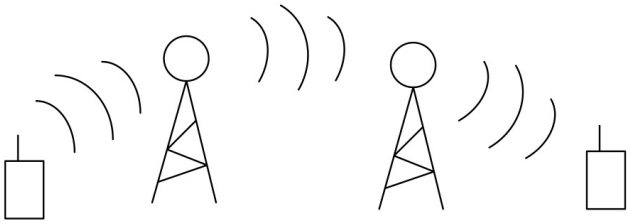
ECE 830 Fall 2010 Statistical Signal Processing

instructor: R. Nowak

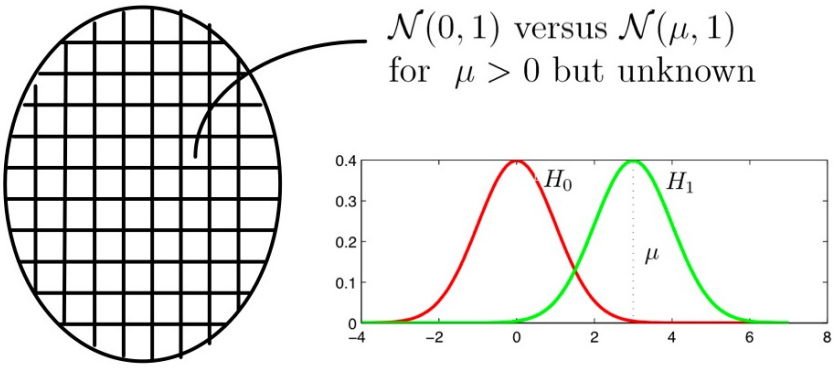
Lecture 10: Composite Hypothesis Testing

In many real world problems, it is difficult to precisely specify probability distributions. Our models for data may involve unknown parameters or other characteristics. Here are a few motivating examples.

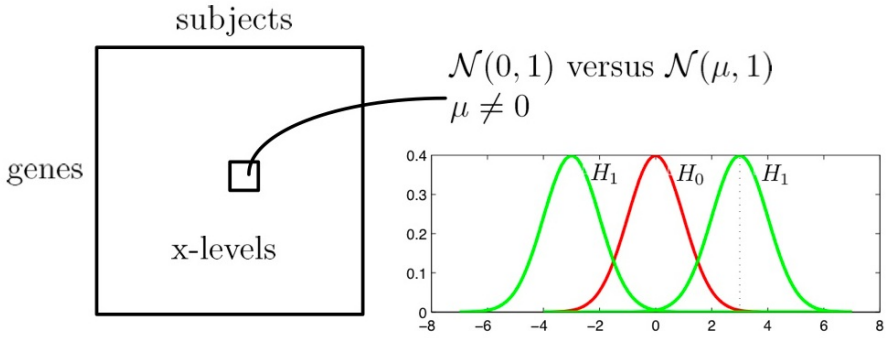
Example 1 *Unknown amplitudes/delays in wireless communications.*



Example 2 *Unknown signal amplitudes in functional brain imaging.*



Example 3 *Unknown expression levels in gene microarray experiments.*



1 Composite Hypothesis Tests

We can represent uncertainty by specifying a collection of possible models for each hypothesis. The collections are indexed by a parameter.

$$\begin{aligned} H_0 : X &\sim p_0(x|\theta_0),\ \theta_0 \in \Theta_0 \\ H_1 : X &\sim p_1(x|\theta_1),\ \theta_1 \in \Theta_1 \end{aligned}$$

In general, the distributions p_0 and p_1 may have different parametric forms. The sets Θ_0 and Θ_1 represent the possible values for the parameters. If a set contains a single element (i.e., a single value for the parameter), then we have a simple hypothesis, as discussed in past lectures. When a set contains more than one parameter value, then the hypothesis is called a *composite* hypothesis, because it involves more than one model. The name is even clearer if we consider the following equivalent expression for the hypotheses above.

$$\begin{aligned} H_0 : X &\sim p_0,\ p_0 \in \{p_0(x|\theta_0)\}_{\theta_0 \in \Theta_0} \\ H_1 : X &\sim p_1,\ p_1 \in \{p_1(x|\theta_1)\}_{\theta_1 \in \Theta_1} \end{aligned}$$

Example 4 *Recall the brain imaging problem.*

$$\begin{aligned} &H_0 : X \sim \mathcal{N}(0,1) \\ &H_1 : X \sim \mathcal{N}(\mu,1),\ \mu > 0 \text{ but otherwise unknown} \\ &\qquad \text{equivalently } X \sim p, p \in \{\mathcal{N}(\mu,1)\}_{\mu>0} \end{aligned}$$

In this example, H_0 is simple and H_1 is composite.

1.1 Uniformly Most Powerful Tests

Let us begin by considering special cases in which the usual likelihood ratio test is computable and optimal. Here is an example.

$$\begin{aligned} H_0 :&\quad x_1, \cdots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \\ H_1 :&\quad x_1, \cdots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu,1),\ \mu > 0 \end{aligned}$$

Log LRT:

$$\log\left(\frac{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-(x_i-\mu)^2/2}}{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-(x_i)^2/2}}\right) = \sum_{i=1}^{n}-\frac{(x_i-\mu)^2}{2}+\frac{x_i^2}{2}$$
$$= \mu\sum_{i=1}^{n}x_i-\frac{n\mu^2}{2}$$

Test statistic:

$$\mu\sum_{i=1}^{n}x_i \underset{H_0}{\overset{H_1}{\gtrless}} \gamma' \quad\equiv\quad \sum_{i=1}^{n}x_i \underset{H_0}{\overset{H_1}{\gtrless}} \gamma'/\mu=\gamma$$

We were able to divide both sides by μ since $\mu > 0$. We do not need to know the exact value of μ in order to compute the test $\sum_i x_i \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$ for any value of γ. Let $t = \sum_{i=1}^{n} x_i$ denote the test statistic. It is easy to determine its distribution(s) under each hypothesis (a composite in the case of H_1).

$$H_0: \quad t \sim \mathcal{N}(0, n)$$
$$H_1: \quad t \sim \mathcal{N}(n\mu, n) \quad \mu > 0 \text{ unknown}$$

Since distribution of t under H_0 is known, we can choose threshold to control P_{FA}.

$$P_{FA} = Q\left(\frac{\gamma}{\sqrt{n}}\right) \Rightarrow \gamma = \sqrt{n}Q^{-1}\left(P_{FA}\right)$$

This is optimal detector (most powerful) according to NP lemma. Several ROC curves corresponding to different values of the unknown parameter $\mu > 0$ are depicted below. We cannot know which curve we are operating on, but we can choose a threshold for a desired P_{FA} and the resulting P_D is the best possible (for the unknown value of μ). In such cases we say that the test is *uniformly most powerful*, that is most powerful no matter what the value of the unknown parameter.

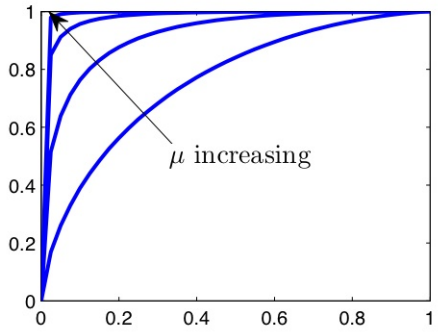


Figure 1: ROC of varied $\mu > 0$ for the simple case.

1.2 Karlin-Rubin Theorem

Theorem 1 *Let t be a scalar test statistic whose density, under both hypotheses, is parameterized by a scalar parameter θ. Assume that the likelihood ratio statistic*

$$\Lambda(t) = \frac{p(t|\theta_1)}{p(t|\theta_0)}$$

is a non-decreasing function of t for every pair $(\theta_0, \theta_1 > \theta_0)$. *We say that t has a monotone likelihood ratio, and the idea is depicted in the figure below. The interpretation is simple: the larger t is, the more probable* H_1 *looks compared to* H_0. *In this case, the threshold test*

$$t \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

is the test that maximizes P_D *for a given* P_{FA} *(both depend on* γ*) for all* $(\theta_0, \theta_1 > \theta_0)$. *We say that this test is uniformly most powerful (UMP) among all tests with this* P_{FA}.

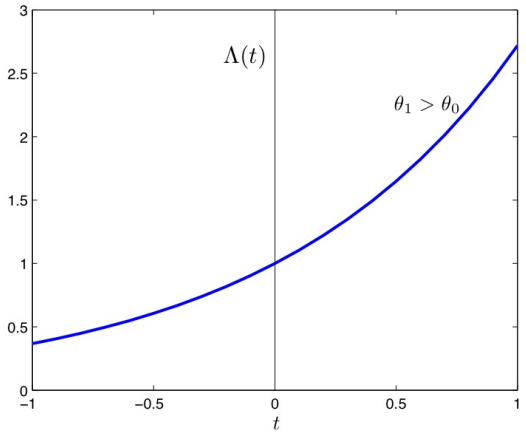


Figure 2: $\Lambda(t)$ as a non-decreasing function of t for a pair $\theta_1 > \theta_0, \theta_0$.

Example 5 *Poisson*

$$\begin{aligned} H_0: &\quad x \sim Poisson(\lambda_0) \\ H_1: &\quad x \sim Poisson(\lambda_1),\ \lambda_1 > \lambda_0 \end{aligned}$$

$$\Lambda(x) = e^{-(\lambda_1 - \lambda_0)} \left(\frac{\lambda_1}{\lambda_0}\right)^x$$

non-decreasing function $\ln x$ *for any* $(\lambda_1 > \lambda_0, \lambda_0)$ *pair.*

Example 6

$$\begin{aligned} H_0: &\quad x = w, \qquad x \sim \mathcal{N}(0, \Sigma) \\ H_1: &\quad x = As + w, x \sim \mathcal{N}(As, \Sigma) \end{aligned}$$

$w \sim \mathcal{N}(0, \sigma)$, *s n-by-1 known waveform,* $A > 0$ *unknown amplitude.*
log LRT:

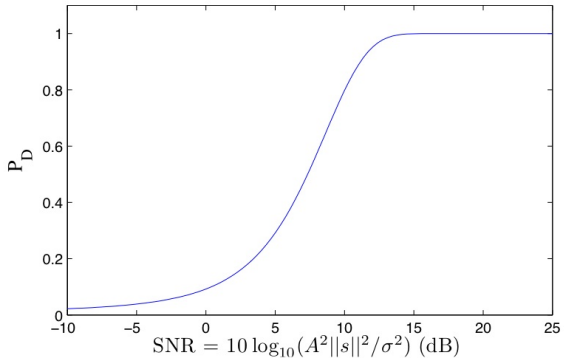
$$\begin{aligned} \log \Lambda(x) &\underset{H_0}{\overset{H_1}{\gtrless}} \gamma' \\ \underbrace{A}_{+} \underbrace{s^T \Sigma^{-1} x}_{t} - \frac{A^2 s^T \Sigma^{-1} s}{2} &\underset{H_0}{\overset{H_1}{\gtrless}} \gamma' \quad \log \Lambda \text{ monotone in } t \\ t(x) = s^T \Sigma^{-1} x &\underset{H_0}{\overset{H_1}{\gtrless}} \gamma, \quad \text{since } A > 0. \end{aligned}$$

$$p(t|A) = \mathcal{N}(As^T\Sigma, s^T\Sigma^{-1}s)$$
$$p(t|0) = \mathcal{N}(0, s^T\Sigma^{-1}s)$$
$$P_{FA} = Q\left(\frac{\gamma}{\sqrt{s^T\Sigma^{-1}s}}\right)$$

Suppose $\Sigma = \sigma^2 I$, $\Rightarrow$ *log LRT*

$$s^T x \gtrless \gamma$$
$$P_{FA} = Q\left(\frac{\gamma}{\sqrt{\sigma^2 s^T s}}\right),\ \gamma = \sqrt{||s||^2\sigma^2}Q^{-1}\left(P_{FA}\right)$$
$$P_D = Q\left(\frac{\gamma - A||s||^2}{\sqrt{\sigma^2||s||^2}}\right) = Q\left(Q^{-1}\left(P_{FA}\right) - \sqrt{\frac{A^2||s||^2}{\sigma^2}}\right)$$

where we term $SNR = A^2||s||^2/\sigma^2$. *If* $s_\ell = A\sin(\frac{2\pi}{10}\ell), \ell = 1, \cdots, 100$, $P_{FA} = 10^{-2}$



1.3 Two-sided Tests

To see how special the UMP condition is, consider the following simple generalization of the testing problems above.

$$H_0: \quad x \sim \mathcal{N}(0,1)$$
$$H_1: \quad x \sim \mathcal{N}(\mu,1),\ \mu \neq 0$$

The log-likelihood ratio statistic is

$$\log\Lambda(x) \ = \ -\frac{(x-\mu)^2}{2} + \frac{x^2}{2} = \mu x - \mu^2/2$$

and the log-LRT has the form

$$\mu\, x - \mu^2/2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma' \ .$$

We can move the term $\mu^2/2$ to the other side and absorb it into the threshold, but this leaves us with a test of the form

$$\mu\, x \underset{H_0}{\overset{H_1}{\gtrless}} \gamma \ ,$$

and since μ is unknown (and not necessarily positive) the test is uncomputable. How can we proceed? Look at two densities in Figure 4. Intuitively the test $|x| \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$ seems reasonable. This is called the *Wald Test*.

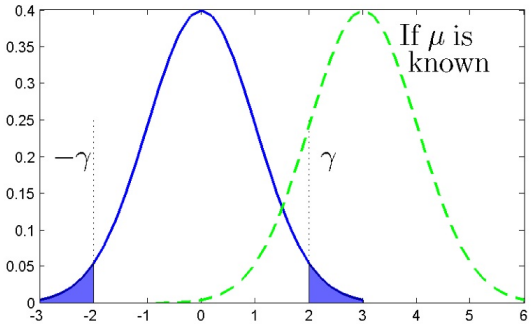


Figure 3

The P_{FA} of the Wald test can be seen from Figure 5.

$$\begin{aligned} P_{FA} &= 2Q(\gamma) \Rightarrow \gamma = Q^{-1}(P_{FA}/2) \\ P_D &= \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2} dx + \int_{-\infty}^{-\gamma} \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2} dx \quad y = x - \mu \\ &= \int_{\gamma-\mu}^{\infty} \mathcal{N}(0,1) dy + \int_{-\infty}^{-\gamma-\mu} \mathcal{N}(0,1) dy \\ &= Q(\gamma - \mu) + Q(\gamma + \mu) \ . \end{aligned}$$

The P_D depends on μ, which is unknown.

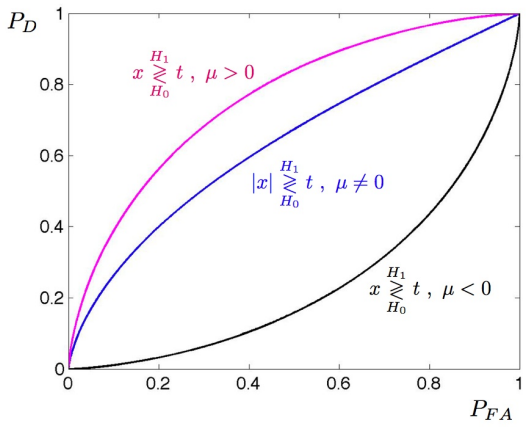


Figure 4

1.4 Two "Derivations" of the Wald Test

Generalized Likelihood Ratio Test (GLRT) Model μ as a deterministic, but unknown, parameter. Estimate μ from the data and plug the estimate into the LRT. Under H_1 the distribution is $X \sim \mathcal{N}(\mu, 1)$, so a natural estimate for μ is $\widehat{\mu} = x$, the observation itself. The plugging this into the likelihood ratio yields

$$\widehat{\Lambda}(x) \ = \ \frac{p(x|\widehat{\mu})}{p(x|0)} \ = \ \frac{\exp(-(x-\widehat{\mu})^2/2)}{\exp(-x^2/2)} \ = \ e^{x^2/2} \ .$$

This is the generalized likelihod ratio. In effect, this compares the *best* fitting model in the composite hypothesis H_1 with the model H_0. Taking the log yields the test

$$\log \widehat{\Lambda}(x) \;=\; x^2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma \;,$$

which is equivalent to the Wald test.

Bayes Factor Model μ as an independent random parameter with prior probability distribution $p(\mu)$. The alternative hypothesis is that $\mu \neq 0$, and with no other prior information it is natural to take $p(\mu)$ to be symmetric about the origin. In particular, the prior probability distirbution $\mu \sim \mathcal{N}(0, \sigma^2)$ is symmetric and models a prior belief that smaller values of μ are more probable that larger values. The Gaussian form is also convenient to use with the Gaussian likelihood. The *Bayes Factor* is the ratio

$$\Lambda_{BF}(x) \;=\; \frac{\int p(x|\mu)p(\mu)d\mu}{p(x|0)} \;.$$

This ratio compares the *average* model in H_1 (with respect to the prior $p(\mu)$) with the H_0 model. The integral in the numerator is easy to compute. Note that $X = \mu + W$, where $W \sim \mathcal{N}(0,1)$. So X is the sum of two independent Gaussian random variables, and its distribution is $X \sim N(0, 1+\sigma^2)$. The Bayes Factor is therefore

$$\Lambda_{BF}(x) \;=\; \frac{\frac{1}{\sqrt{2\pi(1+\sigma^2)}} \exp(-\frac{x^2}{2(1+\sigma^2)})}{\frac{1}{\sqrt{2\pi}} \exp(-x^2/2)} \;.$$

Taking the log and absorbing constant factors and terms into the threshold yields the test

$$x^2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma \;,$$

which again is equivalent to the Wald test.

1.5 GLRT and Bayes Factors

Consider a composite hypothesis test of the form

$$\begin{aligned} H_0 : X &\;\sim\; p_0(x|\theta_0),\; \theta_0 \in \Theta_0 \\ H_1 : X &\;\sim\; p_1(x|\theta_1),\; \theta_1 \in \Theta_1 \end{aligned}$$

The general forms for the GLRT and Bayes Factor are as follows.

GLRT

$$\frac{\max_{\theta_1 \in \Theta_1} p_1(x|\theta_1)}{\max_{\theta_0 \in \Theta_0} p_0(x|\theta_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma \;.$$

Bayes Factor Assume $\theta_0 \sim p(\theta_0)$ and $\theta_1 \sim p(\theta_1)$, two different prior probability distributions. The Bayes Factor is

$$\frac{\int_{\Theta_1} p_1(x|\theta_1)p(\theta_1)d\theta_1}{\int_{\Theta_0} p_0(x|\theta_0)p(\theta_0)d\theta_0} \;.$$

The GLRT compares the best model in H_1 to the best in H_0, and the Bayes Factor compares the average model in H_1 to the average model in H_0, with respect to the specified prior probability distributions.

Example 7

$$\begin{aligned} H_0 :&\quad \mathcal{N}(0, \sigma^2 I), \quad \sigma^2 > 0 \textit{ known}, \theta \in \mathbb{R}^k \textit{ unknown} \\ H_1 :&\quad \mathcal{N}(H\theta, \sigma^2 I), \quad H = [h_1, h_2, \cdots, h_k] \textit{ n-by-k known} \end{aligned}$$

log LR

$$\begin{aligned}\Lambda(x) &= -\frac{1}{2\sigma^2}\left((x-H\theta)^T(x-H\theta)-x^Tx\right)\\ &= -\frac{1}{\sigma^2}(-2\theta^TH^Tx+\theta^TH^TH\theta)\underset{H_0}{\overset{H_1}{\gtrless}}\gamma'\\ &\quad \theta^TH^T\underset{H_0}{\overset{H_1}{\gtrless}}\gamma \text{ (not computable w/o knowledge of } \theta)\end{aligned}$$

Recall that

$$\begin{aligned}H_1: &\quad x\sim\mathcal{N}(H\theta,\sigma^2I),\theta\in\mathbb{R}^k\\ or &\quad x\sim p_1, p_1\in\{\mathcal{N}(H\theta,\sigma^2I)\}_{\theta\in\mathbb{R}^k},\end{aligned}$$

we want to pick p_1 *in* $\{\mathcal{N}(H\theta,\sigma^2I)\}$ *that matches* x *the best.*

$$p(x|\theta,H_1)=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp(-\frac{1}{2\sigma^2}(x-H\theta)^T(x-H\theta))$$

Find θ *that maximizes prob. of observing* x

$$\hat{\theta}=\arg\min_{\theta\in\mathbb{R}^k}\underbrace{(x-H\theta)^T(x-H\theta)}_{||x-H\theta||^2}=(H^TH)^{-1}H^Tx$$

Plugging $\hat{\theta}$ *into the test statistic* θ^TH^Tx*, we have*

$$\begin{aligned}&\hat{\theta}^TH^Tx\\ =\;& x^TH(H^TH)^{-1}H^Tx\underset{H_0}{\overset{H_1}{\gtrless}}\gamma\;\sim\;\mathcal{X}_k^2\end{aligned}$$

This is the so-called Generalized LRT (GLRT) and its distribution is chi-square with k *degrees of freedom (*k *being the dimension of the subspace spanned by the columns of* H*). This distribution is denoted* $\mathcal{X}_k^2$*. In lecture we also showed that using the prior* $\theta\sim\mathcal{N}(0,\alpha^2I)$ *and computing the Bayes Factor yields the same test. The test computes the energy in the* signal subspace *and if the energy is large enough, then* H_1 *is accepted.*