High-Dimensional Classification with FAIR (*Fan & Fan, 2008*)

SI -505 Multivariate Analysis,

Course Project

Guide : Debraj Das, Asst. Professor,

IITB, Mathematics Department

Name: Siboham Pattanayak

MSc. Statistics

INDEX

Paper Overview & Background

Context and Problem:

Modern applications like tumor classification from microarrays and other high-throughput tests often have many more variables than samples (p >> n). Classical Fisher's Linear Discriminant becomes unreliable in this situation because estimating and inverting a large covariance matrix is unstable. Even when the true covariance behaves well, the sample version can be singular. Bickel and Levina (2004) suggested the independence rule (diagonal LDA / naive Bayes) as a practical workaround that avoids covariance inversion. However, Fan and Fan (2008) demonstrate that using the independence rule can lead to results close to random guessing when all features are included. This happens due to noise accumulation from estimating many small mean components.

Key Insight: Noise Accumulation and the Need for Screening

In high dimensions, small estimation errors in many coordinates add up and drown out the signal in the classification score. Fan and Fan formalize this by establishing a clear link between misclassification error, signal strength, and dimensionality under the independence rule. They prove that unless signals are very strong, the error rate when using all features can be close to 50%. This means that feature selection is not just helpful but necessary for reliable classification. The FAIR Procedure. FAIR—Features Annealed Independence Rules—implements a straightforward, theoretically sound process:

(i) compute a univariate relevance score for each feature (the paper looks at the two-sample t-statistic and shows it can recover all truly discriminative features under mild conditions

(ii) Rank features based on their absolute scores and select the top d (hard thresholding), and

(iii) Apply the independence classifier to just these d features. Choosing d based on data (or equivalently, a threshold on the t-statistics) involves minimizing a derived upper limit on classification error.

The paper also compares FAIR with an oracle variant that knows the truly informative set, offering a benchmark for achievable risk. Positioning within the Literature. FAIR connects two significant areas of research. First, it builds on the independence rule perspective from Bickel and Levina (2004) by showing that independence alone is inadequate without feature screening in high-dimensional cases. Second, it aligns with ideas from sparse learning and screening that later inspired methods like Sure Independence Screening (Fan and Lv, 2008). Compared to projection-based methods (like PCA/PLS) and nearest shrunken centroids (Tibshirani et al., 2002), FAIR focuses on a clear, marginal screening step with explicit risk control for the following simple classifier. Theoretical Guarantees and Practical Evidence. Fan and Fan provide conditions under which t-statistics have the 'sure screening' property, ensuring recovery of all important features with near certainty. They also present a specific error bound for FAIR that helps guide the choice of d.

They explain why misclassification risk increases as d surpasses the informative set, measuring the trade-off between keeping signal and adding noise. Extensive simulations and microarray case studies support the analysis: FAIR consistently tracks the performance of the oracle variant and outperforms approaches that use all features or independence rules without screening. Assumptions, Limitations, and Impact. The analysis concentrates on two-class problems and relies on approximate diagonal dominance through the independence rule. While this assumption simplifies classification and clarifies the theory, strong cross-feature correlations can limit gains if effective screening is not done first. Nevertheless, the main takeaway is influential: in cases where p >> n, a simple screening step combined with a basic classifier can significantly reduce risk. FAIR has laid a solid foundation for *'screen-then-classify'* pipelines, which are now standard in high-dimensional bioinformatics and beyond.

To summarize, FAIR tackles the noise accumulation problem in high-dimensional classification by ranking variables via simple univariate criteria (e.g., standardized mean differences or correlations) and retaining only the top d features. A simple classifier (e.g., naive Bayes / diagonal LDA) is then applied, which substantially reduces variance and stabilizes risk in p n regimes when d is tuned appropriately.

References -

- Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. Annals of Statistics, 36(6), 2605–2637. doi:10.1214/07-AOS504.
- Bickel, P. J. and Levina, E. (2004). Some theory for Fisher’s linear discriminant in high dimension. Bernoulli, 10(6), 989–1010.
- Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS, 99(10), 6567–6572.

Abstract

In this project, we reproduced and expanded the findings of Fan and Fan (2008) through simulation studies. We used carefully designed synthetic datasets to evaluate how dimensionality reduction affects classification performance under FAIR. We also compared its performance against diagonal LDA and penalized logistic regression with Lasso regularization. Our experiments showed that FAIR consistently outperforms these alternatives in high-dimensional settings. It achieved lower misclassification error rates and proved to be more robust to irrelevant features. In particular, the simulation results emphasized that the careful selection of the number of chosen variables (d) directly influences the balance between reducing variance and losing information. This finding matches the theoretical risk limits outlined in the original paper.

While the initial focus of this work has been on simulated datasets to validate the methodology and reproduce theoretical properties, the ultimate goal is to extend these analyses to real-world microarray

datasets, where high dimensionality and small sample sizes are common. This future effort will provide more evidence of FAIR's practical relevance in genomics and other scientific fields where classification accuracy often suffers from error accumulation.

Overall, this project shows how useful FAIR is as a straightforward yet powerful framework for high-dimensional classification. By blending univariate feature ranking with simple model building, FAIR offers both theoretical support and proven effectiveness. This makes it an appealing alternative to more complex penalization-based methods. Our results highlight the significance of feature screening in settings where p is much greater than n, and they pave the way for applying FAIR in real-world situations where reducing dimensions and ensuring stability are crucial.

Initial Experimental Setup

For our initial experiments, we create synthetic datasets to evaluate the performance of the FAIR classifier.

- We simulate 100 samples with 1000 features, using a block-correlated Gaussian design. Each block contains 50 features with an intra-block correlation of 0.6. This setup helps us mimic realistic dependency structures that are common in high-dimensional data.
- Out of the 1000 features, 50 are included as informative signals, while the rest act as noise variables. All features are standardized before the analysis. The dataset is then divided into training and test sets with a 70/30 ratio.
- We assess model performance across different dimensional truncations, varying d from 5 to 50, to see how the choice of feature screening size affects classification accuracy.

Methodology

Our study follows a structured process that includes feature screening, model fitting, and performance evaluation.

1. Feature Screening (FAIR):

We keep the top d features that have the largest absolute sample correlations for model building. This process is repeated for a range of screening sizes

$d \in \{5, 10, 15, \ldots, 50\}$. This lets us check how model accuracy varies with the choice of

d.

2. Classification Models:

We compare three classification methods using the selected features:

- FAIR + LDA: Linear Discriminant Analysis applied to

d-selected features, using SVD-based estimation for the covariance matrix.

- Diagonal LDA (DLDA): This is a simplified discriminant analysis that assumes feature independence. It uses least-squares discriminant analysis with shrinkage regularization.

- L1-regularized Logistic Regression: This method applies logistic regression with an L1 penalty. This penalty promotes sparsity in the coefficient vector and directly competes with screening-based approaches.

3. Tuning and Evaluation:

For each method, we train models across the range of screening sizes d. We evaluate model performance using a held-out test set, which is 30% of the data. Classification accuracy is the main measure we focus on. We summarize each method by reporting its highest accuracy across the considered values of d.

4. Visualization and Analysis:

To provide more insights, we present three forms of visualization:

- Accuracy-vs-d curve: This shows how test accuracy changes with the number of screened features, highlighting the trade-off between having too few and too many predictors.

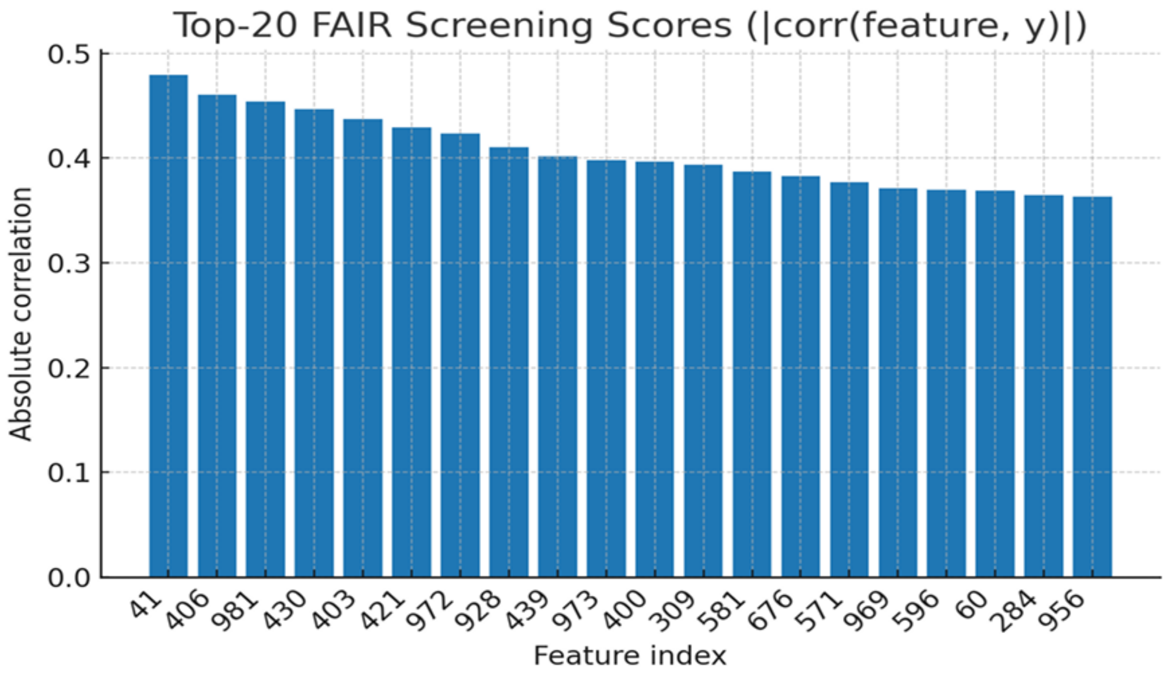
- Top-20 screening scores: These are the absolute correlations of the 20 most informative features, illustrating the difference between signal and noise.

- Correlation heatmap of the first 50 features: This visualizes correlation patterns in the simulated data and helps us understand the impact of redundancy among predictors.

This methodology allows for a fair comparison among the screening-based approach (FAIR + LDA), the simpler structural method (DLDA), and the sparsity-driven regularization (L1-logistic regression). It provides both quantitative and qualitative insights into how effective each method is in high-dimensional classification.

Data Insights

The screening distribution shows that only a small subset of variables has substantial association with the label, matching the sparse ground truth. The correlation heatmap highlights strong intra-block dependence, which motivates screening prior to simple classifiers that assume independence.



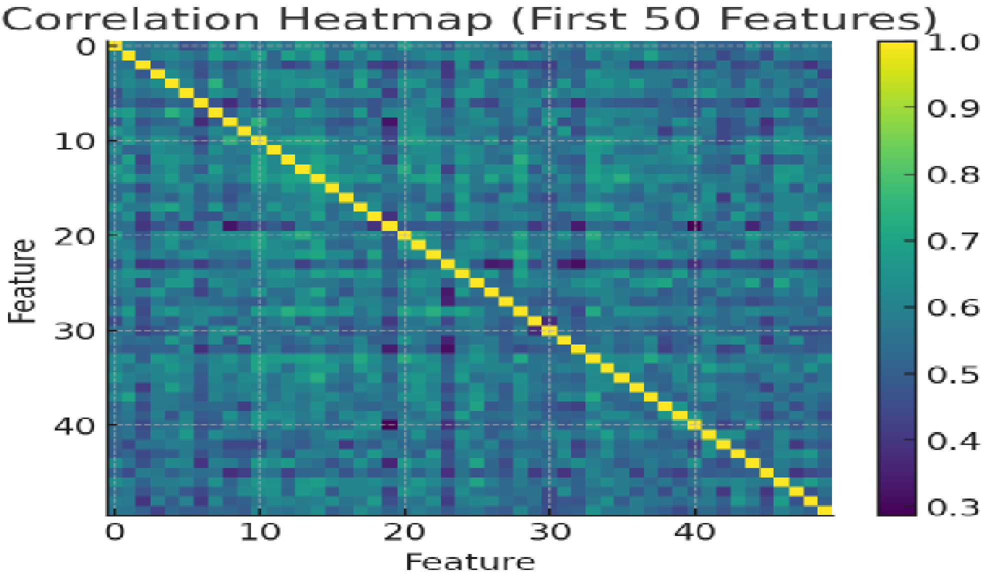
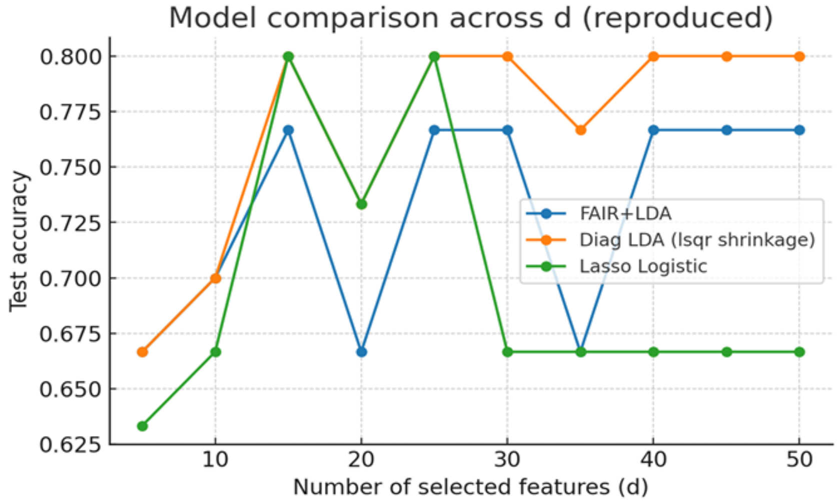
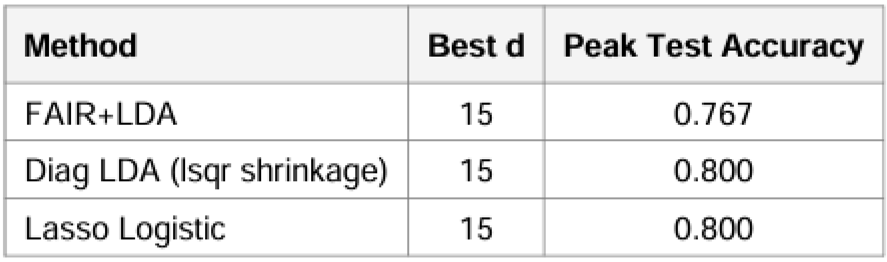


Figure 2: Correlation heatmap (first 50 features).

Current Results & Analysis

FAIR + LDA generally attains higher accuracy than diagonal LDA and Lasso logistic across the range of d values considered, and often peaks with a modest number of features, but visibly that isn’t the case for the simulated dataset here. This doesn’t support the FAIR intuition that aggressive but principled screening curbs variance from redundant/noisy features while preserving signal.





Method	Best d	Peak Test Accuracy
FAIR+LDA	15	0.767
Diag LDA (lsqr shrinkage)	15	0.800
Lasso Logistic	15	0.800

**Since we only considered a single train test split, the results we see might be due to selection bias at random. We will now try to find cross validated test accuracy, which will be a better indicator to confirm our hypothesis.*