# Random Matrix Theory Improvements on the Matched Subspace Classifier

Nicholas Asendorf

July 6, 2011

## 1 Problem Statement

We consider the classification problem where our observed data, $y$, may be one of two classes. We may either observe signal in the presence of noise, or simply noise itself. Our setup is as follows:

$$y = \begin{cases} z & y \in H_0 \\ U_1 x + z & y \in H_1 \end{cases} \tag{1}$$

where $z \sim \mathcal{N}(0, I)$, $U_1 \in \mathbb{C}^{n \times k}$ is unknown with orthonormal columns, $x \sim \mathcal{N}(0, \Sigma_1)$ with $\Sigma_1 = \mathbf{diag}(\sigma_1^2, \ldots, \sigma_k^2)$ with $\sigma_i^2$ unknown. We also assume that $x$ and $z$ are independent.

We are given labeled training data $y_1, \ldots, y_m$, with $m \geq n$ and $y_i \in H_1$ for $i = 1, \ldots, m$. We will use this training data to form estimates $\hat{U}_1$, $\hat{\Sigma}_1$ of our unknown parameters $U_1$, $\Sigma_1$.

We consider the processed data $w = \hat{U}_1^H y \in \mathbb{C}^n$. We are also given unlabeled testing data $y_1, \ldots, y_r$. Our goal is to determine a classifier, $g(w) \to \{0, 1\}$ which solves the following problem for our testing data:

$$\begin{array}{ll} \text{maximize} & P_D = P\left(g(w) = 1 | y \in H_1\right) \\ \text{subject to} & P_F = P\left(g(w) = 1 | y \in H_0\right) \end{array} \tag{2}$$

## 2 Parameter Estimation

We have two unknown parameters, $U_1$, $\Sigma_1$. Using our training data, $\{y_1, \ldots, y_m\}$, we make estimate of these parameters. To do so, we form the matrix $Y = [y_1, \ldots, y_m]$ by stacking the training data as columns in a matrix. Define $S_1 = \frac{1}{m} Y Y^H$ as the sample covariance of our training data. By properties of Gaussian random variables, under $H_1$, $y_i \sim \mathcal{N}(0, U_1 \Sigma_1 U_1^H + I)$. Taking $U_2 = U_1^\perp$ to be the orthogonal complement of $U_1$ we may write this covariance as

$$\begin{aligned} U_1 \Sigma U_1^H + I &= \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_1^H \\ U_2^H \end{bmatrix} + \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} U_1^H \\ U_2^H \end{bmatrix} \\ &= \begin{bmatrix} U_1 & U_2 \end{bmatrix} \left( \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} I_k & 0 \\ 0 & I_{n-k} \end{bmatrix} \right) \begin{bmatrix} U_1^H \\ U_2^H \end{bmatrix} \\ &= \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 + I_k & 0 \\ 0 & I_{n-k} \end{bmatrix} \begin{bmatrix} U_1^H \\ U_2^H \end{bmatrix} \end{aligned} \tag{3}$$

Clearly this is the in the form of an eigenvalue decomposition of our covariance matrix. Therefore if we take the eigenvalue decomposition of the sample covariance matrix, $S$, we can form an estimate of our subspace $U_1$ and our covariances $\sigma_i^2$. Defining the eigenvalue decomposition $S_1 = V\Lambda V^H$ where $\Lambda = \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$ and $V = [v_1, \ldots, v_n]$ such that $\lambda_1 > \lambda_2 > \cdots > \lambda_n$ we have

$$\begin{aligned} \hat{U}_1 &= [v_1 \ldots v_k] \\ \hat{\sigma}_i^2 &= \lambda_i - 1 \text{ for } i = 1, \ldots, k \end{aligned} \tag{4}$$

We also define $\hat{\Sigma}_1 = \mathbf{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_k^2)$.

# 3 Random Matrix Theory Estimates

If we are given a sample covariance matrix, $S = \frac{1}{m}YY^H$, where the columns of $Y$ are drawn from $y \sim \mathcal{N}(0, \Sigma)$ where $\Sigma = \mathbf{diag}(\lambda_1, \ldots, \lambda_k, 1, \ldots, 1) \in \mathbf{R}^n$, Paul's paper tells us that

$$\hat{\lambda}_i \to \begin{cases} \left(1+\sqrt{c}\right)^2 & \text{if } \lambda_i \leq 1+\sqrt{c} \\ \lambda_i\left(1+\frac{c}{\lambda-1}\right) & \text{if } \lambda_i > 1+\sqrt{c} \end{cases} \tag{5}$$

where $c = \frac{n}{m}$ and $\hat{\lambda}_i$ are the eigenvalues of $S$.

Because our sample covariance matrix $S_1$ takes this form, we may apply this theorem to our problem at hand:

$$\begin{aligned} \hat{\sigma}_i^2 + 1 \quad &\to \begin{cases} \left(1+\sqrt{c}\right)^2 & \text{if } \sigma_i^2 + 1 \leq 1+\sqrt{c} \\ (\sigma_i^2+1)\left(1+\frac{c}{\sigma_i^2+1-1}\right) & \text{if } \sigma_i^2+1 > 1+\sqrt{c} \end{cases} \\ &\to \begin{cases} \left(1+\sqrt{c}\right)^2 & \text{if } \sigma_i^2 \leq \sqrt{c} \\ \sigma_i^2 + 1 + c + \frac{c}{\sigma_i^2} & \text{if } \sigma_i^2 > \sqrt{c} \end{cases} \\ \hat{\sigma}_i^2 \quad &\to \begin{cases} 2\sqrt{c} + c & \text{if } \sigma_i^2 \leq \sqrt{c} \\ \sigma_i^2 + c + \frac{c}{\sigma_i^2} & \text{if } \sigma_i^2 > \sqrt{c} \end{cases} \end{aligned} \tag{6}$$

Solving for $\sigma_i^2$ we have obtain our random matrix theory estimate of $\sigma_i^2$

$$\boxed{\tilde{\sigma}_{i_{\text{rmt}}}^2 = \begin{cases} \sqrt{c} & \text{if } \hat{\sigma}_i^2 \leq c + 2\sqrt{c} \\ \frac{\hat{\sigma}_i^2 - c + \sqrt{\left(\hat{\sigma}_i^2 - c\right)^2 - 4c}}{2} & \text{if } \hat{\sigma}_i^2 > c + 2\sqrt{c} \end{cases}} \tag{7}$$

From Paul's paper, we also have that

$$| < v_i, \hat{v}_i > |^2 \to \begin{cases} 0 & \text{if } \lambda_i \leq 1+\sqrt{c} \\ \frac{1-\frac{c}{(\lambda-1)^2}}{1+\frac{c}{\lambda-1}} & \text{if } \lambda_i > 1+\sqrt{c} \end{cases} \tag{8}$$

where $\hat{v}_i$ is the eigenvector of the sample covariance matrix corresponding to the eigenvalue $\lambda_i$ and $v_i$ is the true underlying eigenvalue. Applying this theorem to our problem, we have

$$
\begin{aligned}
|<u_i,\hat{u}_i>|^2 \quad &\to \begin{cases} 0 & \text{if } \sigma_i^2+1 \le 1+\sqrt{c} \\ \frac{1-\frac{c}{(\sigma_i^2+1-1)^2}}{1+\frac{c}{\sigma_i^2+1-1}} & \text{if } \sigma_i^2+1 > 1+\sqrt{c} \end{cases} \\
&\to \begin{cases} 0 & \text{if } \sigma_i^2 \le \sqrt{c} \\ \frac{\frac{\sigma_i^4-c}{\sigma_i^4}}{\frac{\sigma_i^2+c}{\sigma_i^2}} & \text{if } \sigma_i^2 > \sqrt{c} \end{cases} \\
&\to \begin{cases} 0 & \text{if } \sigma_i^2 \le \sqrt{c} \\ \frac{\sigma_i^4-c}{\sigma_i^4+\sigma_i^2 c} & \text{if } \sigma_i^2 > \sqrt{c} \end{cases}
\end{aligned} \tag{9}
$$

We then substitute our expression for $\sigma_i^2$ derived in (7)

$$
\boxed{|<u_i,\hat{u}_i>|^2_{\text{rmt}} \to \begin{cases} 0 & \text{if } \tilde{\sigma}^2_{i_{\text{rmt}}} \le \sqrt{c} \\ \frac{\tilde{\sigma}^4_{i_{\text{rmt}}}-c}{\tilde{\sigma}^4_{i_{\text{rmt}}}+\tilde{\sigma}^2_{i_{\text{rmt}}} c} & \text{if } \hat{\sigma}^2_{i_{\text{rmt}}} > \sqrt{c} \end{cases}} \tag{10}
$$

# 4 Processed Matched Subspace Classifier

By properties of Gaussian random variables, under $H_0$, $y \sim \mathcal{N}(0, I)$ and under $H_1$, $y \sim \mathcal{N}(0, U_1\Sigma_1 U_1^H + I)$. For our processed data, $w = \hat{U}_1^H$, using properties of Gaussian random variables, under $H_0$, $w \sim \mathcal{N}(0, I_k)$ and under $H_1$, $w \sim \mathcal{N}(0, \hat{U}_1^H U_1 \Sigma_1 U_1^H \hat{U}_1 + I)$.

We will consider 3 different classifiers by examining the likelihood ratio test (LRT) for our data $w$. The first is an oracle classifier, which will assume that $U_1$ and $\Sigma_1$ are known. The purpose of this is to give an upper bound on a classifier's performance. The second classifier is a plug-in classifier which will approximate the oracle classifier by simply plugging in our estimates $\hat{U}_1$, $\hat{\Sigma}_1$ for our unknown $U_1$ and $\Sigma_1$. The third classifier uses the results of random matrix theory to form an approximation to the oracle classifier.

## 4.1 Oracle Classifier

Our (LRT) for our processed data $w$, is

$$
\begin{aligned}
\Lambda(w) \quad &= \frac{(2\pi)^{-k/2}|\hat{U}_1^H U_1\Sigma_1 U_1^H \hat{U}_1 + I|^{-1/2} \exp\{-\frac{1}{2} w^H \left[\hat{U}_1^H U_1\Sigma_1 U_1^H \hat{U}_1 + I\right]^{-1} w\}}{(2\pi)^{-k/2}\exp\{-\frac{1}{2}w^H w\}} \\
&= |\hat{U}_1^H U_1\Sigma_1 U_1^H \hat{U}_1 + I|^{-1/2} \exp\{-\frac{1}{2} w^H \left[\left(\hat{U}_1^H U_1\Sigma_1 U_1^H \hat{U}_1 + I\right)^{-1} - I\right] w\}
\end{aligned} \tag{11}
$$

where, defining $\eta = \frac{P(y\in H_0)}{P(y \in H_1)}$ our classifier is

$$
g_{\text{oracle}}(w) = \begin{cases} 0 & \text{if } \Lambda(w) < \eta \\ 1 & \text{if } \Lambda(w) > \eta \end{cases} \tag{12}
$$

We may apply the natural logarithm operator to both sides as it is a monotonic operation. Our statistic becomes

$$\boxed{\Lambda_{\text{oracle}}(w) = w^H \left[ I - \left( \hat{U}_1^H U_1 \Sigma_1 U_1^H \hat{U}_1 + I \right)^{-1} \right] w} \tag{13}$$

and defining a threshold $\gamma = 2 \ln \left( \eta | \hat{U}_1^H U_1 \Sigma_1 U_1^H \hat{U}_1 + I|^{1/2} \right)$ we have the classifier

$$\boxed{g_{\text{oracle}}(w) = \begin{cases} 0 & \text{if } \Lambda_{\text{oracle}}(w) < \gamma \\ 1 & \text{if } \Lambda_{\text{oracle}}(w) > \gamma \end{cases}} \tag{14}$$

## 4.2 Plug-in Classifier

As is the case, $U_1$ and $\Sigma_1$ are not known, and we cannot compute (13) directly. One solution to this problem is to plug in our estimates $\hat{U}_1$ and $\hat{\Sigma}_1$ wherever $U_1$ and $\Sigma_1$ appear respectively. Using our estimates in (4) have the following plug-in classifier statistic:

$$\begin{aligned} \Lambda_{\text{plugin}}(w) &= w^H \left( I - \left[ \hat{U}_1^H \hat{U}_1 \hat{\Sigma}_1 \hat{U}_1^H \hat{U}_1 + I \right]^{-1} \right) w \\ &= w^H \left( I - \left( \hat{\Sigma}_1 + I \right)^{-1} \right) w \\ &= w^H \left( I - \mathbf{diag} \left( \hat{\sigma}_i^2 + 1 \right)^{-1} \right) w \end{aligned} \tag{15}$$

This simplifies to

$$\boxed{\Lambda_{\text{plugin}}(w) = w^H \, \mathbf{diag} \left( \frac{\hat{\sigma}_i^2}{1 + \hat{\sigma}_i^2} \right) w = \sum_{i=1}^{k} \frac{w_i^2 \hat{\sigma}_i^2}{\hat{\sigma}_i^2 + 1}} \tag{16}$$

and our classifier becomes

$$\boxed{g_{\text{plugin}}(w) = \begin{cases} 0 & \text{if } \Lambda_{\text{plugin}}(w) < \gamma \\ 1 & \text{if } \Lambda_{\text{plugin}}(w) > \gamma \end{cases}} \tag{17}$$

## 4.3 Random Matrix Theory Classifier

To utilize our random matrix theory expressions derived in Section 3, we first make a diagonal approximation of (13)

$$\begin{aligned} \tilde{\Lambda}(w) &= w^H \left[ I - \left( \hat{U}_1^H U_1 \Sigma_1 U_1^H \hat{U}_1 + I \right)^{-1} \right] w \\ &\approx w^H \left( I - \left[ \mathbf{diag} \left( | < u_i, \hat{u}_i > |^2 \sigma_i^2 \right) + I \right]^{-1} \right) w \\ &= w^H \left( \mathbf{diag} \left( \frac{| < u_i, \hat{u}_i > |^2 \sigma_i^2}{| < u_i, \hat{u}_i > |^2 \sigma_i^2 + 1} \right) \right) w \end{aligned} \tag{18}$$

However, $\sigma_i^2$ and $| < u_i, \hat{u}_i > |^2$ are unknown and we must use an estimate for them. However, instead of using $\hat{\sigma}_i^2$ and estimating $| < u_i, \hat{u}_i > |^2 = 1$ as the plug-in classifier does, we use expressions derived in

Section 3 which considers the error in estimating the eigenvalues and eigenvectors of our sample covariance matrix.

Using (7) and (10) our random matrix theory statistic becomes

$$\boxed{\Lambda_{\text{rmt}}(w) = w^H \,\textbf{diag}\left(\frac{|<u_i,\hat{u}_i>|^2_{\text{rmt}}\tilde{\sigma}^2_{i_{\text{rmt}}}}{|<u_i,\hat{u}_i>|^2_{\text{rmt}}\tilde{\sigma}^2_{i_{\text{rmt}}}+1}\right) w = \sum_{i=1}^{k} \frac{w_i^2|<u_i,\hat{u}_i>|^2_{\text{rmt}}\tilde{\sigma}^2_{i_{\text{rmt}}}}{|<u_i,\hat{u}_i>|^2_{\text{rmt}}\tilde{\sigma}^2_{i_{\text{rmt}}}+1}} \tag{19}$$

and our classifier beocmes

$$\boxed{g_{\text{rmt}}(w) = \begin{cases} 0 & \text{if } \Lambda_{\text{rmt}}(w) < \gamma \\ 1 & \text{if } \Lambda_{\text{rmt}}(w) > \gamma \end{cases}} \tag{20}$$

# 5 Theoretical ROC for (16)

Under $H_0$ we have that $w \sim \mathcal{N}(0, I)$ so $w_i \sim \mathcal{N}(0,1)$ are i.i.d for $i = 1, \ldots, k$. So $w_i^2 \sim \chi_1^2$ are i.i.d for $i = 1, \ldots, k$. So under $H_0$,

$$\Lambda_{\text{plugin}}(w) = \sum_{i=1}^{k} \left(\frac{\sigma_i^2}{1+\sigma_i^2}\right) \chi^2_{1i} \tag{21}$$

That is, a weighted sum of independent chi-square random variables with 1 degree of freedom.

Now, under $H_1$, we have that $w \sim \mathcal{N}(0, \hat{U}_1^H U_1 \Sigma_1 U_1^H \hat{U}_1 + I)$ so $w_i \approx \mathcal{N}(0, \sigma_i^2|<u_i,\hat{u}_i>|^2+1)$ are i.i.d. Therefore,

$$\frac{w_i^2}{\sigma_i^2|<u_i,\hat{u}_i>|^2+1} \sim \chi_1^2 \tag{22}$$

Therefore, under $H_1$,

$$\Lambda_{\text{plugin}}(w) = \sum_{i=1}^{k} \left(\frac{\sigma_i^2\left(\sigma_i^2|<u_i,\hat{u}_i>|^2+1\right)}{1+\sigma_i^2}\right) \chi^2_{1i} \tag{23}$$

which is also a weighted sum of independent chi-square random variables with 1 degree of freedom.

# 6 Theoretical ROC for (19)

Under $H_0$ we have again that $w \sim \mathcal{N}(0, I)$ so $w_i \sim \mathcal{N}(0,1)$ are i.i.d for $i = 1, \ldots, k$. So $w_i^2 \sim \chi_1^2$ are i.i.d for $i = 1, \ldots, k$. So under $H_0$,

$$\Lambda - \text{rmt}(w) = \sum_{i=1}^{k} \left(\frac{\sigma_i^2|<u_i,\hat{u}_i>|^2_{\text{rmt}}}{1+\sigma_i^2|<u_i,\hat{u}_i>|^2_{\text{rmt}}}\right) \chi^2_{1i} \tag{24}$$

That is, a weighted sum of independent chi-square random variables with 1 degree of freedom.

Now, under $H_1$, we again have that $w \sim \mathcal{N}(0, \hat{U}_1^H U_1 \Sigma_1 U_1^H \hat{U}_1 + I)$ so $w_i \approx \mathcal{N}(0, \sigma_i^2 | < u_i, \hat{u}_i > |^2 + 1)$ are i.i.d. Therefore,

$$\frac{w_i^2}{\sigma_i^2 | < u_i, \hat{u}_i > |^2 + 1} \sim \chi_1^2 \tag{25}$$

Therefore, under $H_1$,

$$\Lambda - \text{rmt}(w) = \sum_{i=1}^{k} \left( \sigma_i^2 | < u_i, \hat{u}_i > |_{\text{rmt}}^2 \right) \chi_{1i}^2 \tag{26}$$

which is also a weighted sum of independent chi-square random variables with 1 degree of freedom.

# 7 Simulation Results

We now demonstrate the performance of the three classifiers derived in Section 4 through numerical simulations. To compare classifiers across all thresholds, $\gamma$, we generate a Receiver Operating Characteristic (ROC) curve for each classifier. ROC curves plot $P_D$ vs. $P_F$ for a classifier. Curves lying in the northwest regime are the best as they operate with a high probability of detection and a low probability of false-alarm.

To test our classifiers first generate a random $U_1$ by taking the first $k$ left singular vectors of a random $n \times n$ matrix. Using the desired $\Sigma_1$ we generate $m$ training points via (1). We then form our parameter estimates (4) and random matrix theory values (7) and (10) to be used in our classifiers.

We then generate $r$ testing points of each class via (1) and process them via $w = \hat{U}_1^H y$. We calculate our statistic for each testing point for each classifier via (13), (16) and (19).

Using algorithm 1 of Fawcett 2005 we calculate the ROC curve of each classifier by sweeping $\gamma$ used in (14), (17) and (20)

We then repeat this process multiple times with a different random orthogonal $U_1$ to generate multiple ROC curves. Using algorithm 4 of Fawcett 2005 we average the ROC curves of each trial to produce a one final ROC curve for each of the three classifiers.

Table 1 provides the parameters for each of the different simulations conducted. The corresponding figures follow and show the empirical ROC curves.

Table 1: Simulation Parameters

| Figure | $n$ | $m$ | $c = n/m$ | $k$ | $r$ | trials | $\Sigma_1$ |
|---|---|---|---|---|---|---|---|
| 1 | 100 | 100 | 1 | 1 | 5000 | 10 | $\mathbf{diag}(10, 1)$ |
| 2 | 100 | 100 | 1 | 1 | 5000 | 10 | $\mathbf{diag}(10, 1)$ |
| 3 | 100 | 100 | 1 | 1 | 5000 | 10 | $\mathbf{diag}(10, 1)$ |
| 4 | 100 | 100 | 1 | 1 | 5000 | 10 | $\mathbf{diag}(10, 1)$ |
| 5 | 100 | 100 | 1 | 1 | 5000 | 10 | $\mathbf{diag}(10, 1)$ |

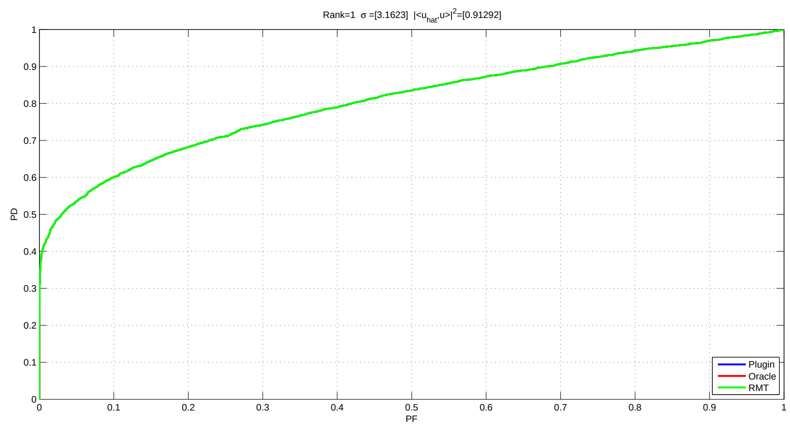


Figure 1: Rank 1

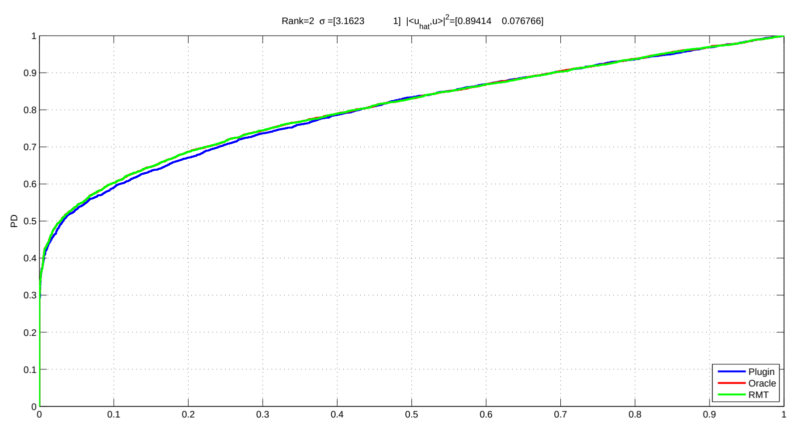


Figure 2: Rank 2