ECE 830 / CS 761 Spring 2016 **Instructors:** R. Willett & R. Nowak

Lecture 3: Likelihood ratio tests, Neyman-Pearson detectors, ROC curves, and sufficient statistics

1 Executive summary

In the last lecture we saw that the likelihood ratio statistic was optimal for testing between two simple hypotheses. The test simply compares the likelihood ratio to a threshold. The "optimal" threshold is a function of the prior probabilities and the costs assigned to different errors. The choice of costs is subjective and depends on the nature of the problem, but the prior probabilities must be known.

In practice, we face several questions:

1. Unfortunately, often the prior probabilities are not known precisely, and thus the correct setting for the threshold is unclear. How should we proceed?
2. What are the tradeoffs among different measures of error (e.g. probability of false alarm, probability of miss, etc.)?
3. Is the LRT still optimal for different error criteria?
4. Do we really need to store all the observed data, or can we get by with some summary statistics?

We will address these questions in these notes.

To explore these questions, we will look at a simple example.

Example: Detecting ET

We observe a sampled radio signal from outer space. Is this just cosmic radiation / noise, or are we receiving a message from extra-terrestrial intelligence (ETI)?

$$\text{null hypothesis (no ETI) } H_0 : x_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),\ i = 1, \dots, n \tag{1}$$

$$\text{alternative hypothesis (ETI) } H_1 : x_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2),\ \mu > 0,\ i = 1, \dots, n. \tag{2}$$

Assume that $\sigma^2 > 0$ is known. The first hypothesis is simple. It involves a fixed and known distribution. The second hypothesis is simple if μ is known. However, if all we know is that $\mu > 0$, then the second hypothesis is the composite of many alternative distributions, i.e., the collection $\{N(\mu, \sigma^2)\}_{\mu>0}$. In this case, H_1 is called a composite hypothesis.

The likelihood ratio test takes the form

$$\frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}x_i^2}} = \frac{\frac{1}{(2\pi\sigma^2)^{n/2}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}}{\frac{1}{(2\pi\sigma^2)^{n/2}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}x_i^2}} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

The inequalities are preserved if we apply a monotonic transformation to both sides, so we can simplify the expression by taking the logarithm, giving us the log-likelihood ratio test

$$\frac{-1}{2\sigma^2}\left(-2\mu\sum_{i=1}^{n} x_i + n\mu^2\right) \underset{H_0}{\overset{H_1}{\gtrless}} \log(\gamma)$$

Assuming $\mu > 0$, this is equivalent to

$$\sum_{i=1}^{n} x_i \underset{H_0}{\overset{H_1}{\gtrless}} \nu \ ,$$

with $\nu = \frac{\sigma^2}{\mu}\ln\gamma + \frac{n\mu}{2}$.

2 Sufficient Statistics

Our test statistic $t := \sum_{i=1}^{n} x_i$ is called the sufficient statistic for the mean of a normal distribution. Let's rewrite our hypotheses in terms of the sufficient statistic:

$$H_0 : t \sim \mathcal{N}(0, n\sigma^2), \tag{3}$$

$$H_1 : t \sim \mathcal{N}(n\mu, n\sigma^2), \qquad \mu > 0 \tag{4}$$

We call t a sufficient statistic because t is sufficient for performing our likelihood ratio test. More formally, a sufficient statistic is defined as follows:

Definition: Sufficient statistic

Let X be an n-dimensional random vector and let θ denote a p-dimensional parameter of the distribution of X. The statistic $t := T(x)$ is a *sufficient statistic* for θ if and only if the conditional distribution of X given $T(X)$ is independent of θ.

More details are available at http://willett.ece.wisc.edu/wp-uploads/2016/02/04-SuffStats.pdf.

If our data is drawn from an exponential family probability model parameterized by θ with the form

$$p_\theta(x) = \exp[a(\theta)b(x) + c(x) + d(\theta)]$$

then $t(x) = \sum_{i=1}^{n} b(x_i)$ is a sufficient statistic. Thus if we are faced with a hypothesis test of the form

$$H_0 : x_i \overset{iid}{\sim} p_{\theta_0}(x) \tag{5}$$

$$H_1 : x_i \overset{iid}{\sim} p_{\theta_1}(x) \tag{6}$$

then our test statistic is a simple function of the data, and we do not need to know the functions a, c, and d to compute it (though we may need these functions to choose an appropriate threshold).

3 Choosing thresholds

Recall that our ideal threshold is $\nu = \frac{\sigma^2}{\mu}\ln\gamma + \frac{n\mu}{2}$. However, in our ETI setting we have no way of knowing prior probabilities π_0, π_1 to set γ, AND we have no way of knowing a good value for μ.

To deal with this, consider an alternative design specification. Let's design a test that minimizes one type of error subject to a constraint on the other type of error. This constrained optimization criterion does not require knowledge of prior probabilities nor cost assignments. It only requires a specification of the maximum allowable value for one type of error, which is sometimes even more natural than assigning costs to the different errors. A classic result due to Neyman and Pearson shows that the solution to this type of optimization is again a likelihood ratio test.

4 Neyman-Pearson Lemma

Assume that we observe a random variable distributed according to one of two distributions.

$$\begin{aligned} H_0 : X &\sim p_0 \\ H_1 : X &\sim p_1 \end{aligned}$$

In many problems, H_0 is consider to be a sort of baseline or default model and is called the *null hypothesis*. H_1 is a different model and is called the *alternative hypothesis*. If a test chooses H_1 when in fact the data were generated by H_0 the error is called a *false-positive* or *false-alarm*, since we mistakenly accepted the alternative hypothesis. The error of deciding H_0 when H_1 was the correct model is called a *false-negative* or *miss*.

Let T denote a testing procedure based on an observation of X, and let R_T denote the subset of the range of X where the test chooses H_1. The probability of a false-positive is denoted by

$$P_0(R_T) \ := \ \int_{R_T} p_0(x)\,dx \ .$$

The probability of a false-negative is $1 - P_1(R_T)$, where

$$P_1(R_T) \ := \ \int_{R_T} p_1(x)\,dx \ ,$$

is the probability of correctly deciding H_1, often called the *probability of detection*.

Consider likelihood ratio tests of the form

$$\frac{p_1(x)}{p_o(x)} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda \ .$$

The subset of the range of X where this test decides H_1 is denoted

$$R_{LR}(\lambda) \ := \ \{x \,:\, p_1(x) > \lambda\, p_0(x)\} \ ,$$

and therefore the probability of a false-positive decision is

$$P_0(R_{LR}(\lambda)) \ := \ \int_{R_{LR}(\lambda)} p_0(x)\,dx \ = \ \int_{\{x:p_1(x)>\lambda p_0(x)\}} p_0(x)\,dx$$

This probability is a function of the threshold λ; the set $R_{LR}(\lambda)$ shrinks/grows as λ increases/decreases. We can select λ to achieve a desired probability of error.

Neyman-Pearson Lemma

Consider the likelihood ratio test

$$\frac{p_1(x)}{p_o(x)} \overset{H_1}{\underset{H_0}{\gtrless}} \lambda$$

with $\lambda > 0$ chosen so that $P_0(R_{LR}(\lambda)) = \alpha$. There does not exist another test T with $P_0(R_T) \leq \alpha$ and $P_1(R_T) > P_1(R_{LR}(\lambda))$. That is, the LRT is the **most powerful test** with probability of false-negative less than or equal to α.

Proof. Let T be any test with $P_0(R_T) = \alpha$ and let NP denote the LRT with λ chosen so that $P_0(R_{LR}(\lambda)) = \alpha$. To simplify the notation we will denote use R_{NP} to denote the region $R_{LR}(\lambda)$. For any subset R of the range of X define

$$P_i(R) := \int_R p_i(x)\, dx,$$

This is simply the probability of $X \in R$ under hypothesis H_i. Note that

$$\begin{aligned} P_i(R_{NP}) &= P_i(R_{NP} \cap R_T) + P_i(R_{NP} \cap R_T^c) \\ P_i(R_T) &= P_i(R_{NP} \cap R_T) + P_i(R_{NP}^c \cap R_T) \end{aligned}$$

where the superscript c indicates the complement of the set. By assumption $P_0(R_{NP}) = P_0(R_T) = \alpha$, therefore

$$P_0(R_{NP} \cap R_T^c) = P_0(R_{NP}^c \cap R_T)\ .$$

Now, we want to show

$$P_1(R_{NP}) \geq P_1(R_T)$$

which holds if

$$P_1(R_{NP} \cap R_T^c) \geq P_1(R_{NP}^c \cap R_T)\ .$$

To see that this is indeed the case,

$$\begin{aligned} P_1(R_{NP} \cap R_T^c) &= \int_{R_{NP} \cap R_T^c} p_1(x)\, dx \\ &\geq \lambda \int_{R_{NP} \cap R_T^c} p_o(x)\, dx \\ &= \lambda\, P_o(R_{NP} \cap R_T^c) \\ &= \lambda\, P_o(R_{NP}^c \cap R_T) \\ &= \lambda \int_{R_{NP}^c \cap R_T} p_o(x)\, dx \\ &\geq \int_{R_{NP}^c \cap R_T} p_1(x)\, dx \\ &= P_1(R_{NP}^c \cap R_T). \end{aligned}$$

□

The probability of a false-positive is also called the probability of false-alarm, which we will denote by P_{FA} in the following examples. We will also denote the probability of detection ($1-$ probability of a false-negative) by P_D. The NP test maximizes P_D subject to a constraint on P_{FA}.

Example: Detecting a DC Signal in Additive White Gaussian Noise

Return to our ETI example from earlier. Assuming $\mu > 0$, our LRT amounts to

$$\sum_{i=1}^{n} x_i \underset{H_0}{\overset{H_1}{\gtrless}} \nu \ ,$$

with $\nu = \frac{\sigma^2}{\mu}\ln\gamma + \frac{n\mu}{2}$, and since γ was ours to choose, we can equivalently choose ν to trade-off between the two types of error.

Let's now determine P_{FA} and P_D for the log-likelihood ratio test.

$$P_{FA} \ = \ \int_{\nu}^{\infty} \frac{1}{\sqrt{2n\pi\sigma^2}} e^{-\frac{t^2}{2n\sigma^2}}\, dt \ = \ Q\left(\frac{\nu}{\sqrt{n\sigma^2}}\right) \ ,$$

where $Q(z) = \int_{u\geq z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du$, the tail probability of the standard normal distribution. Similarly,

$$P_D \ = \ \int_{\nu}^{\infty} \frac{1}{\sqrt{2n\pi\sigma^2}} e^{-\frac{(t-n\mu)^2}{2n\sigma^2}}\, dt \ = \ Q\left(\frac{\nu - n\mu}{\sqrt{n\sigma^2}}\right) \ .$$

In both cases the expression in terms of the Q function is the result of a simple change of variables in the integration. The Q function is invertible, so we can solve for the value of ν in terms of P_{FA}, that is $\nu = \sqrt{n\sigma^2}Q^{-1}(P_{FA})$. Using this we can express P_D as

$$P_D = Q\left(Q^{-1}(P_{FA}) - \sqrt{\frac{n\mu^2}{\sigma^2}}\right),$$

where $\sqrt{\frac{n\mu^2}{\sigma^2}}$ is simply the square root of the signal-to-noise ratio ($\sqrt{SNR}$). Since $Q(z) \to 1$ as $z \to -\infty$, it is easy to see that the probability of detection increases as μ and/or n increase.

Example: Detecting a Change in Variance

Consider the binary hypotheses

$$\begin{aligned} H_0 : \ X_1, \ldots, X_n \ &\overset{iid}{\sim} \ \mathcal{N}(0, \sigma_0^2) \\ H_1 : \ X_1, \ldots, X_n \ &\overset{iid}{\sim} \ \mathcal{N}(0, \sigma_1^2) \ , \ \sigma_1 > \sigma_0 \end{aligned}$$

The log-likelihood ratio test is

$$\frac{n}{2}\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) + \left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right)\sum_{i=1}^{n} x_i^2 \underset{H_0}{\overset{H_1}{\gtrless}} \ln(\gamma) \; .$$

Some simple algebra shows

$$\sum_{i=1}^{n} x_i^2 \underset{H_0}{\overset{H_1}{\gtrless}} \nu$$

with $\nu = 2\left(\frac{\sigma_1^2\sigma_o^2}{\sigma_1^2-\sigma_0^2}\right)(\log(\gamma) + n\ln(\frac{\sigma_1}{\sigma_o}))$. Note that $t := \sum_{i=1}^{n} x_i^2$ is the sufficient statistic for variance of a zero-mean normal distribution.

Now recall that if $X_1, \ldots, , X_n \overset{iid}{\sim} N(0,1)$, then $\sum_{i=1}^{n} X_i^2 \sim \chi_n^2$ (chi-square distributed with n degrees of freedom). Let's rewrite our null hypothesis test using the sufficient statistic:

$$H_0 \;:\; t = \sum_{i=1}^{n} \frac{x_i^2}{\sigma_0^2} \sim \chi_n^2$$

The probability of false alarm is just the probability that a χ_n^2 random variable exceeds ν/σ_0^2. This can be easily computed numerically. For example, if we have $n = 20$ and set $P_{FA} = 0.01$, then the correct threshold is $\nu = 37.57\sigma_0^2$.

5 Receiver Operating Characteristic (ROC) curves

The binary hypothesis test

$$\begin{aligned} H_0 :&\ X = W \\ H_1 :&\ X = S + W \end{aligned}$$

where $W \sim N(0, \sigma^2 I_{n\times n})$ and $S = [s_1, s_2, \ldots, s_n]^T$ is the known signal.

$$\begin{aligned} P_0(X) &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\left(-\frac{1}{2\sigma^2}X^T X\right) \\ P_1(X) &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\left[-\frac{1}{2\sigma^2}(X-S)^T(X-S)\right] \end{aligned}$$

Apply the likelihood ratio test (LRT):

$$\log\Lambda(x) = \log\frac{P_1(X)}{P_0(X)} = -\frac{1}{2\sigma^2}[-2X^T S + S^T S] \underset{H_0}{\overset{H_1}{\gtrless}} \gamma'$$

After simplification, we have

$$X^T S \underset{H_0}{\overset{H_1}{\gtrless}} \sigma^2\gamma' + \frac{S^T S}{2} = \gamma$$

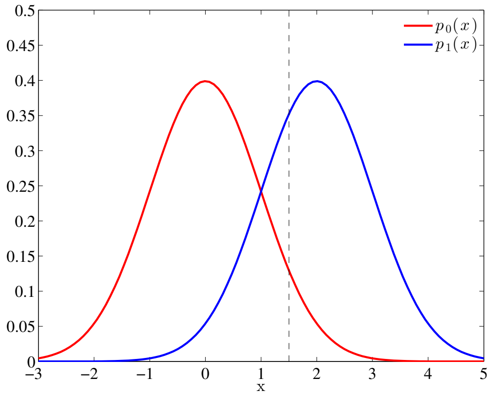
The test statistic $X^T S$ is usually called a "matched filter". The LR detector "filters" data by projecting them onto the signal subspace.

5.1 Quality of the classifier

Question: How can we assess the quality of a detector?

Example:

$$H_0: \; X \sim N(0,1)$$
$$H_1: \; X \sim N(2,1)$$



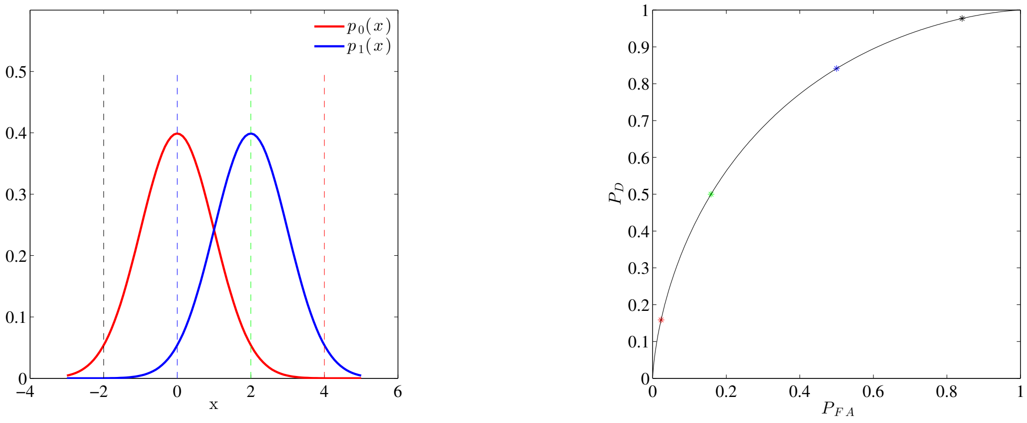
P_{FA} and P_D characterize the performance of the detector. As γ increases, P_{FA} decreases (good) and P_D decreases (bad).

Definition: Receiver Operating Characteristic (ROC) Curve

An ROC curve is a plot that illustrates the performance of a detector (binary classifier) by plotting its P_D vs. P_{FA} at various threshold settings.

First use: In World War II. The ROC curve was first developed by electrical engineers and radar engineers during for detecting aircrafts from radar signals after the attack on the Pearl Harbor.

To compute the ROC curve, vary the threshold level γ and compute P_{FA} and P_D.



The ROC curve

- Starts from $(0,0)$ and ends at $(1,1)$ (unless $p_i(\pm\infty) > 0$).
- The diagonal line from $(0,0)$ to $(1,1)$ corresponds to random guesses.
- Depends on signal strength, noise strength, noise type, etc.

Example: ROC and SNR

$$\begin{aligned} H_0 :&\ X \sim N(0, \sigma^2 I) \\ H_1 :&\ X \sim N(S, \sigma^2 I) \end{aligned}$$

The likelihood ratio test gives

$$X^T S \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

$X^T S$ is also Gaussian distributed. Recall if $X \sim N(\mu, \Sigma)$, then $Y = AX \sim N(A\mu, A\Sigma A^T)$. So we can get

$$\begin{aligned} H_0 :&\ X^T S \sim N(0^T S, S^T \sigma^2 I S) = N(0, \sigma^2 \|S\|_2^2) \\ H_1 :&\ X^T S \sim N(S^T S, S^T \sigma^2 I S) = N(\|S\|_2^2, \sigma^2 \|S\|_2^2) \end{aligned}$$

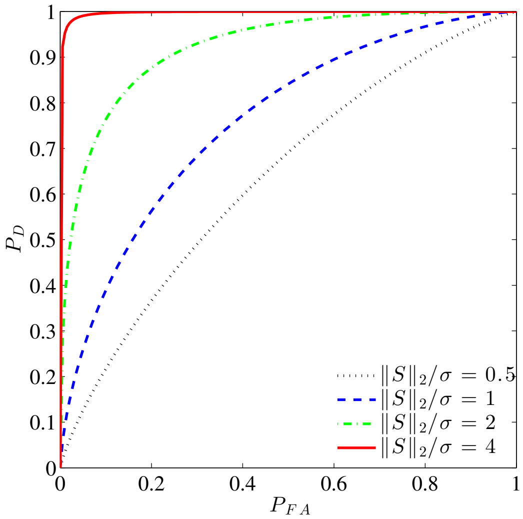
We can also compute

$$\begin{aligned} P_{FA} &= Q(\frac{\gamma - 0}{\sigma \|S\|_2}) \\ P_D &= Q(\frac{\gamma - \|S\|_2^2}{\sigma \|S\|_2}) = Q(\frac{\gamma}{\sigma \|S\|_2} - \frac{\|S\|_2}{\sigma}) \end{aligned}$$

Since Q function is invertible, we can get $\frac{\gamma}{\sigma\|S\|_2} = Q^{-1}(P_{FA})$. Therefore,

$$P_D = Q(Q^{-1}(P_{FA}) - \frac{\|S\|_2}{\sigma}),$$

where $\frac{\|S\|_2}{\sigma}$ is the square root of Signal-to-Noise Ratio($\sqrt{SNR}$).



5.2 The AWGN Assumption

AWGN is gaussian distributed as

$$W \sim N(0, \sigma^2 I)$$

Is real-world noise really additive, white and Gaussian? Noise in many applications (e.g. communication and radar) arise from several independent sources, all adding together at sensors and combining additively to the measurement.

Central Limit Theorem

If $x_1, \ldots, x_n$ are independent random variables with means μ_i and variances $\sigma_i^2 < \infty$,then $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu_i}{\sigma_i} \to N(0,1)$ in distribution as $n \to \infty$.

Thus, it is quite reasonable to model noise as additive and Gaussian in many applications. However, whiteness is not always a good assumption.

5.3 Colored Gaussian Noise

Example: Correlated noise

$W = S_1 + S_2 + \cdots + S_k$, where $S_1, S_2, \ldots S_k$ are interferring signals that are not of interest, and each of them is structured/correlated in time.

$W \sim N(0, \Sigma)$ is called correlated or "colored" noise, where Σ is a structured covariance matrix.

Consider the binary hypothesis test in this case.

$$H_0 : \ X = S_0 + W$$
$$H_1 : \ X = S_1 + W$$

where $W \sim N(0, \Sigma)$ and S_0 and S_1 are know signal. So we can rewrite the hypothesis as

$$\begin{aligned} H_0 :&\ X \sim N(S_0, \Sigma) \\ H_1 :&\ X \sim N(S_1, \Sigma) \end{aligned}$$

The probability density of each hypothesis is

$$P_i(X) = \frac{1}{(2\pi)^{\frac{2}{n}}(\Sigma)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(X - S_i)^T \Sigma^{-1}(X - S_i)\right], i = 0, 1$$

The log likelihood ratio is

$$\begin{aligned} \log\left(\frac{P_1(X)}{P_2(X)}\right) &= -\frac{1}{2}\left[(X - S_1)^T \Sigma (X - S_1) - (X - S_0)^T \Sigma^{-1}(X - S_0)\right] \\ &= X^T \Sigma^{-1}(S_1 - S_0) - \frac{1}{2} S_1^T \Sigma^{-1} S_1 + \frac{1}{2} S_0^T \Sigma^{-1} S_0 \\ &\overset{H_1}{\underset{H_0}{\gtrless}} \gamma' \end{aligned}$$

Equivalently,

$$(S_1 - S_0)^T \Sigma^{-1} X \overset{H_1}{\underset{H_0}{\gtrless}} \gamma' + \frac{S_1^T \Sigma^{-1} S_1}{2} - \frac{S_0^T \Sigma^{-1} S_0}{2} = \gamma$$

Let $t(X) = (S_1 - S_0)^T \Sigma^{-1} X$, we can get

$$\begin{aligned} H_0 :&\ t \sim N((S_1 - S_0)^T \Sigma^{-1} S_0, (S_1 - S_0)^T \Sigma^{-1}(S_1 - S_0)) \\ H_1 :&\ t \sim N((S_1 - S_0)^T \Sigma^{-1} S_1, (S_1 - S_0)^T \Sigma^{-1}(S_1 - S_0)) \end{aligned}$$

The probability of false alarm is

$$P_{FA} = Q\left(\frac{\gamma - (S_1 - S_0)^T \Sigma^{-1} S_0}{[(S_1 - S_0)^T \Sigma^{-1}(S_1 - S_0)]^{\frac{1}{2}}}\right)$$

In this case it is natural to define

$$SNR = (S_1 - S_0)^T \Sigma^{-1}(S_1 - S_0)$$

Example: ROC with colored Gaussian noise

$$\begin{aligned} S_1 &= [\frac{1}{2}, \frac{1}{2}], S_0 = [-\frac{1}{2}, -\frac{1}{2}], \\ \Sigma &= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \Sigma^{-1} = \frac{1}{1 - \rho^2}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}. \end{aligned}$$

The test statistics is

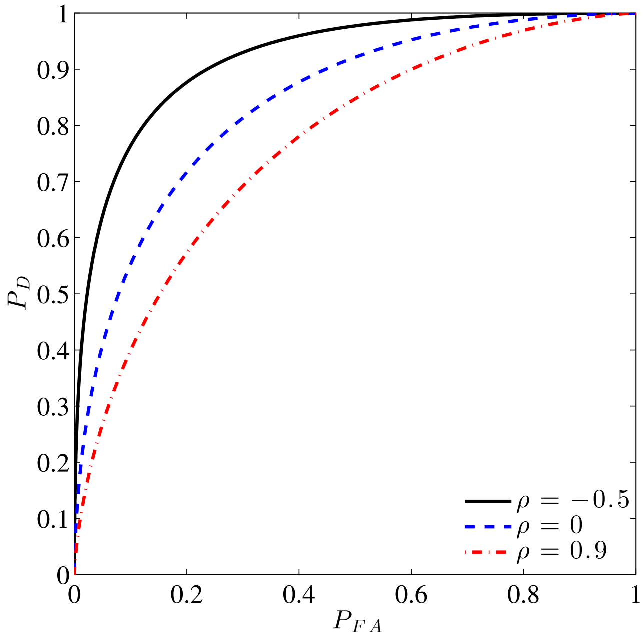
$$\begin{aligned} y &= (S_1 - S_0)^T \Sigma^{-1} X \\ &= [1, 1] \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= \frac{1}{1+\rho}(x_1 + x_2) \end{aligned}$$

The testing problem is equivalent to

$$\begin{aligned} H_0 :\ & y \sim N(-\frac{1}{1+\rho}, \frac{2}{1+\rho}) \\ H_1 :\ & y \sim N(+\frac{1}{1+\rho}, \frac{2}{1+\rho}) \end{aligned}$$

The probabilities of false alarm and detection are

$$\begin{aligned} P_{FA} &= Q(\frac{\gamma + \frac{1}{1+\rho}}{\sqrt{\frac{2}{1+\rho}}}) \\ P_D &= Q(\frac{\gamma - \frac{1}{1+\rho}}{\sqrt{\frac{2}{1+\rho}}}) \end{aligned}$$


P_D
P_{FA}
$\rho = -0.5$
$\rho = 0$
$\rho = 0.9$