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: Unknown amplitudes/delays in wireless communications.**

We don’t always know how many relays a signal will go through, how strong the signal will be at each receiver, the distance between relay stations, etc.

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

\( \mathcal{N}(0,1) \) versus \( \mathcal{N}(\mu,1) \) for \( \mu > 0 \) but unknown

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

\( \mathcal{N}(0,1) \) versus \( \mathcal{N}(\mu,1) \) for \( \mu \neq 0 \)
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.

\[ H_0 : X \sim p_0(x|\theta_0), \theta_0 \in \Theta_0 \]
\[ H_1 : X \sim p_1(x|\theta_1), \theta_1 \in \Theta_1 \]

- In general, the distributions \( p_0 \) and \( p_1 \) may have different parametric forms.
- The sets \( \Theta_0 \) and \( \Theta_1 \) 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.

\[ H_0 : X \sim p_0, \ p_0 \in \{p_0(x|\theta_0)\}_{\theta_0 \in \Theta_0} \]
\[ H_1 : X \sim p_1, \ p_1 \in \{p_1(x|\theta_1)\}_{\theta_1 \in \Theta_1} \]

Example: Brain imaging

Recall the brain imaging problem.

\[ H_0 : X \sim \mathcal{N}(0, 1) \]
\[ H_1 : X \sim \mathcal{N}(\mu, 1), \ \mu > 0 \text{ but otherwise unknown} \]

equivalently \( X \sim p, p \in \{\mathcal{N}(\mu, 1)\}_{\mu > 0} \)

In this example, \( H_0 \) is simple and \( H_1 \) is composite.

2 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.

\[ H_0 : \ x_1, \cdots, x_n \overset{iid}{\sim} \mathcal{N}(0, 1) \]
\[ H_1 : \ x_1, \cdots, x_n \overset{iid}{\sim} \mathcal{N}(\mu, 1), \ \mu > 0 \]

Log LRT:

\[ \log \left( \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\mu)^2/2}}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i)^2/2}} \right) = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2} + \frac{x_i^2}{2} = \mu \sum_{i=1}^{n} x_i - \frac{n\mu^2}{2} \]
Composite Hypotheses and Generalized Likelihood Ratio Tests

Test statistic:

\[ \mu \sum_{i=1}^{n} x_i \overset{H_1}{\gtrless} H_0 \gamma' \iff \sum_{i=1}^{n} x_i \overset{H_1}{\gtrless} H_0 \frac{\gamma'}{\mu} = \gamma \]

We were able to divide both sides by \( \mu \) since \( \mu > 0 \). We do not need to know the exact value of \( \mu \) in order to compute the test \( \sum_i x_i \overset{H_1}{\gtrless} H_0 \gamma \) for any value of \( \gamma \).

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 \( H_1 \)).

\( H_0 : \quad t \sim N(0,n) \)

\( H_1 : \quad t \sim N(n\mu,n) \quad \mu > 0 \) unknown

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

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

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 \( \mu \)). 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 for various \( \mu > 0 \) for the simple case.](image)

**Definition: Uniformly Most Powerful Test**

A uniformly most powerful (UMP) test is a hypothesis test which has the greatest power (i.e. greatest probability of detection) among all possible tests yielding a given false alarm rate regardless of the underlying true parameter(s).
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

\[ \mu x - \frac{\mu^2}{2} \overset{H_1 \sim H_0}{\geq} \gamma'. \]

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

\[ \mu x \overset{H_1}{\geq} \overset{H_0}{\sim} \gamma. \]

Since \(\mu\) is unknown (and not necessarily positive) the test is uncomputable.

**How can we proceed?** Look at two densities in the microarray experiment. Intuitively the test \(\vert x \vert \overset{H_1}{\geq} \overset{H_0}{\sim} \gamma\) seems reasonable. This is called the Wald Test. The \(P_{FA}\) of the Wald test can be seen below.

\[
\begin{align*}
P_{FA} & = 2Q(\gamma) \Rightarrow \gamma = Q^{-1}(P_{FA}/2) \\
P_D & = \int_{-\gamma}^{\gamma} \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2} dx + \int_{-\infty}^{-\gamma} \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2} dx \\ & \quad + \int_{\gamma}^{\infty} \mathcal{N}(0,1) dy + \int_{-\infty}^{-\gamma-\mu} \mathcal{N}(0,1) dy \\ & = Q(\gamma - \mu) + Q(\gamma + \mu).
\end{align*}
\]
The $P_D$ depends on $\mu$, which is unknown.

Model $\mu$ as a deterministic, but unknown, parameter. Estimate $\mu$ 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 $\mu$ is $\hat{\mu} = x$, the observation itself. The plugging this into the likelihood ratio yields

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

This is the generalized likelihood 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

$$
\log \hat{\Lambda}(x) = x^2 \stackrel{H_1}{\gtrless} H_0 \gamma,
$$

which is equivalent to the Wald test.

4 The Generalized Likelihood Ratio Test (GLRT)

Consider a composite hypothesis test of the form

$$
H_0 : X \sim p_0(x|\theta_0), \ \theta_0 \in \Theta_0
H_1 : X \sim p_1(x|\theta_1), \ \theta_1 \in \Theta_1
$$

The parametric densities $p_0$ and $p_1$ need not have the same form.

The generalized likelihood ratio test (GLRT) is a general procedure for composite testing problems. The basic idea is to compare the best model in class $H_1$ to the best in $H_0$, which is formalized as follows.

**Definition: Generalized Likelihood Ratio Test (GLRT)**
The GLRT based on an observation $x$ of $X$ is

$$\hat{\Lambda}(x) = \frac{\max_{\theta_1 \in \Theta_1} p_1(x|\theta_1) \ H_1}{\max_{\theta_0 \in \Theta_0} p_0(x|\theta_0) \ H_0} \gtrless \gamma,$$

or equivalently

$$\log\hat{\Lambda}(x) \overset{H_1}{\gtrless} \frac{H_0}{\gtrless} \gamma.$$

**Example:**

We observe a vector $X \in \mathbb{R}^n$, and consider two hypotheses:

- $H_0 : X \sim \mathcal{N}(0, \sigma^2 I_n)$
- $H_1 : X \sim \mathcal{N}(H\theta, \sigma^2 I_n)$.

where $\sigma^2 > 0$ is known, $\theta \in \mathbb{R}^k$ is unknown, and $H \in \mathbb{R}^{n \times k}$ is known and full rank. The mean vector $H\theta$ is a model for a signal that lies in the $k$-dimensional subspace spanned by the columns of $H$ (e.g., a narrowband subspace, polynomial subspace, etc.). In other words, the signal has the representation

$$s = \sum_{i=1}^k \theta_i h_i, \ H = [h_1, \ldots, h_k].$$

The null hypothesis is that no signal is present (noise only).

The log likelihood ratio is

$$\Lambda(x) = -\frac{1}{2\sigma^2} (x - H\theta)^\top (x - H\theta) - x^\top x$$

$$= -\frac{1}{\sigma^2} (-2\theta^\top H^\top x + \theta^\top H^\top H\theta)$$

Thus we may consider the test

$$\theta^\top H^\top x \overset{H_1}{\gtrless} \frac{H_0}{\gtrless} \gamma,$$

but this test is not computable without knowledge of $\theta$. Recall that

$$H_1 : x \sim \mathcal{N}(H\theta, \sigma^2 I), \ \theta \in \mathbb{R}^k \iff H_1 : x \sim p_1, \ p_1 \in \left\{ \mathcal{N}(H\theta, \sigma^2 I) \right\}_{\theta \in \mathbb{R}^k}.$$

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

$$p(x|\theta, H_1) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - H\theta)^\top (x - H\theta) \right\}$$
Find $\theta$ that maximizes the likelihood of observing $x$:

$$\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^k} (x - H\theta)^\top (x - H\theta).$$

Taking the gradient with respect to $\theta$

$$\frac{\partial}{\partial \theta} (x^\top x - 2\theta^\top H^\top x + \theta^\top H^\top H\theta) = 0$$

$$\Rightarrow 0 - 2H^\top x + 2H^\top H\theta = 0$$

$$\Rightarrow \hat{\theta} = (H^\top H)^{-1} H^\top x$$

Plugging $\hat{\theta}$ into the test statistic $\theta^\top H^\top x$, we have

$$\hat{\theta}^\top H^\top x = x^\top H(H^\top H)^{-1}H^\top x \overset{H_1}{\gtrless} \gamma$$

Recall that the projection matrix onto the subspace is defined as $P_H := H(H^\top H)^{-1}H^\top$

$$\log \widehat{\Lambda}(x) = \frac{1}{2\sigma^2} x^\top P_H x = \frac{1}{2\sigma^2} \|P_H x\|^2.$$ 

Thus the GLRT computes the energy in the signal subspace and if the energy is large enough, then $H_1$ is accepted. In other words, we are taking the projection of $x$ onto $H$ and measuring the energy. The expected value of this energy under $H_0$ (noise only) is

$$\mathbb{E}_{H_0} [\|P_H X\|^2_2] = k\sigma^2,$$

since a fraction $k/n$ of the total noise energy $n\sigma^2$ falls into this subspace.

The performance of the subspace energy detector can be quantified as follows. We choose a $\gamma$ for the desired $P_{FA}$:

$$\frac{1}{\sigma^2} x^\top P_H x \overset{H_0}{\gtrless} \gamma$$

What is the distribution of $x^\top P_H x$ under $H_0$? First use the decomposition

$$P_H = UU^\top$$

where $U \in \mathbb{R}^{n \times k}$ with orthonormal columns spanning columns of $H$, and let $y := U^\top x$. Then

$$\frac{1}{\sigma^2} x^\top P_H x = \frac{1}{\sigma^2} x^\top UU^\top x = \frac{1}{\sigma^2} y^\top y$$

$$y \sim \mathcal{N}(0, \sigma^2 U^\top U) \equiv \mathcal{N}(0, \sigma^2 I_{k \times k})$$

$$y_i/\sigma \sim \mathcal{N}(0, 1), \quad i = 1, \ldots, k$$

$$\Rightarrow \frac{y^\top y}{\sigma^2} \sim \chi_k^2,$$ 

chi-squared with $k$ degrees of freedom
Under $H_0$, 
\[ \frac{1}{\sigma^2} x^T P_H x \sim \chi^2_k \quad \implies \quad P_{FA} = \mathbb{P}(\chi^2_k > \gamma) \]

To calculate the tails on $\chi^2_k$ distributions you can software such as Matlab
(chi2cdf(x,k), chi2inv(γ,k), chi2cdf(x,k)). Remember the mean of a $\chi^2_k$ distribution
is $k$, so we want to choose a $\gamma$ bigger than $k$ to produce a small $P_{FA}$.

## 5 Wilks’ Theorem

**Wilk’s Theorem (1938)**

Consider a composite hypothesis testing problem

\[ H_0 : X_1, X_2, ..., X_n \overset{iid}{\sim} p(x|\theta_0), \]
\[ \text{where } \theta_{0,1}, \ldots, \theta_{0,\ell} \in \mathbb{R} \text{ are free parameters and } \]
\[ \theta_{0,\ell+1} = a_{\ell+1}, \ldots, \theta_k = a_k \text{ are fixed at the values } \]
\[ a_{\ell+1}, \ldots, a_k \]
\[ H_1 : X_1, X_2, ..., X_n \overset{iid}{\sim} p(x|\theta_1), \theta_1 \in \mathbb{R}^k \text{ are all free parameters} \]

and the parametric density has the same form in each hypothesis.

In this case family of models in $H_0$ is a subset of those in $H_1$, and we say that
the hypotheses are nested. (This is a key condition that must hold for this theorem.)

If the 1st and 2nd order derivatives of $p(x|\theta_i)$ with respect to $\theta_i$ exist and if
\[ \mathbb{E} \left[ \frac{\partial \log p(x|\theta_i)}{\partial \theta_i} \right] = 0 \] (which guarantees that the MLE $\hat{\theta}_i \to \theta_i$ as $n \to \infty$), then the
generalized likelihood ratio statistic, based on an observation $X = (X_1, \ldots, X_n)$,

\[ \hat{\Lambda}_n(X) = \frac{\max_{\theta_1} p(x|\theta_1)}{\max_{\theta_0} p(x|\theta_0)} \]

(1)
has the following asymptotic distribution under $H_0$:

\[ 2 \log \hat{\Lambda}(x) \xrightarrow{n \to \infty} \chi^2_{k-\ell} \quad \text{i.e.,} \quad 2 \log \hat{\Lambda}(x) \overset{D}{\to} \chi^2_{k-\ell} \]

**Proof:** (Sketch) under the conditions of the theorem, the log GLRT tends to the log GLRT in a Gaussian setting according to the Central Limit Theorem (CLT).

**Example: Nested Condition**

$H_0 : x_i \overset{iid}{\sim} \mathcal{N}(0, 1)$

$H_1 : x_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), i = 1, 2, \ldots, n, \sigma^2 > 0$ unknown

log LR:

\[ \sum_{i=1}^{n} \left( -\frac{1}{2} \log \sigma^2 - x_i^2 \left( \frac{1}{2\sigma^2} - \frac{1}{2} \right) \right) \]

MLE of $\sigma^2$:

\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \]

log GLR under $H_0$:

\[ 2 \left[ \sum -\frac{1}{2} \log \left( \frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) - \frac{x_i^2}{2} \left( \frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) - 1 \right] \xrightarrow{n \to \infty} \chi^2_1 \]