Searches and Likelihood Ratio Tests

8 Oct 2021

Reed Essick

Learning Objectives

- Basics of Decision Theory (how to build optimal searches)
- Neyman-Pearson Lemma
- Sufficient Statistics
- Marginalization vs. Maximization (max-likelihood estimators)
- Bayes Factors vs. Odds Ratios (Occam Factors)
- Computational Methods
  - Savage-Dickey Density Ratio
  - Thermodynamic Integration (power sampling)
  - Nested Sampling

Decision Theory as classification, hypothesis testing

Consider 2 classes of events, each of which will produce a set of observables x with some probability

$p(x|A)$: prob. of x given event is of type A

$p(x|B)$: prob of x given event is of type B

We wish to determine which type a particular event is based on the observables for that event. What is the "most powerful test" to distinguish between the classes?

The Neyman-Pearson Lemma demonstrates an optimal statistic is the Likelihood Ratio Test.

$$\Lambda_B^A(x) \equiv \frac{p(x|A)}{p(x|B)}$$

This is derived by optimizing the true alarm prob. at a fixed false alarm prob; both defined as integrals over regions of observable space (R)

$$P_{\text{true alarm}}(R) = \int_R dx\, p(x|A)$$

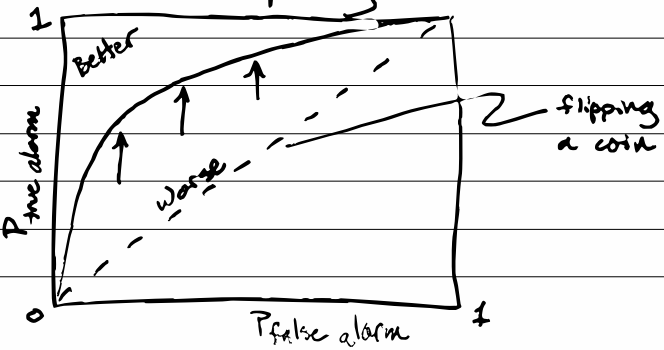
$$P_{\text{false alarm}}(R) = \int_R dx\, p(x|B)$$

Defining regions $\left(R(\Lambda_0) : \Lambda_B^A(x) \geq \Lambda_0\right)$ maximizes

$$\int_R dx\, p(x|A) - \lambda\left(P_0 - \int_R dx\, p(x|B)\right)$$

That is, this maximizes the prob. of correctly identifying an event of class A at a fixed prob. of misidentifying an event of class B.

This is often plotted as a Receiver Operating Characteristic (ROC) curve:
Λ_B^A maximizes the height of this curve @ every point in its domain.



The likelihood ratio test is of fundamental importance because it tells you how to order multi-dimensional observations. It is also a sufficient statistic.

Consider a set of observables: x. We can ask whether adding additional observables y will carry more information. That is, are we better able to distinguish events when we consider $y \oplus x$ than when we only consider x? First, we need a concept of information.

The mutual information is defined as:

$$I(a,b) = \int da\, db\; p(a,b) \log\left[\frac{p(a,b)}{p(a)\,p(b)}\right]$$

w/ $p(a) = \int db\, p(a,b)$; $p(b) = \int da\, p(a,b)$

(KL-Divergence between joint distrib & product of marginal distribs)

larger values of I imply more "correlation" between a & b.

Let us consider some function $c(b)$. The data processing inequality states

$$I(a,b) \geq I(a,c)$$

with inequality iff c is a sufficient statistic for b. That is, sufficient statistics contain all the information available.

The likelihood ratio test maps an N-dim space onto the real line & retains all available information from the original space!

Bayes Factors (ratios of marginal likelihoods) and Odds Factors (ratios of posterior probabilities) are examples of likelihood ratio tests.

$$B^A_B = \frac{p(h \mid H_A)}{p(h \mid H_B)} = \frac{\int ds\, p(h|s)\, p(s|H_A)}{\int ds\, p(h|s)\, p(s|H_B)}$$ evidence ratio

$$O^A_B = \frac{p(H_A|h)}{p(H_B|h)} = B^A_B \frac{p(H_A)}{p(H_B)}$$

$\frac{p(H_A)}{p(H_B)}$ ↑ prior odds

∴ under the assumption that all modeling is correct (noise distrib, signal model, etc) then ordering events by the Bayes Factor (or odds ratio) is an optimal test in that there are sufficient statistics to distinguish btwn hypotheses A & B.

Note, however, that the calculation of B often involves a high-dimensional integral over the signal space. This can be difficult & expensive in practice.

Searches, thus, often define <u>maximum likelihood</u> estimates as their search statistics.

Note! Maximization over θ is not equivalent to marginalization over θ

Max-likelihood estimators are often computationally efficient ways to obtain good detection statistics or estimates of signal parameters. In general, these are defined as

$$\hat{\theta}_{MLE}(h) = \underset{\theta}{\operatorname{argmax}} \{ p(h|\theta) \} = \underset{\theta}{\operatorname{argmax}} \{ \log p(h|\theta) \}$$

$$\lambda_{MLE} = p(h|\hat{\theta}_{MLE}(h))$$

One can then use λ_{MLE} as a detection statistic & $\hat{\theta}_{MLE}$ as an estimate of the parameters. MLE estimators are usually well behaved (unbiased, Fisher efficient).

In general, to build an Optimal Search you should

1. write down the likelihood
2. write down your signal model & prior
   if you don't have a prior, skip to 3b)
3. see if you can efficiently compute Bayes Factors between signal and noise hypotheses
   a) if yes, use the Bayes Factor as detection stat.
   b) if not, define max. likelihood estimate and use that as detection statistic
4. measure distribution of detection statistic under null hypothesis (noise-only model) to determine map from detection-statistic → false alarm prob. (or more generally the full ROC curve)

WORKED EXAMPLE: Matched Filtering in Stationary Gaussian Noise

Let's assume we know the signal's shape (s) perfectly but do not know it's amplitude ($A \in \mathbb{R}$). This means our likelihood is

$$\log p(h|As) = -\frac{1}{2}\left(4\int_0^\infty df\ \frac{|h - As|^2}{S_n}\right)$$

where S_n is the <u>One-Sided Power Spectral Density</u> defined as

$$\langle n(t)n(t+\tau)\rangle = \frac{1}{2}\int df\, e^{2\pi i f\tau}\, S_n(f)$$

NOTE: if we define a prior for A under the signal model then we may be able to define a Bayes Factor for signal ($A>0$) vs. noise ($A=0$).

Let's instead define a MLE for A

$$\frac{\partial \log p}{\partial A} = -2\int_0^\infty df\ \frac{2A|s|^2 - 2\,\mathrm{Re}\{h^* s\}}{S_n} = 0$$

$$\Rightarrow \hat{A} = \frac{4\int_0^\infty df\ \mathrm{Re}\{h^* s\}/S_n}{4\int_0^\infty df\ |s|^2/S_n}$$

(the factors of 4 are conventional)

Now, we are free to choose the scale of the template s and it is convenient to normalize it so that

$$4\int_0^\infty df\,\frac{|s|^2}{S_n} = 1$$

so that $\hat{A} = 4\int_0^\infty df\,\frac{\mathrm{Re}\{h^* s\}}{S_n} \equiv \underbrace{\text{Signal-to-Noise Ratio}}_{(\text{SNR})}$

In matched filtering, the SNR is just the max. likelihood estimate for the normalized signal amplitude.

Let's plug this back into the likelihood

$$\log p(h|\hat{A}s) = -2\int_0^\infty df\,\frac{|h|^2 - 2\mathrm{Re}\{h^* s\}\hat{A} + \hat{A}^2|s|^2}{S_n}$$

$$= -2\int_0^\infty df\,\frac{|h|^2}{S_n} - \frac{1}{2}\hat{A}^2\underbrace{\left(4\int_0^\infty df\,\frac{|s|^2}{S_n}\right)}_{1} + \hat{A}\underbrace{\left(4\int_0^\infty df\,\frac{\mathrm{Re}\{h^* s\}}{S_n}\right)}_{\hat{A}}$$

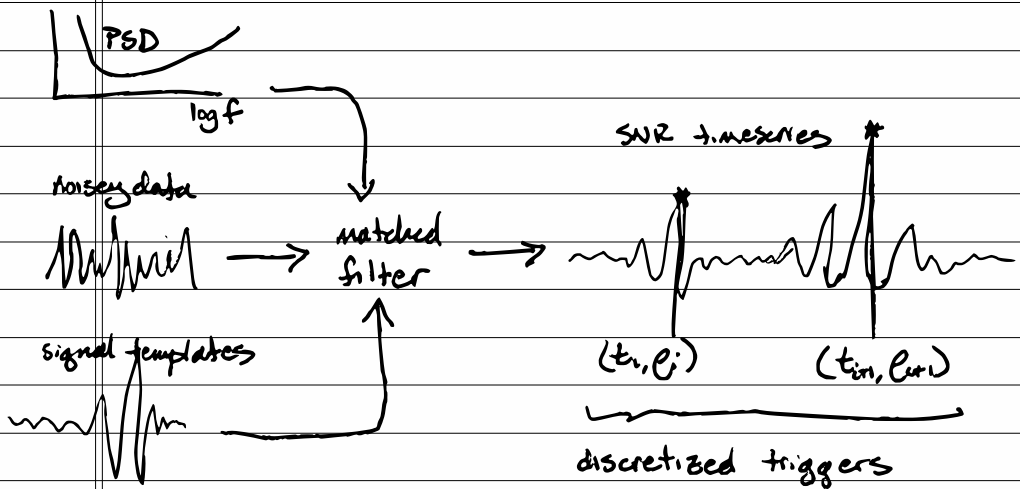
$$\log p(h|\hat{A}s) = -\frac{1}{2}\left(4\int_0^\infty df\,\frac{|h|^2}{S_n}\right) + \frac{1}{2}\hat{A}^2$$

noise-only likelihood → does not depend on signal

$$\log \mathcal{L}_{\text{MLE}} \sim \frac{1}{2}(\text{SNR})^2$$

EXERCISE: derive the max. likelihood estimator for an unmodeled signal ($\hat{s}$) and the corresponding $\mathcal{L}_{\text{MLE}}$

Matched filter searches work by defining a template bank of known signals & then computing the SNR for each template as a function of time. Peaks in the SNR timeseries correspond to higher likelihoods that a signal is present. Selecting local maxima in the SNR timeseries produces a sequence of discretized "triggers," which are the candidate events.



Most (all?) matched filter searches follow this basic approach. Many of the differences come in after triggers have been produced (ie, how one estimates the false alarm prob. or rate from the set of triggers).

Computational Techniques for Bayes Factors

① analytic integration (often not possible)

if you can, compute the marginal likelihood (evidence) analytically. this can often provide insight & be computationally very efficient. Sometimes scaling of $\mathcal{B}$ with SNR can be surprising.

EXERCISE: consider a univariate Gaussian variable

$$x \sim \mathcal{N}(\mu, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\mu)^2\right)$$

w/ noise model: $\mu = 0$
signal model: $\mu \neq 0$

Compute the Bayes factor under signal model where you know the sign & magnitude of μ & when you only know the magnitude of μ (marginalize over the sign).

② Savage-Dickey Density Ratio (SDDR)

For nested models (one hypothesis corresponds to a subset of the alternative hypothesis) so that one model is recovered by a specific set of parameters in the other, we can compute the Bayes Factor from the posterior & the prior.

$$p(\theta | h, H_A) = \frac{p(h|\theta) p(\theta | H_A)}{\int d\theta\, p(h|\theta) p(\theta | H_A)}$$

now, if H_B corresponds to H_A w/ $\theta = \theta_B$, then

$$p(\theta_B | h, H_A) = \frac{p(h|\theta_B) p(\theta_B | H_A)}{\int d\theta\, p(h|\theta) p(\theta | H_A)}$$

$$\Rightarrow \frac{p(h|\theta_B)}{\int d\theta\, p(h|\theta) p(\theta | H_A)} = \frac{p(h|H_B)}{p(h|H_A)} = B_A^B = \frac{p(\theta_B | h, H_A)}{p(\theta_B | H_A)}$$

Savage-Dickey Density Ratio

Because it is often easier to draw samples from $p(\theta | h, H)$ than to estimate $p(h|H)$, this can be very efficient.

(3) Thermodynamic Integration (Power Sampling) and Nested Sampling

More general techniques exist to compute B when we cannot marginalize analytically or we do not have nested models. These tend to be (much) more expensive & opaque. Open source libraries exist in several languages, but BE CAREFUL and make sure you know what you're computing.

Example: Power Sampling

Consider the integral defined for $\beta \in [0,1]$

$$Z_\beta = \int d\theta\, q(\theta) \left(\frac{p(\theta)}{q(\theta)}\right)^\beta$$

now, $\frac{\partial \ln Z_\beta}{\partial \beta} = \frac{1}{Z_\beta} \frac{\partial}{\partial \beta} \int d\theta\, q \left(\frac{p}{q}\right)^\beta$

$$= \int d\theta \left[\frac{q (p/q)^\beta}{Z_\beta}\right] \left(\frac{q}{p}\right)^\beta \frac{\partial}{\partial \beta} \left(\frac{p}{q}\right)^\beta$$

$$= \int d\theta\, \frac{q (p/q)^\beta}{Z} \frac{\partial}{\partial \beta} \beta \log(p/q)$$

$$= \int d\theta\, r(\theta|\beta) \log\left(p(\theta)/q(\theta)\right)$$

where $r(\theta|\beta) \equiv q(\theta) \left(p(\theta)/q(\theta)\right)^\beta / Z_\beta$

We note that $\int_0^1 d\beta \frac{\partial}{\partial \beta} \ln Z_\beta = \ln \int d\theta\, p(\theta) - \ln \int d\theta\, q(\theta)$

and if $p = p(A|\theta) p(\theta|A)$; $q = p(A|\theta) p(\theta|B)$ this becomes

$$\int_0^1 d\beta \frac{\partial}{\partial \beta} \ln Z_\beta = \ln B \Big|_0^1$$

we can estimate $\frac{\partial}{\partial \beta} \ln Z_\beta$ via a Monte Carlo sum

$$\frac{\partial}{\partial \beta} \ln Z_\beta \approx \frac{1}{N} \sum_i^N \log \frac{p(\theta_i)}{q(\theta_i)} \quad | \quad \theta_i \sim r(\theta|\beta)$$

We proceed by defining a "temperature ladder"

$\beta = 1/T \in [0, 1] \Rightarrow T \in [1, \infty)$

and sampling from $q_\beta(\theta)$ separately for each rung on the ladder. If we consider N_T temperatures, we thus obtain N_T sets of samples A can use each to estimate $\frac{\partial}{\partial \beta} \ln Z_\beta$ at that temperature.

With the set of Monte Carlo estimates for $\frac{\partial}{\partial \beta} \ln Z_\beta$, we can then numerically integrate the curve to obtain $\ln B^A_B$.

Note, Thermodynamic Integration is an example of power sampling where

$$p(\theta) = p(d|\theta) p(\theta|H) \quad \text{(signal model)}$$
$$q(\theta) = p(\theta|H) \quad \text{(noise model)}$$

and is often used to estimate B^S_N: Bayes factor in favor of the signal model compared to just noise. B^S_N is a powerful detection statistic.