Elliptically-Contoured Tensor-variate Distributions with Application to Image Learning

CARLOS LLOSA-VITE, Statistical Sciences, Sandia National Laboratories, USA
RANJAN MAITRA, Department of Statistics, Iowa State University, USA

Statistical analysis of tensor-valued data has largely used the tensor-variate normal (TVN) distribution that may be inadequate for data arising from distributions with heavier or lighter tails. We study a general family of elliptically contoured (EC) tensor-variate distributions and derive its characterizations, moments, marginal and conditional distributions. We describe procedures for maximum likelihood estimation from data that are (1) uncorrelated draws from an EC distribution, (2) from a scale mixture of the TVN distribution, and (3) from an underlying but unknown EC distribution, for which we extend Tyler’s robust estimator. A detailed simulation study highlights the benefits of choosing an EC distribution over the TVN for heavier-tailed data. We develop tensor-variate classification rules using discriminant analysis and EC errors and show that they better predict cats and dogs from images in the Animal Faces-HQ dataset than the TVN-based rules. A novel tensor-on-tensor regression and tensor-variate analysis of variance (TANOVA) framework under EC errors is also demonstrated to better characterize gender, age and ethnic origin than the usual TVN-based TANOVA in the celebrated Labeled Faces of the Wild dataset.

1 INTRODUCTION

Elliptically contoured (EC) distributions [22, 23, 30, 38, 48, 56, 64] are a flexible class of symmetric vector-variate distributions that generalize the multivariate normal distribution distribution, and facilitate the modeling of data with heavy or light tails. Similarly, matrix-variate EC distributions [21, 28, 29] extends the matrix-variate normal distribution for the case where the tails are heavier or lighter. The emerging interest in tensor data analysis has led to alternatives to the TVN distribution that do not rely on least squares estimation and have more flexible tail weights. In particular, tensor-valued EC-valued distributions have been studied by [2, 6], where they were characterized and some of their fundamental properties were developed. More recently, [69, 73] studied parameter inference for the tensor-variate t distribution, which is an important member of the EC family with adaptive heavy tails.

In this article we provide a thorough study of the EC tensor-variate family. We explore fundamental properties and propose maximum likelihood estimation techniques that we then study and apply to real data in Section 2.
we review and study properties of EC tensor-variate distributions by characterizing them, their marginal and conditional distributions and moments. Inference is no longer as straightforward under EC errors, so Section 3 provides computationally practical methodology for parameter estimation under three different scenarios, specifically a reduced rank tensor-on-tensor regression (ToTR) and tensor-variate analysis of variance (TANOVA) framework with EC errors, and a robust Tyler estimator for when data arise from an underlying but unknown EC distribution. We evaluate performance of our algorithms in Section 4. Section 5 shows the value of our methodology in two real-data scenarios. First, we develop discriminant rules using EC tensor-variate distributions to compare the predictive performance of the TVN and the tensor-variate-C(TV-C) distributions in classifying cats and dogs from their images in the Animal Faces-HQ (AFHQ) database. In all cases, the flexible tensor-variate-C distribution with estimated degrees of freedom (DF) outperforms the TVN in terms of area under the receiver operating characteristic and precision-recall curves. Our second application demonstrates the ability of our ToTR and TANOVA methodology with EC errors to better characterize the Labeled Faces in the Wild (LFW) dataset in terms of age, ethnic origin and gender, than the TVN-based ToTR and TANOVA. We conclude this article with a discussion on our contributions in this paper and propose potential further generalizations. An appendix, with sections, theorems, proofs, equations and figures, provides additional technical and other details.

2 DEFINITIONS AND CHARACTERIZATIONS

2.1 Background and preliminaries

We define a tensor as a multi-dimensional array of numbers. This article uses \( X, \hat{X} \) and \( X \) to denote deterministic tensors, matrices and vectors, with bold-faced fonts for their random counterparts (e.g., \( \mathbf{X}, \mathbf{X} \) and \( \mathbf{X} \) denote random tensors, matrices and vectors). Further, we assume that \( X \in \mathbb{R}^{r \times r \times \cdots \times r} \) has \((i_1, i_2, \ldots, i_p)\)th element written as \( X(i_1, i_2, \ldots, i_p) \). Tensor reshapings [41] allow us to modify the structure of a tensor while preserving its elements. We can reshape \( X \) into a \((m_1) \times (\prod_{i=1,i \neq k}^p m_i)\)-matrix \( X_{(k)} \) using its \( k \)th mode matricization. The tensor \( X \) can also be reshaped, via its \( k \)th canonical matricization, into a \((\prod_{i=1}^k m_i) \times (\prod_{i=k+1}^p m_i)\)-matrix \( X_{(k)} \), and into a vector \( \text{vec}(X) \) of \( \prod_{i=1}^p m_i \) elements using vectorization – see (26), (27) and (28) in Section A.1 for formal definitions. The \( k \)th mode product between \( X \) and the \( n_k \times m_k \) matrix \( A_k \) multiplies \( A_k \) to the \( k \)th mode of \( X \), resulting in the tensor \( X \times_k A_k \in \mathbb{R}^{m_1 \times \cdots \times m_{k-1} \times m_k \times m_{k+1} \times \cdots \times m_p} \). Applying the \( k \)th mode product with respect to \( A_k \) to every mode of \( X \) results in the Tucker (TK) product \([X; A_1, \ldots, A_p] \in \mathbb{R}^{r \times r \times \cdots \times r} \) [40]. The inner product between two equal-sized tensors \( X \) and \( Y \) is defined as \( \langle \text{vec}(X) \rangle^T \langle \text{vec}(Y) \rangle \) and denoted as \( \langle X, Y \rangle \). If \( B \) is a \((p + q)\)-way tensor of size \( m_1 \times m_2 \times \cdots \times m_p \times h_1 \times h_2 \times \cdots \times h_q \), then the partial contraction [46] between \( B \) and \( X \) denoted as \( \langle X | B \rangle \) is a \( q \)-way tensor of size \( h_1 \times h_2 \times \cdots \times h_q \) with elements

\[
\langle X | B \rangle(j_1, \ldots, j_q) = \sum_{i_1, \ldots, i_p} X(i_1, \ldots, i_p) B(i_1, \ldots, i_p, j_1, \ldots, j_q).
\]

We refer to Section A.1 for the reshaping and product definitions encountered in this section.

A random tensor \( X \) is a tensor whose vectorized form \( \text{vec}(X) \) is a random vector. In many cases \( \text{Var} \{ \text{vec}(X) \} = \sigma^2 \otimes_{k=p} \Sigma_k \) for \( \Sigma_k \in \mathbb{R}^{m_k \times m_k} \), where \( \otimes \) denotes the Kronecker product of Equation (25). In these cases the squared Mahalanobis distance (with respect to the scale matrices \( \sigma^2 \Sigma_1, \ldots, \Sigma_p \)) of \( X \) from its expectation \( \mathbb{E}(X) = M \) is

\[
D^2_{\sigma^2 \Sigma}(X, M) = \frac{1}{\sigma^2} \langle X - M; [X - M; \Sigma_1^{-1}, \ldots, \Sigma_p^{-1}] \rangle.
\]

See [74] for other possible Kronecker structures.

A random vector \( X \) has a spherical distribution if its probability density function (PDF) is invariant to rotations: that is, for every \( \Gamma \in O(\mathbf{h}) \) = \{ \( H \in \mathbb{R}^{h \times h} : HH^T = I_h \} \), we have \( X \overset{d}{=} \Gamma X \), or equivalently \( \mathcal{L}(X) = \mathcal{L}(\Gamma X) \), with \( \overset{d}{=} \)
denoting distributional equality and $\mathcal{L}(X)$ denoting the law of $X$. This notion of sphericity can be extended to a random matrix $X$ if it is invariant to rotations of its rows and/or columns [21]. Also, a random tensor $X \in \mathbb{R}^{k_1 \times \ldots \times k_p}$ follows a tensor-valued spherical distribution if $\text{vec}(X)$ follows a vector-valued spherical distribution. Thus $\text{vec}(X) \sim \Gamma \text{vec}(\mathbf{y})$ for any $\Gamma \in O(\prod_{i=1}^{p} k_i)$, implying that the distribution of $X$ is invariant under the group of transformations $\mathcal{G} = \{ \phi_{\Gamma} : \Gamma \in O(\prod_{i=1}^{p} k_i) \}$, where $\phi_{\Gamma}(X) = \Gamma \text{vec}(X)$ is the matrix product of $\text{vec}(X)$ with the orthogonal matrix $\Gamma$. Since $\langle X, X \rangle$ is maximally invariant under $\mathcal{G}$ (cf. Example 2.11 of [20]), it follows that the PDF $f(\cdot)$ and characteristic function $\psi(\cdot)$ have the form

$$
\psi_{\mathbf{X}}(\mathbf{Z}) = \phi(\langle \mathbf{Z}, \mathbf{Z} \rangle), \quad f_{\mathbf{X}}(\mathbf{X}) = g(\langle \mathbf{X}, \mathbf{X} \rangle)
$$

for some functions $\phi$ and $g$, respectively called the characteristic generator and the probability density generator of $\mathbf{X}$. We write $X \sim S_{b, \ldots, b}(\langle \cdot \rangle)$ if $X$ has the characteristic function in (3). We define EC-distributed random tensors from spherical distributions through the TK product. In the following, we will define EC distributions that are nonsingular, and have positive definite scale matrices.

**Definition 1.** Suppose that $M \in \mathbb{R}^{k \times m \times m}$, and $Q_k \in \mathbb{R}^{m \times h_k}$ are matrices such that $Q_k Q_k^T = \Sigma_k$ is positive definite for all $k = 1, \ldots, p$. Then if

$$
Y \overset{d}{=} M + [X; Q_1, \ldots, Q_p],
$$

(4)

for some $X \sim S_{b, \ldots, b}(\langle \cdot \rangle)$, we say that $Y$ has an EC distribution with mean $M$, scale matrices $\Sigma_1, \Sigma_2, \ldots, \Sigma_p$, and characteristic generator $\phi$, written as $Y \sim EC_m(M, \Sigma_1, \Sigma_2, \ldots, \Sigma_p, \phi)$, where $m = (m_1, m_2, \ldots, m_p)^T$.

For the remainder of this article we will use $m = (m_1, \ldots, m_p)$, $m = \prod_{i=1}^{p} m_i$, $m = \frac{m}{m_k}$, $\Sigma = \otimes_{i=p}^{1} \Sigma_i, \Sigma_{-k} = \otimes_{i=p, i \neq k} \Sigma_i$. For $\phi(x) = \exp(-x/2)$, $Y$ follows the TVN distribution $\mathcal{N}_m(M, \Sigma_1, \ldots, \Sigma_p)$ with PDF

$$
f_{\mathbf{Y}}(\mathbf{Y}) = \left[ 2\pi \left( \prod_{k=p}^{1} \Sigma_k \right)^{-\frac{1}{2}} \right] \exp \left( -\frac{1}{2} D_2(X, M) \right).
$$

(5)

Scale mixtures of TVN are an important sub-family of EC distributions, and are defined as

$$
\mathcal{L}(\mathbf{Y}|Z = z) = \mathcal{N}_m(M, z^{-1} \Sigma_1, \Sigma_2, \ldots, \Sigma_p)
$$

(6)

for some non-negative random variable $Z$, which needs to be multiplied to any one $\Sigma_k$, so we choose $\Sigma_1$ to be this $\Sigma_k$. This distribution is the tensor-variate extension [6] of the vector-variate case studied in Andrews and Mallows [5], Chu [13], Yao [76] and the matrix-valued case investigated in Gupta and Nagar [28], Gupta and Varga [29]. A commonly-used mixing distribution in (6), for the vector-variate case, uses $Z \sim \text{Gamma}(a/2, b/2)$: doing so in the tensor-variate case yields the PDF

$$
f_{\mathbf{Y}}(\mathbf{Y}) = \left[ \pi b \left( \prod_{k=p}^{1} \Sigma_k \right)^{-\frac{1}{2}} \frac{1}{\Gamma((m + a)/2)} \right] \left( 1 + \frac{D_2(X, M)}{b} \right)^{-\frac{m+a}{2}}
$$

(7)

that corresponds to the TV-$t$ distribution with $v$ degrees of freedom when $a = b = v$, the tensor-variate Cauchy distribution when additionally $v = 1$, and a tensor-variate Pearson Type VII distribution (cf. Page 450 of [60]) with parameter $q$ when $a = m$ and $b = q$. Recall in Equation (7) that $D_2(X, M)$ denotes the squared Mahalanobis distance of Equation (2).

We now derive some useful properties of the EC tensor-variate distributions.
2.2 Characterizing the EC family of TV distributions

Defining EC distributions in terms of (4) allows us to extend results from the spherical family of distributions to its EC counterpart. For instance, from (3) and (4), we can write the characteristic function of \( \mathbf{y} \) in terms of its characteristic generator \( \varphi \) as

\[
\psi_\mathbf{y}(\mathbf{z}) = \exp(i \langle \mathbf{z}, \mathbf{M} \rangle) \varphi(\langle \mathbf{z}, [\Sigma_1, \Sigma_2, \ldots, \Sigma_p] \rangle).
\]

(8)

See (30) for a detailed derivation of (8).

Similarly, if \( \mathbf{x} \sim \mathcal{S}_{h_1, h_2, \ldots, h_p}(\varphi) \) has the PDF in (3), then the transformation induced in (4) has Jacobian determinant \( |\bigotimes_{k=p}^1 \Sigma_k|^{-1/2} \), so that if \( \mathbf{y} \) possesses a density, it must be of the form

\[
f_\mathbf{y}(\mathbf{y}) = \left| \bigotimes_{k=p}^1 \Sigma_k \right|^{-1/2} g(D^2_\mathbf{y}(\mathbf{y}, \mathbf{M})),
\]

where \( D^2(\cdot) \) is the squared Mahalanobis distance of (2). Table 1 defines some EC distributions through their density generators \( g(\cdot) \) (see [7, 25] for additional specifications).

Table 1. Some common EC distributions, defined using (9) with their corresponding probability density generators \( g(\cdot) \), given up to their constant of proportionality.

| Distribution       | Additional parameters | \( g(\chi) \)               |
|--------------------|-----------------------|-----------------------------|
| Normal             | –                     | \( \exp(-\chi/2) \)         |
| Student’s-t        | \( q > 0 \)           | \( (1 + q^{-1}x)^{-q(m+1)/2} \) |
| Pearson Type VII   | \( q > 0 \)           | \( (1 + x/q)^{-m} \)         |
| Kotz Type          | \( q > 0 \)           | \( x^{m-1} \exp(-qx) \)     |
| Logistic           | –                     | \( \exp(-x)/(1 + \exp(-x))^2 \) |
| Power exponential  | \( q > 0 \)           | \( \exp(-x^2/2) \)          |

The following theorem allows us to express EC tensor-variate distributions in terms of vector-variate EC distributions that have been studied extensively in the literature.

**Theorem 2.** With the reshappings of (26), (27) and (28), and for all \( k = 1, 2, \ldots, p \), the following statements are equivalent, with \( n_k = \prod_{i=1}^k m_i \):

1. \( \mathbf{y} \sim \mathcal{E}_{C_k}(\mathbf{M}, \Sigma_1, \Sigma_2, \ldots, \Sigma_p, \varphi) \).
2. \( \mathbf{y}_{(k)} \sim \mathcal{E}_{C_k, m_k}(\mathbf{M}_{(k)}, \Sigma_k, \Sigma_{-k}, \varphi) \).
3. \( \mathbf{y}_{<k> \sim} \mathcal{E}_{C_{<k>}}(\mathcal{M}_{<k>}, \bigotimes_{i=k}^1 \Sigma_i, \bigotimes_{i=p}^{k+1} \Sigma_i, \varphi) \).
4. \( \text{vec}(\mathbf{y}) \sim \mathcal{E}_{C_{m}}(\text{vec}(\mathbf{M}), \Sigma, \varphi) \).

**Proof.** See Section B.1. \( \square \)

Spherical distributions derive their name from the observation that if \( \mathbf{x} \sim \mathcal{S}_{h_1, h_2, \ldots, h_p}(\varphi) \), then \( \mathbf{x} = R \mathbf{u} \), where the magnitude \( R = \| \mathbf{x} \| \) is independent of \( \mathbf{u} \), and \( \text{vec}(\mathbf{u}) \) is uniformly distributed on the surface of the \( \prod_{i=1}^p h_k \)-dimensional unit sphere [10, 39, 64]. Based on the characteristic function (3) and the independence of \( R \) and \( \mathbf{u} \), \( \psi_\mathbf{x}(\mathbf{z}) = \mathbb{E}\{\psi_\mathbf{u}(R\mathbf{z})\} \). Furthermore, since \( \mathbf{u} \) is also spherically distributed, the characteristic generator of \( \mathbf{x} \) can be written, in terms of the characteristic generator of \( \mathbf{u} \), as

\[
\varphi(\mathbf{u}) = \mathbb{E}\{\varphi_\mathbf{u}(R^2\mathbf{u})\}.
\]

(10)
The distribution of $R$ determines the distribution of $\mathbf{X}$. For instance, if $R^2 \sim \chi^2_h$ where $h = \prod_{k=1}^p h_k$, then $\phi(u) = \exp(-u/2)$ and $\mathbf{X}$ follows a TVN distribution. Further, if $\mathbf{Y} \sim EC_m(0, \sigma^2 \Sigma_1, \ldots, \Sigma_p, \varphi)$ as in (4) and $\mathbf{X} = R^T \mathbf{U}$ as above, then $Z = \mathbf{Y} / ||\mathbf{Y}||$ does not depend on $R$, meaning that the distribution of $Z$ does not depend on the original EC distribution of $\mathbf{Y}$. The distribution of $Z = \text{vec}(Z)$ is called the elliptical angular distribution [51, 75], and is denoted as $Z \sim AC_m(\Sigma)$, with PDF
\[
 f_Z(z) = \frac{\Gamma(m/2)}{2^{m/2} \pi^{m/2}} |\Sigma|^{-1/2} (z^T \Sigma^{-1} z)^{-m/2}.
\]
The above PDF does not depend on $\sigma^2$. In Section 3.5, we derive a robust tensor-variate Tyler estimator that exploits the fact that $Z$ is the same for any EC distribution.

### 2.3 Marginal and conditional distributions

Our next theorem shows that the TK product of an EC-distributed random tensor is also EC-distributed.

**Theorem 3.** Let $\mathbf{Y} \sim EC_m(M, \Sigma_1, \ldots, \Sigma_p, \varphi)$, $n = [n_1, \ldots, n_p]$ and $A_k \in \mathbb{R}^{n_k \times m_k}$, $\forall k = 1, \ldots, p$. Then
\[
 [([\mathbf{Y}; A_1, \ldots, A_p]) \sim EC([M; A_1, \ldots, A_p], A_1 \Sigma_1 A_1^T, \ldots, A_p \Sigma_p A_p^T, \varphi).]
\]

**Proof.** See Section B.2. $\square$

As a corollary, the marginal distributions follow from Theorem 3 for appropriate choices of $A_1, \ldots, A_p$, as detailed in Section A.2. We now derive conditional distributions for tensor-variate EC random tensors.

**Theorem 4.** Suppose that $\mathbf{Y} \sim EC_m(M, \Sigma_1, \ldots, \Sigma_p, \varphi)$, where $m = n_1 + \ldots + n_p > 1$ for some $n_1, \ldots, n_p \in \mathbb{N}$, and $\Sigma_p$ is a $2 \times 2$ block matrix with $(i, j)$th block $\Sigma_{ij} \in \mathbb{R}^{n_i \times n_j}$. Partition $\mathbf{Y}$ and $M$ over the $p$th mode with subtensors $\mathbf{Y}_1, M_1 \in \mathbb{R}^{m_1 \times \ldots \times m_{p-1} \times n_1}$ and $\mathbf{Y}_2, M_2 \in \mathbb{R}^{m_2 \times \ldots \times m_p \times n_p}$. Then
\[
 \mathbf{Y}_1 | (\mathbf{Y}_2 = \mathbf{y}_2) \sim EC_m(M_{1(1)}(2), \Sigma_1, \ldots, \Sigma_{p-1}, \Sigma_{p,11}, \varphi_q(y_2)),
\]

where $M_{1(1)}(2) = M_1 + (\mathbf{y}_2 - \mathbf{M}_2) \times_p (\Sigma_{p,12} \Sigma_{p,12}^{-1})$, $\Sigma_2 = \Sigma_{p,11}$, $\varphi_q(y_2)(u) = \mathbb{E} [R_{y_2, y_2}(u) | \mathbf{y}_2 = \mathbf{y}_2]$, $R_{y_2, y_2} = D_{\Sigma_2}^T (\mathbf{Y}_1, M_1)$, $m_1 = [m_1, \ldots, m_{p-1}, n_1]$, and $\Sigma_{p,12} = \Sigma_{p,11} - \Sigma_{p,12} \Sigma_{p,22}^{-1} \Sigma_{p,22}$.

**Proof.** See Section B.3. $\square$

The characteristic generator $\varphi_q(y_2)$ in Theorem 4 is a conditional moment, just like in the more general case of (10). Although Theorem 4 applies only for the conditional distribution that results from partitioning the last mode of the random tensor, we can find conditional distributions of any subtensor by applying Theorem 4 multiple times, as demonstrated in Section A.4.

### 2.4 Moments

The moments of EC tensor-variate distributions are found by differentiating (8). We provide the first four moments in Theorem 5 and then use them in Corollary 6 to find moments of other special forms.

**Theorem 5.** Suppose $\mathbf{Y} \sim EC_m(M, \Sigma_1, \ldots, \Sigma_p, \varphi)$, and let $i = (i_1, \ldots, i_p)^T$, $j = (j_1, \ldots, j_p)^T$, $k = (k_1, \ldots, k_p)^T$ and $l = (l_1, \ldots, l_p)^T$ be sets of indices such that $i_q, j_q, k_q, l_q \in \{1, \ldots, m_q\}$ for $q = 1, \ldots, p$. Further, denote $\mathbf{Y}((i_1, \ldots, i_p) = \mathbf{Y}_i$, $M((i_1, \ldots, i_p) = m_i$ and $\sigma_{ij} = \prod_{q=1}^p \Sigma_q(i_q, j_q)$. Then
\[
\begin{align*}
(1) \quad & \mathbb{E}(Y_i) = m_i, \\
(2) \quad & \mathbb{E}(Y_i Y_j) = m_i m_j - 2 \varphi(0) \sigma_{ij}, \\
(3) \quad & \mathbb{E}(Y_i Y_j Y_k) = m_i m_j m_k - 2 \varphi(0) (m_i \sigma_{kj} + m_j \sigma_{ik} + m_k \sigma_{ij}).
\end{align*}
\]
Suppose that we observe independent identically distributed realizations of the TVN distribution, and show how these estimation algorithms can extend the ToTR framework of [46] to the case where the algorithms for finding MLEs from independent identically distributed realizations of a scale TVN mixture distribution. Next, we provide expectation conditional maximization (ECM) [17, 55] and ECM Either (ECME) [43] algorithms. This section derives MLEs from EC tensor-variate distributed data under four different scenarios. We first derive these estimators from uncorrelated draws that have common and joint EC tensor-variate distributions. This allows us to derive several moments of EC tensor-variate distributions. A few derivations follow.

\[ \mathbb{E}(Y_i Y_j Y_k Y_l) = m_1 m_2 m_3 m_4 + 4\phi''(0) \left( \sigma_{kl} \sigma_{ij} + \sigma_{ij} \sigma_{kl} + \sigma_{il} \sigma_{jk} - 2\phi'(0) \left( m_1 m_2 \sigma_{ij} + m_2 m_3 \sigma_{il} + m_3 m_4 \sigma_{jk} \right) \right). \]

Here, the \( k \)th statement of the theorem requires \( \phi^{(k)}(0) < \infty \), for \( k = 1, 2, 3, 4 \).

**Proof.** See Section B.4.

Theorem 5 allows us to derive several moments of EC tensor-variate distributions. A few derivations follow.

**Corollary 6.** Let \( \mathcal{Y} \sim EC_m(\mathcal{M}, \Sigma_1, \Sigma_2, \ldots, \Sigma_p, \varphi) \), where \( \varphi''(0) < \infty \). Further, let \( A_k \) and \( B_k \) be \( n_k \times m_k \) and \( h_k \times m_k \) matrices for \( k = 1, 2, \ldots, p \), and define \( D(A)(X) = \langle X, [[X; A_1, A_2, \ldots, A_p]] \rangle \) whenever each \( A_k \) is a square matrix. Then,

1. \( \mathbb{E}(\mathcal{Y}) = \mathcal{M} \).
2. \( \text{Var}(\text{vec}(\mathcal{Y})) = -2\phi'(0) \sum_{k=1}^{p} \Sigma_k \).
3. If \( n_k = m_k \) for all \( k = 1, 2, \ldots, p \), then
   \[ \mathbb{E}\{D(A)(\mathcal{Y})\} = D(A)(\mathcal{M}) - 2\phi'(0) \left[ \prod_{k=1}^{p} \text{tr}(\Sigma_k A_k^\top) \right]. \]
4. If \( \mathcal{V} \) is of size \( n_1 \times n_2 \times \ldots \times n_p \), then
   \[ \mathbb{E}(\langle \mathcal{V}, [[\mathcal{Y}; A_1, \ldots, A_p]] \rangle, \mathcal{Y}) = \langle \mathcal{V}, [[\mathcal{M}; A_1, \ldots, A_p]] \rangle M - 2\phi'(0) \left[ \prod_{k=1}^{p} \text{tr}(\Sigma_k A_k^\top) \right]. \]
5. If \( n_k = m_k \) for all \( k = 1, 2, \ldots, p \), then
   \[ \mathbb{E}\{D(A)(\mathcal{Y})D(B)(\mathcal{Y})\} = 4\phi''(0) \left[ \prod_{k=1}^{p} \text{tr}(A_k \Sigma_k B_k \Sigma_k) \right] + \left[ \prod_{k=1}^{p} \text{tr}(A_k \Sigma_k B_k \Sigma_k) \right]. \]

**Proof.** See Section B.5.

From Parts 3 and 6 of Corollary 6, for \( \mathcal{Y} \) specified as in Definition 1, \( \mathbb{E}\{D(A)(\mathcal{Y}, \mathcal{M})\} = -2\phi'(0) \times m_1 \) and \( \text{Var}(D(A)(\mathcal{Y}, \mathcal{M})) = 4 \times [\phi''(0) \times (m_1^2 + 2m_1) - (\phi'(0) \times m_1^2)]. \)

### 3 Maximum Likelihood Estimation

This section derives MLEs from EC tensor-variate distributed data under four different scenarios. We first derive these estimators from uncorrelated draws that have common and joint EC tensor-variate distributions. This procedure depends on identifiable MLEs from the TVN distribution, and so we also describe how to obtain such estimates. Next, we provide expectation conditional maximization (ECM) [17, 55] and ECM Either (ECME) [43] algorithms for finding MLEs from independent identically distributed realizations of a scale TVN mixture distribution, and show how these estimation algorithms can extend the ToTR framework of [46] to the case where the error distribution is a scale mixture of TVNs. Finally, we derive a robust estimation algorithm for the scale matrices in the spirit of Tyler’s estimation [70–72] for unknown EC distributions.

### 3.1 Estimation in the TVN model

Suppose that we observe independent identically distributed realizations \( \mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_p \) from the \( \mathcal{N}_m(0, \varphi \Sigma_1, \Sigma_2, \ldots, \Sigma_p) \) distribution. We assume without loss of generality here that the mean is zero because its estimate is...
the sample mean and is separable from the scale estimation process. The loglikelihood is
\[ \ell(\theta) = -\frac{n}{2} \log |\sigma^2 \Sigma| - \frac{1}{2\sigma^2} \sum_{i=1}^{n} D_{ik}(Y_i, 0). \] (12)
The scale matrices \( \Sigma_1, \Sigma_2, \ldots, \Sigma_p \) are not identifiable since \( \Sigma_k \otimes (a\Sigma_k) = (a\Sigma_k) \otimes \Sigma_l \) for any \( a \in \mathbb{R} \). Moreover, each \( \tilde{\Sigma}_k \) has no closed-form solution but depends on the other \( \tilde{\Sigma}_k \)'s as \( \tilde{\Sigma}_k = \Sigma_k/(nm_k\sigma^2) \), where \( \Sigma_k = \sum_{i=1}^{n} Y_{ik}(\Sigma_{-k}^{-1} Y_{ik}') \), with \( \Sigma_{-k} \) as defined in the paragraph following Definition 1. The existing methodology [18, 19, 63, 67] accounts for the indeterminacy of the \( \tilde{\Sigma}_k \)'s by scaling each \( \tilde{\Sigma}_k \) such that the trace, determinant or (1, 1)th element is unity. Instead, we use the ADJUST procedure of Glanz and Carvalho [27] to optimize the loglikelihood with respect to \( \Sigma_k \) under the constraint \( \Sigma_k(1, 1) = 1 \). The ADJUST procedure simply solves the constrained optimization problem
\[ \text{ADJUST}(n, \sigma^2, S) = \arg \max_{\Sigma(1,1)=1} \left\{ \frac{n}{2} \log |\Sigma^{-1}| - \frac{1}{2\sigma^2} \text{tr}(\Sigma^{-1}S) \right\}. \] (13)
The exact procedure in our case is outlined in greater detail in Section B.7. From our estimation procedure therefore, we obtain the estimator \( \hat{\Sigma}_k = \text{ADJUST}(nm_k, \sigma^2, \Sigma_k) \), with \( \sigma^2 \) inexpensively estimated after obtaining \( \hat{\Sigma}_k \), as \( \hat{\sigma}^2 = \text{tr}(\Sigma_{-k}^{-1})/(nm) \). These minor modifications to the MLE algorithm of Akdemir and Gupta [2], Hoff [32], Manceur and Dutilleul [50], Ohlson et al. [57] will be used in the following sections.

3.2 Estimation from uncorrelated EC data
Consider the \((m_1 \times m_2 \times \cdots \times m_p \times n)\)-sized EC-distributed tensor
\[ \mathbf{Y} \sim EC_{[m,n]}(M_n, \sigma^2 \Sigma_1, \Sigma_2, \ldots, \Sigma_p, I_n, \phi), \] (14)
where \( \Sigma_k \) is constrained to have \( \Sigma_k(1, 1) = 1 \) for all \( k = 1, 2, \ldots, p \), and \( \sigma^2 \) captures the overall proportionality of \( \Sigma \), while \( M_n \) contains \( n \) copies of the \( m_1 \times m_2 \times \cdots \times m_p \) tensor \( M \) (that is, \( M_n \times_{p+1} e_i^T = M \) for all \( i = 1, 2, \ldots, n \), where \( e_i^T \in \mathbb{R}^n \) is a unit basis vector with 1 as the \( i \)th element and 0 elsewhere). In the next theorem, we first show that the \( n \) sub-tensors \( \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n \) of \( \mathbf{Y} \) (where \( \mathbf{Y}_i = \mathbf{Y} \times_{p+1} e_i^T \)) are uncorrelated and identically distributed. We then show how to obtain MLEs from such a sample, in a tensor-variate extension of the theorem of Anderson et al. [4].

**Theorem 7.** Let \( \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n \) jointly follow the distribution in (14). Then, we have
1. \( \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n \) are identically distributed as \( EC_{m_n}(M_n, \sigma^2 \Sigma_1, \Sigma_2, \ldots, \Sigma_p, \phi) \), and are mutually uncorrelated, since \( \forall i \neq j, \mathbb{E} \left( \text{vec}(\mathbf{Y}_i - M_n) \text{vec}(\mathbf{Y}_j - M_n)^T \right) = 0 \).
2. Suppose that the density generator \( g(\cdot) \) of \( \mathbf{Y} \) is such that \( h(d) = d^{nm/2}g(d) \) has a finite positive maximum \( d_g \). Also, let \( (\hat{\sigma}^2, \hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_p) \) be the TVN MLEs described in Section 3.1. Then the MLE for \((\sigma^2, \Sigma_1, \Sigma_2, \ldots, \Sigma_p)\) is \((\hat{d}_g, \hat{\sigma}^2, \hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_p)\).

**Proof.** See Section B.6.

It follows that \( d_g = mn \) in the TVN case, \( d_g = nmb/a \) when \( g(\cdot) \) is of the form (7), \( d_g = (mn/s)^{1/s} \) in the power exponential case, and \( d_g \) is the solution to \( mn/(2d) = \tanh (d/2) \) in the logistic case [22, 25].

3.3 Estimation from TVN scale mixture realizations
Suppose that we observe \( n \) independent identically distributed realizations \( \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n \) from the scale mixture distribution of (6). The complete-data loglikelihood is
\[ \ell_c(\theta) = -\frac{n}{2} \log |\sigma^2 \Sigma| - \frac{1}{2\sigma^2} \sum_{i=1}^{n} z_i D_{ik}(Y_i, M) \] (15)
where $\theta_2 = (\sigma^2, (\text{vec } \Sigma_1)^T, (\text{vec } \Sigma_2)^T, \ldots, (\text{vec } \Sigma_p)^T)^T$ and $\theta = ((\text{vec } M)^T, \theta_1^T)^T$. We now use $\ell_t$ to propose an EM algorithm with the following expectation (E) and conditional maximization (CM) steps:

**E-step:** Here, we obtain the conditional expectation of (15), given the observed data and evaluated at the $t$th parameter iteration $\theta^{(t)}$, to get

$$Q(\theta; \theta^{(t)}) = -\frac{n}{2} \log |\sigma^2 \Sigma| - \frac{1}{2\sigma^2} \sum_{i=1}^n z_i^{(t)} D_{\mathbb{Z}_i}^2 (Y_i, M),$$

where $z_i^{(t)} = \mathbb{E}(Z_i | Y_i = ; Y_i; \theta^{(t)})$. This conditional expectation depends on the distribution of $Z_i$. Whenever $Z_i \sim \text{Gamma}(a/2, b/2)$, $Y_i$ has PDF given by (7) and then $z_i^{(t)} = (m+a)/(D_{\mathbb{Z}_i}^2 (Y_i, M)^{t} + b)$.

**CM-steps:** The CM steps maximize (16), but with respect to subsets of parameters. The first CM block maximizes (16) with respect to $M$ and is obtained from its total differential of $M$ as

$$\partial Q(M) = (\sigma^2)^{-1} \partial M \cdot [S; \Sigma_1^{-1}, \Sigma_2^{-1}, \ldots, \Sigma_p^{-1}])],$$

where $S = \sum_{i=1}^n z_i^{(t)} (M \sum_{i=1}^n z_i^{(t)}$). Hence, $\partial Q(M) = 0$ implies $S = 0$, or $\hat{M}^{(t+1)} = (\sum_{i=1}^n z_i^{(t)} Y_i)/(\sum_{i=1}^n z_i^{(t)})$. Setting $M = \hat{M}^{(t+1)}$ in (16) profiling out $M$ and yields the TVN loglikelihood (12), but with $Y_i = \sum_{i=1}^n z_i^{(t)} (Y_i - \hat{M}^{(t+1)})$ instead of $Y_i$. Hence, the next $(p+1)$ CM block (corresponding to $\sigma^2, \Sigma_1, \Sigma_2, \ldots, \Sigma_p$) are obtained similarly as for the TVN model of Section 3.1. Specifically, $\hat{X}_k = \text{ADJUST}(nm, k, \sigma^2, \Sigma_k)$, followed by $\hat{\sigma}^2 = \text{tr}(S_k \Sigma_k^{-1})/(nm)$, where $S_k = \sum_{i=1}^n Y_i \Sigma_k^{-1} Y_i$. We initialize the algorithm by running the TVN model of Section 3.1 for a few iterations (taken to be ten in all the experiments and applications in this paper). The resulting ECM algorithm is summarized in the following steps:

1. **Initialization:** Set initial values and $t = 0$.
2. **E-step:** Find $z_i^{(t)} = \mathbb{E}(Z_i | Y_i = ; Y_i; \theta^{(t)})$ for $i = 1, \ldots, n$.
3. **CM-Step 1:** Find $\hat{M}^{(t+1)} = (\sum_{i=1}^n z_i^{(t)} Y_i)/(\sum_{i=1}^n z_i^{(t)}$).
4. **CM-Step 2:** For each $k = 1, 2, \ldots, p$, estimate $\hat{X}_k$.
5. **CM-Step 3:** Estimate $\hat{\sigma}^2$.
6. **Iterate or stop:** Return to the E-step and set $t \leftarrow t + 1$, or stop if convergence is met.

**Alternative EM approaches for estimating $\sigma^2$:** The above algorithm is an ECM algorithm, since it estimates $\sigma^2$ once for each iteration. A faster alternative is to estimate $\sigma^2$ right after each $\hat{X}_k$, making our algorithm an AECM algorithm [54]. An even faster alternative estimates $\sigma^2$ directly from the loglikelihood $\ell(\sigma^2) = -mn \log(\sigma^2)/2 + \sum_{i=1}^n \alpha_i \log g(D_{\mathbb{Z}_i}^2 (Y_i, M)^{t})/\sigma^2)$ making the algorithm an ECME algorithm [43]. Our experiments showed our ECME algorithm to be considerably more computationally efficient for heavy-tailed EC distributions.

In some settings, such as when we need to estimate the degrees of freedom parameter in the $t$ distribution, the mixing random variable $Z$ in (6) has extra parameters that show up in the last term of (15), and may be estimated in additional CM steps [44, 53]. These extra parameters can also be optimized directly from the loglikelihood as an Either step [43]. For the TV-$t$ distribution with an unknown degrees of freedom $v$ (the PDF of the TV-$t$ distribution is provided in (7) whenever $a = b = v$), we estimate $v$ as the solution to

$$0 = 1 + \frac{1}{n} \sum_{j=1}^n \left[ \log z_j^{(k+1)} (v) - z_j^{(k+1)} (v) \right] - \psi \left( \frac{v}{2} \right) + \psi \left( \frac{v + m}{2} \right) - \log \left( \frac{n + m}{2} \right)$$

(17)
at each iteration, where $\hat{z}_{ij}^{(k+1)}(\nu) = \frac{\nu^v}{\nu^v + m} e^{\nu^v + m} Y^{(t)}_{ij} M_{ij}^{(t)}$, and $\psi$ is the digamma function. We thus extend the vector-variate results of Liu and Rubin [44] to the tensor-variate case.

3.4 ToTR with scale TVN mixture errors

We extend the algorithm in Section 3.3 to when $Y_i, \ldots, Y_n$ are independent realizations of the ToTR model

$$Y_i = (X_i | B) + E_i, \quad i = 1, 2, \ldots, n,$$

where $X_i$ is the $i$th tensor-variate regression covariate, $E_i$ is the $i$th regression random error following the scale mixture distribution of (6) for $M = 0$, $B$ is the tensor-variate regression coefficient, and $\langle \cdot | \cdot \rangle$ denotes partial contraction as per (1). We assume that $X_i \in \mathbb{R}^{h_1 \times h_2 \times \cdots \times h_l}$ and $Y_i \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_R}$, which means that $B \in \mathbb{R}^{h_1 \times h_2 \times \cdots \times h_l \times m_1 \times m_2 \times \cdots \times m_R}$. ToTR was proposed by Hoff [33] under TVN errors and outer matrix product (OP) factorization of $B$, by Lock [47] under canonical polyadic or CANDECOMP/PARAFAC (CP) [11, 31, 79] factorization of $B$, by [24] under the TK factorization, and by Liu et al. [45] under the TT format [58] that is the tensor-ring (TR) format [78] when one of the TR ranks is set to 1. The CP, TR, TK and OP factorizations on $B$ allow for substantial parameter reduction without affecting prediction or discrimination ability of the regression model. For more recent developments in ToTR see [49, 62]. The more general case of ToTR can be found under the CP, OP, TR and TK formats and TVN errors in Llosa-Vite and Maitra [46], and here we extend those results to the case where the errors follow a scale mixture of TVN distribution.

Maximum Likelihood Estimation. Following the same steps as in Section 3.3, the E-step is performed by calculating the expected complete loglikelihood

$$Q(\theta; \theta^{(t)}) = -\frac{n}{2} \log |\Sigma| - \frac{1}{2\sigma^2} \sum_{i=1}^{n} D_2^c(Y_{wi}, \langle X_{wi}^{(t)} | B \rangle)$$

with $\theta = (B, \sigma^2, \Sigma)$, and $(Y_{wi}^{(t)}, X_{wi}^{(t)}) = (\sqrt{z_i^{(t)}} Y_i, \sqrt{z_i^{(t)}} X_i)$ for $z_i^{(t)} = \mathbb{E}(Z_i | Y_i; \theta^{(t)})$. For $Z_i \sim Gamma(a/2, b/2)$, we have $z_i^{(t)} = \frac{\sigma^2}{D_2^c(Y_{wi}, \langle X_{wi}^{(t)} | B \rangle)^{1/2}}$.

The CM-steps are performed by sequentially optimizing (19) with respect to the parameter factors involved in $\theta$. This equation is identical to the loglikelihood for the ToTR model under TVN errors (see Eqn. (16) of [46]), and so the CM-steps are performed in one iteration of their block-relaxation algorithm, but with $\{(Y_{wi}^{(t)}, X_{wi}^{(t)}), i = 1, 2, \ldots, n \}$ as the responses and covariates. A block-relaxation algorithm [16] partitions the parameter space into multiple blocks, and sequentially optimizes each block while fixing the remaining blocks. This approach is very general as it works for any of the low-rank formats studied in Llosa-Vite and Maitra [46].

EM algorithm: The ECM algorithm iterates the E- and CM-steps until convergence. The low-rank structure of $B$ will typically lead to a multi-modal loglikelihood, and hence our EM algorithms will converge to a local maxima. Multi-modal objective functions are ubiquitous across most optimization algorithms that solve low-rank tensor-regression models, and tensor decompositions. We determine convergence based on the relative difference in $|B| + |\Sigma|^2$. We initialize $\Sigma_1, \ldots, \Sigma_p$ by fitting the to the least-squares residuals, the TVN algorithm of Section 3.1 for a few iterations (10 in our experiments). The ECM algorithm for maximum likelihood estimation is summarized as follows:

1. **Initialization**: Set initial values and $t = 0$.
2. **E-Step**: Find $\hat{z}_{i}^{(t)} := \mathbb{E}(Z_i | Y_i = Y_i; \theta^{(t)})$ and $(Y_{wi}^{(t)}, X_{wi}^{(t)}) \forall i = 1, 2, \ldots, n$.
3. **CM-Step**: Perform one iteration of the respective ToTR algorithm of Llosa-Vite and Maitra [46], but with $\{(Y_{wi}^{(t)}, X_{wi}^{(t)}), i = 1, 2, \ldots, n \}$ as the responses and covariates.
4. **Iterate or stop**: Return to the E-step and set $t \leftarrow t + 1$, or stop if convergence is met.
Section 5.1 shows that the proposed ToTR framework with EC errors outperforms ToTR with TVN errors in terms of classification performance on color bitmap images of cats and dogs from the AFHQ database.

3.5 Robust tensor-variate Tyler’s estimator

While we proposed several methods for maximum likelihood estimation, they all require the form of the underlying EC distribution to be known. In this section, we derive constrained MLEs that are robust to the type of EC distribution in the spirit of Tyler’s estimator [66, 70–72]. Consider $Y_i \sim EC_m(0, \sigma^2 \Sigma_1, \Sigma_2, \ldots, \Sigma_p, \varphi)$ for $i = 1, 2, \ldots, n$. As described in Section 2.2, if we let $Z_i = Y_i/\|Y_i\|$ for all $i = 1, 2, \ldots, n$, then vec$(Z_i) \sim AC_m(\Sigma)$, and does not depend on the underlying type of EC distribution. Using this result along with the PDF in (11) means that, regardless of $\varphi$, we can write the loglikelihood for $\Sigma_1, \Sigma_2, \ldots, \Sigma_p$ as

$$
\ell(\Sigma_1, \Sigma_2, \ldots, \Sigma_p) = -\frac{n}{2} \log \left( \prod_{k=p}^{1} \Sigma_k \right) - \frac{m}{2} \sum_{i=1}^{n} \log D_k^2(Z_i, 0). 
$$

We now optimize the loglikelihood function in (20) using a block-relaxation algorithm [16]. In this context, each block corresponds to $\Sigma_k$ under the constraint $\Sigma_k(1, 1) = 1$ for all $k = 1, 2, \ldots, p$, and is derived in the following theorem.

**Theorem 8.** For a fixed value of $\Sigma_{-k}$ and for $S_{ik} = Y_{i(k)} \Sigma_{-k}^{-1} Y_{i(k)}^\top$, a fixed-point algorithm for the maximum likelihood estimation of $\Sigma_k$ updates $\Sigma_{k,(t)}$, under the constraint that $\Sigma_k(1, 1) = 1$, to

$$
\Sigma_{k,(t+1)} = \text{ADJUST}\left( n/m_k, 1, \frac{\sum_{i=1}^{n} S_{ik}}{\text{tr}(\Sigma_{k,(t)}^{-1} S_{ik})} \right).
$$

**Proof.** See Section B.7. □

3.5.1 Estimation algorithm. For each $k$, the fixed-point iteration algorithm of Theorem 8 depends on $\Sigma_{-k}$ only through $S_{ik}$. Obtaining $S_{ik}$ for all $i = 1, 2, \ldots, n$ and for any $k = 1, 2, \ldots, p$ has $O(nmm)$ computational complexity (where $m_M = \max(m_1, m_2, \ldots, m_p)$) while one iteration of (21) is usually considerably cheaper with a complexity of at most $O(m_M^3)$. Therefore, in most cases we suggest performing $n_t$ iterations of (21) before going from $\Sigma_k$ to $\Sigma_{k+1}$. While all values of $n_t$ will eventually converge, we set $n_t = 12$ based on numerical experiments. The final estimation algorithm is:

1. **Initialization:** Fit the TVN model of Section 3.1.
2. For $k = 1, 2, \ldots, p$
   (a) Obtain $S_{ik}$ from Theorem 8 $\forall i = 1, \ldots, n$
   (b) Update $\Sigma_{k,(t+1)}$ as per (21) for $n_t$ iterations.
3. **Iterate or stop:** Stop if converged, or return to Step 2.

3.5.2 Robust estimation of the location parameter. When $M$ is also unknown, Tyler [71] and Maronna [52] proposed for the vector-variate case the fixed-point iteration algorithm that we extend to the tensor-variate case as

$$
\hat{M}_{(t+1)} = \frac{\sum_{i=1}^{n} X_i D_{\Sigma}^{-1}(X_i, \hat{M}_{(t)})}{\sum_{i=1}^{n} D_{\Sigma}^{-1}(X_i, \hat{M}_{(t)})}.
$$

Tyler [71] was unable to show the joint existence of $\Sigma$ and $M$ due to discontinuities in the objective function, and so we do not integrate this result into our estimation algorithm. However, it can be easily integrated to estimate $M$ at each iteration and then used to center the data before estimating $\Sigma_1, \Sigma_2, \ldots, \Sigma_p$. Although to our
knowledge, no theoretical guarantees exist for this approach even in the vector-variate case, we have found this algorithm to converge and produce stable estimates (see Figures 1 and 2).

3.6 Summary of algorithms presented in this section

In this section we presented five different algorithms for maximum likelihood estimation under various assumptions on the data. In Table 2 we compare these algorithms across the five different settings under which they are applicable.

| Table 2. Settings of the five algorithms we presented in this section |
|---------------------------------------------------------------|
| setting | section | 3.1 | 3.2 | 3.3 | 3.4 | 3.5 |
|------------------------|--------|-----|-----|-----|-----|-----|
| independent sample | ✓ | ✓ | ✓ | ✓ | ✓ |
| non-Gaussian tails | ✓ | ✓ | ✓ | ✓ | ✓ |
| unknown tail weight | ✓ | ✓ | ✓ | ✓ | ✓ |
| covariates/low-rank mean | ✓ | ✓ | ✓ | ✓ | ✓ |
| undefined EC distribution | ✓ | ✓ | ✓ | ✓ | ✓ |

First, recall that the algorithm of Section 3.2 is for an uncorrelated (but dependent) sample of any EC distribution with finite PDF, and that it is obtained from fitting the TVN model (which assumes independent Gaussian data) of Section 3.1 and scaling \( \hat{\sigma}^2 \). Hence, the algorithm of Section 3.2 trades the benefit of more flexible tail assumptions at the cost of dependence between the sample data. In Table 2, a benefit is named adaptive tail weights, which refers to the case where the algorithm adapts to the appropriate tail weight observed in the data. The algorithms of Sections 3.3 and 3.4 have the TV-\( t \) distribution as a special case, which adapts to the observed tail weights through the estimated degrees of freedom parameter \( \nu \). The algorithm of Section 3.5 also has adaptive tails, since it works for any EC distribution, which does not need to be specified. While the algorithms of Sections 3.3 and 3.4 have the same error assumptions, they differ in the mean structure. Indeed, the development of Section 3.4 allows for covariates in the mean structure, and a low-rank regression parameter. When \( X = 1 \), both algorithms learn from an independent and identically distributed sample, with the difference that the methods in Section 3.4 learns a low-rank tensor mean. If additionally \( n = 1 \), then the algorithm of Section 3.4 is decomposing a tensor under EC errors.

In the previous paragraph, we highlighted some differences between the algorithms in the previous sections, but they all share similarities, as they are all iterative algorithms that assume uncorrelated data. Moreover, if \( Z_i \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_p} \) for \( i = 1, 2, \ldots, n \), then all the algorithms are dominated by the operation \( Z_i(k) \Sigma^{-1} Z_i(k) ^\top \), which after computing for all \( (i, k) \), has a computational cost of \( O(nm \Sigma_m^p m_k) \). This is the same complexity of the TVN algorithm that we improve upon by improved modeling in this paper.

4 PERFORMANCE EVALUATIONS

This section evaluates performance of the maximum likelihood estimation procedures proposed in Section 3 by fitting them to realizations \( Y \) from the TVN and from gamma scale TVN mixture (GSM) distributions. We study the \( EC_m(M; \sigma^2 \Sigma_1, \ldots, \Sigma_6) \) distribution with PDF as in (7) for \((a, b) = (3, 15)\). This implies that \( \text{Var} \{ \text{vec}(Y) \} = 15 \sigma^2 \Sigma \), and hence we compare it against the \( N_m(M; 15 \sigma^2 \Sigma_1, \ldots, \Sigma_6) \) TVN distribution, with both sets of parameters chosen to be of comparable overall variability. We assume that \( i = 1, 2, \ldots, n = 100 \) and \( m = (m_1, m_2, m_3, m_4, m_5, m_6) = (7, 9, 3, 23, 7, 3) \), and compare the performance of different estimation algorithms at varying signal-to-noise ratios by keeping \( M \) fixed and setting \( \sigma \in \{2, 6, 10\} \).

For \( k = 1, 2, \ldots, 6 \) we generated \( \Sigma_k \) from the \( W_{m_k}(100 m_k, I_{m_k}) \) Wishart distribution before constraining its condition numbers to be at most 50 and scaling each entry by its \((1, 1)\)th element. The tensor-valued \( M \) was chosen
such that when rearranged in a 3-way tensor of size $(7 \times 9 \times 3) \times (23 \times 7) \times 3$, it corresponds to the three RGB channel images of the $189 \times 161$ true color image of the Lion Capital of Ashoka, shown in the top left display of Figure 1, and that is publicly available at https://upload.wikimedia.org/wikipedia/ur/1/19/Emblem_of_India_1947-1950.png. Figure 1 (top right) displays simulated realizations at each setting: as expected, increasing $\sigma^2$ produces noisier images. Figure 1 (bottom block) displays TVN-, GSM- and robust-estimated $\hat{M}$s, estimators obtained using the algorithm described in Sections 3.1, 3.3, and 3.5 respectively. For Tyler’s robust estimator we used the location estimation procedure of Section 3.5.2. For all cases, estimation performance worsens with increasing $\sigma$. For TVN-simulated data, there is little difference in the quality of $\hat{M}$ obtained by the three methods. But for heavy-tailed GSM-simulated data, the GSM or robust $\hat{M}$s recover $M$ better than the TVN-estimated $\hat{M}$.

Our illustration in Figure 1 was only on one realization per setting, so we performed a larger study under a similar framework. We generated TVN and GSM data for $\sigma = 2, 6, 10$, and 100 different realizations of $(\Sigma_1, \Sigma_2, \ldots, \Sigma_6)$, each drawn as before. Figure 2 displays the relative Frobenius norms of the difference between the true $(M, \Sigma_1, \Sigma_2, \ldots, \Sigma_6)$ and the estimated $(\hat{M}, \hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_6)$ parameters, obtained under the GSM, TVN and Tyler models, and for 100 independent samples realized under GSM and TVN assumptions. Specifically, the relative Frobenius normed difference is $R_d = \frac{||\theta - \hat{\theta}_{TVN}||}{||\theta - \hat{\theta}||}$, where $\theta$ generically denotes the parameter being compared, $\hat{\theta}_{TVN}$ corresponds to the estimated parameter under the TVN model and $\hat{\theta}$ is the the parameter estimate obtained under the assumed GSM or Tyler model. Therefore, values larger than unity indicate better performance of the GSM (or Tyler) estimator over that obtained under TVN assumptions. The converse holds for values below unity. Our display in Figure 2 is grouped by parameters of similar dimensionality, with the first graphic displaying results for $M$ and $\Sigma_4 \in \mathbb{R}^{23 \times 23}$, the second for $\Sigma_1, \Sigma_5 \in \mathbb{R}^{7 \times 7}$ and $\Sigma_2 \in \mathbb{R}^{9 \times 9}$, and the third display of values of the $3 \times 3$ matrices $\Sigma_3$ and $\Sigma_6$. For TVN-simulated data, there is not much to choose from between the TVN-assumed estimates and those obtained using GSM or Tyler estimators. On the other hand, for GSM-simulated data and the higher-dimensional parameters $M$ and $\Sigma_4$, the robust GSM and Tyler models...
Fig. 2. Relative Frobenius norms of the differences in parameters (generically denoted by \( \theta \)) for \( \sigma \in \{ 2, 6, 10 \} \), TVN- and GSM-simulated data, and estimates obtained under Tyler’s robust and GSM model assumptions. Each plot is obtained over 100 realizations from the GSM and TVN models with different values of \( \Sigma \), themselves independent Wishart realizations. All points are concentrated around 1 for TVN-simulated data, indicating that the GSM and Tyler estimates perform at least as well as the TVN model when the data is TVN-simulated. For the higher dimensional parameters \( \Sigma_1 \) and \( \Sigma_2 \in \mathbb{R}^{23 \times 23} \) and for GSM-simulated data, nearly all points are above 1, indicating that the GSM and Tyler models outperform the TVN model in this scenario. The same pattern holds for medium-sized parameters \( \Sigma_1, \Sigma_2 \in \mathbb{R}^{7 \times 7} \), and to a lesser extent to smaller \( 3 \times 3 \) matrices \( \Sigma_3 \) and \( \Sigma_6 \). In conclusion, the GSM and Tyler models perform at least as well as the TVN model for TVN-simulated data. Moreover, the GSM and Tyler models outperform the TVN model for GSM-simulated data, and their relative performance increase with larger covariance matrices. Finally, we note that both plots show similar performance, and this is expected since both are robust estimators for heavy-tailed data.

outperform the TVN model, since nearly all values exceed unity. The performance of the TVN model is still worse for \( \Sigma_1, \Sigma_2, \Sigma_5 \) when compared to the GSM and Tyler models, since \( R_d \) is then mostly above 1. However, for smaller \( 3 \times 3 \) covariance matrices (\( \Sigma_3 \) and \( \Sigma_6 \)), the TVN model performs more like the GSM and Tyler models. In conclusion, the GSM and Tyler fits perform similarly to the TVN model for TVN-simulated data, but the GSM and Tyler models outperform the TVN model when the data are GSM-simulated, and their improvements relative to the TVN increase with larger covariance matrices. Our simulations show the importance of considering an EC distribution that is more general than the TVN for heavier-tailed data. It also shows the benefit of using Tyler’s robust estimator when the exact EC distribution is unknown.

5 APPLICATION TO IMAGE LEARNING

We apply our EC tensor-variate estimation methodology to two learning applications involving images. The first application is in the context of improved prediction of an image class using tensor-variate discriminant analysis and classification while the second application characterizes the distinctiveness of face images in terms of gender, age and ethnic origin. In each case, we demonstrate that the EC distribution with heavier tails (here, the TV-t distribution) has better results than its TVN counterpart.

5.1 Discriminant analysis for image classification

In this section, we provide a general linear (LDA) and quadratic (QDA) discriminant analysis framework for the classification of tensor-valued data, in a similar manner as done for vector- [3] and matrix-variate [68] data.
We also illustrate how our maximum likelihood estimation methods of Section 3 are incorporated into the LDA and QDA frameworks, and the benefit of using a more general EC tensor-variate distribution than the TVN in predicting images of dogs or cats in a two-class digital image classification problem. The development of linear and quadratic discriminant analysis is not a key contribution of this article, so we refer to Section C.1 for details. However, we note that when compared to traditional discriminant analysis, our covariance matrices are invertible with a much smaller sample size because their Kronecker structure. The methods in Section 3 are used to estimate the parameters in each class population. Our framework can extend other discriminant analysis methods that considered the normal distribution [39].

The Animal Faces-HQ (AFHQ) dataset [12] has 15,000 512×512 digital photographs of wild and domestic felines and canines. In this paper, we use our developed methodology to classify the 5,653 cat and 5,239 dog images from the AFHQ dataset, of which 500 images of each animal are test datasets [12]. Figure 3a shows images of six randomly selected dogs and cats from the database. Despite the attempts at alignment, the images have dissimilar lighting conditions and viewing angles, so we serially applied the Radon [61] and 2D discrete wavelet transforms (DWT) [14] to the images in each of the RGB channels and extracted the low-frequency (LL) components [35, 36], yielding a $22 \times 22 \times 3$ array of features for each training and test image. Figure 6a displays the sequentially applied Radon and DWT transforms that correspond to the cat and dog images of Figure 3a. The processing extracts the local and directional features in the images, with the Radon transform improving the low-frequency components [35] that are then extracted by the wavelet transform [36].

We use QDA and LDA to build discriminant rules from these transformed array data, by modeling each population of transformed images via the TV-t distribution with unknown degrees of freedom, and for comparison, also the TVN distribution. In particular, we fit

$$Y_i = \langle x_i | B \rangle + \mathcal{E}_i, \quad \mathcal{E}_i \sim t_{22,22,3}(v; 0, \Sigma_1, \Sigma_2, \Sigma_3),$$  \hspace{1cm} (22)

where $Y_i \in \mathbb{R}^{22 \times 22 \times 3}$ is the $i$th transformed image and $x_i$ is $[0, 1]'$ or $[1, 0]'$ depending on whether $Y_i$ is a transformed image of a cat or a dog. The tensor-variate errors $\mathcal{E}_i$ are set to follow a TV-t distribution with unknown ($v$) degrees of freedom, which corresponds to the GSM distribution of (7) for $a = b = v$. Based on our model (22), the densities $f_{\text{cat}}$ and $f_{\text{dog}}$ involved in the classification rule of (38) correspond to the EC distributions $t_{22,22,3}(v_{\text{cat}}; B_{\text{cat}}, \Sigma_1, \Sigma_2, \Sigma_3)$ and $t_{22,22,3}(v_{\text{dog}}; B_{\text{dog}}, \Sigma_1, \Sigma_2, \Sigma_3)$, where $B_{\text{cat}} = B \times_1 [1, 0]$ and $B_{\text{dog}} = B \times_1 [0, 1]$ are the mean parameters of the transformed cat and dog image populations. Similarly for QDA, we estimate $f_{\text{cat}}$ from $t_{22,22,3}(v_{\text{cat}}; B_{\text{cat}}, \Sigma_1, \Sigma_2, \Sigma_3)$ and $f_{\text{dog}}$ from $t_{22,22,3}(v_{\text{dog}}; B_{\text{dog}}, \Sigma_1, \Sigma_2, \Sigma_3)$ using the ToTR methodology of (22) with $x_i \equiv 1$ always.

These models (one for LDA and two for QDA) were fit to the training set images of 5153 cats and 4739 dog images using the ECM-algorithm of Section 3.4. The degrees of freedom $v$ were estimated using the either step of (17) to be $v = 3.33$ for the LDA case, and $v_{\text{cat}} = 3.05$ and $v_{\text{dog}} = 3.51$ for the QDA case. We specified $\Sigma_1$ and $\Sigma_2$ to be AR(1) correlation matrices to capture spatial context in the transformed images, and $\Sigma_3$ to have an equicorrelation structure to denote correlation between the three RGB channels. This autocorrelation was estimated to be 0.94 for the LDA case and the same for the cat transformed images but marginally lower (0.93) for the dogs in the QDA fits. We also imposed low-rank CP formats on $B$, with ranks chosen from among five candidates using cross validation (CV) on the training data, and both misclassification rate and area under the curves (AUC) as our CV decision criteria. For comparison, we also performed QDA and LDA under the TVN model.

The classification rule of (38) for a transformed image $Y$ is specified in terms of the loglikelihood ratio (LLR) $\log(f_{\text{cat}}(Y)) - \log(f_{\text{dog}}(Y)) + \log(\eta_1 / \eta_2)$. A large LLR value indicates high posterior probability that $Y$ is a cat image, with small LLR values conversely indicating high posterior probability that $Y$ is the image of a dog. A zero value indicates complete uncertainty in the classification. In our application, we set $\eta_1 = \eta_2 = 0.5$. We obtained posterior logit-probabilities for the 500 cat and 500 dog images in the test set, and used them to
Fig. 3. (a) Sample dog and cat images from the AFHQ dataset, and (b) Precision-recall (PR) and receiver operating characteristic (ROC) curves for the tensor-variate-$t$ (TV-$t$) model and the TVN model, and across both LDA and QDA classification frameworks. The TV-$t$ model corresponds to higher curves, and hence it outperforms the TVN model in terms of ROC and PR.

calculate the precision-recall (PR) curves [15] and the receiver operating characteristic (ROC) curves [8, 9] of Figure 3b. We observe that for both LDA and QDA, PR curves are higher with the TV-$t$ than with the TVN models, indicating higher precision at all possible recall values using the TV-$t$ model than with the TVN model. In the QDA PR curves, the TVN model sometimes has higher precision than the TV-$t$ model. However, this happens for thresholds with recall values of less than 0.2, which are not practical since they lead false negatives (that is, cat images being classified as dog images) to be four times more numerous than true positives (correctly classified cat images). Similarly, in all cases the ROC curve is higher for the TV-$t$ model than for the TVN model. Higher ROC curves mean that all possible false positive rates have higher sensitivity (true positive rate) for the TV-$t$ than for the TVN case. Further, the degree of freedom parameter $\nu$ was estimated to be below 3.6 in all cases, meaning that our tensor-variate data are quite heavy-tailed. This explains the better classification performance of TV-$t$ relative to TVN distribution in this real-data example, since the TV-$t$ can better accommodate the heavier tails of the transformed image data. We finish this section noting that our Figure 3b uses the specific training/test split provided by the data authors. In Figure 6b we display ROC and PR curves corresponding to multiple combinations.
of training/test sets, resulting in a smoother curve than that of 3b, since it is discretized at 10,000 points instead of 1,000. In Section C.2 we provide more details on the derivation of this Figure.

5.2 Distinguishing facial characteristics

The LFW database is commonly used in the development and testing of facial recognition methods [34], and consists of over 13,000 250 \times 250 color images. Each image has the labeled attributes of ethnic origin, age group and gender [1, 42]. We use our ToTR methodology of Section 3.4 to distinguish the visual characteristics of different attributes. Specifically, we perform a 3-way tensor-variate analysis of variance (TANOVA) model to distinguish faces across the three factors of gender (male or female), continent of ethnic origin (African, European or Asian, as specified in the database), and cohort (child, youth, middle-aged or senior). Of the more than 13,000 images in the LFW database, we use the 605 images with unambiguous genders, age group and ethnic origin, and with at most 33 images for each factor combination that were selected and made available by Llosa-Vite and Maitra [46]. The authors also cropped the images to a central region of size 151 \times 151 each, and logit-transformed them to match the statespace of the normal distribution. In this application, we evaluate whether using a TV-t distribution for the errors can improve model fit.

We fit the 3-way TANOVA model with TV-t errors:

$$Y_{ijkl} = \langle X_{ijk} | B \rangle + \varepsilon_{ijkl}, \quad \varepsilon_{ijkl} \sim t_{m_i}(\nu; 0, \sigma^2 \Sigma_1, \Sigma_2, \Sigma_3).$$

Table 3 displays the BICs across the six (two error distributions across three formats) assumptions. The chosen CP rank was 70, with 19,175 unconstrained parameters in B, or a 98.4% dimension reduction relative to the unconstrained B that contains more than 1.2 million parameters. The TT rank was chosen to be g = (1, 2, 4, 8, 4, 2), which resulted in a total of 8,450 unconstrained parameters in B, or a 99% dimension reduction relative to the unconstrained B. Table 3 displays the BICs across the six (two error distributions across three B formats) assumptions.

Table 3. BICs for different fits, with smaller values indicating better model fit. The TV-t fits in this table and this paper have a selected degrees of freedom parameter \(\nu\) of 2.01, which in all cases outperformed their TVN counterparts.
From Table 3, we observe that the TV-$t$ distribution always outperforms its TVN counterpart, which indicates the benefit of considering a model for the heavy tails. Moreover, within the TVN or TV-$t$ models, the CP and TT formats always outperform the model with unconstrained $B$, with the CP always performing better than the TT. This shows the benefit of using a low-rank format that allows massive parameter reductions while also preserving model accuracy. Importantly in the context of this paper, we find that the TV-$t$ distributions with constrained or unconstrained $B$ outperform all the TVN models, including the TT and CP formats. Our findings show that performing parameter reduction through the CP or TT low-rank formats should be accompanied with the more appropriate model for the tail weights. Indeed, the best performance is achieved when the low-rank format is used in combination with the heavy-tailed TV-$t$ distribution.

Figures 4a and 4b display the estimated $\hat{B}$ for the TV-$t$ ($v = 2.01$) and TVN models. In both cases, the CP format preserves vital visual characteristics regarding ethnic origin, gender, and age group. However, the TVN-formatted $B$ results in more exaggerated facial expressions, while the fits under the TV-$t$ model are sometimes noisier, but more accurately reflect the diversity of facial characteristics in each sub-group. Therefore, we contend that the TV-$t$ model not only provides a quantitatively better fit to the TANOVA model, but also more accurately displays the heterogeneity in the dataset.

6 DISCUSSION

In this article, we introduced, defined, characterized and studied in detail the family of spherical and EC tensor-variate distributions that generalize the well-studied TVN distributions. We derived characteristic functions, PDFs, exact distributions of linear reshapings, and linear transformations, derived the moments of different orders as well as of special forms. These properties arise not only from the vectorized elliptical form of the random tensor, but also from its intrinsic tensor-variate structure, which in many cases makes estimation and inference possible. We provided algorithms for maximum likelihood estimation under different assumptions on the observed EC data. First, we suggested a modification to the maximum likelihood estimation algorithm for TVN data that makes its parameter estimates identifiable. We then used this algorithm to propose a maximum likelihood estimation algorithm for uncorrelated EC tensor-valued observations. We also derived EM algorithm variants for maximum likelihood estimation under independent identically distributed draws from a scale mixture of the TVN distribution, and a novel, robust and constrained Tyler-type algorithm for maximum likelihood estimation when the underlying EC tensor-variate distribution is unknown. We further proposed a ToTR framework with EC errors that extends the work of Llosa-Vite and Maitra [46] to the case where the tensor-response regression has EC tensor-variate distributed errors under various low-rank format assumptions on the regression coefficient. We studied the performance of our estimation algorithms through a simulation experiment on 6-way tensor-valued data that showed the limitations of the TVN distribution when the data have heavier tails than can be modeled by the TVN. Indeed, we demonstrated that fitting a GSM or Tyler robust estimating model to such data results in improved performance when the data are not Gaussian. Further, we provided methodology that allows us to use our maximum likelihood methodology to perform LDA and QDA on tensor-data classification and discrimination, and applied it to predict images of cats or dogs in the AFHQ dataset. We demonstrated that the AFHQ data is quite heavy-tailed, resulting in better performance for the TV-$t$ model relative to the TVN model in terms of ROC and PR curves. Finally, we used our ToTR methodology to perform a 3-way TANOVA of facial images from the LFW database. We demonstrated that the database is exceptionally heavy-tailed (with a selected degree of freedom of $v=2.01$), since the TV-$t$ models outperform their TVN counterparts in all scenarios.

Two co-editors-in-chief (co-EICs) have asked whether our methodology, and its use in discriminant analysis or TANOVA, has advantages over nonlinear deep learning methods, especially in the context of the applications of Section 5. We agree that deep learning is a competitor in a predictive classification framework, however, our methodology provides more natural interpretation in discriminant analysis. Moreover, our EC tensor-variate
Fig. 4. Estimated CP-factorized $\mathcal{B}$ under a 3-way factor ANOVA of the LFW dataset and the (a) TV-$t$ distribution and (b) TVN distribution. The results are compressed mean images across 24 factors of age group, gender and ethnic-origin, while providing more than 98% parameter reduction.
distributions can potentially enhance deep neural networks, by incorporating a more sophisticated likelihood function that incorporates spatial context between the image pixels, and that can be used as a loss function while training and choosing a neural network. For instance, as pointed out by a co-EIC, there are many methods that use multivariate Gaussian scoring for the match between data and model, and others use multivariate Gaussians in variational methods in the form of variational autoencoders. Our development in this paper provides an additional option by allowing for the use of EC tensor-variate distributions in place of these multivariate or tensor-variate Gaussians. In the context of the application in Section 5.1, our generalized model-based discriminant methodology has smaller compute demands than neural networks. Also, deep learning generally requires substantial amounts of training data. For the LFW dataset of Section 5.2, we have factor combinations that are also severely imbalanced in size, with one factor combination having around 2000 replications, and some others having only a handful of images. Further, although not presented in this article, we can more directly develop inference tools for our methodology that, as in [46], can be easily interpreted. These benefits are not immediate with deep learning methodology. Finally, selecting the right deep learning model requires knowledge of the loss function and training method, and calls for considerable skill and expertise. In contrast, our setup allows for an objective tool (BIC) for choosing the correct model.

There are several other possible extensions and generalizations of our work. For instance, we can define spherical distributions in multiple ways. The class of distributions \( \mathcal{F}_k = \{ \mathbf{X} : \mathbf{X} \in \{ X_{i_1}, \ldots, X_{i_k}, I_{m_1}, \ldots, I_{m_p} \} \}, I_j \in \mathcal{O}(m_j) \forall j \in \{ 1, \ldots, k \} \) is such that \( \mathcal{F}_k \subseteq \mathcal{F}_{k-1} \subseteq \cdots \subseteq \mathcal{F}_1 \), where \( \mathcal{F}_k \) is the family we defined in this article. All these families have Kronecker-separable covariance structures if the family of EC random tensors is defined using the Tucker product, as in (4). The Kronecker-separable structure makes it practical for us to explore the parameter space of the ultra-high dimensional covariance matrix. However, other types of variance structures can also be considered. We have studied in detail the family of tensor-variate elliptical distributions that are extensions of symmetric distributions. Another possible avenue of work is to extend the TVN distribution to skewed distributions and study their properties, similar to [26, 77]. Finally, the sampling distributions of \( B \) in our ToTR with EC errors are unknown, and may be of interest in applications where hypothesis and significance assessment are important. Thus, we see that though we have developed theory and estimation methodology for EC tensor-variate distributions, there remain a number of extensions and generalizations worthy of investigation.

ACKNOWLEDGMENTS

We thank two co-EICs, an anonymous Associate Editor and two anonymous reviewers for their insightful comments on an earlier draft of this paper that greatly improved its content. A version of this paper won C. Llosa-Vite an award at the 2022 Statistical Methods in Imaging Student Paper Competition, organized by the American Statistical Association’s Section on Statistics in Imaging. This research was supported in part by the National Institute of Justice (NIJ) under Grants No. 2018-R2-CX-0034 and 15PNIJ-21-GG-04141-RESS. The research of the second author was also supported in part by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health (NIH) under Grant R21EB034184, and the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) Hatch project IOW03717. The content of this paper is however solely the responsibility of the authors and does not represent the official views of the NIJ, the NIBIB, the NIH, the NIFA or the USDA.

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC (NTESS), a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy’s National Nuclear Security Administration (DOE/NNSA) under contract DE-NA0003525. This written work is authored by an employee of NTESS. The employee, not NTESS, owns the right, title and interest in and to the written work and is responsible for its contents. Any subjective views or opinions that might be expressed in the written work do not necessarily represent the views of the U.S.
Government. The publisher acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this written work or allow others to do so, for U.S. Government purposes. The DOE will provide public access to results of federally sponsored research in accordance with the DOE Public Access Plan.

REFERENCES

[1] Mahmoud Afifi and Abdelrahman Abdelhamied. 2019. AFIF4: Deep gender classification based on AdaBoost-based fusion of isolated facial features and foggy faces. *Journal of Visual Communication and Image Representation* 62 (2019), 77 – 86.

[2] Deniz Akdemir and Arjun K. Gupta. 2011. Array variate random variables with multivariate Kronecker delta covariance matrix structure. *Journal of Algebraic Statistics* 2, 1 (April 2011), 98–113.

[3] T. W. Anderson and R. R. Bahadur. 1962. Classification into two Multivariate Normal Distributions with Different Covariance Matrices. *The Annals of Mathematical Statistics* 33, 2 (June 1962), 420–431. https://doi.org/10.1214/aoms/1177704568

[4] T. W. Anderson, Huang Hsu, and Kai-Tai Fang. 1986. Maximum-likelihood estimates and likelihood-ratio criteria for multivariate elliptically contoured distributions. *Canadian Journal of Statistics* 14, 1 (March 1986), 55–59. https://doi.org/10.2307/3315036

[5] D. F. Andrews and C. L. Mallows. 1974. Scale Mixtures of Normal Distributions. *Journal of the Royal Statistical Society. Series B (Methodological)* 36, 1 (1974), 99–102.

[6] M. Arashi. 2017. Some theoretical results on tensor elliptical distribution. arXiv:1709.00801 [math, stat] (Sept. 2017). arXiv: 1709.00801.

[7] M. Arashi and S. M. M. Tabatabaey. 2010. A note on classical Stein-type estimators in elliptically contoured models. *Journal of Statistical Planning and Inference* 140, 5 (May 2010), 1206–1213. https://doi.org/10.1016/j.jspi.2009.11.001

[8] Leonidas E. Bantis and Ziding Feng. 2016. Comparison of two correlated ROC curves at a given specificity or sensitivity level. *Statistics in Medicine* 35, 24 (2016), 4352–4367.

[9] Andrew P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30, 7 (July 1997), 1145–1159.

[10] Stamatis Cambanis, Steel Huang, and Gordon Simons. 1981. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis* 11, 3 (Sept. 1981), 368–385.

[11] J. Douglas Carroll and Jih-Jie Chang. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of “Eckart-Young” decomposition. *Psychometrika* 35, 3 (Sept. 1970), 283–319.

[12] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[13] Kai-Ching Chu. 1973. Estimation and decision for linear systems with elliptical random processes. *IEEE Trans. Automat. Control* 18, 5 (1973), 499–505.

[14] Ingrid Daubechies. 1992. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

[15] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning (ICML ’06)*. Association for Computing Machinery, 233–240.

[16] Jan de Leeuw. 1994. Block-relaxation Algorithms in Statistics. (1994), 308–324.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *The Annals of Mathematical Statistics* 39, 1 (Sept. 1977), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

[18] Pierre Dutilleul. 1999. The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation* 64 (09 1999), 105–123.

[19] Pierre Dutilleul. 2018. Estimation and testing for separable variance-covariance structures. *Wiley Interdisciplinary Reviews: Computational Statistics* 10, 4 (March 2018)

[20] Morris L. Eaton. 1989. Group Invariance Applications in Statistics. *Regional Conference Series in Probability and Statistics* 1 (1989), i–133.

[21] Kaitai Fang and Handeng Chen. 1984. Relationships among classes of spherical matrix distributions. *Acta Mathematicae Applicatae Sinica* 1, 2 (Dec. 1984), 138–148. https://doi.org/10.1007/BF01669674

[22] Kai-Tang Fang and T. W. Anderson. 1990. *Statistical inference in elliptically contoured and related distributions*. Allerton Press, New York. OCLC: 20490516.

[23] Gabriel Frahm. 2004. *Generalized Elliptical Distributions: Theory and Applications*. Doctoral Thesis. Universität zu Köln.

[24] Mostafa Reisi Gahrooei, Hao Yan, Kamran Paynabar, and Jianjun Shi. 2021. Multiple Tensor-on-Tensor Regression: An Approach for Modeling Processes With Heterogeneous Sources of Data. *Technometrics* 63, 2 (April 2021), 147–159. https://doi.org/10.1080/00401706.2019.1708463

[25] Manuel Galea, Marco Riquelme, and Gilberto A. Paula. 2000. diagnostic methods in elliptical linear regression models. *Brazilian Journal of Probability and Statistics* 14, 2 (2000), 167–184.
As in Section 2.1, we assume $X$ is a $p$-way tensor of size $m_1 \times m_2 \times \ldots \times m_p$ and $e^m_i \in \mathbb{R}^m$ is a unit-basis vector with 1 as the $i$th element and 0 elsewhere. The vector outer product of a set of vectors $(z_1, z_2, \ldots, z_p)$, where $z_j \in \mathbb{R}^{m_j}$ for all $j = 1, 2, \ldots, p$, is written as $\otimes_{j=1}^p z_j$ and results in a $p$-way tensor of size $n_1 \times n_2 \times \ldots \times n_p$ such that $(\otimes_{j=1}^p z_j)(i_1, i_2, \ldots, i_p) = \prod_{j=1}^p z_j(i_j)$. This product is useful for expressing $X$ as

$$\sum_{i_1=1}^{m_1} \ldots \sum_{i_p=1}^{m_p} X(i_1, i_2, \ldots, i_p) (\otimes_{q=1}^p e^m_{i_q}),$$

(24)
and the Tucker product between $X$ and $A_1, A_2, \ldots, A_p$ as
\[
[X; A_1, A_2, \ldots, A_p] = \sum_{i_1, \ldots, i_p} X(i_1, i_2, \ldots, i_p)(\bigotimes_{q=1}^p A_q(:, i_q)),
\]
where $A_q(:, i_q)$ is the $i_q$th column of $A_q$. The Kronecker product $\otimes$ between two matrices $A$ and $B$, where $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times q}$, results in the following block matrix
\[
A \otimes B = \begin{bmatrix}
A(1, 1)B & \cdots & A(1, m)B \\
\vdots & \ddots & \vdots \\
A(n, 1)B & \cdots & A(n, m)B
\end{bmatrix}.
\]

Tensor reshapings are obtained by manipulating the vector outer product. We define the vectorization, $k$th mode matricization and $k$th canonical matricization of $X$ respectively for any $k = 1, 2, \ldots, p$ as
\[
\vec{X} = \sum_{i_1=1}^{m_1} \cdots \sum_{i_p=1}^{m_p} X(i_1, i_2, \ldots, i_p)(\bigotimes_{q=p}^m e_{i_q}^{m_q}),
\]
\[
X_{(k)} = \sum_{i_1=1}^{m_1} \cdots \sum_{i_p=1}^{m_p} X(i_1, i_2, \ldots, i_p) e_{i_k}^{m_k} (\bigotimes_{q=p,q\neq k}^m e_{i_q}^{m_q})^\top,
\]
\[
X_{\lessdot k} = \sum_{i_1=1}^{m_1} \cdots \sum_{i_p=1}^{m_p} X(i_1, i_2, \ldots, i_p) (\bigotimes_{q=p}^m e_{i_q}^{m_q}) (\bigotimes_{q=p}^m e_{i_q}^{m_q})^\top.
\]

### A.2 Marginal distributions

The marginal distributions follow from Theorem 3, by choosing $A_1, A_2, \ldots, A_p$ appropriately. As an example, consider the 3-way random tensor $Y \sim EC_{m_1, m_2, m_3} (M, \Sigma_1, \Sigma_2, \Sigma_3, \varphi)$, which will be split into eight sub-tensors as
\[
Y_{ijk} = [[Y; A_{i1}, A_{i2}, A_{i3}]], \quad M_{ijk} = [[M; A_{i1}, A_{i2}, A_{i3}]],
\]
where $A_{ij} = [l_{ij}; 0]$ and $A_{2l} = [0; l_{2l}-n_{2l}]$ are block matrices of size $n_l \times m_l$ and $(m_l-n_l) \times m_l$ respectively, and $n_l < m_l$ for all $l = 1, 2, 3$ and $i, j, k = 1, 2$ (here $I_h$ denotes an $h \times h$ identity matrix and $0$ is a matrix of zeroes). Figure 5 displays $Y$ along one of its eight sub-tensors $Y_{111}$. Using Theorem 3 and (29), we obtain its marginal distribution as
\[
Y_{111} \sim EC_{n_1, n_2, n_3} (M_{111}, A_{111} \Sigma_1 A_{111}^\top, A_{121} \Sigma_2 A_{121}^\top, A_{131} \Sigma_3 A_{131}^\top, \varphi),
\]
where $A_{1k} \Sigma_k A_{1k}^\top$ is a sub-matrix corresponding to the first $n_k$ columns and rows of $\Sigma_k$. 

![Fig. 5. A random third order tensor $Y$ along with one of its subtensors $Y_{111}$.](image-url)
A.3 Derivation of Equation (8)

From the definition of characteristic function, and the property that \( \langle \mathcal{A}, [\mathcal{B}; C_1, C_2, \ldots, C_p] \rangle = \langle [\mathcal{A}; C_1', C_2', \ldots, C_p'], \mathcal{B} \rangle \), we have that

\[
\psi_Y(Z) = E(\exp(i\langle Z, M \rangle)) = \exp(i\langle Z, M + [\mathcal{A}; Q_1, Q_2, \ldots, Q_p] \rangle) \quad \text{(per Equation (4))}
\]

\[
= \exp(i\langle Z, M \rangle)E(\exp(i\langle X, [\mathcal{A}; Q_1', Q_2', \ldots, Q_p'] \rangle)) \quad \text{(upon rearranging)}
\]

\[
= \exp(i\langle Z, M \rangle)\psi_X ([\mathcal{A}; Q_1', Q_2', \ldots, Q_p']) \quad \text{(from the definition of } \psi_X \text{)}
\]

\[
= \exp(i\langle Z, M \rangle)\varphi (\langle [\mathcal{A}; Q_1', Q_2', \ldots, Q_p'], [\mathcal{A}; Q_1', Q_2', \ldots, Q_p'] \rangle) \quad \text{(from Equation (3))}
\]

\[
= \exp(i\langle Z, M \rangle)\varphi (\langle Z, [\mathcal{A}; \Sigma_1, \Sigma_2, \ldots, \Sigma_p] \rangle) \quad \text{(since } Q_k' = \Sigma_k \text{)}.
\]

A.4 Conditional distributions along multiple modes

Here we demonstrate how to find the conditional distribution along multiple modes. As an example, consider the subtensor \( Y_{111} \) of \( Y \sim \mathcal{E}C_{r_1, r_2, m_1} (M, \Sigma_1, \Sigma_2, \Sigma_3, \varphi) \), both shown in Figure 5 and let

\[
\Sigma_i = \begin{bmatrix} \Sigma_{i,11} & \Sigma_{i,12} \\ \Sigma_{i,21} & \Sigma_{i,22} \end{bmatrix}, \quad \Sigma_{i,11} \in \mathbb{R}^{n_i \times n_i},
\]

and \( \Sigma_{i,11} = \Sigma_{i,12} \Sigma_{22}^{-1} \Sigma_{12} \) for \( i = 1, 2, 3 \). Then

\[
\mathcal{L}(Y_{111} | Y_{111} = Y_{111}) = \mathcal{E}C_{r_1, r_2, m_1} (M_{111}|_{Y_{111}}, \Sigma_{111}, \Sigma_{22}, \varphi_{111}),
\]

where the event \( Y_{111} = Y_{111} \) means that \( Y_{112} = Y_{112}, Y_{121} = Y_{121}, Y_{122} = Y_{122}, Y_{211} = Y_{211}, Y_{212} = Y_{212}, Y_{221} = Y_{221}, \) and \( Y_{222} = Y_{222}. \)

\[
M_{111} = \{ ([Y_{222} - M_{222}); \Sigma_{12}^{-1}, \Sigma_{22}^{-1}, \Sigma_{12}^{-1}, \Sigma_{22}^{-1}, \Sigma_{12}^{-1}, \Sigma_{22}^{-1}] \}
\]

\[
+ \{ (Y_{212} - M_{212}) \Sigma_{12}^{-1}, \Sigma_{22}^{-1}, \Sigma_{12}^{-1}, \Sigma_{22}^{-1}, \Sigma_{12}^{-1}, \Sigma_{22}^{-1}, \Sigma_{12}^{-1}, \Sigma_{22}^{-1}) \}
\]

\[
\varphi_{111}(u) = E \left( \left( \begin{array}{c} \mathcal{L}(Y_{111} | Y_{111} = Y_{111}) \end{array} \right) \right) \text{ with } R_{111}^2 = D_{111}^2 (\Sigma_{111} \otimes \Sigma_{111} \otimes \Sigma_{111} \otimes \Sigma_{111} \otimes \Sigma_{111} \otimes \Sigma_{111}), \quad (M_{111}, \Sigma_{111}), \text{ If } \varphi(u) = \exp(-u/2), \text{ then the conditional distribution of } R_{111}^2 \text{ given } Y_{111} = Y_{111} \text{ is } \chi_{n_1, n_2, n_3} \text{ and } \varphi_{111}(u) = \exp(-u/2).
\]

B PROOFS OF THEOREMS AND COROLLARIES IN SECTIONS 2 AND 3

B.1 Proof of Theorem 2

\textbf{Proof.} From the invariance of the Mahalanobis distance under tensor reshapings, \( Y, Y_{(k)}, Y_{<k>} \) and \( \text{vec}(Y) \) all have the same characteristic function, equivalently written for each case, as

\[
\psi_Y(Z) = \exp(i\langle Z, M \rangle)\varphi (\langle Z; [\Sigma_1, \Sigma_2, \ldots, \Sigma_p] \rangle)
\]

\[
= \exp(i\langle Z, M_{(k)} \rangle)\varphi (\langle Z_{(k)}; [\Sigma_k, \Sigma_{-k}] \rangle)
\]

\[
= \exp(i\langle Z_{<k>}, M_{<k>} \rangle)\varphi (\langle Z_{<k>}; [\Sigma_{<k>}; \Theta_{q=1}^k \Sigma_q, \Theta_{q=k+1}^p \Sigma_q] \rangle)
\]

\[
= \exp(i\langle \text{vec}(Z), \text{vec}(M) \rangle)\varphi (\langle \text{vec}(Z); [\Sigma_{\text{vec}}] \rangle).
\]
B.2 Proof of Theorem 3

Proof. The characteristic function of \([\mathbf{Y}; A_i, \ldots, A_p]\) can be expressed in terms of the characteristic function of \(\mathbf{Y}\) as \(\psi_{\mathbf{Y}}(\mathbf{Z}; A_1, \ldots, A_p)\), and using (8) yields

\[
\psi_{\mathbf{Y}}(\mathbf{Z}; A_1, \ldots, A_p) = \exp(i\langle Z, [M; A_1, \ldots, A_p]\rangle)\varphi(\langle Z, [\Sigma_1 A_1 \ldots, A_p \Sigma_1 A_1 \ldots, A_p]\rangle).
\]

The rest of Part 1 of the proof follows from Corollary 5 in [10] and Part 1 of Theorem 2, after noticing that

\[
\Box
\]

B.3 Proof of Theorem 4

Proof. From Part 4 of Theorem 2, we can write the distribution of \(\text{vec}(\mathbf{Y})\) as

\[
\text{vec}(\mathbf{Y}) = [\text{vec}(\mathbf{Y}_1) \quad \text{vec}(\mathbf{Y}_2)] \sim EC_{m_p \times (n_1+n_2)} \left( \begin{bmatrix} \text{vec}(\mathbf{M}_1) \\ \text{vec}(\mathbf{M}_2) \end{bmatrix} ; \begin{bmatrix} \Sigma_{11} \otimes \Sigma_{-p} & \Sigma_{12} \otimes \Sigma_{-p} \\ \Sigma_{21} \otimes \Sigma_{-p} & \Sigma_{22} \otimes \Sigma_{-p} \end{bmatrix}, \varphi \right).
\]

The remainder of the proof follows from Corollary 5 in [10] and Part 1 of Theorem 2, after noticing that \(\text{vec}(\mathbf{M}(1)_{(21)}) = \text{vec}(\mathbf{M}) + (\Sigma_{12} \otimes \Sigma_{-p}) (\Sigma_{22} \otimes \Sigma_{-p})^{-1} (\text{vec}(\mathbf{Y}_2) - \text{vec}(\mathbf{M}_2))\), and \(\Sigma_{p_{(11)}} \otimes \Sigma_{-p} = (\Sigma_{11} \otimes \Sigma_{-p}) - (\Sigma_{12} \otimes \Sigma_{-p}) (\Sigma_{22} \otimes \Sigma_{-p})^{-1} (\Sigma_{21} \otimes \Sigma_{-p})\). □

B.4 Proof of Theorem 5

The characteristic function of \(\mathbf{Y}\) in (8) is \(\psi_{\mathbf{Y}}(\mathbf{Z}) = \exp(ih(\mathbf{Z}))\varphi(g(\mathbf{Z}))\), where \(h(\mathbf{Z}) = \langle \mathbf{Z}, \mathbf{M} \rangle\) and \(g(\mathbf{Z}) = \langle \mathbf{Z}, \Sigma_1, \ldots, \Sigma_p \rangle\). Denote \(\mathbf{Z}_i = \mathbf{Z}(i_1, \ldots, i_p)\), the first derivative of \(h(\mathbf{Z})\) as \(\partial h(\mathbf{Z})/\partial \mathbf{Z}_i = m_i\), the first derivative of \(g(\mathbf{Z})\) as \(g_i = \partial g(\mathbf{Z})/\partial \mathbf{Z}_i\), and the second derivative of \(g(\mathbf{Z})\) as \(g_{ij} = \partial^2 g(\mathbf{Z})/\partial \mathbf{Z}_i \partial \mathbf{Z}_j = 2\sigma_{ij}\). All higher order derivatives are zero. Then

1. The first moment is obtained from \(E(Y_i) = i^{-1} \frac{\partial \psi_{\mathbf{Y}}}{\partial \mathbf{Z}_i}(0)\), where

\[
\frac{\partial \psi_{\mathbf{Y}}}{\partial \mathbf{Z}_i}(0) = \exp(ih(\mathbf{Z}))(g_i \varphi'(g(\mathbf{Z}))) + m_i \varphi(g(\mathbf{Z})).
\]

2. The second moment uses \(E(Y_i Y_j) = i^{-2} \frac{\partial^2 \psi_{\mathbf{Y}}}{\partial \mathbf{Z}_i \partial \mathbf{Z}_j}(0)\), where \(\frac{\partial^2 \psi_{\mathbf{Y}}}{\partial \mathbf{Z}_i \partial \mathbf{Z}_j}(\mathbf{Z})\) is

\[
\exp(ih(\mathbf{Z})) \left[ \varphi''(g(\mathbf{Z}))(g_i g_j) + \varphi'(g(\mathbf{Z}))(g_{ij} + ig_i m_j + ig_j m_i) + \varphi(g(\mathbf{Z}))(i^2 m_i m_j) \right].
\]

3. The third moment is obtained using \(E(Y_i Y_j Y_k) = i^{-3} \frac{\partial^3 \psi_{\mathbf{Y}}}{\partial \mathbf{Z}_i \partial \mathbf{Z}_j \partial \mathbf{Z}_k}(0)\), where \(\frac{\partial^3 \psi_{\mathbf{Y}}}{\partial \mathbf{Z}_i \partial \mathbf{Z}_j \partial \mathbf{Z}_k}(\mathbf{Z})\) is

\[
\exp(ih(\mathbf{Z})) \left[ \varphi'''(g(\mathbf{Z}))(g_i g_j g_k) + \varphi(g(\mathbf{Z}))(i^3 m_i m_j m_k) + \varphi''(g(\mathbf{Z}))(g_{ijk} g_i + g_{ik} g_j + g_{j i} g_k + ig_{ij} g_k m_i + ig_{jk} g_m j + ig_{k j} g_m k) + \varphi'(g(\mathbf{Z}))(i g_{ijk} m_k + i g_{j k} m_i + i g_{i j} m_k + i^2 m_j m_k g_i + i^2 m_i m_k g_j + i^2 m_i m_j g_k) \right].
\]

4. We get the fourth moment from \(E(Y_i Y_j Y_k Y_l) = i^{-4} \frac{\partial^4 \psi_{\mathbf{Y}}}{\partial \mathbf{Z}_i \partial \mathbf{Z}_j \partial \mathbf{Z}_k \partial \mathbf{Z}_l}(0)\), where, upon ignoring terms involving the first derivative of \(g(\cdot)\) because they are proportional to elements in \(\mathbf{Z}\) that will be set to 0,
Proof of Corollary 6

**Proof.** In these proofs, LHS and RHS refer to the left and right hand sides of an expression.

1. The proof follows directly from Part 1 of Theorem 5.
2. Let $X \sim EC_{m_1, \ldots, m_p}(0, \Sigma_1, \Sigma_2, \ldots, \Sigma_p, \varphi)$. Then, $\text{Var}(\text{vec}(Y)) = \text{Var}(\text{vec}(X))$, and from Part 2 of Theorem 5 and (26), we have

   $$\text{Var}(\text{vec}(X)) = \sum_{i_1, \ldots, i_p, j_1 \ldots, j_p} \mathbb{E}[X(i)X(j)] (\bigotimes_{q=p}^{1} e_{iq}^{m_q})(\bigotimes_{q=p}^{1} e_{jq}^{m_q})' = -2\varphi'(0) \bigotimes_{q=p}^{1} \Sigma_q.$$ 

   The following moment follows from Parts 1 and 2 of this theorem, and is used in Parts 3 and 4.

   $$\mathbb{E}[(\text{vec}(Y)(\text{vec}(Y)')] = (\text{vec}(M)(\text{vec}(M)'] - 2\varphi'(0))(\bigotimes_{k=p}^{1} \Sigma_k).$$ (31)

3. We can write the LHS as $\text{tr} \{ (\bigotimes_{k=p}^{1} A_k) \mathbb{E}[(\text{vec}(Y)(\text{vec}(Y)')] \}$. The result follows from (31) after using $\text{tr}(\bigotimes_{k=p}^{1} A_k) = \prod_k \text{tr}(A_k)$.

4. The vectorization of the LHS can be expressed as $\mathbb{E}[(\text{vec}(Y)(\text{vec}(Y)'] (\bigotimes_{k=p}^{1} A_k)(\text{vec}(V))$. Using (31) results in

   $$\langle V, [M; A_1, A_2, \ldots, A_p] \rangle \text{vec}(M) - 2\varphi'(0) \text{vec}[V; \Sigma_1 A_1', \ldots, \Sigma_p A_p'] \].$$

   We have proved that the vectorizations of the LHS and the RHS are the same, and since they are tensors of the same size, they must be the same.

5. Let $X \sim SC_{m_1, m_2, \ldots, m_p}(\varphi)$, and then the first and third moments of $X$ are zero, and using $Y = X + \mathcal{M}$, the LHS can be expressed as

   $$\langle M, [M; A_1, A_2, \ldots, A_p] \rangle M + \mathbb{E}(\langle X, [X; A_1, \ldots, A_p] \rangle) + \mathbb{E}(\langle M, [X; A_1, \ldots, A_p] \rangle X) + \mathbb{E}(\langle M, [X; A_1', \ldots, A_p'] \rangle X).$$

   The statement of this part of the theorem follows from Parts 3 and 4.

6. If $X \sim S_{m_1, m_2, \ldots, m_p}(\varphi)$ and $(C_k, D_k) = (2^{1/2} A_k \Sigma_k^{1/2}, \Sigma_k^{1/2} B_k \Sigma_k^{1/2})$ for all $k = 1, 2, \ldots, p$, where $(.)^{1/2}$ is the symmetric square root, then the LHS can be written as

   $$\sum_{i,j,k,l} \mathbb{E}(X_i X_j X_k X_l)(c_{ijkl}),$$ (32)

   ACM Trans. Probab. Mach. Learn.
where $X_i = X(i_1, i_2, \ldots, i_p)$ and $c_{ij} = \prod_{q=1}^p C_q(i_q, j_q)$. Using Part 4 of Theorem 5, we know that $E(X_i X_k X_l) = 4\phi''(0)(1_k1_l1_j + 1_j1_k1_i + 1_i1_l1_k)$, where $1_k = 1$ only if $i_k = j_k$ for all $k = 1, \ldots, p$, and it is 0 otherwise. From these, we can write (32) based on

$$\sum_j (c_{i1}d_{jj} + c_{ij}d_{ij} + c_{jj}d_{ii}) = \prod_{k=1}^p \left[ \text{tr}(C_k) \text{tr}(D_k) \right] + \prod_{k=1}^p \text{tr}(C_k D_k) + \prod_{k=1}^p \text{tr}(C_k D_k^T).$$

The statement follows after replacing $(C_k, D_k)$ with their original expressions.

\[\square\]

B.6 Proof of Theorem 7

**Proof.** We prove the two parts of the theorem in turn:

1. For any $i = 1, 2, \ldots, n$, the distribution of each $Y_i$ is obtained from Theorem 3 with $Y_i = Y \times e_i e_i^T$. They are uncorrelated using Part 2 of Theorem 5 because whenever $i_{p+1} \neq j_{p+1}$,

$$E \left[ (Y - M)(i_1, i_2, \ldots, i_{p+1}) \right] = 0.$$

2. We have from Part 3 of our Theorem 2 that $Y_{<p>} \sim EC_{m,n}(M_{<p>}, \sigma^2 \Sigma L_n, \varphi)$, where the $m \times n$ matrix $M_{<p>} = [\text{vec}(M) \ldots \text{vec}(M)]$. Therefore, from Theorem 1 of [4] we have that the maximum likelihood estimator of $\text{vec}(M)$ is $\text{vec}(\tilde{M})$, and that of $\sigma^2 I_n \otimes \left[ \bigotimes_{k=p}^1 \Sigma_k \right]$ is $(nm/d) \sigma^2 I_n \otimes \left[ \bigotimes_{k=p}^1 \Sigma_k \right]$. This means that in our case, the maximum likelihood estimator of the covariance matrix is the same as in the TVN case but up to a constant of proportionality, and since the scale matrices $\Sigma_k$ are proportionally constrained as $\Sigma_k(1, 1) = 1$ for all $k = 1, 2, \ldots, p$, it follows that the maximum likelihood estimator of $\Sigma_k$ is $\tilde{\Sigma}_k$, and the proportionality constant $nm/d_k$ is absorbed by $\sigma^2$.

\[\square\]

B.7 Proof of Theorem 8

As described in Section 3, the ADJUST procedure [27] solves the constrained optimization problem of Equation (13). For $S \in \mathbb{R}^{n \times m}$, denote $S_{-1,-1}$ is the submatrix $S$ after excluding the first row and column, and $S_{1,-1}$ is the first row after excluding the first entry. Then the ADJUST procedure ADJUST($n, \sigma^2, S$) is as follows

1. Let $\eta = S_{1,1}/(n\sigma^2)$.
2. Scale the whole matrix: $S \leftarrow S/S_{1,1}$
3. $S_{-1,-1} \leftarrow \eta S_{-1,-1} + (1 - \eta) S_{1,-1} S_{1,-1}^T$.

After having described the ADJUST procedure, we can now prove Theorem 8.

**Proof.** First, let $Z_{ik} = S_{ik}/|Y_i|^2$. Then the loglikelihood in (20) can be written in terms of $\Sigma_k$ only as

$$\ell(\Sigma_k) = \frac{nm-k}{2} \log |\Sigma_k^{-1}| - \frac{m}{2} \sum_{i=1}^n \log \text{tr}(\Sigma_k^{-1}Z_{ik}),$$

and has matrix derivative

$$\frac{\partial \ell}{\partial \Sigma_k^{-1}} = \frac{nm-k}{2} \Sigma_k - \frac{m}{2} \sum_{i=1}^n \frac{S_{ik}}{\text{tr}(\Sigma_k^{-1}S_{ik})}.$$
where the equality follows because $S_{ik}/\text{tr}(\Sigma_{k}^{-1}S_{ik}) = Z_{ik}/\text{tr}(\Sigma_{k}^{-1}Z_{ik})$. Setting (34) to 0 leads to the following fixed-point iterative procedure for obtaining $\Sigma_{k}$:

$$
\hat{\Sigma}_{k}^{(i+1)} = \frac{m_k}{n} \sum_{i=1}^{n} \frac{S_{ik}}{\text{tr}(\hat{\Sigma}_{k}^{(i)}S_{ik})}.
$$

However, we are interested in the constrained optimization under $\Sigma_{k}(1, 1) = 1$. To do this, we write the equality constraint function $g$ and its matrix derivative as

$$
g(\Sigma_{k}) = e_{m_{1}}^{m_{k}}\Sigma_{k}e_{i}^{m_{k}} - 1, \quad \frac{\partial g}{\partial \Sigma_{k}} = -\sigma \sigma' \tag{36}
$$

where $\sigma$ is the first column of $\Sigma_{k}$ and $e_{i}^{m_{k}} \in \mathbb{R}^{m_{k}}$ has 1 at position 1 and zero everywhere else. Our equality constraint function is zero only if $\Sigma_{k}(1, 1) = 1$. With $\ell(\cdot)$ and $g(\cdot)$ as in (33) and (36) we write the Lagrange multiplier as $L(\Sigma_{k}, \lambda) = \ell(\Sigma_{k}) - \frac{\lambda}{2} g(\Sigma_{k})$, and its derivative with respect to $\Sigma_{k}$ follows directly from (34) and (36). Based on this derivative, the constrained optimization must satisfy

$$
\frac{nm_{k} - m}{2} \Sigma_{k} - \frac{m}{2} \sum_{i=1}^{n} \frac{S_{ik}}{\text{tr}(\Sigma_{k}^{-1}S_{ik})} + \lambda \sigma \sigma' = 0. \tag{37}
$$

In (34) and (36), we ignored the symmetric structure of $\Sigma_{k}$ when taking derivatives because it is simpler and leads to the same root in (37). This equation is of the form of Equation (B.1) in Glanz and Carvalho [27], and therefore we know that it is satisfied for (21).

C ADDITIONAL DETAILS FOR SECTION 5

C.1 Linear and quadratic tensor-variate classification

Consider $G$ tensor-valued populations $\pi_{1}, \pi_{2}, \ldots, \pi_{G}$. A Bayes-optimal rule that minimizes the total probability of misclassification assigns an observed tensor $X$ to group $\pi_{i}$ if $i^* = \arg\max_{i=1,\ldots,G} \eta_{i}f_{i}(X)$, where $f_{i}(X)$ is the PDF of $\pi_{i}$ evaluated at $X$, and $\eta_{i}$ is the prior probability that a member of $\pi_{i}$ gets correctly classified [3, 68].

We first briefly visit the two-class problem. Here, we consider two tensor-valued populations $\pi_{1}$ and $\pi_{2}$. For $i, j = 1, 2$ let $\eta_{i}$ be the prior probability that a member of $\pi_{i}$ gets correctly classified to $\pi_{i}$, and also let $\mathbb{P}(i|j)$ be the probability that a member of $\pi_{j}$ gets misclassified to $\pi_{i}$ ($i \neq j$). The total probability of misclassification (TPM) in this case is $\mathbb{P}(2|1)\eta_{1} + \mathbb{P}(1|2)\eta_{2}$. Now, suppose we are interested in classifying one tensor-valued observation $X$ to $\pi_{1}$ or $\pi_{2}$. In this case, a Bayes-optimal rule that minimizes the TPM assigns $X$ to $\pi_{1}$ if $\eta_{1}f_{1}(X) \geq \eta_{2}f_{2}(X)$, where $f_{i}(X)$ is the PDF of group $\pi_{i}$ evaluated at $X$ [3, 68].

The extension of this rule to the case where there are $G$ groups $\pi_{1}, \pi_{2}, \ldots, \pi_{G}$ assigns $X$ to $\pi_{i}$ if

$$
i^* = \arg\max_{i=1,\ldots,G} \eta_{i}f_{i}(X),
$$

where $f_{i}$ and $\eta_{i}$ correspond to the PDF and prior probability that a member of $\pi_{i}$ gets correctly classified, respectively.

Suppose that $\pi_{i}$ has the $\mathcal{E}G_{m}(\mathcal{M}_{i}, \Sigma_{i,1}, \ldots, \Sigma_{i,3}, \phi)$ distribution with PDF (9). Then an observation $Y$ is assigned to $\pi_{1}$ over $\pi_{2}$ if

$$
\eta_{1}|\Sigma_{1}|^{-1/2}g(D_{\Sigma_{1}}^{2}(Y, \mathcal{M}_{1})) \geq \eta_{2}|\Sigma_{2}|^{-1/2}g(D_{\Sigma_{2}}^{2}(Y, \mathcal{M}_{2})), \tag{38}
$$

where $\Sigma_{i} = \bigotimes_{k=p}^{1} \Sigma_{i,k}$. This rule is greatly simplified for the TVN case where $g(u) = \pi^{-m/2} \exp(-u/2)$ as

$$
\log\left(\frac{\eta_{1}^{2}}{\eta_{2}^{2}}\right) - \log\left(\frac{|\Sigma_{1}|}{|\Sigma_{2}|}\right) - D_{\Sigma_{1}}^{2}(Y, \mathcal{M}_{1}) + D_{\Sigma_{2}}^{2}(Y, \mathcal{M}_{2}) > 0. \tag{39}
$$
(39) is a quadratic discrimination analysis (QDA) classification rule, with the usual reduction to linear discriminant analysis (LDA) for homogeneous covariance structures in the case of GSM distributions with PDF as in (7).

The parameters involved in (38) can be estimated from a training dataset using the methodology of Section 3. Moreover, our ToTR methodology of Section 3.4 allows us to estimate the mean parameters $\mu_1$ and $\mu_2$ under low-rank formats, thereby facilitating substantial parameter reduction in many imaging applications. Maximum likelihood estimates under the homogeneity of variance can be obtained by fitting the TANOVA model

$$Y_i = \langle x_i | B \rangle + \epsilon_i, \quad \epsilon_i \sim \mathcal{EC}_m(0, \sigma^2 \Sigma_1, \ldots, \Sigma_p, \varphi),$$

(40)

where $x_i = [0, 1]'$ or $x_i = [1, 0]'$ accordingly as whether $Y_i \in \pi_1$ or $Y_i \in \pi_2$. With population-specific covariances, we estimate the parameters in (38) by fitting separate EC models to the training data from each group using the methods of Section 3. For the ToTR under low-rank format case, fitting models to separate groups can be thought as the special case of (40) when $x_i$ is the scalar 1. We now illustrate our methodology on images from the AFHQ database.
C.2 Details on the classification of cats and dogs

In section 5.1 we display the sequentially applied Radon and DWT transforms of the cat and dog images displayed in Figure 3a. These transformed figures are displayed in Figure 6a and are examples of the $Y$s that are used in the calculation of the decision rule and in the prediction.

While the ROC and PR curves of Figure 3b demonstrate an improved classification for the TV-$t$ distribution when compared to the TVN, it is on the specific set of training (500 cats and 500 dogs) and test (5153 cats and 4739 dogs) data provided by the data authors. To demonstrate that the results hold for other combinations, we recreated the settings that generated Figure 3b under ten different training/test data combinations. We combined the training/test sets, and for each of the 10 combinations we chose 500 cats and 500 dogs at random without replacement for training, and used the remaining for testing. For each of these training/testing combinations we performed classification the same way it was described in Section 5.1, and used the classification scores to display the PR and ROC curves in Figure 6b. This figure is similar to Figure 6b, but is smoother, since it is discretized at 10,000 values instead of 1,000. The similarity of the two figures indicates that all the training/testing data combinations that we chose at random behave similarly to those used to generate Figure 3b.

Received 10 October 2023; revised 14 May 2024; accepted 11 June 2024