Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Comput Graph Stat. Author manuscript; available in PMC 2021 March 12.

Published in final edited form as:

J Comput Graph Stat. 2020 ; 29(3): 668–674. doi:10.1080/10618600.2019.1696208.

Classification with the matrix-variate-*t* distribution

Geoffrey Z. Thompson, **Ranjan Maitra**, **William Q. Meeker**, **Ashraf F. Bastawros**

Iowa State University, Ames, Iowa, USA.

Abstract

Matrix-variate distributions can intuitively model the dependence structure of matrix-valued observations that arise in applications with multivariate time series, spatio-temporal or repeated measures. This paper develops an Expectation-Maximization algorithm for discriminant analysis and classification with matrix-variate *t*-distributions. The methodology shows promise on simulated datasets or when applied to the forensic matching of fractured surfaces or the classification of functional Magnetic Resonance, satellite or hand gestures images.

Keywords

BIC; ECME; fMRI; fracture mechanics; LANDSAT; supervised learning

1 Introduction

Matrix-variate distributions (Gupta and Nagar, 1999) can conveniently model matrix-valued observations that arise, for instance, with multivariate time series or spatial datasets or when we observe p-variate responses at q different settings, yielding a $p \times q$ matrix of responses for each unit of observation. The matrix-variate normal distribution (abbreviated in this paper by MxVN) is also helpful for inference, but is sometimes inadequate for modeling populations where the matrix-variate-t distribution (henceforth MxVt) may be a better fit.

There exist discriminant analysis and classification methods for MxVN (Viroli, 2011b; Anderlucci and Viroli, 2015) mixtures but for many applications, the MxV*t* distribution may model each group better. However, parameter estimation for the MxV*t* distribution requires special care because, unlike in the normal case, it can not be viewed as simply a rearrangement of its vector-multivariate cousin (Dickey, 1967) for which several variants of the Expectation-Maximization (EM) algorithm exist (Dempster et al., 1977; Meng and Rubin, 1993; Liu and Rubin, 1994).

Corresponding author Ranjan Maitra maitra@iastate.edu.

Supplementary Materials

Supplement:
Provides the derivations for the EM and ECME algorithms presented in this paper.

Code:
Includes the R scripts and data to reproduce the results in Section 3.

MixMatrix:
An R package implementing the methods presented in this paper as well as some additional functions for working with matrix variate distributions.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

This paper develops, in Section 2, methodology for parameter estimation in the MxV*t* distribution and extends it to discriminant analysis and classification using MxV*t* mixtures. Our methods are evaluated on simulated and real-life datasets in Section 3. This paper concludes with some discussion. An online supplement explicitly detailing the derivations of our algorithm, with sections referenced using the prefix "S-", and an R (R Core Team, 2019) package MixMatrix (Thompson, 2019) that implements the methodology are also included.

2 Methodology

2.1 Background and Preliminary Development

2.1.1 The Matrix-variate Normal Distribution

Definition 1.: A random matrix **X** of p rows and q columns has the MxVN distribution with parameters **M**, **Σ** and **Ω** if it has the probability density function (PDF)

$$f(\boldsymbol{X};\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Omega}) = \frac{\exp\left(-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}^{-1}(\boldsymbol{X}-\mathbf{M})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\mathbf{M})\right]\right)}{(2\pi)^{pq/2}|\boldsymbol{\Omega}^{p/2}|\boldsymbol{\Sigma}|^{q/2}},$$

where $|\cdot|$ denotes the determinant, **M** is a p × q matrix that is the mean of **X**, and **Σ** and **Ω**>0 describing the covariances between, respectively, each of the p rows and the q columns of X. We write $\boldsymbol{X}\sim\mathcal{N}_{p,q}(\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Omega})$ For identifiability, we set the first element of **Σ** to be unity.

The MxVN distribution can be considered, after rearranging into a vector (denoted by vec(***X***)), to be from a multivariate normal (MVN) distribution with a Kronecker product covariance structure (Gupta and Nagar, 1999). So, if $\boldsymbol{X}\sim\mathcal{N}_{p,q}(\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Omega}),$ then $\mathrm{vec}(\boldsymbol{X})\sim\mathcal{N}_{pq}(\mathrm{vec}(\mathbf{M}),\boldsymbol{\Omega}\otimes\boldsymbol{\Sigma})$. This reformulation allows us to readily obtain the maximum likelihood (ML) estimates. If $\boldsymbol{X}i,\ i=1,2,\ldots,n$ are independent identically distributed (IID) random matrices from the $\mathcal{N}_{p,q}(\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Omega})$, then **M**,**Σ**, and **Ω** have ML estimators $\mathbf{M}=n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i, \boldsymbol{\Sigma}=(np)^{-1}\sum_{i=1}^{n}(\boldsymbol{X}-\mathbf{M})\boldsymbol{\Omega}^{-1}(\boldsymbol{X}-\mathbf{M})^T$, and **Ω** under the $=(nq)^{-1}\sum_{i=1}^{n}(\boldsymbol{X}-\mathbf{M})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\mathbf{M})$ constraint for identifiability that the (1, 1)th element of **Σ** is set to unity. The matrices **Σ** and **Ω** are obtained iteratively after initialization with any positive definite matrices, by default using the identity matrix ***I***. The ML estimates exist and are unique almost surely if $n>p/q/+q/p+2$ (Soloveychik and Trushin, 2016).

2.1.2 The Matrix-variate *t*-distribution

Definition 2.: A random p × q matrix **X** has a MxVt distribution with parameters (**M,Σ,Ω**) of similar order as in Definition 1 (with **Σ** and **Ω**>0) and degrees of freedom (df) v≥1 if its PDF is

$$f(\boldsymbol{X};\nu,\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Omega}) = \frac{\Gamma_p\left(\frac{\nu+p+q-1}{2}\right)}{(\pi)^{\frac{pq}{2}}\Gamma_p\left(\frac{\nu+p-1}{2}\right)}|\boldsymbol{\Omega}|^{-\frac{p}{2}}|\boldsymbol{\Sigma}|^{-\frac{q}{2}}|\mathbf{I}_p+\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\mathbf{M})\boldsymbol{\Omega}^{-1}(\boldsymbol{X}-\mathbf{M})^{\mathrm{T}}|^{\frac{\nu+p+q-1}{2}}.$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We use the notation $\boldsymbol{X}\sim t_{p,q}(\nu,\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Omega})$ to indicate that X has this density.

Properties: We mention some properties of the MxVt distribution relevant to this paper.

1. For $p = 1$ and $\boldsymbol{\Sigma} \equiv \nu$(or $q = 1$ and $\boldsymbol{\Omega} \equiv \nu$), the MxV*t* distribution reduces to its vector-multivariate *t* (MVT) cousin. However, this reduction does not generally hold so additional development is needed for inference. We provide methods to do so in the next section.

2. Let the random matrix $\boldsymbol{S}\sim\mathscr{W}_p\left(\upsilon+p-1,\boldsymbol{\Sigma}^{-1}\right)$, where $\mathscr{W}_p(\kappa,\boldsymbol{\Psi})$ is the p × *p*-dimensional Wishart distribution with d.f. $\boldsymbol{\kappa}$ and scale matrix $\boldsymbol{\Psi}$. If $\boldsymbol{X}\,|\,S\sim\mathscr{N}_{p,q}\left(\mathbf{M},S^{-1},\boldsymbol{\Omega}\right)$, then $\boldsymbol{X}\sim t_{p,q}(\nu,\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Omega})$ (see Gupta and Nagar, 1999, p. 135). Further, $\boldsymbol{S}\,|\,\boldsymbol{X}\sim\mathscr{W}_p\left(\upsilon+p+q-1,\left[(\boldsymbol{X}-\mathbf{M})\boldsymbol{\Omega}^{-1}(\boldsymbol{X}-\mathbf{M})^T+\boldsymbol{\Sigma}\right]^{-1}\right)$ (Iranmanesh et al., 2010).

2.2 ML Estimation of the MxV*t* parameters

The MxV*t* distribution does not have closed-form ML estimators so we provide an Expectation/Conditional Maximization Either (ECME) algorithm (Liu and Rubin, 1994) in a manner that is similar to that used to find ML parameter estimates in the MVT distribution, with the main contribution being the extension to the matrix variate case by deriving the estimates in terms of a matrix variate normal mixture with a Wishart distribution rather than a multivariate normal mixture with a chi-squared distribution.

Let $\boldsymbol{X}_i, i = 1,2,\ldots,n$ be IID realizations from $t_{p,q}(\nu,\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Omega})$ Write $\boldsymbol{\Theta} \equiv \{\upsilon,\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Omega}\}$. For each $i = 1,2,\ldots,n$, let $\boldsymbol{S}_i$ be (unobserved) Wishart-distributed random matrices that are as per Property 2.

From the detailed development and derivations provided in Section S-1.2, we get the expectation step (E-step) updates at the current value $\boldsymbol{\Theta}^{(t)}$ of $\boldsymbol{\Theta}$ by taking the expected values of the $\boldsymbol{S}_i$ given the current value of $\boldsymbol{\Theta}^{(t)}$:

$$\boldsymbol{S}_i^{(t+1)} \doteq \mathbb{E}_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{S}_i\,|\,\boldsymbol{X}_i) = \left(\upsilon^{(t)}+p+q-1\right)\left[\left(\boldsymbol{X}_i-\mathbf{M}^{(t)}\right)\boldsymbol{\Omega}^{(t)^{-1}}\left(\boldsymbol{X}_i-\mathbf{M}^{(t)}\right)^T+\boldsymbol{\Sigma}^{(t)}\right]^{-1}$$

$$\mathbb{E}_{\boldsymbol{\Theta}^{(i)}}(\log|\boldsymbol{S}_i|\,|\,\boldsymbol{X}_i) = \psi_p\left(\frac{\upsilon^{(t)}+p+q-1}{2}\right)+p\log 2+\log\left|\frac{\boldsymbol{S}_i^{(t+1)}}{\upsilon^{(t)}+p+q-1}\right|$$

where $\psi_p(\cdot)$ is the p-variate digamma function – that is, $\psi_p(x) = \frac{d}{dx}\log\Gamma_p(x)$. Further computational and notational reductions are possible by defining and storing the updates in terms of the expected sufficient statistics

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$S_S^{(t+1)} \doteq \sum_{i=1}^{n} S_i^{(t+1)},$$

$$S_{SX}^{(t+1)} \doteq \sum_{i=1}^{n} \mathbb{E}_{\Theta^{(t)}}(S_i X_i \mid X_i) = \sum_{i=1}^{n} S_i^{(t+1)} X_i,$$

$$S_{XSX}^{(t+1)} \doteq \sum_{i=1}^{n} \mathbb{E}_{\Theta^{(t)}}\left(X_i^T S_i X_i \mid X_i\right) = \sum_{i=1}^{n} X_i^T S_i^{(t+1)} X_i,$$

$$S_{|S|}^{(t+1)} \doteq \mathbb{E}_{\Theta^{(t)}}\left[\sum_{i=1}^{n} \log|S_i| \mid X_i\right].$$

These statistics can be expressed with $\left(\upsilon^{(t)} + p + q - 1\right)$ factored out, and for convenience may be computed and stored as such when ν needs to be estimated. These quantities can be computed in $O\left(npq^2\right) + O\left(np^2q\right) + O\left(np^3\right)$ flops.

The M-step updates are split into two conditional maximization steps, one updating ($\mathbf{M}$,$\mathbf{\Sigma}$,$\mathbf{\Omega}$,) and one updating ν. The first step is conceptually immediate and maximizes Equation (S-2) with respect to $\mathbf{\Theta}$ to yield $\mathbf{\Theta}(t+1)$ and with updates of ($\mathbf{M}$,$\mathbf{\Sigma}$,$\mathbf{\Omega}$) given $\upsilon^{(t)}$ as follows:

$$\mathbf{M}^{(t+1)} = \left(\sum_{i=1}^{n} S_i^{(t+1)}\right)^{-1} \sum_{i=1}^{n} S_i X_i = S_S^{(t+1)-1} S_{SX}^{(t+1)}$$

$$\mathbf{\Omega}^{(t+1)} = \frac{1}{np} \sum_{i=1}^{n} \left(X_i - \mathbf{M}^{(t)}\right)^T S_i^{(t+1)} \left(X_i - \mathbf{M}^{(t)}\right) = \frac{1}{np}\left(S_{XSX}^{(t+1)} - S_{SX}^{(t+1)T} S_S^{(t+1)-1} S_{SX}^{(t+1)}\right)$$

$$\mathbf{\Sigma}_{(t+1)}^{-1} = \frac{1}{n\left(\upsilon^{(t)} + p - 1\right)_i} \sum_{i=1}^{n} S_i^{(t+1)} = \frac{S_s^{(t+1)}}{n\left(\upsilon^{(t)} + p - 1\right)}.$$

These quantities can be computed in $O\left(pq^2\right) + O\left(p^2q\right) + O\left(p^3\right)$ flops. As discussed in Section S-1.2.1, if the previous $\mathbf{\Sigma}^{(t)}$ and $\mathbf{\Omega}^{(t)}$ were positive definite, then the updates exist and $\mathbf{\Sigma}$ is positive definite, though the necessary or sufficient conditions for $\mathbf{\Omega}$ to be positive definite (a.s.) are not known.

The conditional maximization of ν given $\left(\mathbf{M}^{(t+1)}, \mathbf{\Sigma}^{(t+1)}, \mathbf{\Omega}^{(t+1)}\right)$ can be sped up substantially by maximizing it instead over the observed log-likelihood function given $\left(\mathbf{M}^{(t+1)}, \mathbf{\Sigma}^{(t+1)}, \mathbf{\Omega}^{(t+1)}\right)$. We get the ML estimating equation (MLEE):

$$n\frac{d}{d\upsilon}\log\Gamma_p((\upsilon + p - 1)/2) - \frac{1}{2}(S_{|S|} - np\log 2 + n\log|\mathbf{\Sigma}|) = 0. \quad (1)$$

Writing $\kappa = \upsilon + p + q - 1$ for notational compactness, we have,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$0 = n\psi_p((\upsilon + p - 1)/2) - \left\{ n\psi_p\left(\frac{\kappa}{2}\right) + \sum_{i=1}^{n} \log\left|\frac{\boldsymbol{S}_i^{(t+1)}}{\kappa}\right| - n\log\left|\frac{\boldsymbol{S}_S^{(t+1)}}{n(\upsilon + p - 1)}\right| \right\}$$

$$^{*} = \psi_p((\nu + p - 1)/2) - \left\{ \psi_p\left(\frac{\kappa}{2}\right) + \frac{1}{n}\sum_{i=1}^{n} \log|\boldsymbol{Z}_i^{(t+1)}| + p\log\frac{n(\nu + p - 1)}{\kappa} - \log|\boldsymbol{Z}_S^{(t+1)}| \right\},$$

where $\boldsymbol{Z}_*^{(t+1)}$ is the appropriate $\boldsymbol{S}_*^{(t+1)}$ statistic with $\left(\upsilon^{(t)} + p + q - 1\right)$ factored out. The MLEE can be solved using a one-dimensional search, yielding an ECME algorithm with the steps:

1. **E-step:** Update $\boldsymbol{S}_i$ weights and statistics based on $\boldsymbol{\Theta}^{(t)}$ and $\boldsymbol{X}_i$.
2. **CME-step:** Update $\boldsymbol{\Theta}_1^{(t+1)} = \left(\mathbf{M}^{(t+1)}, \boldsymbol{\Sigma}^{(t+1)}, \boldsymbol{\Omega}^{(t+1)}\right)$.
3. **CME-step:** Update $\boldsymbol{\Theta}_2^{(t+1)} = \nu^{(t+1)}$ using the observed log-likelihood given $\left(\mathbf{M}^{(t+1)}, \boldsymbol{\Sigma}^{(t+1)}, \boldsymbol{\Omega}^{(t+1)}\right)$.

Repeat these steps until convergence.

As explained in Section S-1.2.1, each iteration of this algorithm takes $\boldsymbol{O}\left(npq^2\right) + \boldsymbol{O}\left(np^2q\right) + \boldsymbol{O}\left(np^3\right)$ flops in addition to the number of iterations required to estimate ν in the second CME step. This suggests that the orientation of the matrices should be chosen such that the row dimension $p < q$, a suggestion that is also experimentally verified in Section 3.1.

We conclude here by noting that restrictions on the parameter set (Roy and Khattree, 2005), such as imposing an mth-order auto-regressive structure (AR(m)) on either $\boldsymbol{\Sigma}$ or $\boldsymbol{\Omega}$ or both as in our applications in Section 3 can be easily incorporated within our algorithm (see Section S-1.2.2).

2.3 Discrimination and Classification

Linear (LDA) and Quadratic Discriminant Analysis (QDA) for matrix-variate populations follow a similar approach as for the multivariate case, with the MxVN (but not MxVt) cases affording substantial reductions in the computations. We provide here the general framework for matrix-variate distributions and then discuss reductions for the cases of the MxVN models.

Suppose that there are two populations $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$, with prior probabilities η_1 and η_2 for an observation belonging to either. Let $\mathbb{P}(1|2)$ be the probability of classifying a member of $\boldsymbol{\pi}_2$ to $\boldsymbol{\pi}_1$ (and vice versa). As usual, the *total probability of misclassification* (TPM) is defined to be $\mathbb{P}(2|1)\eta_1 + \mathbb{P}(1|2)\eta_2$. A Bayes optimal classification rule that minimizes the TPM assigns a matrix-valued observation $\boldsymbol{X}$ to $\boldsymbol{\pi}_1$ if $\frac{f_1(\boldsymbol{X})}{f_2(\boldsymbol{X})} \geq \frac{\eta_2}{\eta_1}$, is the PDF for group $\boldsymbol{\pi}_i$ evaluated

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

at $\boldsymbol{X}$ (Anderson and Bahadur, 1962). The classification rule can be easily extended to the case when there are G groups $\pi_1, \pi_2, \ldots, \pi_G$, each with prior probabilities of membership $\eta_1, \eta_2, \ldots, \eta_G$ and densities $f_1, f_2, \ldots, f_g$. Then the Bayes optimal classification for a matrix-valued observation $\boldsymbol{X}$ is $\text{argmax}_{i \in \{1,2,\ldots,g\}} R_i(\boldsymbol{X})$, where the cost function $R_i(\boldsymbol{X})$ is defined as $\log \eta_i f_i(\boldsymbol{X})$.

Unlike for the MxVt distributions, the MxVN case has closed-form solutions analogous to that of LDA or QDA in multivariate statistics. For the MxVN populations, the closed-form classification rule assigns $\boldsymbol{X}$ to the gth group where $g = \text{argmax}_{i=1,2,\ldots,G} R_i(\boldsymbol{X})$, with

$$R_i(\boldsymbol{X}) = \text{trace}\left\{-\frac{1}{2}\left(\boldsymbol{\Omega}_i^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}\right) + \boldsymbol{\Omega}_i^{-1}\mathbf{M}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X} - \frac{1}{2}\boldsymbol{\Omega}_i^{-1}\mathbf{M}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{M}_i\right\}$$
$$= \frac{1}{2}\left(p\log|\boldsymbol{\Sigma}_i| + q\log|\boldsymbol{\Omega}_i|\right).$$

The first and last term disappear when the G MxVN populations have common covariances, yielding a linear decision rule. Many adaptations (Inoue and Urahama, 2006; Inoue et al., 2011; Li and Yuan, 2005; Lu et al., 2009; Mahanta and Plataniotis, 2014; Yan et al., 2007; Ye et al., 2005; Zhao et al., 2012; Zheng et al., 2008; Molstad and Rothman, 2019) of LDA exist for homogeneous MxVN populations, but our development provides a natural and direct approach that is also flexible enough to include a range of assumptions. Assuming homogeneity does not yield a linear rule for MxVt populations where we still get a quadratic rule. Finally, in all cases, the parameters in $R(\boldsymbol{X})$ can be estimated using ML on the training set (with the ECME methodology of Section 2.2 for MxVt populations) and incorporated into the decision rule.

3. Performance Evaluations

This section evaluates performance of the ECME algorithm in recovering the MxVt parameters and also classification performance of our methodology on some real-life datasets.

3.1 Simulation Study

Our simulation study generated 200 datasets from the $t_{5,3}(\upsilon, \mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$ distribution with $\upsilon = 5, 10, 20$ and $n \in \{35, 50, 100\}$, with the smallest n chosen to be larger than the number of parameters to be estimated, which was also large enough for all but one of the simulations to converge, and the larger sample sizes were chosen to give an idea of consistency of parameter estimation. The ECME algorithm in Section 2.2, with unconstrained ($\mathbf{M}$,$\boldsymbol{\Sigma}$ $\boldsymbol{\Omega}$) was used to estimate the parameters. Figure 1 summarizes the estimated $\hat{\upsilon}$ over the 200 samples for each ν. (We constrain $\hat{\upsilon}$ to be in (2, 1000) in the Either step of the ECME algorithm.)

As expected, higher n improves both accuracy and precision of the estimates. For all nine cases, the peak of the distribution of $\hat{\upsilon}$ was close to the true ν value. Lower values for ν were more easily estimated in the sense that for any n, the $\hat{\upsilon}$ values are closer to the true ν (Figure 1). This may be because for larger true ν, the distributions are similar in a wider

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

range (after all, as $\upsilon \to \infty$, the distribution reduces to the MxVN). For one aberrant sample with $\nu = 20$ and $n = 35$, the optimizer attained the upper bound and did not converge to an interior point. On the whole, however, our simulation results indicate good performance of the ECME algorithm in recovering the MxVt parameters. Additional information about convergence and the recovery of the center and scatter parameters is contained in Section S-1.3.1.

In another simulation study, the speed of the algorithm was demonstrated for $p = 5, 25, 100, q = 5, 25, 100,$ and $N = 100, 500$ with 0 mean and identity spread parameters for $\nu = 5$. Figure 2 summarizes the results of the simulation. As the derivation in Section 2.2 suggests, the row dimension p dominates when determining the speed of the computation. Increasing column dimension q for a given row dimension p seems to hasten convergence to some extent, likely because they act to effectively increase the sample size, which reduces the number of iterations the algorithm needs to run. This suggests the orientation should be chosen so that $p < q$.

Section S-1.3.1 also details results from a a simulation study showing the performance of the method when the type of model (MxVt or MxVN) or degrees of freedom ν are misspecified. We see that models that are fit with ν closer to the true value perform better than those that do not. Also, as expected, the performance using MxVt more closely approaches that under the MxVN model as ν increases.

3.2 Classification Examples

We evaluate MxVt classification and discrimination on four different datasets.

3.2.1 Matching Fractured Surfaces—

Our first example is on the potential ability of our classification algorithm to distinguish between pairs of fractured surfaces into matches or non-matches, with implications in forensics to decide on, say, whether a knife blade fragment found at a crime scene is a match to something that visually appears to be the remainder of the blade. Because of the novelty of this application, we discuss it at some length here. Our investigation is a formal proof-of-concept conducted in the lab where a set of 38 stainless steel knives had their blades broken under similar conditions, resulting in each of them having a base and a tip. The cross-sectional fractured surfaces were then scanned using a standard non-contact 3D optical interferometer at 9 regularly-spaced locations to get 9 successive 1024 × 1024 images (with 75% overlap, in order to get a reasonable number of replications while also imaging the entire length of the exposed surfaces.

Cross-correlations between matching knife base-tip image pairs in the $5 - 10\ \mu m^{-1}$ and $10 - 20\ \mu m^{-1}$ two-dimensional (2D) Fourier frequencies were computed, yielding, for each knife, a 2 × 9 matrix of measurements describing the similarity of the base of the knife to the tip (2 measured cross-correlations per image and 9 images). Similar cross-correlations between all possible knife base-tip pairs (regardless of origin) yielded a sample from the population of similarity matrices coming from known matching (KM) and known non-matching (KNM) base-tip pairs.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Figure 3 shows the scatterplot of the Fisher's Z-transformed (Fisher, 1915) cross-correlation data to be fairly elliptical. The two classes are almost but not completely separated when looking at individual image pairs. Classification using only one pair of images per surface rather than a set of multiple images is potentially ambiguous. We remove this potential ambiguity by considering multiple images on each surface. These multiple sets of images on each knife are not independent and have a natural multivariate repeated measures (*i.e.* matrix-variate) structure because of the 75% overlap between successive images so a model incorporating such structure may improve classification accuracy.

We model each match/non-match dataset in terms of the MxVt distribution with group-specific mean matrix and matrix dispersion structures, with an AR(1) correlation structure for the Fourier domain correlations at the same frequency band between successive (overlapping) image pairs. The AR(1) structure is appropriate because of the overlap between successive images: this correlation structure also has the best Bayesian Information Criterion (BIC) among the correlation structures tested on the data (Schwarz, 1978). The mean across the images for each frequency band was constrained to be constant. Because there are only 9–10 observations for the cases where the knife tip-base have the same origin, we forgo estimating ν and instead investigate classification with the MxVt distribution with $\nu = 5$ and $\nu = 10$ (in addition to the MxVN).

Figure 4 displays the distribution of the log-odds of being a match for the models based on each of the four training sets. The models trained on each set were then tested on the data from all four sets of surfaces. In this figure, a positive log-odds indicates a higher probability of being a KM and a negative log-odds indicates a higher probability of being a KNM.

With equal priors, there is a 0% false exclusion (false negative) rate and a 0.003% false identification (false positive) rate (1 FP). The only FP is from the MxVN model, which is also overly confident about the matches it produces. It predicts some surfaces being a match with log-odds greater than 200, which is extremely implausible. The MxVt distribution accounts for uncertainty better and results in more plausible log-odds ratios. This is because the normal distribution is much more thin in its tails than the t-distribution, and increasing the dimensionality as occurs in the matrix variate case multiplies this effect. This means the MxVN will penalize observations far from the center of the class more than the MxVt. Perfect discrimination is attained with MxVt for all four training sets, suggesting that the results generalize well to out-of-sample data despite the relatively small sample size.

For comparison, we also obtained predictions using the penalized likelihood approach of Molstad and Rothman (2019) which works only when at least two sets of knives are used as training sets, two sets as a validation set for the tuning parameters, and the rest as the test set. We were able to obtain perfect classification for all permutations of the six sets when an appropriate grid of tuning parameters is used (two additional sets of images taken from one set of knives were used to make a total of six sets). However, the method forced the scatter matrices to be diagonal, which is unlikely to be reasonable given the 75% overlap between successive images.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

3.2.2 Finger-tapping Experiment—Maitra et al. (2002) provided 12 functional Magnetic Resonance Imaging (fMRI) scans of the brain of a right-hand dominant male subject during a right-hand finger-thumb opposition activity and 12 similar scans using the left-hand, with each pair of scan collected at regular intervals over a 2-month period. We restrict attention to the 20th slice of the image volume, with 128 × 128 pixels, that previous work (Maitra, 2009, 2010) indicated as adequate to distinguish activation between the left- and right-hand finger-tapping. With only 12 observations per class, we are limited in the types of correlation matrices that we may consider, so we selected a 20 × 20 section of the 20th slice having the left-topmost pixel at (33, 67), which was the 20 × 20 section of the slice displaying the highest average activation in the left-hand activation images as determined by Almodóvar-Rivera and Maitra (2019)'s FAST-fMRI algorithm. We then trained and tested the classifiers using the leave-one-out method with an AR(1) covariance structure and a compound symmetry covariance structure in the MxVN and MxVt distributions with $\nu = 5$ or $\nu = 10$ (for the MxVt). The BIC on the fitted models indicated that a compound symmetry covariance structure was the best model. In all cases, except that of the MxVt distribution with $\nu = 10$ and an AR(1) covariance structure, 23 out 24 images were correctly classified. The one mislabeled case was the same one that was previously identified by Maitra (2010) as an outlier. Using the MxVt distribution with $\nu = 10$ had one more misclassification. The number of cases for this reduced dataset is not enough for MatrixLDA to estimate the correlation structure so we forgo that comparison here.

3.2.3 Landsat Satellite Data—Multi-spectral satellite imagery allows for multiple observations over a spatial grid, yielding matrix-valued observations. We examine a set of satellite images (Dua and Graff, 2017) that are in two visible and two infrared bands. The subset of images under consideration (Viroli, 2011a) consists of a training and a test set of 3 × 3 pixel segments labeled according to the terrain type (961 gray soil, 415 damp gray soil, and 470 soil with vegetation stubble segments, for 1846 total observations in the training set and 397, 311, 237, and 845 total in the test set) of their middle pixel. Each observation, then, is a 9-pixel segment with a label according to soil type, and the problem is to predict the soil type from the pixel values. Regarding the data as a 4 × 9 matrix and with an MxVN classifier and unconstrained covariance matrices yielded an error rate of 0.116 (Viroli, 2011a), while MatrixLDA with tuning parameters selected by 5-fold CV (Molstad and Rothman, 2019) yielded a 0.118 error rate. Our MxVN and MxVt models (the latter with ν = 10 and 20) with unconstrained covariance matrices and prior probabilities equal to the class representation in the training set yielded error rates of 0.126, 0.116, and 0.109, in line with previous results. BIC indicated that using unconstrained covariance matrices and means constrained to be equal within rows as a better model, with error rates of 0.123, 0.121, and 0.107.

3.2.4 Cambridge Hand Gestures Data—We tested our method using leave-one-out cross-validation (LOOCV) on the set of 80 images extracted from the Cambridge hand gestures database (Kim et al., 2007) as processed by Molstad and Rothman (2019) into 80 × 60 pixel gray-scale images. There are four classes in this problem: the images show a hand gesture in one of two shapes and one of two orientations: in each image, the hand is either in a flat or "V" shape and is located either in the center of the image or to the left side of the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

image. We fit models with an AR(1) structure on both dimensions, compound symmetric structure on both, and an unconstrained covariance structure, with 5 and 10 degrees of freedom for the MxV*t* distribution. The AR(1) structure provided the best fit according to BIC, and by using it we were able to obtain a 100% classification rate using LOOCV on the dataset. Molstad and Rothman (2019) report a 90% correct classification rate using LOOCV on this dataset.

4 Conclusions

We have provided an ECME method for fitting the parameters of the MxV*t* distribution that can be used on three-way data sets such as multivariate repeated measures, image or spatial data, and have demonstrated the method on simulation datasets and on classification and discrimination in four real-world applications where the new method using the MXV*t*-distribution outperforms that using the MxVN. The ECME algorithm and the discriminant analysis are implemented in the R package MixMatrix. The package also includes functions for sampling from and computing the density of the MxVN and MxV*t* distributions and includes the datasets used in this paper.

Our model can be extended beyond supervised learning to mixture model-based clustering and can be made to accommodate more specialized covariance structures such as those described in Fraley and Raftery (2002) and Andrews et al. (2018). It may also be readily extended to cases with incomplete records. Determining the existence, convergence and uniqueness properties would also be desirable. For instance, we know how many observations are required to have unique ML estimates of the parameters in the MxVN distribution with unconstrained mean and covariance matrices but such results may be useful to develop for the MxV*t* or the constrained MxVN. Nevertheless, the EM algorithm is guaranteed to converge to a local stationary point, provided it is initialized where the log-likelihood function is finite (Wu, 1983). Finally, another area that could benefit from further development is the extension of MatrixLDA to include the MxV*t* distribution, where we believe our development in this paper will be helpful.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank the anonymous reviewer and the associate editor whose detailed comments and suggestions greatly improved the quality of this article.

Funding Information

This research was supported in part by the National Institute of Justice (NIJ) under Grants No. 2015-DN-BX-K056 and 2018-R2-CX-0034. 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 R21EB016212, and the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) Hatch project IOW03617. 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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

References

Almodóvar-Rivera I and Maitra R (2019), "Fast adaptive smoothing and thresholding for improved activation detection in low-signal fMRI," IEEE Transactions on Medical Imaging, 1–1. DOI: 10.1109/TMI.2019.2915052.

Anderlucci L and Viroli C (2015), "Covariance pattern mixture models for the analysis of multivariate heterogeneous longitudinal data," The Annals of Applied Statistics, 9, 777–800. DOI: 10.1214/15-AOAS816.

Anderson TW and Bahadur RR (1962), "Classification into two Multivariate Normal Distributions with Different Covariance Matrices," The Annals of Mathematical Statistics, 33, 420–431. DOI: 10.1214/aoms/1177704568.

Andrews J, Wickins J, Boers N, and McNicholas P (2018), "teigen: An R Package for model-based clustering and classification via the multivariate t distribution," Journal of Statistical Software, Articles, 83, 1–32. DOI: 10.18637/jss.v083.i07.

Dempster AP, Laird NM, and Rubin DB (1977), "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society. Series B (methodological), 1–38. DOI: 10.1111/j.25176161.1977.tb01600.x.

Dickey JM (1967), "Matricvariate generalizations of the multivariate t distribution and the inverted multivariate t distribution," The Annals of Mathematical Statistics, 38, 511–518. DOI: 10.1214/aoms/1177698967.

Dua D and Graff C (2017), "UCI Machine Learning Repository," http://archive.ics.uci.edu/ml.

Fisher RA (1915), "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population," Biometrika, 10, 507–521. DOI: 10.1093/biomet/10.4.507.

Fraley C and Raftery AE (2002), "Model-based clustering, discriminant analysis, and density estimation," Journal of the American Statistical Association, 97, 611–631. DOI: 10.1198/016214502760047131.

Gupta AK and Nagar DK (1999), Matrix Variate Distributions, vol. 104, CRC Press.

Inoue K, Hara K, and Urahama K (2011), "Non-iterative symmetric two-dimensional linear discriminant analysis," IEICE Transactions on Information and Systems, E94.D, 926–929. DOI: 10.1587/transinf.E94.D.926.

Inoue K and Urahama K (2006), "Non-Iterative Two-Dimensional Linear Discriminant Analysis," in 18th International Conference on Pattern Recognition (ICPR'06), 540–543. DOI: 10.1109/ICPR.2006.860.

Iranmanesh A, Arashi M, and Tabatabaey S (2010), "On conditional applications of matrix variate normal distribution," Iranian Journal of Mathematical Sciences and Informatics. DOI: 10.7508/ijmsi.2010.02.004.

Kim T, Wong S, and Cipolla R (2007), "Tensor canonical correlation analysis for action classification," in 2007 IEEE Conference on Computer Vision and Pattern Recognition, 1–8. DOI: 10.1109/CVPR.2007.383137.

Li M and Yuan B (2005), "2D-LDA: A Statistical Linear Discriminant Analysis for Image Matrix," Pattern Recognition Letters, 26, 527–532. DOI: 10.1016/j.patrec.2004.09.007.

Liu C and Rubin DB (1994), "The ECME Algorithm: A Simple Extension of EM and ECM with Faster Monotone Convergence," Biometrika, 81, 633–648. DOI: 10.1093/biomet/81.4.633.

Lu H, Plataniotis KN, and Venetsanopoulos AN (2009), "Uncorrelated multilinear discriminant analysis with regularization and aggregation for tensor object recognition," IEEE Transactions on Neural Networks, 20, 103–123. DOI: 10.1109/TNN.2008.2004625. [PubMed: 19095532]

Mahanta MS and Plataniotis KN (2014), "Ranking 2DLDA features based on Fisher discriminance," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 8307–8311. DOI: 10.1109/ICASSP.2014.6855221.

Maitra R (2009), "Assessing certainty of activation or inactivation in test-retest fMRI studies," Neuroimage, 47, 88–97. DOI: 10.1016/j.neuroimage.2009.03.073. [PubMed: 19361562]

—(2010), "A re-defined and generalized percent-overlap-of-activation measure for studies of fMRI reproducibility and its use in identifying outlier activation maps," NeuroImage, 50, 124–135. DOI: 10.1016/j.neuroimage.2009.11.070. [PubMed: 19963068]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Maitra R, Roys SR, and Gullapalli RP (2002), "Test-retest reliability estimation of functional MRI data," Magnetic Resonance in Medicine, 48, 62–70. DOI: 10.1002/mrm.10191. [PubMed: 12111932]

Meng X-L and Rubin DB (1993), "Maximum likelihood estimation via the ECM algorithm: A general framework," Biometrika, 80, 267–278. DOI: 10.1093/biomet/80.2.267.

Molstad AJ and Rothman AJ (2019), "A Penalized Likelihood Method for Classification With Matrix-Valued Predictors," Journal of Computational and Graphical Statistics, 28, 11–22. DOI: 10.1080/10618600.2018.1476249.

R Core Team (2019), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

Roy A and Khattree R (2005), "On discrimination and classification with multivariate repeated measures data," Journal of Statistical Planning and Inference, 134, 462–485. DOI: 10.1016/j.jspi.2004.04.012.

Schwarz G (1978), "Estimating the Dimension of a Model," The Annals of Statistics, 6, 461–464. DOI: 10.1214/aos/1176344136.

Soloveychik I and Trushin D (2016), "Gaussian and robust Kronecker product covariance estimation: Existence and uniqueness," Journal of Multivariate Analysis, 149, 92–113. DOI: 10.1016/j.jmva.2016.04.001.

Thompson G (2019), MixMatrix: Classification with Matrix Variate Normal and t Distributions, R package version 0.2.2.

Viroli C (2011a), "Finite mixtures of matrix normal distributions for classifying three-way data," Statistics and Computing, 21, 511–522. DOI: 10.1007/s11222-010-9188-x.

Viroli C (2011b), "Model based clustering for three-way data structures," Bayesian Analysis, 6, 573–602. DOI: 10.1214/11-BA622.

Wu CFJ (1983), "On the Convergence Properties of the EM Algorithm," The Annals of Statistics, 11, 95–103. DOI: 10.1214/aos/1176346060.

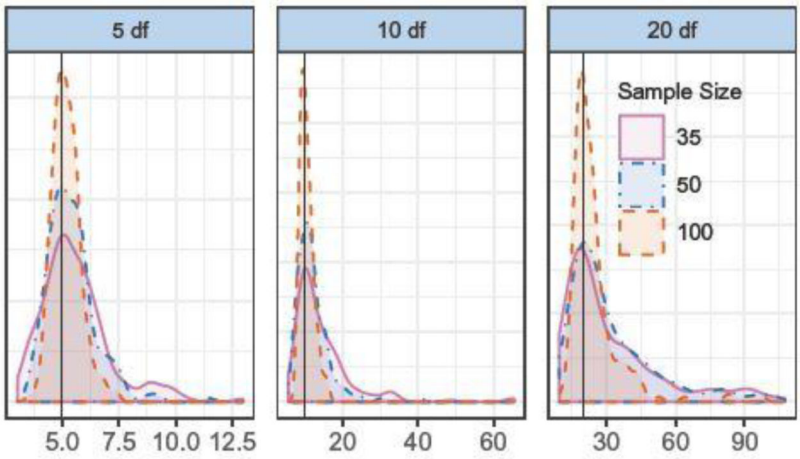
Yan S, Xu D, Yang Q, Zhang L, Tang X, and Zhang H-J (2007), "Multilinear Discriminant Analysis for Face Recognition," IEEE Transactions on Image Processing, 16.

Ye J, Janardan R, and Li Q (2005), "Two-dimensional linear discriminant analysis," in Advances in Neural Information Processing Systems 17 - Proceedings of the 2004 Conference, NIPS 2004, Neural information processing systems foundation.

Zhao J, Yu PL, Shi L, and Li S (2012), "Separable linear discriminant analysis," Computational Statistics and Data Analysis, 56, 4290–4300. DOI: 10.1016/j.csda.2012.04.003.

Zheng W-S, Lai J-H, and Li SZ (2008), "1D-LDA vs. 2D-LDA: When is vector-based linear discriminant analysis better than matrix-based?" Pattern Recognition, 41, 2156–2172.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



(a)

ν	n	Range	Median	Mean	SD
5	35	(2.62, 15.12)	5.24	5.45	1.43
	50	(3.46, 12.78)	5.32	5.45	1.28
	100	(3.76, 7.01)	5.14	5.18	0.56
10	35	(5.41, 395.12)	11.57	16.28	28.26
	50	(5.28, 106.72)	10.44	11.94	7.91
	100	(6.99, 18.10)	10.19	10.56	1.84
20	35	(9.93, 999.83)	29.94	89.01	149.12
	50	(11.38, 495.67)	24.45	46.97	73.09
	100	(12.68, 147.67)	21.98	25.52	14.91

(b)

Fig. 1.

(a) Density plots and (b) numerical summaries of $\hat{v}$ for datasets of size $n = 35, 50, 100$ with true $v = 5,10,20$ (vertical line).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

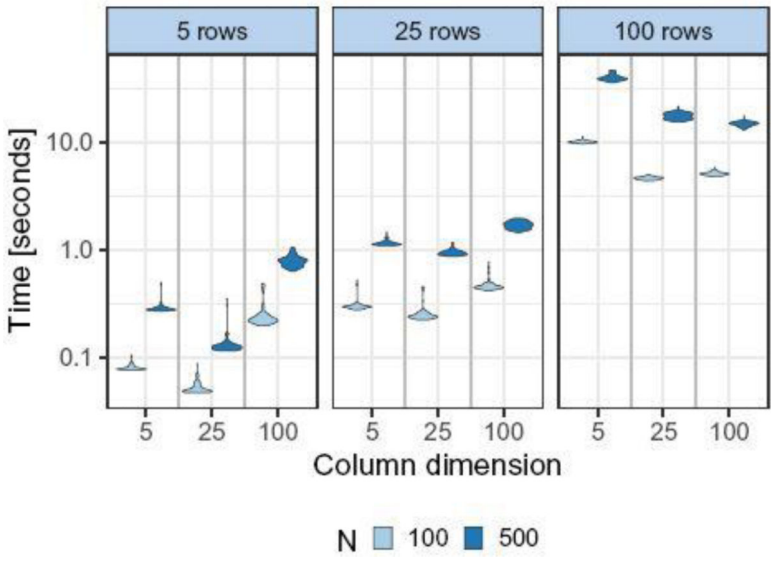


Fig. 2.

Run times for 100 repetitions of the proposed MxV*t* estimation procedure for $p = 5, 25, 100$, $q = 5, 25, 100$, and $N = 100, 500$ with $\nu = 5$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

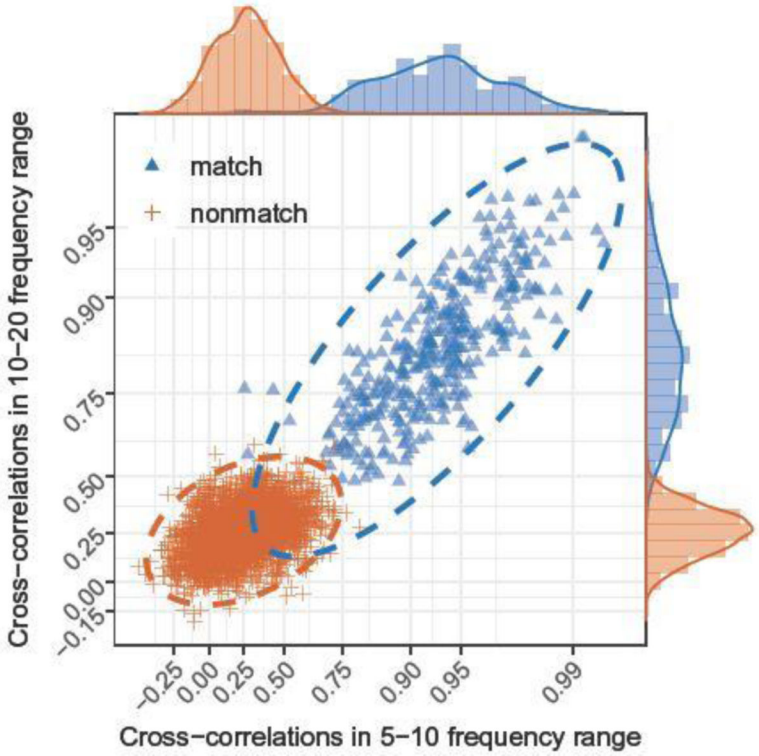


Fig. 3.

Cross-correlations for individual images, along with 99% confidence ellipses under bivariate normal assumptions, in known matching (KM) and known non-matching (KNM) surfaces on Fisher-transformed axes. Known matches and known non-matches can be distinguished, but not perfectly, in this example by features in these two frequency ranges.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

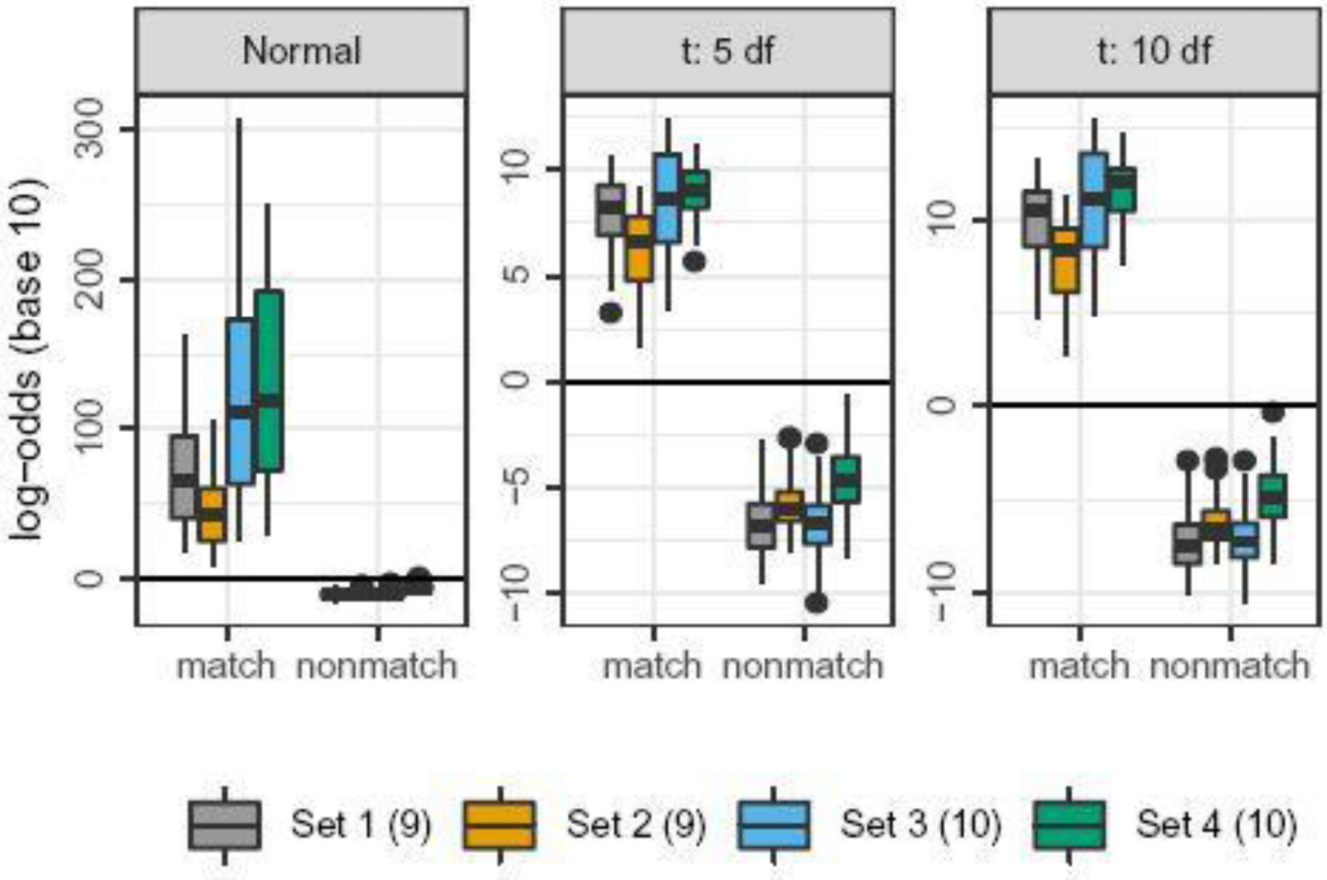


Fig. 4.

Positive results indicate match is more probable than non-match. There are 9 known matches and 72 known non-matches in Set 1 and Set 2 and 10 known matches and 90 known non-matches in Sets 3 and 4. The results are from training on the indicated set and testing on all four sets.