

1 **Short Letter on the Semiparametric**
2 **Transformation Discriminant Analysis**

3 Jing-Hao Xue^{a,*}, D. Michael Titterington^a

4 ^a*Department of Statistics, University of Glasgow, Glasgow G12 8QQ, UK*

5 **Abstract**

6 The purpose of this letter is to comment that, for the semiparametric transformation
7 discriminant analysis (Lin and Jeon, 2003), the approximate method-of-moments es-
8 timators or approximate maximum likelihood estimators are not necessary to esti-
9 mate the mean, and for the linear discriminant case the variance, of the transformed
10 negative class.

11 *Key words:* Normal copulas; Semiparametric transformation discriminant analysis
12 (TDA)

13 **1 Estimation Methods**

14 Lin and Jeon (2003) proposed a semiparametric, normal-based transforma-
15 tion discriminant analysis (TDA), which performed normal-based linear and
16 quadratic discriminant analysis after transforming the continuous-valued data

* Corresponding author. Tel.: +44 141 330 2474; fax: +44 141 330 4814.

Email addresses: jinghao@stats.gla.ac.uk (Jing-Hao Xue),
mike@stats.gla.ac.uk (D. Michael Titterington).

1 to become normally distributed. The data are denoted by a d -variate ran-
2 dom vector $\mathbf{X} = (X_1, \dots, X_d)^T$; the transformation, if it exists, is a vector
3 of univariate strictly monotone functions $\mathbf{g} = (g_1, \dots, g_d)^T$ such that $\mathbf{g}(\mathbf{X}) =$
4 $(g_1(X_1), \dots, g_d(X_d))^T$ is normally distributed with mean $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)^T$,
5 correlation matrix $Q = (\rho_{jk})$ and $\text{var}\{g_j(X_j)\} = \sigma_j^2$, $j, k = 1, \dots, d$. Subse-
6 quently, as shown in Equation (6) in Lin and Jeon (2003), the log density
7 $\log p(\mathbf{X} = \mathbf{x})$ can be written as

$$\log p(\mathbf{X} = \mathbf{x}) = \sum_{j=1}^d \log f_j(x_j) - \frac{1}{2} \mathbf{z}^T (Q^{-1} - I) \mathbf{z} - \frac{\log \det(Q)}{2}, \quad (1)$$

8 where $\mathbf{x} = (x_1, \dots, x_d)^T$, \mathbf{z} is the corresponding value of a random vector
9 $\mathbf{Z}(\mathbf{X}) = (Z_1, \dots, Z_d)^T$ at $\mathbf{X} = \mathbf{x}$ such that

$$\mathbf{z} = (z_1, \dots, z_d)^T = \left(\frac{g_1(x_1) - \mu_1}{\sigma_1}, \dots, \frac{g_d(x_d) - \mu_d}{\sigma_d} \right)^T, \quad (2)$$

10 and the log marginal density $\log f_j(x_j)$ is

$$\log f_j(x_j) = -\frac{\log 2\pi}{2} - \log \sigma_j + \log |g'_j(x_j)| - \frac{1}{2} \left(\frac{g_j(x_j) - \mu_j}{\sigma_j} \right)^2. \quad (3)$$

11 The two classes are named the positive class $Y = 1$ and the negative class
12 $Y = -1$. As assumed in Lin and Jeon (2003), the same functions \mathbf{g} transform
13 $\mathbf{X}_+ = \{\mathbf{X}|Y = 1\}$ and $\mathbf{X}_- = \{\mathbf{X}|Y = -1\}$ to be normally distributed with pa-
14 rameters (μ_+, σ_+, Q_+) and (μ_-, σ_-, Q_-) , respectively. Therefore, for a new ob-
15 servation \mathbf{x} , the discriminant function $\lambda(\mathbf{x}) = \log \frac{p(Y=1|\mathbf{x})}{p(Y=-1|\mathbf{x})} = \log \frac{w_+ p(\mathbf{x}|Y=1)}{w_- p(\mathbf{x}|Y=-1)}$,
16 where $w_+ = p(Y = 1)$ and $w_- = p(Y = -1)$, can be written, by the cancella-
17 tion of common terms, as

$$\lambda(\mathbf{x}) = \log \frac{w_+}{w_-} - \sum_{j=1}^d \log \frac{\sigma_{+j}}{\sigma_{-j}} - \frac{\mathbf{z}_+^T Q_+^{-1} \mathbf{z}_+ - \mathbf{z}_-^T Q_-^{-1} \mathbf{z}_-}{2} - \frac{1}{2} \log \frac{\det(Q_+)}{\det(Q_-)}, \quad (4)$$

18 where \mathbf{z}_+ and \mathbf{z}_- are \mathbf{Z} evaluated at $\mathbf{X} = \mathbf{x}$ for the positive class $Y = 1$ and
19 the negative class $Y = -1$, respectively.

1 The classical normal-based linear and quadratic discriminant analysis (LDA
 2 and QDA) can be viewed as special cases of normal-based TDA, where $g_j(x_j) =$
 3 x_j , for all $j = 1, \dots, d$.

4 From Equation (4), we observe that the parameters to be estimated are $\frac{w_+}{w_-}$,
 5 $\frac{\sigma_{+j}}{\sigma_{-j}}$, Q_+ , Q_- , \mathbf{z}_+ and \mathbf{z}_- .

6 In general, $\frac{w_+}{w_-}$ is estimated by the ratio of proportions $\frac{n_+}{n_-}$, where n_+ and n_-
 7 are the numbers of observations for the positive and negative classes in the
 8 training set, respectively.

9 Lin and Jeon (2003) set $\mu_+ = (0, \dots, 0)^T$ and $\sigma_+ = (1, \dots, 1)^T$ such that
 10 $g_j = \Phi^{-1} \circ F_{+j} = (\Phi^{-1} \circ F_{-j})\sigma_{-j} + \mu_{-j}$, where $j = 1, \dots, d$, \circ denotes the
 11 composition of functions, and F_{+j} and F_{-j} are cumulative distribution func-
 12 tions (CDFs) of X_{+j} and X_{-j} , respectively (Lin and Jeon, 2003). Based on the
 13 empirical CDF \tilde{F}_{+j} , they proposed robust, computationally fast, approximate
 14 method-of-moments estimators (A-MME) and approximate maximum likeli-
 15 hood estimators (A-MLE) to estimate μ_- and σ_- , and then used probability
 16 quantiles of a normal mixture distribution to estimate $g_j(x_j)$.

17 Given that $Z_j = \frac{g_j(X_j) - \mu_j}{\sigma_j} \sim \mathcal{N}(0, 1)$, it follows that, for the new observation
 18 \mathbf{x} ,

$$z_{+j} = \Phi^{-1}(\Phi(z_{+j})) = \Phi^{-1}(F_{+j}(x_j)) , \text{ and } z_{-j} = \Phi^{-1}(F_{-j}(x_j)) , \quad (5)$$

19 and thus straightforward estimators of z_{+j} and z_{-j} are

$$\hat{z}_{+j} = \Phi^{-1}(\tilde{F}_{+j}(x_j)) , \hat{z}_{-j} = \Phi^{-1}(\tilde{F}_{-j}(x_j)) . \quad (6)$$

20 The correlation matrices Q_+ and Q_- in Equation (4) are to be estimated from
 21 the training set, which includes observations $\{\mathbf{x}_+^i\}_{i=1}^{n_+}$ from the positive class

1 and $\{\mathbf{x}_-^i\}_{i=1}^{n_-}$ from the negative class. Lin and Jeon (2003) proposed a robust
 2 semiparametric estimator of Q_+ and Q_- based on the estimation of μ_- , σ_-
 3 and $g_j(x_j)$.

4 Given that, for all $j, k = 1, \dots, d$,

$$\rho_{+jk} = \text{corr}\{g_j(X_{+j}), g_k(X_{+k})\} = \text{corr}\{Z_{+j}, Z_{+k}\}, \quad (7)$$

5 Q_+ can be estimated by the sample correlation matrix of \mathbf{Z}_+ using $\{\mathbf{x}_+^i\}_{i=1}^{n_+}$.
 6 Similarly, $\{\mathbf{x}_-^i\}_{i=1}^{n_-}$ can be used to estimate \mathbf{z}_- , and then Q_- .

7 In summary, in order to obtain the discriminant function $\lambda(\mathbf{x})$, given the esti-
 8 mates of \mathbf{z} for the positive, negative and new observations, it is not necessary
 9 to estimate μ_+ , μ_- and $g_j(x_j)$. Also, for the linear discriminant case where
 10 $\sigma_- = \sigma_+$, the estimation of σ_- or σ_+ is always ignored. Here, we denote by
 11 zLDA such a normal-based TDA for linear discrimination that is based on the
 12 estimate of \mathbf{z} .

13 In practice the parameter estimation for zLDA can be simple, but, the predic-
 14 tive performance of zLDA is subject to the accuracy of estimates of F_{+j} and
 15 F_{-j} , and thus to the size of training sets. In other words, the out-of-sample
 16 predictive performance of zLDA can be inferior to that of the TDA of Lin
 17 and Jeon (2003). Lin and Jeon (2003) estimated F_{+j} , which is the CDF of the
 18 class with more training observations amongst the two classes, and the CDF
 19 of a mixture of normals; without being subject to the estimation inaccuracy
 20 for F_{-j} from fewer training observations, it may offer more robust prediction
 21 than zLDA.

1 2 Relation with Normal Copulas

2 As pointed out by Lin and Jeon (2003), the transformation by \mathbf{g} is equivalent
 3 to the use of multivariate normal copulas to describe the dependence structure
 4 between the marginal distributions of the X_j . When \mathbf{X} is continuous-valued,
 5 according to Sklar's theorem (Joe, 1997; Nelsen, 1999), the joint CDF $F(\mathbf{x})$
 6 can be represented using copula \mathcal{C} as

$$F(\mathbf{x}) = \mathcal{C}(F_1(x_1), \dots, F_d(x_d)), \quad (8)$$

7 and, when the CDFs are differentiable, the joint density $p(\mathbf{x})$ can be written
 8 as

$$p(\mathbf{X} = \mathbf{x}) = c(F_1(x_1), \dots, F_d(x_d)) \prod_{j=1}^d f_j(x_j), \quad (9)$$

9 where $c(F_1(x_1), \dots, F_d(x_d))$ is called the copula density. A multivariate normal
 10 copula is defined as

$$\mathcal{C}(F_1(x_1), \dots, F_d(x_d); Q) = \Phi_Q(\Phi^{-1}(F_1(x_1)), \dots, \Phi^{-1}(F_d(x_d))), \quad (10)$$

11 and its density can be written as

$$c(F_1(x_1), \dots, F_d(x_d); Q) = (\det(Q))^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \mathbf{z}^T (Q^{-1} - I) \mathbf{z} \right\}, \quad (11)$$

12 where $\mathbf{z} = (\Phi^{-1}(F_1(x_1)), \dots, \Phi^{-1}(F_d(x_d)))^T$ (Nelsen, 1999). If we use the nor-
 13 mal copula, it follows that the log density $\log p(\mathbf{X} = \mathbf{x})$ is the same as that in
 14 Equation (1).

15 As Lin and Jeon (2003) pointed out, the normal-based TDA can be seen as
 16 a generalisation of the additive log density model such as the naïve Bayes
 17 classifier, which assumes conditional independence of the X_j given the class
 18 label Y and thus only uses the first term in Equation (1). The TDA regu-

1 larises the naïve Bayes classifier with terms for dependence structure, which
2 corresponds to the multivariate normal copula. From the perspective of copu-
3 las, other copulas can also be tried to model the dependence structure, which
4 may correspond to TDAs that transform the data to follow other distributions
5 before using the Bayes' classification criterion.

6 A copula-based parametric estimator called multi-stage MLE, which is used to
7 estimate the parameters of the marginal distributions, such as μ_- and σ_- , and
8 the parameters of the copula, such as Q , may be found in Patton (2006), where
9 the asymptotic efficiency is justified under certain conditions. Both Lin and
10 Jeon (2003) and Patton (2006) estimated parameters through a multi-stage
11 approach.

12 **3 Empirical Studies**

13 Lin and Jeon (2003) provided empirical studies on two simulated and six real
14 datasets; the performance of the normal-based TDA is encouraging, compared
15 to other classifiers such as the widely-used normal-based linear discriminant
16 analysis (LDA). In this section, we use the same datasets to compare the
17 performance of four linear discrimination approaches, where $\sigma_+ = \sigma_-$ and
18 $Q_+ = Q_-$ are assumed, based on four different estimation methods, respec-
19 tively.

20 The first approach (denoted by $TDA_1.MM-1$ hereafter) is to use the robust,
21 semiparametric, approximate method-of-moments estimators of μ_- , σ_- , $g_j(x_j)$
22 and then Q_+ and Q_- , which assumes $\mu_+ = (0, \dots, 0)^T$ and $\sigma_+ = (1, \dots, 1)^T$.
23 It uses Equation (10) in Lin and Jeon (2003) as the discriminant function.

1 The second approach (denoted by TDA₁.MM-2 hereafter) is even more robust
 2 than TDA₁.MM-1 since the former re-estimate μ_- , σ_- , μ_+ and σ_+ (and hence
 3 the assumptions that $\mu_+ = (0, \dots, 0)^T$ and $\sigma_+ = (1, \dots, 1)^T$ are not neces-
 4 sarily satisfied and then Equation (10) in Lin and Jeon (2003) is not used),
 5 which is the default method used by Lin and Jeon (2003) for the reported
 6 results according to their source code.

7 The third approach (zLDA) is to use estimators of \mathbf{z}_+ , \mathbf{z}_- and then Q_+ and
 8 Q_- , which provides, as mentioned in Section 1 and 2, a normal-copula-based
 9 estimator. As with that in Lin and Jeon (2003), the estimation of the CDFs
 10 F_{+j} and F_{-j} is performed through linearly interpolating the empirical CDFs
 11 \tilde{F}_{+j} and \tilde{F}_{-j} , where $\tilde{F}_{-j}(x_{-j}) = \frac{k-0.5}{n_-}$, in which the observation x_{-j} is the k -th
 12 smallest amongst the training observations of the negative class, $\tilde{F}_{-j}(x_j) =$
 13 $\frac{n_-+0.5}{n_-+1}$ for a new observation x_j larger than the largest training observation
 14 and $\tilde{F}_{-j}(x_j) = \frac{0.5}{n_-+1}$ for x_j smaller than the smallest training observation.
 15 Similarly, $\tilde{F}_{+j}(x_{+j})$ is obtained.

16 Since the performance of the third approach is subject to the accuracy of esti-
 17 mation of the CDFs, we suggest the fourth approach (denoted by zLDA-Kernel
 18 hereafter): it is all the same as the third approach, except that a kernel-based
 19 estimator of F_{+j} and F_{-j} is used instead. We use kernel density estimates with
 20 a Gaussian Kernel and normal reference bandwidth; the smoothing bandwidth
 21 is determined by the rule of thumb in Silverman (1986), the default in an R
 22 function *density* from the R package **stats**. Then the probabilities of observa-
 23 tions are obtained through linear interpolation.

24 For the last two approaches, the estimate $\hat{Q} = \frac{n_+\hat{Q}_++n_-\hat{Q}_-}{n_++n_-}$ is the pooled esti-
 25 mate of the correlation matrix.

1 The data in both simulated datasets arise from two 7-variate normal distribu-
2 tions, one representing the positive class $Y = 1$ and the other representing the
3 negative class $Y = -1$. The two multivariate normal distributions have the
4 same covariance matrices but different means. One dataset (denoted by S-100
5 hereafter) includes 100 training observations and the other one (denoted by
6 S-1K hereafter) includes 1000 training observations; for each dataset, there
7 are 1000 test observations and two-thirds of the observations are from $Y = 1$.

8 The six real datasets are named BCW, BCW+, BLD, BLD+, PID and PID+,
9 amongst which BCW, BLD and PID are datasets from the UCI machine learn-
10 ing repository (Newman et al., 1998) and BCW+, BLD+ and PID+ are corre-
11 sponding noisy versions with different numbers of independent noise predictor
12 variables added (see Lin and Jeon (2003) for details of the datasets).

13 As was done in Lin and Jeon (2003), for the simulated datasets, misclassi-
14 fication error rates are averaged over 100 simulations; for the real datasets,
15 error rates are averaged over 100 realizations of ten-fold cross-validation. The
16 mean error rates are listed in Table 1. Besides the above four approaches under
17 study, Table 1 also lists the results reported by Lin and Jeon (2003) for LDA,
18 QDA and the naïve Bayes classifier (NB) as reference points. As a result of
19 random simulation and random partition for the ten-fold cross-validation, the
20 results of TDA₁.MM-2 are slightly different from those reported in Lin and
21 Jeon (2003).

22 From Table 1, we observe the following.

23 (1) For the simulated datasets, zLDA and zLDA-Kernel perform worse than
24 TDA₁.MM-2 and TDA₁.MM-1; their predictive performance is similar to
25 those of QDA and NB when the training set is small; all the classifiers

Table 1

The mean error rates obtained from different linear discrimination methods on two simulated and six real datasets. (*: results from Lin and Jeon (2003) for LDA, QDA and NB)

Dataset	TDA ₁ .MM-1	TDA ₁ .MM-2	zLDA	zLDA-Kernel	LDA*	QDA*	NB*
S-100	0.242	0.241	0.313	0.284	0.246	0.282	0.312
S-1K	0.225	0.224	0.243	0.236	0.222	0.225	0.281
BCW	0.147	0.027	0.037	0.048	0.040	0.049	0.038
BCW+	0.150	0.029	0.041	0.048	0.040	0.049	0.038
BLD	0.310	0.274	0.269	0.267	0.319	0.401	0.354
BLD+	0.337	0.284	0.303	0.307	0.322	0.378	0.370
PID	0.312	0.228	0.229	0.224	0.221	0.240	0.227
PID+	0.312	0.222	0.232	0.224	0.219	0.236	0.229

- 1 improve their performance when the size of training set increases.
- 2 (2) For the datasets BCW and BCW+, the performance of zLDA and zLDA-
- 3 Kernel is similar to those of LDA, QDA and NB. Methods zLDA and
- 4 zLDA-Kernel perform worse than TDA₁.MM-2 but much better than
- 5 TDA₁.MM-1.
- 6 (3) For the datasets BLD, PID and PID+, zLDA and zLDA-Kernel perform
- 7 comparably to TDA₁.MM-2 and much better than TDA₁.MM-1.

1 **References**

- 2 Joe, H., 1997. *Multivariate Models and Multivariate Dependence Concepts*.
3 London : Chapman & Hall.
- 4 Lin, Y., Jeon, Y., 2003. Discriminant analysis through a semiparametric
5 model. *Biometrika* 90 (2), 379–392.
- 6 Nelsen, R. B., 1999. *An Introduction to Copulas*. New York: Springer.
- 7 Newman, D. J., Hettich, S., Blake, C. L., Merz, C. J., 1998. UCI
8 Repository of machine learning databases. University of Cal-
9 ifornia, Irvine, Dept. of Information and Computer Sciences,
10 <http://www.ics.uci.edu/~mllearn/MLRepository.html>.
- 11 Patton, A. J., 2006. Estimation of multivariate models for time series of pos-
12 sibly different lengths. *Journal of Applied Econometrics* 21 (2), 147–173.
- 13 Silverman, B. W., 1986. *Density Estimation for Statistics and Data Analysis*.
14 London : Chapman & Hall.