We consider classification in a supervised regime in conjunction with dimensionality reduction. Specifically, we analyse the performance of Fisher’s Linear Discriminant (FLD) classifier in randomly projected data spaces. Existing results in the literature typically consider dimensionality reduction in isolation from the classifier which will subsequently work with the data. For random projections (RP), one may then prove performance guarantees using approximate preservation of data geometry, via the Johnson-Lindenstrauss lemma (JLL). However, such an approach has the unfortunate side effect that in those bounds the projection dimension, \( k \), is logarithmically dependent on the number of observations, \( N \).

In [1] we take a different tack: In order to leverage the class structure inherent in the problem we consider RP and FLD together. This allows us to focus on the preservation of certain key distances (for FLD, the separation of classes are modelled as Gaussians with identical covariance matrices and a query point \( x \) is assigned to the class with the nearest mean under the Mahalanobis metric.

We assume that the observations \( x \) are drawn from one of \( m+1 \) multivariate Gaussian classes \( D_k = \{ x \mid N(\mu_k, \Sigma) \} \) with the unknown parameters \( \mu_k \), a vector of means and \( \Sigma \) a full-rank covariance matrix that we must estimate from training data.

Random projection is a probabilistic method of dimensionality reduction. To randomly project a vector observation \( x \in \mathbb{R}^d \) to a \( k \)-dimensional subspace \( R(\mathbb{R}^d) \cong \mathbb{R}^k \) one left multiplies \( x \) by the random projection matrix \( R \). We construct \( R \) by drawing entries \( r_{ij} \sim N(0, 1/\lambda) \) and then orthonormalising the rows. Our results also hold for other forms of \( R \), e.g. [2].

The random projection method was originally motivated by high probability guarantees that such projections approximately preserve distances and angles between points uniformly [3]. We extend this method by considering random projection as a general preprocessing step for reducing the data dimensionality, and give guarantees on classification performance without worrying too much about approximate geometry preservation.

### Summary

Fisher’s Linear Discriminant

FLD is a successful and widely used generative classifier that seeks to model, given training data \( T_{\text{tr}} \), the optimal decision boundary between classes. In the classical version, the classes are modelled as Gaussians with identical covariance matrices and a query point \( x \) is assigned to the class with the nearest mean under the Mahalanobis metric.

Random Projection

Random projection is a probabilistic method of dimensionality reduction. To randomly project a vector observation \( x \in \mathbb{R}^d \) to a \( k \)-dimensional subspace \( R(\mathbb{R}^d) \cong \mathbb{R}^k \) one left multiplies \( x \) by the random projection matrix \( R \). We construct \( R \) by drawing entries \( r_{ij} \sim N(0, 1/\lambda) \) and then orthonormalising the rows. Our results also hold for other forms of \( R \), e.g. [2].

The random projection method was originally motivated by high probability guarantees that such projections approximately preserve distances and angles between points uniformly [3]. We extend this method by considering random projection as a general preprocessing step for reducing the data dimensionality, and give guarantees on classification performance without worrying too much about approximate geometry preservation.

### Bound on Error of RP FLD

Let \( x_q \sim D_q = N(\mu_q, \Sigma) \). Let \( \hat{F} \) be an instance of FLD functions and let \( \hat{h} \) be the instance learned from the training data \( T_{\text{tr}} \). Let \( R \in \mathbb{R}^{k \times d} \) be a random projection matrix with entries drawn i.i.d. from the univariate Gaussian \( N(0, 1/d) \).

Then the estimated misclassification error \( \hat{P}_{T_{\text{tr}}}^h [y | R(x_q)] \neq y \) is bounded above by:

\[
\frac{1}{1 + g(\Sigma)} \left( \frac{1}{\max(\lambda_k)} \right)^{1/2}
\]

with \( \mu_q \) the mean of the class from which \( x_q \) was drawn, estimated class means \( \hat{\mu}_0 \) and \( \hat{\mu}_q \), model covariance \( \Sigma \), and \( \lambda_k \) the user-specified tolerance of misclassification probability, \( k \) the projection dimensionality and the factor \( \max(\lambda_k) \) is positive definite, be as given in [3].

### Sufficient Dimensionality

Let \( k, d, \Sigma \) be as in theorem (1). Then, for an \( m+1 \)-class problem, in order that the probability of misclassification in the RP space remains below \( \delta \) it is sufficient to take:

\[
k > 8 \left( \frac{\min_j \| \mu_j - \hat{\mu}_j \|^2}{\| \Sigma \|} \right) \log(m/\delta)
\]

### Covariance Misspecification

Let \( g(Q), Q \) is positive definite, be as given in theorem (1). Then, for any fixed \( k \times d \) matrix \( R \) with orthonormal rows:

\[
g(\Sigma R \Sigma^T)^{-1} (\Sigma R \Sigma^T) > g(\Sigma)
\]

We see the quality of fit is poor in high dimensions and improves dramatically in the projected space, approaching the best value as \( k \) decreases.

### References


An algorithmic theory of learning.