PARAMETER ESTIMATION OF LINEAR CLASSIFIERS UNDER HIGH-DIMENSIONAL ASYMPTOTICS – Supplementary Materials

Mengyi Zhang

I. Unbalanced number of samples

The scenario that the number of samples of class 1 is different from that of class 2 is considered. Here it is assumed that $\frac{n_1}{n_2} = \frac{\pi_1}{\pi 2} = 2$. The other settings are the same as Section 4.1 of the paper. The results are plotted in Figure S1.

Let $p = 60$, $n_2 = 30$, $n_1 = 2n_2$, $\Delta^2 = 6$, $\rho \in [0,\ 3]$, the performance of true error, proposed error estimator, and conventional plug-in error estimator is plotted in Figure S1a. It can be seen that the conventional plug-in error estimator with three types of $\mathbf{R}_0$ underestimate the error seriously for all values of ρ. The true error with $\rho = 0$ corresponds to the error of traditional LDA classifier, hence it can be seen that the traditional LDA is suboptimal in this scenario. With each value of ρ, the proposed error estimator is close to the corresponding true error for all the three types of $\mathbf{R}_0$. Let $p = 60$, $\Delta^2 = 6$, $\rho = 1$, $n \in [60,\ 120]$, the performance of true error, proposed error estimator, and conventional plug-in error estimator is plotted in Figure S1b. It can be seen that the conventional plug-in error estimator with three types of $\mathbf{R}_0$ underestimate the error seriously for all values of n. For each type of $\mathbf{R}_0$, the proposed error estimator is close to the corresponding true error, and the gaps become even smaller as n increases.

The different discriminant approaches with different covariance matrix estimators and optimal shrinkage parameter are then compared. It can be seen in Figure S1c that as the number of samples increases, the misclassification error of all the classifiers decrease. The proposed GLDA classifier outperforms the other classifiers, and its performance is very close to the lower bound of CLV classifier. In Figure S1d, the proposed GLDA classifier outperforms the other classifiers for all values of p, and the gaps between the error of GLDA classifier and the error of CLV classifier become smaller as p increases. Therefore, the proposed covariance matrix estimator with general shrinkage matrix is asymptotically as good as the clairvoyant estimator in the scenario of unbalanced number of samples.

II. Effect of Gaussian AR parameter

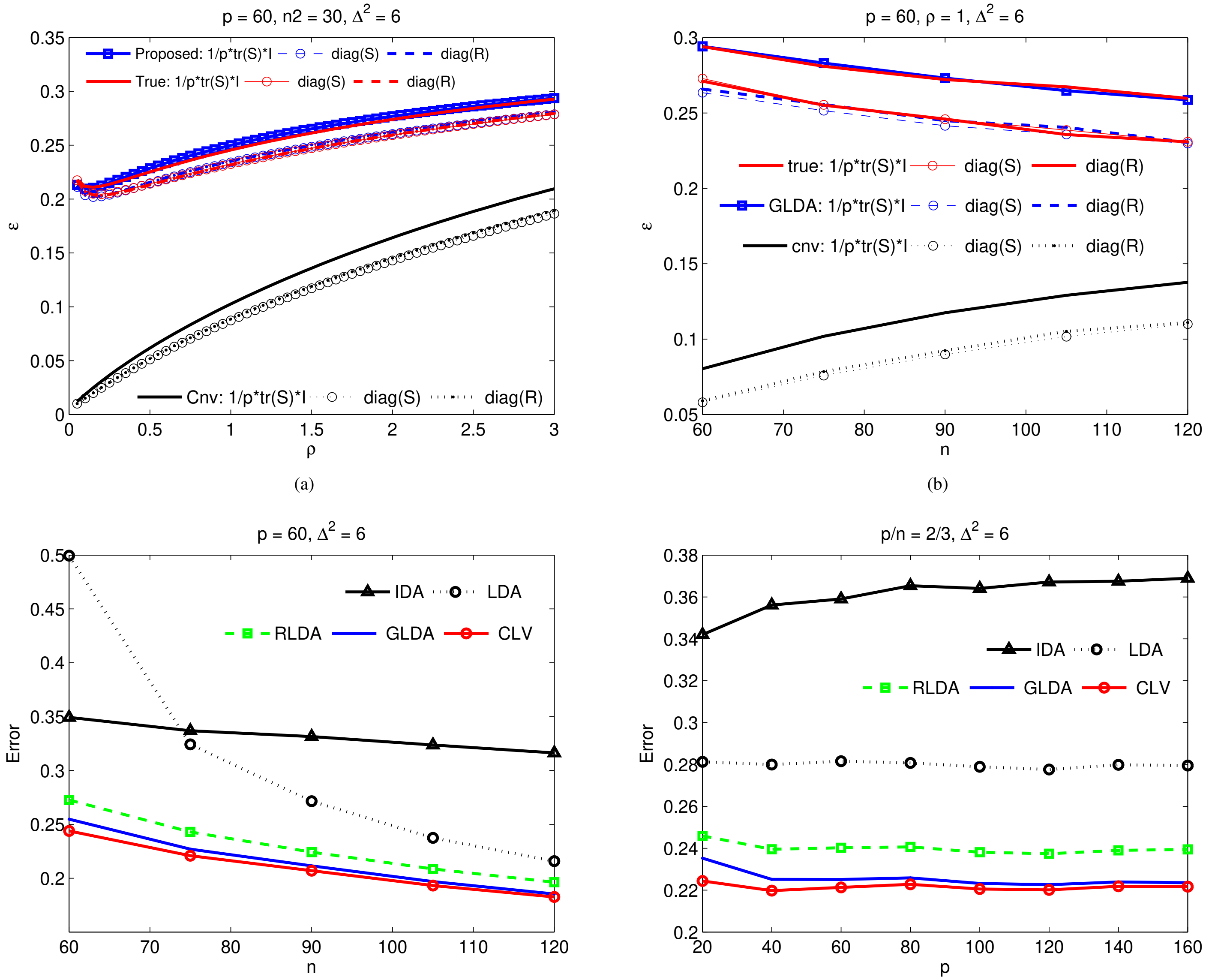
The effect of Gaussian AR parameter α is studied in [1] for IDA approaches. Basically, the performance of IDA increase when α decreases. Note that our stochastic shrinkage matrix $\mathsf{diag}(\mathbf{S}_{pooled})$ coincides with the covariance matrix estimator for IDA, therefore, the changes of α will also have effect on our covariance matrix estimators.

In the first set of experiments, $\alpha = 0.3$, and the other settings are the same as that for Fig.1 in the paper. Let $p = 60$, $n_1 = n_2 = 40$, $\Delta^2 = 6$, $\rho \in [0,\ 3]$, the performance of true error, proposed error estimator, and conventional plug-in error estimator is plotted in Figure S2a. It can be seen that the conventional plug-in error estimator with three types of $\mathbf{R}_0$ underestimate the error seriously for all values of ρ. When ρ increases, both the true error and the error with proposed covariance estimator decreases. The reason is that shrinkage matrix $\mathsf{diag}(\mathbf{S}_{pooled})$ outperforms the sample covariance matrix $\mathbf{S}_{pooled}$ in covariance matrix estimation of discriminant analysis with a small α. Let $p = 60$, $\Delta^2 = 6$, $\rho = 1$, $n \in [50,\ 120]$, the performance of true error, proposed error estimator, and conventional plug-in error estimator is plotted in Figure S2b. It can be seen that the conventional plug-in error estimator with three types of $\mathbf{R}_0$ underestimate the error seriously for all values of n. For each type of $\mathbf{R}_0$, the proposed error estimator is close to the corresponding true error, and the gaps become even smaller as n increases.

In the second set of experiments, first let α vary in $[0.1,\ 0.9]$ and then fix $\alpha = 0.3$. The different discriminant approaches with different covariance matrix estimators and optimal shrinkage parameter are then compared. It can be seen in Figure S2c that as α increases, the misclassification error of IDA increases seriously. When $\alpha \in [0.1,\ 0.3]$, the error of IDA, GLDA, and CLV classifiers are close. However, when $\alpha \in [0.4,\ 0.9]$, the proposed GLDA classifier outperforms the IDA classifier and its performance is close to that of CLV classifier. In Figure S2d with $\alpha = 0.3$, the performance of IDA, GLDA, and CLV classifiers are close. The error of GLDA is slightly larger than IDA when $n \leq p$, and smaller when $n > p$.

III. Robustness to the Gaussian distribution

Finally the robustness to the Gaussian distribution is studied. In practice, there are situations where the tails of the data may be heavier than those of the normal distribution and usually Student's t distribution can be used to characterize the distribution



(a)

(b)

(c)

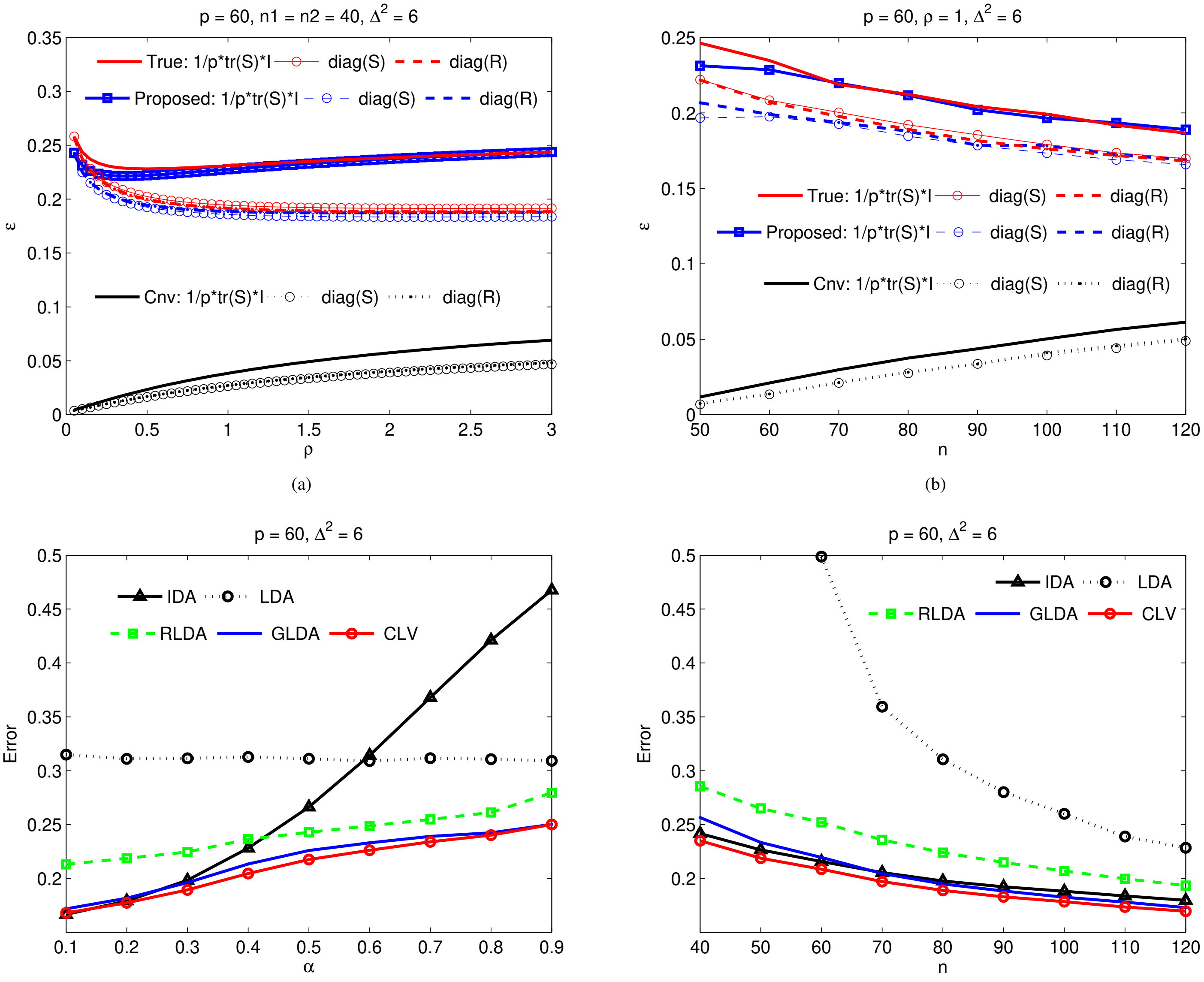
(d)

Figure S1. The error with unbalanced number of samples $n_1 = 2n_2$ versus: (a) $\rho \in [0,\ 3]$; (b) $p = 60$, $n \in [60,\ 120]$; (c) optimal ρ, $n \in [60,\ 120]$; (d) optimal ρ, $p \in [20,\ 160]$

with heavy tails [2]. The synthetic data are generated according to (A1)-(A3), except that the entries of $\mathbf{Z}_1$ and $\mathbf{Z}_2$ follow t distribution with mean 0, variance 1, and degree of freedom ν. (When $\nu \to \infty$, t distribution reduces to the case of Gaussian distribution.) The other settings are the same as Section 4.1 of the paper.

In the first set of experiments, we fix $\nu = 5$, and the other settings are the same as that for Fig.1 in the paper. Let $p = 60$, $n_1 = n_2 = 40$, $\Delta^2 = 6$, $\rho \in [0,\ 3]$, the performance of true error, proposed error estimator, and conventional plug-in error estimator is plotted in Figure S3a. It can be seen that the conventional plug-in error estimator with three types of $\mathbf{R}_0$ underestimate the error seriously for all values of ρ. With each value of ρ, the proposed error estimator is close to the corresponding true error for all the three types of $\mathbf{R}_0$. Let $p = 60$, $\Delta^2 = 6$, $\rho = 1$, $n_1 = n_2$, $n \in [50,\ 120]$, the performance of true error, proposed error estimator, and conventional plug-in error estimator is plotted in Figure S3b. It can be seen that the conventional plug-in error estimator with three types of $\mathbf{R}_0$ underestimate the error seriously for all values of n. For each type of $\mathbf{R}_0$, the proposed error estimator is close to the corresponding true error, and the gaps become even smaller as n increases.

In the second set of experiments, we first let ν vary in $[5,\ 30]$ and then fix $\nu = 5$. The different discriminant approaches with different covariance matrix estimators and optimal shrinkage parameter are then compared. It can be seen in Figure S3c that the misclassification error of all the classifiers are not sensitive to the change of ν. In Figure S3d with $\nu = 5$, the proposed GLDA classifier outperforms the other classifiers and its performance is close to that the CLV classifier. Therefore, the proposed GLDA classifier with covariance matrix estimator is robust to the Gaussian distribution.



(a)

(b)

(c)

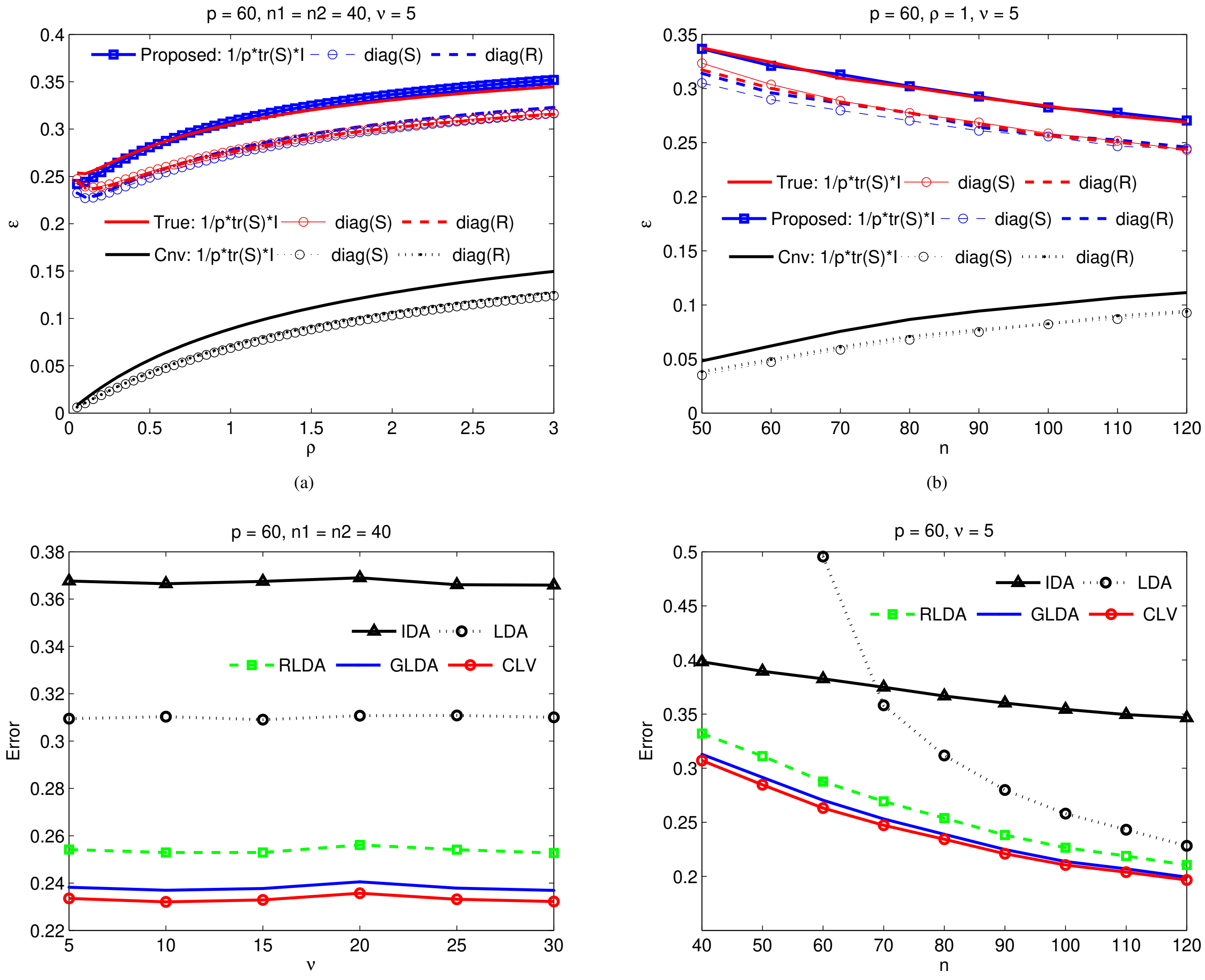
(d)

Figure S2. The effect of different Gaussian AR parameter α and corresponding error versus: (a) $\rho \in [0,\ 3]$; (b) $n \in [50,\ 120]$; (c) optimal ρ, $\alpha \in [0.1,\ 0.9]$; (d) optimal ρ, $\alpha = 0.3$, $p \in [40,\ 120]$

References

[1] P. J. Bickel and E. Levina, “Some theory for Fisher’s linear discriminant function,’naive bayes’, and some alternatives when there are many more variables than observations,” *Bernoulli*, pp. 989–1010, 2004.

[2] C. W. Dunnett and M. Sobel, “A bivariate generalization of student’s t-distribution, with tables for certain special cases,” *Biometrika*, vol. 41, no. 1-2, pp. 153–169, 1954.



(a)

(b)

(c)

(d)

Figure S3. The error with non-Gaussian distributions versus: (a) $\rho \in [0,\ 3]$; (b) $n \in [50,\ 120]$; (c) optimal ρ, $\nu \in [5,\ 30]$; (d) optimal ρ, $\nu = 5$, $p \in [40,\ 120]$