \Psi)^{-1}(R_i X_i - \mu)(R_i X_i - \mu)^\top (\Lambda\Lambda^\top + \Psi)^{-1} \right\}.$$ 

In practice, the ML estimates of the parameters $\mu$, $\Lambda$ and $\Psi$ are plugged into the above. Further, $\mathbb{E}(R_i | X_i)$ and $\mathbb{E}(R_i^2 | X_i)$ at the ML estimates are available as byproducts of the AECM algorithm so these calculations are very easily obtained in the course of the FADS calculations.

III. PERFORMANCE EVALUATIONS

We evaluated our FADS algorithms through simulation experiments that assessed estimation accuracy and consistency as well as computing speed for a range of $(n, p, q)$ settings.

A. Experimental setup

We simulated 100 datasets for each $(n, p) \in \{(300, 30), (1000, 100)\}$ and $q \in \{3, 5\}$. Our strategy for each dataset was to set the diagonal elements of $\Psi$ as i.i.d. $U(0, 0.8)$ and entries in $\Lambda$ as i.i.d. $N(0, 1)$ pseudo-random deviates. Entries of $\mu$ were simulated independently from $N(0, 1)$ and $l_{2}$-normalized. A separate evaluation – to mimic the $p \gg n$ setting of the application in Section I-A4 – used 100 simulated datasets with $(p, q) = (5123, 12)$ and $n \in \{380, 500\}$, with the true $\Psi, \Lambda, \mu$ set to the ML estimates obtained upon fitting FADS-P to the TCGA dataset. In all cases, we stopped FADS-P and FADS-D when the improvement in the observed log-likelihood did not exceed $10^{-4}$ and $\|\nabla Q_{P}\|_{\infty} < \sqrt{e_0}$, where $e_0 \approx 2.2 \times 10^{-16}$ is the machine tolerance, or if the number of iterations exceeded $10^4$. For each simulated dataset, we fitted models with $k = 1, 2, \cdots, 2q$ factors and chose the number of factors using eBIC. All experiments were run on a workstation with Intel E5-2640 v3 CPU clocked @2.60 GHz and 64GB RAM.

B. Results

In our experiments, eBIC always correctly picked $q$. Since factor loadings and uniquenesses are better interpreted on the correlation scale [26], we considered the parameters $\Lambda_\Theta = \text{diag}(\sigma)^{-1/2} \Lambda$, $\Psi_\Theta = \text{diag}(\sigma)^{-1} \Psi$, $\Upsilon_\Theta = \Lambda_\Theta \Lambda_\Theta^\top$ and $\Theta = \Upsilon_\Theta + \Psi_\Theta$, where $\sigma$ is the vector of marginal standard deviations in $\Sigma$. We compared the FADS-D and FADS-P estimates using the relative Frobenius distances: $d_\Theta = \|\Theta - \Theta_\text{true}\| / \|\Theta_\text{true}\|$, $d_{\Theta_\text{F}} = \|\Theta_\text{F} - \Theta_\text{true}\| / \|\Theta_\text{true}\|$, $d_{\Theta_\text{G}} = \|\Theta_\text{G} - \Theta_\text{true}\| / \|\Theta_\text{true}\|$, $d_{\Theta_\text{E}} = \|\Theta_\text{E} - \Theta_\text{true}\| / \|\Theta_\text{true}\|$, $d_{\Theta_\text{D}} = \|\Theta_\text{D} - \Theta_\text{true}\| / \|\Theta_\text{true}\|$, $d_{\Theta_\text{C}} = \|\Theta_\text{C} - \Theta_\text{true}\| / \|\Theta_\text{true}\|$, $d_{\Theta_\text{B}} = \|\Theta_\text{B} - \Theta_\text{true}\| / \|\Theta_\text{true}\|$, $d_{\Theta_\text{A}} = \|\Theta_\text{A} - \Theta_\text{true}\| / \|\Theta_\text{true}\|$, $d_{\Theta_\text{I}} = \|\Theta_\text{I} - \Theta_\text{true}\| / \|\Theta_\text{true}\|$.

Fig. 2a presents the relative speed of FADS-P to FADS-D for the randomly simulated cases – see Section S2.2 for average CPU time. (Our reported compute times include the common initialization time for both algorithms.) For $(n, p) \in \{(300, 30), (1000, 100)\}$, FADS-P was faster than FADS-D, with maximum speedup occurring at true $q$. FADS-P also generally needed far fewer iterations than FADS-D to converge. In particular, FADS-D did not converge even within $10^4$ iterations for four $(n, p, q) = (1000, 100, 5)$ cases. In contrast, FADS-P always converged even though it required more iterations in the over-fitted models than when the fitted $q$ did not exceed the true $q$ (Section S2.1). Therefore, the speedup of FADS-P relative to FADS-D was slightly underestimated because of the censored reports from FADS-D.

For the data-driven cases where $n \ll p$, FADS-D never converged within $10^4$ iterations, but FADS-P always converged successfully. So we only report FADS-P estimates for $n \ll p$, and calculate and compare estimation accuracy of FADS-D and FADS-P for the $n > p$ cases where both of them converged. (The four $n > p$ cases where FADS-D did not converge had terminating results that were not significantly different in estimation accuracy from those from FADS-P.) FADS-P and FADS-D yielded identical values of $\hat{\Theta}, \hat{\Gamma}, \hat{\Upsilon}_\Theta$ and $\hat{\Psi}_\Theta$ under the best-fitted models, so the relative estimation errors (Figure 2b) were also identical. The relative errors for $\mu$ and $\sigma$ were also similar for FADS-P and FADS-D. With $n \ll p$, the relative errors of FADS-P estimates decreased with increasing sample size (Table 1 and Figs. 2b and S1).

The results of our experiments show that FADS-P and FADS-D (when it converges) can, along with the use of eBIC, accurately recover factor structures underlying our PN model. Overall, FADS-P provides faster and more reliable estimates, so we use it in our applications.
### IV. REAL DATA APPLICATIONS

We applied FADS-P to the four datasets of Section I-A. For each dataset, we first discuss how its processing places the observations on the unit sphere and then fit, analyze and interpret our factor models. We also provide diagnostic checks to verify the preference for our PN factor model over the existing (Langevin/Fisher-von Mises) model. The Langevin distribution has hitherto been the only one routinely used to directly model observations on unit spheres of as high dimensions as our applications, so our diagnostic is in the form of a hypothesis test, where the null hypothesis is that the alternative hypothesis specifies each $X_i$ with mean direction vector $\nu$ and concentration parameter $\kappa$. The alternative hypothesis specifies each $X_i \sim \mathcal{PN}_{p}(\mu, \Lambda \Lambda^\top + \Psi)$, where $\mu^\top \mu = 1$, $\Lambda$ is a $p \times q$ matrix and $\Psi$ is a diagonal matrix. Our test statistic is the difference in maximized log-likelihood ($\ell(\nu, \kappa)$) and concentration parameter $\kappa$. The exact Monte Carlo method to calculate the $p$-value of our test statistic in each application. Under this framework, the alternative model is favored by a large $-\text{value}$ of our test statistic in each application. We use [68]'s exact Monte Carlo method to calculate the $p$-value of our test statistic in each application. Under this framework, the alternative model is favored by a large $-\text{value}$ of our test statistic in each application.

#### A. Characterizing major themes underlying the #MeToo tweets

1) Data collection and preprocessing: We collected all 43,247 original tweets (excluding so-called retweets) in English that contained the #MeToo hashtag between December 03, 2018 at 02:06:15 (UTC) and December 13, 2018 at 01:36:20 (UTC). These tweets originally had 1,165,762 unique words. Standard text preprocessing put the dataset through (1) unitization and tokenization, (2) standardization and cleansing, (3) stop word removal, and (4) stemming and lemmatization [70]. In addition,

| $(n, p, q)$ | $R$ | $\Gamma$ | $\Psi_R$ | $\mu$ | $\sigma$ |
|-------------|-----|----------|----------|-------|---------|
| (380,7123,12) | 0.020 | 0.336 | 0.020 | 0.186 | 1.181 | 0.093 |
| (500,7123,12) | 0.020 | 0.326 | 0.020 | 0.184 | 1.188 | 0.093 |
infrequently-occurring words (appearing in less than 0.25% of tweets) were removed from the lexicon, following [10]. We also removed one-word tweets because they contribute no information on relationships between terms. The resulting collection had \( n = 31,385 \) unique tweets and \( p = 721 \) unique words (see Section S3.1). (Henceforth, we use tweets to refer to this reduced dataset.) We applied the commonly used term frequency-inverse document frequency \((tf-idf)\) weighting [70] to modulate the relative frequency of each word by tweet length. This procedure weighted \( d_{ij} \), or the raw frequency of the \( j \)th word, in the \( i \)th tweet by \( \log(n/d_{ij}) \) where \( d_{ij} = \sum_{i=1}^{n} d_{ij} \) is the number of tweets with the \( j \)th word. We therefore have a \( tf-idf \) weighted document-term matrix \((DTM)\). Each tweet vector was then \( l_2 \)-normalized, placing it on the unit sphere. The \( l_2 \)-normalization mitigates the effect of differing tweet lengths [11] and also allows for the use of cosine similarity that performs better than the Euclidean distance metric on sparse text data [71]. The \( l_2 \)-normalized data \( X_1, X_2, \ldots, X_n \) is our dataset for analysis.

2) Results and analysis: We applied FADS-P, with \( q = 0, 1, 2, \ldots, 15 \), to the processed dataset of Section IV-A1. The eBIC chose the eight-factor model, which was significantly preferred over the Langevin model \((p\text{-value} < 0.001)\). Fig. 3a displays the estimated uniquenesses for the fitted model, and is similar to the relative frequency of words contributing to the main discussion topics in the #MeToo movement during the given time period. We analyzed the eight factor loadings from the fitted model using quartimax rotation that simplifies interpretation by minimizing both the number of heavily loaded features on each factor and the number of factors needed to explain a feature [45]. Figs. 3b–i provide word clouds of the eight factor loadings, ordered by the proportion of variance in \( \Sigma \) explained. The word clouds provide insight into the dominant words that characterize each factor and help describe the variability in the #MeToo tweets. We now discuss and interpret the eight factors.

The first, fourth and sixth factors pertain to words appearing in tweets discussing different aspects of sexual harassment in the workplace. For instance, the first factor is dominated by words that appear in tweets discussing the December 3, 2018 Bloomberg article [72] titled “Wall Street Rule For The #MeToo Era: Avoid Women At All Cost” that outlines exclusionary strategies reportedly adopted by male Wall Street employees to avoid allegations in the #MeToo era, but that simultaneously raise obstacles for professional women and exacerbate gender segregation. We notice similar trends in the sixth factor that heavily weights words appearing in tweets discussing dilemmas faced by women in the workplace as their male colleagues limit professional contact \( (e.g., \) by excluding them from dinner outings, solo meetings, sitting together while flying) to avoid accusations of professional sexual misconduct. So, though the words and topics are similar in these two factors, the first factor is focused on the Bloomberg article and avoidance strategies adopted by men on Wall Street while the sixth factor relates to avoidance issues in the general workplace. The related fourth factor is a contrast between the dominant terms “wall” and “street” on one hand and “cost”, “avoid”, “era”, “rule” and “women” on the other. Tweets with these words are generally about the December 5, 2018 Bloomberg article “NYC officials blast Wall Street’s ice-out of women in wake of #MeToo” [73] in response to the article associated with the first factor. These tweets mainly criticise Wall Street men’s reported strategy of avoiding women in the workplace, and instead school them on appropriate behaviours in a professional setting.

The second factor prominently loads words that are in tweets reacting to the cancellation by several radio stations, in the last week of November 2018, of the 1940s classic Christmas song “Baby It’s Cold Outside”, in response to listener complaints about its lyrics describing inappropriate and manipulative male behaviors toward women. Tweets with these words generally support the ban, though some others argue that the song was written in a different era.

The third factor primarily weights words conveying sexual assault, harassment, abuse and violence, and, to a lesser extent, “women”, “allege”, “survive” and “workplace”. We interpret this factor to reflect the main objectives of the #MeToo movement to expose sexual assault and harassment by men in power, especially in the workplace, and to provide support to survivors.

The remaining factors relate to politics and movements. The fifth factor has terms related to politics or policies, mostly appearing in tweets discussing the just concluded 2018 midterm elections in the United States, where female voters, female candidates and the #MeToo movement were generally viewed to have played a major role. The seventh factor is dominated by words that are in tweets related to movements for social justice and reform. Examples are tweets related to gender discrimination and harassment issues across generations and professions \((\text{farm workers, veterans, indigenous people, etc})\) or discussions on socio-political matters or allied movements like “Black Lives Matter” or “March for our Lives”, where millennials have been quite vocal and active. Conversely, the eighth factor is overwhelmingly weighted by words related to right-wing, misogynistic, white supremacist and conspiracy theory-based groups \((\text{#MAGA, QAnon, #Blexit, #MenToo, Men Going Their Own Way, etc})\) and their views on #MeToo, “Black Lives Matter” and other anti-fascist and social justice movements.

Z. Zhu has asked why we do not simply fit a (generative) Gaussian factor model to the tweets dataset before \( l_2 \)-normalization. The reason for not doing so is that the unnormalized tweet vectors cannot be assumed to all come from the same normal distribution. The processed tweets, after routine \( l_2 \)-normalization, are assumed to be i.i.d. from a PN factor model. The only other currently realistic model for such data is the Langevin distribution, but the hypothesis test reported on earlier fit our projected PN factor model significantly better than the Langevin model for this dataset. Our fitted model had eight factors that, upon quartimax rotation, provided insight into the main contributors of variability in the tweets.

B. Characterizing the pre-teen brain at rest

1) Data preprocessing: The data are from resting-state measurements [74] of 33 typically developing children between 8 and 12 years of age. Here, participants were instructed to relax and fixate on a crosshair while fMRI scans were acquired over
Fig. 3: (a) Estimated uniquenesses (with magnitude proportional to shading and size of word font) from the eight-factor PN model fitted to the #MeToo dataset and (b)-(i) the fitted factor loadings after quartimax rotation (here, word color corresponds to the sign and font size and shading correspond to magnitude of factor loadings). We characterize each factor loading through a descriptive sub-caption.
seven minutes, at intervals of 2.5s. The first 30s of data were discarded to let the scanner magnetization achieve a steady state, yielding 156 time points. For each registered, motion-corrected dataset, the AFNI software package [75], [76] used standard preprocessing techniques to regress out nuisance parameters. The residual time series for each subject at each of 75,589 in-brain voxels was normalized by its respective temporal mean and standard deviation [77], resulting in normalized time series on the surface of a 156-dimensional sphere. The normalized time series at each voxel were then summarized over the 33 pre-teens as the mean direction time series vector at each voxel. At this stage, we have a dataset \( D \) of \( n = 75,589 \) 156-dimensional observations on the unit sphere. However, regression (in this case, of the nuisance parameters) yields a singular matrix of residuals, so \( D \) is projected to a 69-dimensional space by means of a matrix of orthogonal rows spanning the null space of the regression hat matrix. (Any lower-dimensional orthogonal projection of a unit norm vector is also a unit norm vector in lower dimensional space if the projection vector is orthogonal to the null space of the vector being projected.) We thus have a dataset on the 69-dimensional unit sphere of \( n = 75,589 \) observations.

2) Results and analysis: FADS-P with eBIC chose the 2-factor model from among \( q = 0, 1, 2, \ldots, 10 \). The PN factor model fit the data significantly better (\( p \)-value < 0.001) than its Langevin cousin. We backprojected the factor loadings to the original 156-dimensional space and used oblimin rotation for interpretable results that we now discuss.

Fig. 4a displays the time series of the oblimin-rotated factor loadings in 156-dimensional space, while Figs. 4b,c plot the auto-correlation function (ACF) and partial ACF (PACF) along with 95% confidence intervals (broken lines). The first factor loadings show larger amplitude fluctuations between the tenth and fortieth time points than elsewhere. The second factor loadings fluctuate more frequently than the first, with a clearer seasonal pattern spanning approximately 15 time-points (37.5s)
in both the time series (Fig. 4a) and the ACF (Fig. 4b). The large spikes in the first few lags of the ACF for both factor loadings, followed by a cycle of decreasing (in magnitude) alternating negative and positive correlations, indicate a higher-order autoregression structure and a strong long-term seasonal component for the second factor loading. The PACF (Fig. 4c) also suggests presence of moving average terms, long-term dependencies and non-stationary components in both factor loadings. In summary, it appears that the variability over time of the in-brain voxels of the pre-teen brain at rest can be summarized in terms of two temporal components, both nonstationary and periodic, but one having longer-range dependence than the other. We now discuss the factor scores.

Figs. 4d,e provide voxel-wise displays of the two factor scores estimated as per Section II-D. The first factor scores are high in magnitude in the visual and sensory cortex, in areas of emotional and higher mental function (concentration, planning, judgment, emotional expression, creativity and inhibition) and in the motor function areas of the cerebellum that coordinate movement and balance. The second factor scores have non-negligible values in areas of the cerebral cortex involving visual, higher mental function, sensory association, olfactory and motor functions that initiate and orient the voluntary movement of muscles and the eye [78]. We see that the factor scores represent fairly functionally distinctive brain regions.

C. Handwritten digits

1) Preprocessing: As per Section I-A3, the dataset is of 70,000 784-pixel bitmap images. Fig. 5a displays 20 random specimens of each digit and shows variation in handwriting between (and within) digits. It is this variability that we want to characterize by our PN factor model. We removed pixels with the same value across all coordinates because they contribute no information on variability. Further, because of the discretized (bitmapped) values in the dataset, we jittered each of them by adding i.i.d. $N(0, 0.1^2)$ pseudo-random deviates. The jittered images are essentially indistinguishable from the originals (Fig. S6) and were $l_2$-normalized for use in our analysis.

2) Results and Analysis: The PN factor model was fit using FADS-P for $q = 0, 1, \ldots, 15$: eBIC chose the 12-factor model that was found to fit the data significantly better ($p$-value < 0.001) than the Langevin. The factor loadings after quartimax rotation are displayed, in decreasing order of their contribution to the variance, in Fig. 5b (first column) – dropped pixels were set to zero for the factor loadings. For interpretability, we also used absolute correlation to assign each handwritten digit to its closest factor loading. Fig. 5b displays the relative-frequency-weighted mean of the assigned images for each factor-digit combination. We now discuss and interpret the factors.

The first factor in Fig. 5b (upper left) is largely an overall contrast between foreground and background pixels over all the digits, and consequently is strongly correlated with samples from each digit. The remaining 11 factor loadings are contrasts between different sets of pixels in the imaging regions. Indeed, each of them are correlated with different kinds of digits (see rows, and caption of Fig. 5b). For example, the second factor loading has high association primarily with “4” and “9” and to a lesser extent “6”, “7” and “0”. Images having high correlations with this factor loading have high values at many of the same pixel locations. Also, the 12th factor loading is primarily associated with large handwritten “7”s spanning the entire image. Some digits, such as “2” and “6”, are strongly associated with more factor loadings than others, probably because there are more handwriting styles (in terms of pixels used) for these digits. D. M. Ommen has pointed out that the third and eleventh factors are, strikingly, largely associated with upright-written digits, potentially indicating left-handed writers. (The MNIST dataset does not have information on handedness, so we are unable to verify this conjecture.)

D. Nettleton has asked about using traditional Gaussian factor analysis on this dataset, that is, essentially ignoring the constraints imposed by the digitization of a bitmap image. Doing so yielded poor results, unlike FADS that not only accounts for the imaging data acquisition process but also identified interpretable factors that capture the variability in handwriting styles both between and within digits. We remark that our methodology has not accounted for spatial context or pixel neighborhood. Incorporating these aspects may improve our results further.

D. TCGA dataset

1) Data collection and preprocessing: We use an mRNA sequence (RNA-seq) dataset comprised of 775 primary breast cancer (BC) tumors from the TCGA effort [79], prepared and released in the Bioconductor package curatedTCGADatasets [80]. RNA-seq leverages high-throughput next-generation sequencing technology [81] to examine the presence and quantity of mRNA from each gene in a biological sample by sequencing random deoxyribonucleic acid fragments derived from the sampled mRNA [82]. During data preparation, the sampled sequence fragments were mapped to 20,502 genes, producing a count of fragments for each gene. Since the total number of fragments sequenced (coverage) is not biologically meaningful, we scaled the counts in each experiment to sum to unity, and then screened out low expression genes, by discarding genes with mean scaled expressions below the third quartile, producing our dataset on 5,123 genes. After rescaling the 5,123 surviving gene proportions to sum to unity in each of the 775 experiments, we applied a square root transformation to place the compositional data on the unit sphere, as in [24].
2) Results and analysis: FADS with eBIC selected a 12-factor model when fitting up to 15 factors. The first eight of these factors explained 95% of the variation in the data. This model also significantly ($p$-value < 0.001) out-performed the Langevin model in fitting the processed data. We used quartimax rotation for further analysis. To determine if any factor could be prognostic, we fitted the Cox proportional-hazards model [83] to the survival time of the 775 observations using each factor score in turn as the predictor. A likelihood ratio test indicated significant association of the sixth factor with survival (Holm's corrected $p$-value of 0.005). Fig. 6a displays survival curves with the 95% confidence intervals, which are estimated using the Kaplan–Meier method [84], and obtained after thresholding on the median value (1.09925) of the sixth factor scores. The two survival curves are significantly different ($p$-value of 0.00069 via a logrank test [85]). Further, high values of the sixth factor correspond to increased survival, especially between years 2–8.

The sixth factor can be associated with genes to identify possible prognostic biomarkers or biological processes to hypothesize a mechanism for altered survival. In particular, it is common to identify the biological processes associated with gene-level statistics using gene set enrichment analysis (GSEA) [86]. We performed GSEA using the sixth factor loadings, on the correlation scale, as the gene-level statistic and an enrichment $p$-value was calculated for each of the 50 hallmark gene sets in the MSigDB collections, version 5 [87]. GSEA found the sixth factor enriched for 13 of the 50 hallmark gene sets with false discovery rate (FDR) adjusted $p$-values < 0.05. Figure 6b presents the normalized enrichment scores (NES) and FDR-adjusted $p$-values of these thirteen selected gene sets.

We examine the five gene sets with FDR-adjusted $p$-values below $10^{-8}$, comparing our findings to the literature. The INTERFERON_GAMMA_RESPONSE and INTERFERON_ALPHA_RESPONSE gene sets are immunity-related and substantially overlapped hallmark gene sets (Jaccard index 0.33). Though the role of the immune response in BC is far from clear [88], the constellation of genes activated in response to the interferon-γ cell-signaling cytokine have recently been associated with poor prognosis in another cohort of BC patients [89]. The top three loaded genes in both gene sets are IFI27, ISG15, and RSAD2.
10
8000
6000
4000
2000
+ + + +
N
+ + + +

This paper develops methodology for exploratory factor analysis of data on a sphere, with a view to explaining the variability in the coordinates in terms of a few underlying (latent) factors. Such datasets arise in applications involving text documents, time series image volumes of the brain, digitized images, or compositional gene expression data, that we investigate in this paper. Our approach uses the PN distribution [4] which is the distribution of a normalized \( N_p(\mu, \Sigma) \) random vector, with the restriction that \( \mu \in S_{p-1} \) for identifiability. Our factor model is introduced through a decomposition of the generative MVN model. We develop computationally fast estimation methods for the parameters using two versions of the AECM algorithm. The first, called FADS-D, uses the latent factors and lengths of the random vectors of our generative model as the missing variables, while the second approach, called FADS-P, profiles out the latent factors in the generative model. We also provided uncertainty estimators for our models. Our methods are implemented in the developmental version of our \texttt{fad} package in R [108] that is available at \url{https://github.com/somakd/fad}. Simulation results show fast and excellent performance, especially of FADS-P with the number of factors well-estimated using eBIC. Analysis of datasets from four motivating applications provided interpretable insights into the underlying mechanisms governing variability in each case.

There are a number of points that could benefit from further attention. For instance, our PN factor model (and indeed the MVN factor model itself) could benefit from results on consistency of the eBIC. Further, our model could also be used in regression and classification settings, replacing, for instance, the Fisher distribution, in the development of the spherical regression model of [109], with appropriate modifications. Such an approach would allow us to characterize resting state brains in autistic and typically developing children, for example. Other extensions to be investigated include Bayesian methods for estimation, as well as including sparsity structures on \( \Lambda \). It may also be of interest to incorporate the PN model in settings

IFN27 physically associates with estrogen receptor \( \alpha \), an important mediator of proliferation in BC [90], ISG15 modifies other proteins through ISGylation and correlates with lymphovascular invasion and poor prognosis [91], [92], and RSAD2, known to inhibit virus replication [93], associates with tumor stage and lymph node metastases [94]. The \texttt{E2F\_TARGETS} and \texttt{G2M\_CHECKPOINT} gene sets are involved in cell proliferation, a commonly upregulated function in tumor samples [95]. E2F transcription factors participate in cell cycle control, and several have recently been associated with BC prognosis [96], while the entire \texttt{G2M\_CHECKPOINT} hallmark gene set has been deemed a prognostic biomarker of metastasis in one type of BC [97]. Transcription factor MYBL2 is the most (negatively) loaded gene on the sixth factor in both gene sets, while \texttt{TOP2A} is the second most (negatively) loaded gene among E2F targets and the third most (negatively) loaded G2M checkpoint gene. MYBL2 directly targets several kinesin motor proteins that are correlated with poor BC prognosis [98]. TOP2A is a DNA topoisomerase whose expression has been correlated with poor BC prognosis [95], [99]. In fact, it is a direct target of some chemotherapies, including a trial where its expression was successfully used to guide treatment [100]. The gene set \texttt{MTORC1\_SIGNALING} consists of genes upregulated through activation of MTORC1, a protein complex responsible for protein synthesis that has also been targeted by chemotherapeutic agents for BC [101]. MTORC1 appears frequently in the BC literature as several prognostic markers are in MTORC1-related gene sets [102], [103], [104], [105]. Here, SQLE is the most (negatively) loaded gene on the MTORC1 gene set and the second most (negatively) loaded gene on the G2M gene set. It is an hallmark gene set has been deemed a prognostic biomarker of metastasis in one type of BC [97]. Transcription factor MYBL2 is the most (negatively) loaded gene on the sixth factor in both gene sets, while \texttt{TOP2A} is the second most (negatively) loaded gene among E2F targets and the third most (negatively) loaded G2M checkpoint gene. MYBL2 directly targets several kinesin motor proteins that are correlated with poor BC prognosis [98]. TOP2A is a DNA topoisomerase whose expression has been correlated with poor BC prognosis [95], [99]. In fact, it is a direct target of some chemotherapies, including a trial where its expression was successfully used to guide treatment [100]. The gene set \texttt{MTORC1\_SIGNALING} consists of genes upregulated through activation of MTORC1, a protein complex responsible for protein synthesis that has also been targeted by chemotherapeutic agents for BC [101]. MTORC1 appears frequently in the BC literature as several prognostic markers are in MTORC1-related gene sets [102], [103], [104], [105]. Here, SQLE is the most (negatively) loaded gene on the MTORC1 gene set and the second most (negatively) loaded gene on the G2M gene set. It is an enzyme involved in cholesterol synthesis, with an impact on proliferation, that has been associated with poor prognosis [106], [107]. Thus, FADS finds a single, highly interpretable, apparently proliferation-related, axis of variation significantly associated with BC survival.

V. DISCUSSION

This paper develops methodology for exploratory factor analysis of data on a sphere, with a view to explaining the variability in the coordinates in terms of a few underlying (latent) factors. Such datasets arise in applications involving text documents, time series image volumes of the brain, digitized images, or compositional gene expression data, that we investigate in this paper. Our approach uses the PN distribution [4] which is the distribution of a normalized \( N_p(\mu, \Sigma) \) random vector, with the restriction that \( \mu \in S_{p-1} \) for identifiability. Our factor model is introduced through a decomposition of the generative MVN model. We develop computationally fast estimation methods for the parameters using two versions of the AECM algorithm. The first, called FADS-D, uses the latent factors and lengths of the random vectors of our generative model as the missing variables, while the second approach, called FADS-P, profiles out the latent factors in the generative model. We also provided uncertainty estimators for our models. Our methods are implemented in the developmental version of our \texttt{fad} package in R [108] that is available at \url{https://github.com/somakd/fad}. Simulation results show fast and excellent performance, especially of FADS-P with the number of factors well-estimated using eBIC. Analysis of datasets from four motivating applications provided interpretable insights into the underlying mechanisms governing variability in each case.

There are a number of points that could benefit from further attention. For instance, our PN factor model (and indeed the MVN factor model itself) could benefit from results on consistency of the eBIC. Further, our model could also be used in regression and classification settings, replacing, for instance, the Fisher distribution, in the development of the spherical regression model of [109], with appropriate modifications. Such an approach would allow us to characterize resting state brains in autistic and typically developing children, for example. Other extensions to be investigated include Bayesian methods for estimation, as well as including sparsity structures on \( \Lambda \). It may also be of interest to incorporate the PN model in settings

![Kaplan-Meier survival plot](image)

\textbf{Fig. 6:} (a) Kaplan-Meier survival plot split by median value of the sixth factor. Orange cases are above and blue are below the median. The horizontal and vertical dashed lines indicate the median survival times for the two groups. (b) Enrichment plot. Gene sets are labeled on y-axis.
that involve a combination of \( p \) linear variables and \( q \) variables on \( S_{q-1} \). From an applications perspective, we may also be interested in using factor scores in genome-wide association studies to account for population structure. Therefore, we see that there are many directions for further investigation.

**ACKNOWLEDGMENTS**

The research was supported in part by the United States Department of Agriculture (USDA)/National Institute of Food and Agriculture (NIFA), Hatch projects IOW03617 and IOW03717. The content of this paper however is solely the responsibility of the authors and does not represent the official views of the USDA.

**REFERENCES**

[1] G. S. Watson, *Statistics on Spheres*, ser. University of Arkansas Lecture Notes in the Mathematical Sciences. Wiley Interscience, 1983.

[2] N. Fisher, T. Lewis, and B. Embleton, *Statistical Analysis of Spherical Data*. Cambridge University Press, 1993.

[3] S. R. Jammalamadaka and A. SenGupta, *Topics in Circular Statistics*. World Scientific, 2001.

[4] K. V. Mardia and P. E. Jupp, *Directional Statistics*. New York: Wiley, 2000.

[5] R. von Mises, „über die “ganzzahligkeit” der atomgewichte und verwandte fragen,” *Physikalische Zeitung*, vol. 19, pp. 490–500, 1918.

[6] R. A. Fisher, “Dispersion on a sphere,” *Proc. R. Soc. Lond. A*, vol. 217, pp. 295–305, 1953.

[7] A. SenGupta and R. Maitra, “On best equivariance and admissibility of simultaneous MLE for mean direction vectors of several Langevin distributions,” *Annals of the Institute of Statistical Mathematics*, vol. 50, no. 4, pp. 715–27, 1998.

[8] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information Processing and Management*, vol. 27, no. 5, pp. 513–523, 1988.

[9] T. G. Kolda, “Limited-memory matrix methods with applications,” Ph.D. dissertation, The Applied Mathematics Program, University of Maryland, College Park, Maryland, 1997.

[10] I. S. Dhillon and D. S. Modha, “Concept decompositions for large sparse text data using clustering,” *Machine Learning*, vol. 42, pp. 143–175, 2001.

[11] A. Singhal, C. Buckley, M. Mitra, and G. Salton, “Pivoted document length normalization,” in *ACM SIGIR* 96, 1996, pp. 21–29.

[12] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, “Clustering on the unit hypersphere using von Mises-Fisher distributions,” *Journal of Machine Learning Research*, vol. 6, pp. 1345–1382, 2005.

[13] B. Biswal, F. Z. Yetkin, V. M. Haughton, and J. S. Hyde, “Functional connectivity in the motor cortex of resting human brain using echo-planar MRI,” *Magnetic Resonance in Medicine*, vol. 34, no. 4, pp. 537–541, 1995.

[14] B. B. Biswal, “Resting state fMRI: A personal history,” *NeuroImage*, vol. 62, no. 2, pp. 938 – 944, 2012.

[15] J. W. Belliveau, D. N. Kennedy, R. C. McKinstry, B. R. Buchbinder, R. M. Weisskoff, M. S. Cohen, J. M. Vevea, T. J. Brady, and B. R. Rosen, “Functional mapping of the human visual cortex by magnetic resonance imaging,” *Science*, vol. 254, pp. 716–719, 1991.

[16] K. K. Kwong, J. W. Belliveau, D. A. Chesler, I. E. Goldberg, R. M. Weisskoff, B. P. Poncelet, D. N. Kennedy, B. E. Hoppel, M. S. Cohen, R. Turner, H.-M. Cheng, T. J. Brady, and B. R. Rosen, “Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, pp. 5675–5679, 1992.

[17] P. A. Bandettini, A. Jesmanowicz, E. C. Wong, and J. S. Hyde, “Processing strategies for time-course data sets in functional mri of the human brain,” *Magnetic Resonance in Medicine*, vol. 30, pp. 161–173, 1993.

[18] M. A. Lindquist, “The statistical analysis of fMRI data,” *Statistical Science*, vol. 23, no. 4, pp. 439–464, 2008.

[19] N. A. Lazar, *The Statistical Analysis of Functional MRI Data*. Springer, 2008.

[20] M. D. Fox, A. Z. Snyder, J. L. Vincent, M. Corbetta, D. C. Van Essen, and M. E. Raichle, “The human brain is intrinsically organized into dynamic, anticorrelated functional networks,” *Proceedings of the National Academy of Sciences*, vol. 102, no. 27, pp. 9673–9678, 2005.

[21] D. Cole, S. Smith, and C. Beckmann, “Advances and pitfalls in the analysis and interpretation of resting-state fMRI data,” *Frontiers in Systems Neuroscience*, vol. 4, p. 8, 2010.

[22] R. Maitra and I. P. Ramler, “A \( k \)-mean-directions algorithm for efficient clustering of data on a sphere,” *Journal of Computational and Graphical Statistics*, vol. 19, no. 2, pp. 379–390, 2010.

[23] Y. LeCun, C. Cortes, and C. J. C. Burges, “The mnist database,” 2019. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[24] X. Dai and H.-G. Müller, “Principal component analysis for functional data on Riemannian manifolds and spheres,” *Annals of Statistics*, vol. 46, no. 6B, pp. 3334–3361, 12 2018.

[25] K. Bose and P. Chaudhari, *Unravelling Cancer Signaling Pathways: A Multidisciplinary Approach*. Springer, 2019.

[26] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate analysis*. New York: Academic Press, 1979.

[27] J. Dortet-Bernadet and N. Wicker, “Model-based clustering on the unit sphere with an illustration using gene expression profiles,” *Biosciences Reports*, vol. 9, no. 1, pp. 66–80, 2008.

[28] C. Bingham, “An antipodally symmetric distribution on the sphere,” *Annals of Statistics*, vol. 2, pp. 1201–1225, 1974.

[29] J. T. Kent, “The Fisher–Bingham distribution on the sphere,” *Journal of the Royal Statistical Society, Series B*, vol. 44, pp. 71–80, 1982.

[30] J. T. Kent, “The Fisher–Bingham distribution and related distributions,” in *Advances and papers in theoretical statistics*, ser. ASTM, 1983, pp. 143–165.

[31] J. T. Kent, “The Fisher–Bingham distribution and related distributions,” *Annals of Statistics*, vol. 33, pp. 1643–1665, 05 2005.

[32] B. Kim, S. Huckemann, J. Schulz, and S. Jung, “Small sphere distributions for directional data with application to medical imaging,” *Scandinavian Journal of Statistics*, vol. 43, pp. 444–445, 07 2016.

[33] F. Wang and A. E. Gelfand, “Directional data analysis under the general projected normal distribution,” *Statistics and Computing*, vol. 50, no. 4, pp. 715–27, 1998.

[34] D. N. Lawley, “The estimation of factor loadings by the method of maximum likelihood,” *Annals of Mathematical Statistics*, vol. 20, pp. 588–591, 1949.

[35] R. Varadhan and C. Roland, “Simple and globally convergent methods for accelerating the convergence of any EM algorithm,” *Scandinavian Journal of Statistics*, vol. 9, pp. 117–131, 01 1982.

[36] R. Varadhan and C. Roland, “Simple and globally convergent methods for accelerating the convergence of any EM algorithm,” *Scandinavian Journal of Statistics*, vol. 9, pp. 117–131, 01 1982.

[37] T. G. Kolda, “Limited-memory matrix methods with applications,” Ph.D. dissertation, The Applied Mathematics Program, University of Maryland, College Park, Maryland, 1997.

[38] T. G. Kolda, “Limited-memory matrix methods with applications,” Ph.D. dissertation, The Applied Mathematics Program, University of Maryland, College Park, Maryland, 1997.
A. Derivation of the projected normal density function

**Theorem S1.** Suppose $Y \sim N_p(\mu, \Sigma)$, with mean vector $\mu$ and covariance matrix $\Sigma$. Then, $X = Y / \|Y\|$ follows a projected normal distribution $P_{N_p}(\mu, \Sigma)$ and the density function of $X$ at $x = y / \|y\|$ is given by,

$$f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{\|x - \mu\|^2}{2\Sigma} + \frac{m^2}{2v} \right\} \int_0^R R^{p-1} \exp \left\{ -\frac{1}{2v} (R - m)^2 \right\} dR,$$

(S1)

where $R = \|y\|$, $m = x^T \Sigma^{-1} \mu / x^T \Sigma^{-1} x$, $v = 1 / x^T \Sigma^{-1} x$.

**Proof.** Let $y = (y_1, y_2, \ldots, y_p)$ with $R = \|y\|$ and $x = (x_1, x_2, \ldots, x_p) = y / \|y\|$ so that $\|x\| = 1$. Define the representation of $x$ in the spherical coordinate system as $\theta = (\theta_1, \theta_2, \ldots, \theta_{p-1})$ so that

$$\begin{align*}
x_1 &= \cos \theta_1; \quad x_j = \prod_{k=1}^{j-1} \sin \theta_k \cos \theta_j, \quad \text{for} \ j = 2, 3, \ldots, p-1; \quad x_p = \prod_{k=1}^{p-1} \sin \theta_k,
\end{align*}$$

(S2)
with the Jacobian of transformation as \( J_x = \prod_{j=1}^{p-2} \sin^{(p-1-j)} \theta_j \). Then, we denote the transformation from \( R, \theta \) to \( y = y(R, \theta) \) which is given by,

\[
y_1 = R \cos \theta_1; \quad y_j = R \prod_{k=1}^{j-1} \sin \theta_k \cos \theta_j, \text{ for } j = 2, 3, \ldots, p - 1; \quad y_p = R \prod_{k=1}^{p-1} \sin \theta_k,
\]

(S3)

with the Jacobian of transformation as \( J_y = R^{p-1} \prod_{j=1}^{p-2} \sin^{(p-1-j)} \theta_j \). Now, by the change of variables, the density function of \( \theta \) is given by,

\[
f(\theta; \mu, \Sigma) = \int_{0}^{\infty} J_y (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (y(R, \theta) - \mu)^\top \Sigma^{-1} (y(R, \theta) - \mu) \right\} dR.
\]

(S4)

Then, by applying the change of variables again, the density function of \( x \) is given by,

\[
f(x; \mu, \Sigma) = \int_{0}^{\infty} J_y / J_x (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (Rx - \mu)^\top \Sigma^{-1} (Rx - \mu) \right\} dR
\]

\[(= (2\pi)^{-\frac{p}{2}} \Sigma^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \mu^\top \Sigma^{-1} \mu + m^2 \right\} \int_{0}^{\infty} R^{p-1} \exp \left\{ -\frac{1}{2v} (R - m)^2 \right\} dR, \]

where \( m = x^\top \Sigma^{-1} \mu / x^\top \Sigma^{-1} x, \; v = 1 / x^\top \Sigma^{-1} x \).

\square

B. Numerical computation of the MLE of \( \mu \)

**Theorem S2.** Let \( B \) be a \( p \times p \) positive definite symmetric matrix, \( b \in \mathbb{R}^p \). Then, the unique global minimizer \( x_0 \) that

\[
\text{minimize } \text{Tr } B^{-1} (b - x)(b - x)^\top \text{ subject to } x^\top x = 1 \tag{S6}
\]

is given by \( x_0 = (I + \lambda_0 B)^{-1} b \), where \( \lambda_0 \) lies between

1. \( (0, b^\top B^{-1} b / 2] \), if \( b^\top b - 1 > 0 \),
2. \( \lambda = 0 \), if \( b^\top b - 1 = 0 \),
3. \( ((u_m^2 b / 2 - 1) / e_m, 0] \), if \( b^\top b - 1 < 0 \).

Where \( e_m, u_m \) are the largest eigenvalue and corresponding eigenvector of \( B \).

**Proof.** The Lagrangian associated with (S6) is given by,

\[
L(x, \lambda) = \text{Tr } B^{-1} (b - x)(b - x)^\top + \lambda (x^\top x - 1). \tag{S7}
\]

with a Lagrange multiplier \( \lambda \). Setting \( \nabla L = 0 \) yields \( x_0 = (I + \lambda B)^{-1} b \) and \( b^\top (I + \lambda B)^{-2} b = 1 \).

Let \( f(\lambda) = b^\top (I + \lambda B)^{-2} b - 1 \), \( B = UMU^\top \), \( M = \text{diag}(e_j), j = 1, 2, \ldots, p \) and \( U^\top b = (c_1, c_2, \ldots, c_p)^\top = c \), then,

\[
f(\lambda) = c^\top (I + \lambda M)^{-2} c - 1 = \sum_{j=1}^{p} \frac{c_j^2}{(1 + \lambda e_j)^2} - 1
\]

and

\[
f'(\lambda) = -2 \sum_{j=1}^{p} \frac{c_j^2 e_j}{(1 + \lambda e_j)^3}, \quad f''(\lambda) = 6 \sum_{j=1}^{p} \frac{c_j^2 e_j^2}{(1 + \lambda e_j)^4}.
\]

For \( \lambda \geq 0 \), \( f(\lambda) \) is a continuous function with \( f'(\lambda) < 0 \) and

\[
f(\lambda) = \sum_{j=1}^{p} \frac{c_j^2}{(1 + \lambda e_j)^2} - 1 \leq \sum_{j=1}^{p} \frac{c_j^2}{2 \lambda e_j} - 1 = \frac{1}{2\lambda} b^\top B^{-1} b - 1.
\]

So,

\[
\lambda > \frac{b^\top B^{-1} b}{2} \Rightarrow f(\lambda) < 0
\]

If \( f(0) = b^\top b - 1 > 0 \), we can then find the root of \( f(\lambda) = 0 \) in \( [0, b^\top B^{-1} b / 2] \).

If \( b^\top b - 1 < 0 \), no positive root exists. Then let \( \lambda < 0 \), we have

\[
f(\lambda) = \sum_{j=1}^{p} \frac{c_j^2}{(1 + \lambda e_j)^2} - 1 \geq \frac{c_j^2}{(1 + \lambda e_j)^2} - 1.
\]
For each \( j = 1, 2, \ldots, p \). So,
\[
-\sqrt{\frac{c^2}{e_j^2} - 1} < \lambda < \sqrt{\frac{c^2}{e_j^2} - 1} \implies f(\lambda) > 0.
\]

According to the Hessian matrix \( H_x = B^{-1} + \lambda I \), the unique global minimizer is obtained when \( \lambda > -\frac{1}{e_m} \), where \( e_m = \max\{e_1, e_2, \ldots, e_p\} \). Let \( u_m \) denote the corresponding eigenvector and \( e_m = u_m^T b \) in \( c \), then \( -\frac{1}{e_j} \leq -\frac{1}{e_m} < \frac{0.5\sqrt{\sigma_m^2} - 1}{e_m} \), for \( j = 1, 2, \ldots, p \) and when \( \lambda \geq \frac{0.5\sqrt{\sigma_m^2} - 1}{e_m} \), \( f(\lambda) \) is continuous. Hence, we can find an negative root of \( f(\lambda) = 0 \) in
\[
\left[ \frac{0.5\sqrt{\sigma_m^2} - 1}{e_m}, 0 \right].
\]

C. Proof of Proposition 3

From Section II-B2.1, for fixed \( \mu_{t+1} \), \( Q_P \) is equivalent to
\[
Q_P(\mu_{t+1}, \Lambda; \Psi; \mu_{t+1}, \Lambda_t; \Psi_t) = c - \frac{n}{2} \{ \log \det(\Lambda \Lambda^T + \Psi) + \text{Tr} (\Lambda \Lambda^T + \Psi)^{-1} S_t \},
\]
where \( S_t = \frac{1}{m} \sum_{i=1}^n \mathbb{E}_{X_t} \left\{ (R_t X_t - \mu_{t+1})(R_t X_t - \mu_{t+1})^T \right\} \).

From (S8), the ML estimators of \( \Lambda \) and \( \Psi \) are obtained by solving the score equations
\[
\begin{cases}
\Lambda(I_q + \Lambda^T \Psi^{-1} \Lambda) = S_t \Psi^{-1} \Lambda \\
\Psi = \text{diag}(S_t - \Lambda \Lambda^T)
\end{cases}
\]
(S9)

From \( \Lambda(I_q + \Lambda^T \Psi^{-1} \Lambda) = S_t \Psi^{-1} \Lambda \), we have
\[
\Psi^{-1/2} \Lambda(I_q + (\Psi^{-1/2} \Lambda)^T \Psi^{-1/2} \Lambda) = \Psi^{-1/2} S_t \Psi^{-1/2} \Lambda.
\]
(S10)

Suppose that \( \Psi^{-1/2} S_t \Psi^{-1/2} = V D V^T \) and that the diagonal elements in \( D \) are in decreasing order with \( \theta_1 \geq \theta_2 \geq \cdots \geq \theta_p \).

Let \( D = \begin{bmatrix} D_q & 0 \\ 0 & D_m \end{bmatrix} \) with \( m = p - q \) and \( D_q \) containing the largest \( q \) eigenvalues \( \theta_1 \geq \theta_2 \geq \cdots \geq \theta_q \). The corresponding \( q \) eigenvectors form columns of the matrix \( V_q \) so that \( V = [V_q, V_m] \). Then, if \( D_q > I_q \), (S10) shows that
\[
\Lambda = \Psi^{1/2} V_q (D_q - I_q)^{1/2}.
\]
(S11)

Hence, conditional on \( \Psi \), \( \Lambda \) is maximized at \( \hat{\Lambda} = \Psi^{1/2} V_q \Delta \), where \( \Delta \) is a diagonal matrix with elements \( \max(\theta_i - 1, 0)^{1/2}, i = 1, \cdots, q \).

From the construction of \( V_q \) and \( V_m \), we have \( V_q^T V_q = I_q, V_m^T V_m = I_m, \quad V_q V_q^T + V_m V_m^T = I_p, \quad V_q^T V_m = 0 \) and hence, \( (V_q D_q V_q^T + V_m V_m^T)(V_q D_q^T V_q^T + V_m V_m^T) = I_p \).

Let \( A = V_q \Delta^2 V_q^T \). Then \( \Lambda \Lambda = V_q \Delta^4 V_q^T \) and
\[
|A + I_p| = |(A + I_p)A|/|A|
= |V_q (\Delta^4 + \Delta^2 V_q^T V_q)|/|V_q \Delta^2 V_q^T|
= |\Delta^2 + I_q| = \prod_{j=1}^q \theta_j
\]
(S12)

and
\[
(A + I_p)^{-1} = (V_q \Delta^2 V_q^T + V_q V_q^T + V_m V_m^T)^{-1}
= (V_q (\Delta^2 + I_q) V_q^T + V_m V_m^T)^{-1}
= (V_q D_q V_q^T + V_m V_m^T)^{-1}
= V_q D_q^{-1} V_q^T + V_m V_m^T.
\]
(S13)
Based on (S12) and (S13) and (9), the profile log-likelihood is given by,

\[ Q_p(\Psi) = c' - \frac{n}{2} \log |\hat{\Lambda}^{\top} + \Psi| - \frac{n}{2} \text{Tr} (\hat{\Lambda}^{\top} + \Psi)^{-1} \hat{S}_t \]

\[ = c' - \frac{n}{2} \left\{ \log |\Psi^{1/2}(V_q \Delta \Psi V_q^{\top} + I_p)\Psi^{1/2}| + \text{Tr} (\Psi^{1/2}(V_q \Delta^{2} V_q^{\top} + I_p)\Psi^{1/2})^{-1} \hat{S}_t \right\} \]

\[ = c' - \frac{n}{2} \left\{ \log \det \Psi + \log |V_q \Delta^{2} V_q^{\top} + I_p| + \text{Tr} (V_q \Delta^{2} V_q^{\top} + V_m V_{\nu}^{\top})^{-1} \hat{S}_t \right\} \]

\[ = c' - \frac{n}{2} \left\{ \log \det \Psi + \sum_{j=1}^{q} \log \theta_j + \text{Tr} D_{q}^{-1} V_q^{\top} V D V^{\top} \Psi + \text{Tr} V_m V D V^{\top} \Psi \right\} \]  \hspace{1cm} (S14)

\[ = c' - \frac{n}{2} \left\{ \log \det \Psi + \sum_{j=1}^{q} \log \theta_j + \text{Tr} D_{q}^{-1} D_{q} + \text{Tr} D_{m} \right\} \]

\[ = c' - \frac{n}{2} \left\{ \log \det \Psi + \sum_{j=1}^{q} \log \theta_j + \text{Tr} \Psi^{-1} \hat{S}_t - \sum_{j=1}^{q} \theta_j \right\} , \]

where \( c' \) is a constant that does not depend on \( \Psi \).

### D. Recursive formulas

Let \( \mathcal{I}_k = \int_0^\infty R^k \exp \left\{ -\frac{1}{2v} (R - m)^2 \right\} dR \) with \( m \in \mathbb{R}, v > 0, k \in \mathbb{N} \), then,

\[ \mathcal{I}_k = \frac{R^{k+1}}{k+1} \exp \left\{ -\frac{1}{2v} (R - m)^2 \right\} \bigg|_0^\infty - \int_0^\infty \frac{R^{k+1}}{k+1} \exp \left\{ -\frac{1}{2v} (R - m)^2 \right\} \left( -\frac{R}{v} + \frac{m}{v} \right) dR \]

\[ = \frac{m}{v(k+1)} \mathcal{I}_{k+2} - \frac{m}{v(k+1)} \mathcal{I}_{k+1}. \]

That is, \( \mathcal{I}_{k+2} = m \mathcal{I}_{k+1} + (k+1)v \mathcal{I}_k \). Then, \( \frac{\mathcal{I}_{k+2}}{\mathcal{I}_k} = m + (k+1)v, \frac{\mathcal{I}_{k+1}}{\mathcal{I}_k} = (k+1)v + m \frac{\mathcal{I}_{k+1}}{\mathcal{I}_k} \).

### E. FADS-D: E- and CM-step computation

The conditional expectations of \( Z_i, Z_i Z_i^{\top} \) and \( R_i, Z_i \) given \( X \) are

\[ \mathbb{E}_{Z_i | X_i}(Z_i) = \mathbb{E}_{R_i | X_i}(Z_i), \mathbb{E}_{Z_i | X_i}(Z_i Z_i^{\top}) = \mathbb{E}_{R_i | X_i}(Z_i Z_i^{\top}) \]

The joint density function of \( R_i, X_i, Z_i \) is given by,

\[ f_{R_i, X_i, Z_i}(R_i, X_i, Z_i) = f_{Y_i | Z_i}(Y_i | Z_i) | J_i | \cdot f_{Z_i}(Z_i) \]

\[ \propto \exp \left\{ -\frac{1}{2} (R_i X_i - \mu - \Lambda Z_i)\Psi^{-1}(R_i X_i - \mu - \Lambda Z_i) - \frac{1}{2} x_i^{\top} Z_i \right\}. \]  \hspace{1cm} (S15)

Then, by the conditional normal distribution, we have

\[ \mathbb{E}_{Z_i | X_i}(Z_i) = \Lambda X_i \Sigma^{-1} \mathbb{E}_{R_i | X_i}(R_i) X_i - \mu, \]

\[ \mathbb{E}_{Z_i | X_i}(R_i Z_i^{\top}) = (R_i \Sigma^{-1} R_i^{\top} + \mu \mu^{\top}) \Sigma^{-1} \Lambda, \]

\[ \mathbb{E}_{Z_i | X_i}(Z_i Z_i^{\top}) = (I + \Lambda^{\top} \Lambda)^{-1} + \Lambda^{\top} \Sigma^{-1} \mathbb{E}_{R_i | X_i}(R_i^{2} X_i^{\top} - R_i \mu^{\top} \Sigma^{-1} \Lambda, \]

Then, conditional on \( \hat{\mu}_{t+1} \) and \( \hat{\Psi}_t \), the CM-step for \( \Lambda_{t+1} \) is given by,

\[ \Lambda_{t+1} = \left\{ \sum_{i=1}^{n} X_i \mathbb{E}_{Z_i | X_i}(R_i Z_i^{\top}) - \hat{\mu}_{t+1} \sum_{i=1}^{n} \mathbb{E}_{Z_i | X_i}(Z_i^{\top}) \right\} \left( \sum_{i=1}^{n} \mathbb{E}_{Z_i | X_i}(Z_i Z_i^{\top}) \right)^{-1}. \]  \hspace{1cm} (S16)

Conditional on \( \hat{\mu}_{t+1} \) and \( \Lambda_{t+1} \), the CM-step for \( \Psi_{t+1} \) is given by,

\[ \Psi_{t+1} = \frac{1}{n} \text{diag} \left\{ \sum_{i=1}^{n} \mathbb{E}_{R_i | X_i}(R_i^{2}) X_i X_i^{\top} - 2 \hat{\mu}_{t+1} \sum_{i=1}^{n} \mathbb{E}_{R_i | X_i}(R_i) X_i^{\top} - 2 \Lambda_{t+1} \sum_{i=1}^{n} \mathbb{E}_{Z_i | X_i}(R_i Z_i) X_i^{\top} + 2 \hat{\mu}_{t+1} \sum_{i=1}^{n} \mathbb{E}_{Z_i | X_i}(Z_i^{\top}) \Lambda_{t+1}^{\top} + \hat{\Lambda}_{t+1} \sum_{i=1}^{n} \mathbb{E}_{Z_i | X_i}(Z_i Z_i^{\top}) \Lambda_{t+1}^{\top} + n \hat{\mu}_{t+1} \hat{\mu}_{t+1}^{\top} \right\} . \]  \hspace{1cm} (S17)
The non-positive values in $\Psi_{t+1}$ are truncated to a small positive number [110].

## S2. Supplement to Section III

### A. Average iterations and convergence rates

**TABLE S1**: Average iterations (in converged cases for both FADS-P and FADS-D) and convergence rates (in %) of FADS-P and FADS-D algorithms applied to randomly simulated datasets where true $(n, p, q) = (300, 30, 3)$.

|        | 1 | 2 | 3 | 4 | 5 | 6 |
|--------|---|---|---|---|---|---|
| FADS-P | 109 | 67 | 48 | 83 | 107 | 115 |
|        | 100 | 100 | 100 | 100 | 100 | 100 |
| FADS-D | 257 | 341 | 595 | 885 | 1438 | 2732 |
|        | 100 | 100 | 100 | 92 | 82 | 70 |

**TABLE S2**: Average iterations (in converged cases for both FADS-P and FADS-D) and convergence rates (in %) of FADS-P and FADS-D algorithms applied to randomly simulated datasets where true $(n, p, q) = (300, 30, 5)$.

|        | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 |
|--------|----|----|----|----|----|----|----|----|----|----|
| FADS-P | 127 | 88 | 70 | 62 | 33 | 76 | 89 | 102 | 111 | 147 |
|        | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| FADS-D | 378 | 467 | 570 | 820 | 1304 | 1952 | 2312 | 3225 | 3296 | 4091 |
|        | 91 | 94 | 98 | 100 | 100 | 82 | 55 | 38 | 28 | 24 |

**TABLE S3**: Average iterations (in converged cases for both FADS-P and FADS-D) and convergence rates (in %) of FADS-P and FADS-D algorithms applied to randomly simulated datasets where true $(n, p, q) = (1000, 100, 3)$.

|        | 1  | 2  | 3  | 4  | 5  | 6  |
|--------|----|----|----|----|----|----|
| FADS-P | 683 | 208 | 123 | 141 | 275 | 402 |
|        | 100 | 100 | 100 | 100 | 100 | 100 |
| FADS-D | 1411 | 2075 | 3156 | 3120 | 3864 | 4042 |
|        | 100 | 100 | 100 | 100 | 100 | 100 |

**TABLE S4**: Average iterations (in converged cases for both FADS-P and FADS-D) and convergence rates (in %) of FADS-P and FADS-D algorithms applied to randomly simulated datasets where true $(n, p, q) = (1000, 100, 5)$.

|        | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 |
|--------|----|----|----|----|----|----|----|----|----|----|
| FADS-P | 789 | 231 | 275 | 233 | 141 | 204 | 307 | 456 | 475 | 601 |
|        | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| FADS-D | 2281 | 2411 | 3581 | 4498 | 5818 | 6185 | 6971 | 7269 | 7206 | 7189 |
|        | 100 | 100 | 100 | 100 | 96 | 79 | 72 | 65 | 64 | 59 |
B. Average CPU time

TABLE S5: Average CPU time (in seconds) of FADS-P and FADS-D algorithms applied to randomly simulated datasets (Converged cases).

|                  | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   |
|------------------|------|------|------|------|------|------|------|------|------|------|
| \((n, p, q) = (300, 30, 3)\) |      |      |      |      |      |      |      |      |      |      |
| FADS-P           | 7.570| 5.640| 3.451| 6.826| 9.241| 10.239| -    | -    | -    | -    |
| FADS-D           | 10.411| 13.965| 24.603| 56.619| 59.461| 113.415| -    | -    | -    | -    |
| \((n, p, q) = (300, 30, 5)\) |      |      |      |      |      |      |      |      |      |      |
| FADS-P           | 7.983| 6.248| 5.901| 4.594| 2.171| 5.698| 7.014| 8.718| 10.044| 13.119|
| FADS-D           | 15.229| 18.929| 23.348| 33.641| 54.090| 81.192| 97.207| 134.969| 138.628| 168.619|
| \((n, p, q) = (1000, 100, 3)\) |      |      |      |      |      |      |      |      |      |      |
| FADS-P           | 134.633| 47.134| 26.230| 36.970| 76.216| 111.779| -    | -    | -    | -    |
| FADS-D           | 219.185| 331.254| 503.038| 498.516| 620.895| 644.078| -    | -    | -    | -    |
| \((n, p, q) = (1000, 100, 5)\) |      |      |      |      |      |      |      |      |      |      |
| FADS-P           | 145.243| 46.902| 60.470| 49.188| 27.735| 49.616| 78.007| 121.077| 131.134| 172.879|
| FADS-D           | 354.947| 381.501| 570.623| 710.967| 927.920| 1007.009| 1147.169| 1189.592| 1184.188| 1178.299|

C. Data-driven cases: Relative Frobenius distances of FADS-P estimates

![Graph showing data-driven cases: Relative Frobenius distances of FADS-P for different settings.](image)

Fig. S1: Data-driven cases: Relative Frobenius distances of FADS-P for \(\mathcal{R}, \Gamma, \Psi, \sigma, \mu\).

S3. Supplement to Section IV

A. List of words in #MeToo tweets

Figures S2–S5 together present the 721 words (in alphabetical order along columns) for the final data matrix of the #MeToo tweet data that were collected and analyzed in Section IV-A.
| abc | cbs | drop | harass | marchforourlives | process | smear | wave |
|-----|-----|------|--------|-----------------|---------|-------|------|
| absolute | cbsnews | due | hard | mark | produce | snowflake | wear |
| abuse | celeb | dumb | harm | market | professional | social | week |
| accept | ceo | dunham | hart | marry | professor | society | weinstein |
| account | challenge | easy | hat | matter | progress | solute | white |
| accuse | chance | economy | hate | media | project | son | wife |
| act | change | educate | head | meet | proof | song | win |
| activism | charge | effect | heal | mental | protect | sound | witchhunt |
| actor | check | elect | hear | mention | protest | space | women |
| actress | child | employ | learnetoo | mentoo | proud | speak | womenmarch |
| actual | china | employee | heart | mentor | prove | special | word |
| add | choice | employer | hill | message | public | speech | worker |
| address | church | empower | hillary | mgtow | publish | staff | workplace |
| admit | churchtoo | encourage | hire | michaeljackson | pull | stand | world |
| adopt | claim | enjoy | history | milano | punish | standard | worry |
| advance | class | entertain | hit | millennials | push | star | worse |
| advice | classic | entire | hold | mind | qanon | start | worth |
| advocate | clean | environment | holiday | misconduct | question | stay | writ |
| affect | cleveland | episode | hollywood | misogyny | quiet | step | write |
| afraid | climate | epstein | home | miss | quote | stop | wrong |
| against | clinton | equal | honor | mjakabar | racism | store | youtube |
| age | close | era | hope | model | radio | story |
| agenda | club | event | horror | mom | raise | strategy |
| agree | cnn | evidence | host | moment | rape | street |
| aide | cold | evil | house | money | rapist | strong |
| air | colleague | excuse | huge | month | reach | struggle |
| allege | college | exist | human | moonves | read | student |
| amaze | comment | expect | hurt | moral | real | stuff |
| anger | commit | experience | husband | morn | reality | stupid |
| answer | common | explain | hypocrisy | mote | realize | subject |
| antifa | community | exploit | idea | move | reason | sue |
|--------|-----------|---------|------|------|--------|-----|
| anymore| company  | expose  | idiot | movement | receive | suffer |
| apology| compassion | extreme | ignore | movie | recent | suggest |
| apparent| complain | eye | illegal | msnbc | relate | suit |
| apply | complete | facebook | imagine | mueller | religion | support |
| argue | concern | fail | impact | murder | remain | suppose |
| arrest | congress | fair | inappropriate | music | republic | surprise |
| art | consequence | fake | include | nation | republican | survey |
| article | conservative | fall | india | neiltyson | resign | survive |
| assault | continue | false | indigenous | newyork | resist | suspend |
| attack | control | family | individual | nfl | resistance | system |
| attempt | controversy | fan | industry | nice | respect | tact |
| attention | conversation | farm | info | night | response | taking |
| attitude | convict | fbi | innocent | note | responsible | talk |
| attorney | cord | fear | inspire | nyt | result | taranaburke |
| australia | corporate | feel | interview | obama | reveal | target |
| avenatti | correct | female | investigate | ocasio | rich | taxpayer |
| avoid | corrupt | feminism | involve | offend | ridiculous | teach |
| award | cosby | fend | issue | offensive | right | team |
| aware | cost | fight | jail | officer | risc | teen |
| baby | count | figure | job | official | risk | tell |
| backfire | country | file | joebiden | opinion | role | term |
| backlash | courage | final | join | opportunity | ruin | text |
| bad | court | finance | joke | organize | rule | thehill |
| ban | cover | fine | journalism | origin | run | theview |
| base |crap | fire | jndge | oscars | sad | thread |
| basic | crazy | fit | jump | outrage | safe | threat |
| bate | create | fix | justice | panel | santa | time |
| baylor | credit | flight | kamalaharris | parent | save | timesup |
| beat | creep | flirt | kavanaugh | partner | scandal | tire |

Fig. S3: Word list for #MeToo tweet data, part 2
| beauty | crime | focus | kick | party | scar | top |
|--------|-------|-------|------|-------|------|-----|
| begin  | crisis| folk  | kid  | pass  | school| topic|
| behave | cry   | follow| kill | passe | science| total|
| believe| culture| force| kiss | past  | scream| touch|
| bias   | current| ford | lack | patriarchy| search| toxic|
| big    | dad   | forget| lady | pay   | season| train|
| bit    | daily | form | laugh| pedophile| secret| trauma|
| black  | damage| forward| law | peace | seek | treat|
| blacklivesmatter| date | found| lawsuit| people | segregate| trend|
| blame  | daughter| founder| lawyer| perfect| sell| trial|
| blexit | day   | fox  | lead | perpetrator| senator| trouble|
| block  | daysofactivism| france| learn| person | send| trump|
| blog   | deal | free | lease| personal| senior| trust|
| bloomberg| death | friend| leave| photo | sense | truth|
| bluewave| board| decade| front| left | physical| sentence| tv|
| body   | decision| fund | legal| pick | serve | tyson|
| book   | deep | funny| level| piece | service| unaided|
| boss   | defense| fuse | lgbt | plan | set | understand|
| boy    | demand| future| lgbtq| platform| settle| unintend|
| boycott| democrat| gain | liberal| play | settlement| university|
| brave  | deny | game | lie  | playlist| sex | usa|
| break  | dependent| gay | life | podcast| sexism| verse|
| brexit | destroy| gender| light| police | sexual| veteran|
| bring  | die   | generate| line | policy | shame| victim|
| brown  | difference| girl | link | poor | share| video|
| build  | difficult| glad | list | port | shock| view|
| bully  | dinner| global| listen| position| shoot| vinoddua|
| bunch  | direct| govern| literal| post | short| violate|
| bush   | director| government| live| potential| shut| violence|
| busy   | discriminate| grope| lose | power | sick| virgin|

Fig. S4: Word list for #MeToo tweet data, part 3
Table: Word list for #MeToo tweet data, part 4

| busy    | discriminate | grope | lose | power | sick | virgin |
|---------|--------------|-------|------|-------|------|--------|
| buy     | discussion   | grow  | love | powerful | silence | voice |
| california | diversity   | guess | low  | practice | simple | vote  |
| call    | dnc          | guilt | lyric | predator | sing | wait  |
| campaign | domestic     | gun   | mad  | president | single | wake  |
| canada  | door          | gust  | maddow | press     | sister | walk  |
| car     | doubt         | hand  | maga | prevent | sit  | wall  |
| care    | dress         | handle | major | prison    | situation | war |
| career  | drink         | happen | male | private | sleep | washingtonpost |
| caught  | drive         | happy | manage | privilege | smart | watch |

Fig. S5: Word list for #MeToo tweet data, part 4

B. Handwritten digits

Fig. S6: (a) Original and (b) Jittered images of 20 handwritten digits. The two sets are essentially indistinguishable.