@ Probability Review

Matrix Operation

$$E[AX+BY+c] = A\,E[X] + B\,E[Y] + c$$
$$\mathrm{cov}(AX+BY+c) = A\,\mathrm{cov}(X)A^T + B\,\mathrm{cov}(Y)B^T + A\,\mathrm{cov}(X,Y)B^T + B\,\mathrm{cov}(Y,X)A^T$$
$$\mathrm{cov}(X,Y) = E\{(X-E[X])(Y-E[Y])^T\}$$
$$\mathrm{cov}(X) = E[\mathrm{cov}(X\mid Y)] + \mathrm{cov}(E[X\mid Y])$$
$$\mathrm{cov}(X,Y) = E[\mathrm{cov}(X,Y\mid Z)] + \mathrm{cov}(E[X\mid Z], E[Y\mid Z])$$

Transformation

Let Y=g(X), then
$$p_Y(y) = p_X(h_1(y_1),\dots,h_n(y_n))\cdot|J|$$
where h(·) is the unique inverse of g(·) and
$$J = \left|\frac{\partial x}{\partial y^T}\right| = \begin{bmatrix} \partial x_1/\partial y_1 & \dots & \partial x_1/\partial y_n \\ \dots & \dots & \dots \\ \partial x_n/\partial y_1 & \dots & \partial x_n/\partial y_n \end{bmatrix}$$

Estimator Performance

$$b(\hat\theta) = E_x[\hat\theta] - \theta$$
$$\begin{aligned}\mathrm{MSE}(\hat\theta) &= E_x[(\hat\theta - E_x[\hat\theta] + E_x[\hat\theta] - \theta)^2] \\ &= \underbrace{E_x[(\hat\theta - E_x[\hat\theta])^2]}_{\mathrm{var}(\hat\theta)} + \underbrace{(E_x[\hat\theta]-\theta)^2}_{b(\theta)^2} \\ &\quad \underbrace{+2E_x[(\hat\theta - E_x[\hat\theta])\cdot \underbrace{(E_x[\hat\theta]-\theta)}_{\text{const}}]}_{0}\end{aligned}$$

MVU estimator: *unbiased* and *min(variance)*

@ Sufficient Statistics

1. Def. $p(x|T(x);\theta) = p(x|T(x))$
2. **[Factorization Thm]**: $p(x;\theta) = g(T(x),\theta)\cdot h(x)$

- T(x) is *minimally sufficient* if it is *sufficient* and provides a greater reduction of data than any other sufficient statistic S(x).

@ Cramer Rao Bound

Single measurement to *Fisher Information* $(0 \leqslant I(\theta) \leqslant \infty)$
$$\mathcal{I}(\theta) = E_X\left\{\left(\frac{\partial}{\partial\theta}\log p(X;\theta)\right)^2\right\} = -E_{p(x;\theta)}\left\{\frac{\partial^2}{\partial\theta^2}\log p(X;\theta)\right\}$$
For $X' = [X_1,\dots,X_n]^T$: $I'(\theta) = n\,I(\theta)$

Δ: I(θ) does not depend on other parameter

Mean and variance of *Score Function*
$$E_X\left(\frac{\partial}{\partial\theta}\log p(X;\theta)\right) = 0 \text{ and, thus, } \mathcal{I}(\theta) = \mathrm{var}_X\left(\frac{\partial}{\partial\theta}\log p(X;\theta)\right)$$

■ Using matrix notation:
$$E_{p(x;\theta)}\left[\frac{\partial}{\partial\boldsymbol\theta}\log p(X;\boldsymbol\theta)\right] = 0$$
$$\mathcal{I}(\boldsymbol\theta) = \mathrm{cov}_{p(x;\theta)}\left[\frac{\partial}{\partial\boldsymbol\theta}\log p(X;\boldsymbol\theta)\right]$$
$$\mathrm{var}(b^T X) = b^T\,\mathrm{var}(X)\,b$$
$$\frac{\partial x^T S x}{\partial x} = 2Sx \text{ if } S \text{ is symmetric } \frac{\partial a^T x}{\partial x} = a.\ \frac{\partial x^T a}{\partial x} = a$$

Cauchy Schwartz Inequality

$$[\mathrm{cov}(P,Q)]^2 \le \mathrm{var}(P)\cdot\mathrm{var}(Q)$$

Information Inequality
$$\mathrm{var}_{p(x;\theta)}[T(X)] \ge \frac{|\psi'(\theta)|^2}{\mathcal{I}(\theta)} \text{ where } \psi'(\theta) = \frac{d\psi(\theta)}{d\theta}$$
$$\psi(\theta) = E_{p(x;\theta)}[T(X)]$$

■ Using Matrix Notation:
For $T(X) = [T_1(X),\dots,T_n(X)]^T$:
$$\psi(\boldsymbol\theta) = E_{p(x;\theta)}[T(X)] = [\psi_1(\boldsymbol\theta),\dots,\psi_d(\boldsymbol\theta)]^T$$
$$\mathrm{cov}_{p(x;\theta)}[T(X)] \ge \frac{\partial\psi(\boldsymbol\theta)}{\partial\boldsymbol\theta^T}\mathcal{I}(\boldsymbol\theta)^{-1}\frac{\partial\psi(\boldsymbol\theta)^T}{\partial\boldsymbol\theta}.$$

■ Matrix Trick

$a(X) = [a_1(X),\dots,a_m(X)]^T$; $\theta = [\theta_1,\dots,\theta_d]^T$
$$\frac{\partial a(\theta)^T}{\partial\theta} = \begin{bmatrix} \partial a_1(X)/\partial\theta_1 & \dots & \partial a_1(X)/\partial\theta_d \\ \dots & \dots & \dots \\ \partial a_m(X)/\partial\theta_1 & \dots & \partial a_1(X)/\partial\theta_d \end{bmatrix}$$
$$\partial a(\theta)^T/\partial\theta = [\partial a(\theta)/\partial\theta^T]^T$$

■ Matrix Trick
$$|\mathrm{cov}(P,Q) = \begin{vmatrix} \mathrm{var}(P) & \mathrm{cov}(P,Q) \\ \mathrm{cov}(P,Q) & \mathrm{var}(Q) \end{vmatrix}$$
$$\mathrm{var}[aP+bQ] = [a,b]\,\mathrm{cov}(P,Q)\,[a,b]^T$$

If we view T(X) as a (generally *biased*) estimator of θ, then,
$$E_{p(x;\theta)}[T(X)] = \psi(\theta) = \theta + \mathrm{bias}(\theta)$$
$$\mathrm{var}_{p(x;\theta)}[T(X)] = |1 + bias(\theta)|^2 /\ I(\theta)$$

By **Minimum Square Error**,
$$\mathrm{MSE}[T(X)] = \mathrm{var}_{p(x;\theta)}[T(X)] + bias(\theta)^2 \ge |1+bias(\theta)|^2 /\ I(\theta) + bias(\theta)^2$$

Δ Remark: 1/I(θ) is often referred to as the *CRB* on *variance* of an *unbiased* estimator of θ

Efficient Estimator

- An *unbiased* estimator of θ that attains *CRB* for θ for all θ in parameter space Θ is *efficient*.
- Under certain regularity conditions, *MLE* attain *CRB* asymptotically (i.e. for large n); hence they are *asymptotically efficient*.
- *Efficiency* ⇒ *MVU* estimator

Exponential Family
$$p(x;\theta) = h(x)\exp\{\eta(\theta)T(x) - A(\theta)\}$$

Δ: T(X) achieves *CRB* and thus is *MVU* estimator of $E_X[T(X)]$. Hence, T(X) is an efficient estimator of $E_X[T(X)] = \psi(\theta)$.

Gaussian CRB

Suppose that X has a n-variate Gaussian distribution, X ~ N(μ(θ),C(θ)) that is,
$$p(x;\boldsymbol\theta) = \frac{1}{\sqrt{(2\pi)^n|C|}}\exp\left[-\tfrac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\right]$$
Then, the (i, k)th element of FIM is given by
$$\mathcal{I}_{i,k} = \frac{\partial\mu^T}{\partial\theta_i}C^{-1}\frac{\partial\mu}{\partial\theta_k} + \frac{1}{2}\cdot\mathrm{tr}\left(C^{-1}\frac{\partial C}{\partial\theta_i}C^{-1}\frac{\partial C}{\partial\theta_k}\right)$$

MGF Trick of Gaussian: X~N(0, σ²)

$E(X^{2k}) = \sigma^{2k}(2k-1)!!$; $E(X^{2k+1}) = 0$.

Trace Operator: A is N x N square matrix

- Def. $\mathrm{tr}(A) = \sum_{i=1}^n a_{ii}$
- Frobenius norm: $||A||^2 = \mathrm{tr}(A^TA)$

1. $\mathrm{tr}(cA + dB) = c\,\mathrm{tr}(A) + d\,\mathrm{tr}(A)$
2. $\mathrm{tr}(A) = \mathrm{tr}(A^T)$;
3. $\mathrm{tr}(B^TA) = \mathrm{tr}(A^TB) = \sum_{i,j} A_{ij}B_{ij}$
4. $\mathrm{tr}(I_n) = n$
5. $\mathrm{tr}(P^{-1}AP) = \mathrm{tr}(APP^{-1}) = \mathrm{tr}(A)$
6. $E[\mathrm{tr}(X)] = \mathrm{tr}[E(X)]$

One parameter *Canonical Exponential Family* for *sufficient* statistics (*unbiased* & *efficient*)
$$p(x;\eta) = h(x)\exp[\eta T(x) - A(\eta)]$$
where
$$A(\eta) = \log\int\cdots\int h(\chi)\exp[\eta T(\chi)]\,d\chi \quad \text{for a pdf } p(x;\eta)$$
$$A(\eta) = \log\sum_{\chi} h(\chi)\exp[\eta T(\chi)] \quad \text{for a pmf } p(x;\eta).$$

If we can compute normalizing term A(η) in a simple form, then it is easy to find *mean* and *variance* of T(X):
$$E_{p(x;\eta)}[T(X)] = \frac{dA(\eta)}{d\eta}, \quad \mathrm{var}_{p(x;\eta)}[T(X)] = \frac{d^2A(\eta)}{d\eta^2}.$$
T(X) is an *efficient* estimator of $E_X[T(X)]$ = dA(η)/ dη; we can easily compute *variance* of this estimator as well:
$$\mathcal{I}(\eta) = \mathrm{var}_{p(x;\eta)}\underbrace{\left(T(X) - \frac{dA(\eta)}{d\eta}\right)}_{\text{score function}} = \mathrm{var}_{p(x;\eta)}(T(X)) = \frac{d^2A(\eta)}{d\eta^2}$$
and this *variance* is equal to *CRB* of dA(η)/ dη:
$$\mathrm{CRB}\left(\frac{dA(\eta)}{d\eta}\right) = \left[\mathcal{I}\left(\frac{dA(\eta)}{d\eta}\right)\right]^{-1} = \mathcal{I}(\eta)$$
$E_X[T(X)] = \theta$. Suppose now that we pick θ = θ(η) = dA(η)/dη ; then, clearly, T(X) is an *efficient* estimator of $E_X[T(X)]$=dA(η)/dη=θ:
$$\underbrace{\frac{dA(\eta)}{d\eta}}_{\theta} = \underbrace{\frac{dB(\theta)}{d\theta}\bigg|_{\theta=\frac{dA(\eta)}{d\eta}}\cdot\underbrace{\frac{d\theta(\eta)}{d\eta}}_{\mathrm{CRB}(\theta)=[\mathcal{I}(\theta)]^{-1}}}_{\text{chain rule}} \implies \frac{dB(\theta)}{d\theta} = \theta\cdot\mathcal{I}(\theta)$$
With I(θ)= dη/dθ = $[d\theta/d\eta]^{-1}$, we have another convenient form to examine *efficiency*
$$\frac{\partial\log p(x;\theta)}{\partial\theta} = \underbrace{\mathcal{I}(\theta)}_{\text{Fisher information for }\theta}\ [T(x) - \theta]$$

@ Linear Model

General model : x = Hθ + w; where x is N ×1 vector and H is a known *deterministic* N × p matrix, with N ⩾ p. (*overestimated*)

We wish to estimate *unknown* parameter vector θ. Assume w ~ N (0, σ²I).

MVU estimator of θ regardless of whether σ² is known or not.
$$\hat{\boldsymbol\theta} = (H^TH)^{-1}H^Tx.$$
$cov(\hat\theta)$ attains *CRB* for all $\theta \in \mathbb{R}^p$:
$$C_{\hat\theta} = E[(\hat\theta-\theta)(\hat\theta-\theta)^T] = \sigma^2(H^TH)^{-1}.$$
$\hat\theta$ coincides with *Least-Squares* estimate of θ
$$\hat\theta = \arg\min_{\theta}||x - H\theta||^2$$

MVUE For *Color Noise* **with** w ~ N (0, C)
$$\hat\theta = (H^TC^{-1}H)^{-1}H^TC^{-1}x$$
$cov(\hat\theta)$ attains *CRB*:
$$C_{\hat\theta} = (H^TC^{-1}H)^{-1}.$$
$\hat\theta$ is a weighted *LS* estimate:
$$\hat\theta = \arg\min_{\theta}||x - H\theta||^2_W = \arg\min_{\theta}(x - H\theta)^T W(x - H\theta)$$
The "optimal weight matrix" W = C⁻¹, prewhitens the residuals.

Best Linear Unbiased Estimator

Given **x = Hθ + w**, w ~ N (0, C),
Assume estimator to be linear (i.e. of form $\hat\theta = A^Tx$) and unbiased; and minimize its variance.
The *BLUE* of θ is
$$\hat\theta = (H^TC^{-1}H)^{-1}H^TC^{-1}x$$
and its covariance matrix is
$$C_{\hat\theta} = (H^TC^{-1}H)^{-1}$$

General MVU Estimation

1.**[THm Rao-Blackwell]** if $\hat\theta_1(x)$ is any unbiased estimator and $T(x)$ is a sufficient statistic: Define $\hat\theta_2(x) = E[\hat\theta_1(x)|T]$, a function of T. Then $\hat\theta_2(x)$ is also unbiased for θ and $Var(\hat\theta_1) \le Var(\hat\theta_2)$

- Def. T(x) is *complete sufficient statistic* if only one estimator $\hat\theta = g(T(x))$ is unbiased.

2. *Corollary*. If $T(x)$ is *complete sufficient statistic*, then the unique unbiased estimate $\hat\theta = g(T(x))$ is MVUE.

ML ESTIMATION

$$\hat\theta = \arg\max_{\theta} p(x;\theta).$$
Assume that certain regularity conditions hold and let $\hat\theta$ be **ML estimate**. Then, as N → ∞,
$$\hat\theta \to \theta_0 \quad \text{(with probability 1)} \quad \text{(consistency)}$$
$$\sqrt{N}(\hat\theta - \theta_0) \xrightarrow{d} \mathcal{N}(0, N\mathcal{I}^{-1}(\theta_0)) \quad \text{(asymptotic efficiency)}$$
ΔRemark: if an *efficient* (in finite samples) estimate exists, it is given by the **ML estimate**.

[**ML Invariance Principle**]

· If $\hat\theta_{MLE}$ exists, $\hat a_{MLE}$ is **MLE** for $a = g(\theta)$

Delta method

· If $\hat\theta$ is consistent, so is $\hat a$; asymptotic **MSE** matrices $C_{\hat\theta}$ and $C_{\hat a}$ are related by
$$C_{\hat a} = \frac{\partial g}{\partial\theta^T}C_{\hat\theta}\frac{\partial g^T}{\partial\theta}.$$
· If $T(x)$ is *minimal* & *complete*, **MLE** is unique.
· If the *ML estimate* is unbiased, then it is **MVU**.

@ Bayesian Inference

Bayes rule
$$p(\theta|x) = \frac{p(x,\theta)}{p(x)} \propto p(x,\theta) = p(x|\theta)\pi(\theta)$$

Posterior Distribution = Likelihood x Prior

Type of Prior

1. Conjugate: 先验后验同分布

Important Case for EXP Family:
If F ≡ exponential family ⇒ distributions in F have natural conjugate priors.
$$p(x_i|\theta) = h(x_i)\,g(\theta)\,\exp[\phi(\theta)^T t(x_i)], \quad i = 1,2,\dots,N$$
Δ **beta** is conjugate prior for Normal & Bin

2. Proper prior: Integrates to 1
$$\pi(\theta) \ge 0 \quad \forall\theta, \qquad \int\pi(\theta)\,d\theta = 1.$$
If a prior is proper, so is the posterior

3. Improper prior : Integrates to ∞

A prior π(θ) is called improper if
$$\pi(\theta) \ge 0 \quad \forall\theta, \qquad \int\pi(\theta)\,d\theta = \infty$$
If a prior is improper, the posterior may not be proper

4. Jeffrey's Prior (*Improper Noninformative*)

⋆Reparametrization: Chain rule⋆
$$\pi_\phi(\phi) = \{\pi_\theta(\theta)\cdot|d\theta/d\phi|\}\big|_{\theta=h^{-1}(\phi)} = \{\pi_\theta(\theta)\cdot|h'(\theta)|^{-1}\}\big|_{\theta=h^{-1}(\phi)}$$
$$\mathcal{I}_\eta(\eta) = \mathcal{I}_\theta(\theta(\eta))\left(\frac{d\theta}{d\eta}\right)^2$$

5. Informative Prior: based off of actual information about the parameters

6. Noninformative Prior: chosen as uniform distribution or as a conjugate prior

⋈ Sequential Bayesian Inference

Suppose that we have observed x_1 and x_2 where x_1 comes first (set it fixed):
$$p(x_2,\theta|x_1) = p(x_2|\theta,x_1)\cdot p(\theta|x_1).$$
We wish to do inference about θ:
$$p(\theta|x_1,x_2) \propto p(x_2|\theta,x_1)\cdot p(\theta|x_1).$$

∝ Operator

p(θ | x₁, x₂) ∝ p(θ, x₁, x₂) ∝ p(θ, x₁| x₂) ∝ p(θ, x₂ | x₁).

⋈ Posterior predictive distribution

Suppose that we have observed x_1 and wish to predict x_2. We use $p(x_2 | x_1)$
A good way to think of *predictive distribution* is as follows. Given x_1:
$$p(x_2,\theta|x_1) = p(x_2|\theta,x_1)\cdot p(\theta|x_1).$$
Then, marginalize this pdf with respect to the unknown parameter θ (i.e. integrate θ out)
$$p_{x_2|x_1}(x_2|x_1) = \int p(x_2,\theta|x_1)\,d\theta = \int p(x_2|\theta,x_1)\cdot p(\theta|x_1)\,d\theta$$

Bayesian Sufficient Statistics

- If a statistic T(X) is sufficient for a parameter θ, then p(θ|T(x)) = p(θ|x).

Bayesian MMSE Estimation

Define *estimation error* ε = θ − $\hat\theta$ and assign a *loss* (cost) L(ε). We may choose $\hat\theta$ to minimize *Bayes (preposterior) risk*:
$$E_{x,\theta}[L(\epsilon)] = E_{x,\theta}[L(\theta - \hat\theta)]$$
but this is equivalent to minimizing the *posterior expected loss*:
$$\rho(\hat\theta|x) = \int L(\theta - \hat\theta)\,p(\theta|x)\,d\theta$$
But this is what *true Bayesians* do.
$$E_{x,\theta}[L(\theta-\hat\theta)] = E_x\{\underbrace{E_{\theta|x}[L(\theta-\hat\theta)]}_{\rho(\hat\theta|x)}\}.$$

Loss Function

1. L(ε) = ε² (*squared-error loss*, accurate);
2. L(ε) = |ε| (robust to outliers),
3. $L(\epsilon) = \begin{cases} 0, & |\epsilon| \le \Delta/2 \\ 1, & |\epsilon| > \Delta/2 \end{cases}$ (0-1 loss, tractable)

Mean-square error measures [popular]

Classical
$$\mathrm{MSE}(\hat\theta) = \int(\hat\theta-\theta)^2 p(x;\theta)\,dx$$
posterior expected squared loss: truly *Bayesian*
$$\rho(\hat\theta|x) = \underbrace{\int\left(\hat\theta - E_{\theta|x}[\theta|x]\right)^2 p(\theta|x)\,d\theta}_{(\hat\theta - E_{\theta|x}[\theta|x])^2} + \int\left(\theta - E_{\theta|x}[\theta|x]\right)^2 p(\theta|x)\,d\theta$$
⇒ **the optimal** $\hat\theta$ satisfies:
$$E_{\theta|x}[\theta|x] = \arg\min_{\hat\theta}\rho(\hat\theta|x)$$
1'.*"Bayesian" MSE* (*Preposterior*, not truly *Bayesian*)
$$\mathrm{BMSE}(\hat\theta) = \int\int(\hat\theta-\theta)^2 p(x|\theta)\pi(\theta)\,dx\,d\theta = E_\theta[\mathrm{MSE}(\hat\theta)]$$
$$\mathrm{BMSE}(\hat\theta) = E_{x,\theta}[(\theta-\hat\theta)^2] = E_x\Big\{\underbrace{E_{\theta|x}[(\theta-\hat\theta)^2|x]}_{\rho(\hat\theta|x)}\Big\}$$
ΔRemark: a.k.a. *posterior mean* BMSE = MMSE
⇒ **the optimal** $\hat\theta$ satisfies:
$$\hat\theta = E_{\theta|x}[\theta|x]$$
2'.With absolute value as loss function, we get *Posterior median* ⇒ **the optimal** $\hat\theta$ satisfies:
$$\int_{-\infty}^{\hat\theta} p(\theta|x)\,d\theta = \int_{\hat\theta}^{\infty} p(\theta|x)\,d\theta = 0.5$$
3'.With 0-1 loss, we get ***MAP*** *(maximum a posteriori) estimator* a.k.a. *Posterior mode* ⇒ **the optimal** $\hat\theta$ satisfies:
$$\arg\max_{\hat\theta} p_{\theta|x}(\hat\theta|x)$$

MAP Estimation

$$\hat\theta_{MAP} = \arg\max_{\boldsymbol\theta}[\log p(x|\theta) + \log\pi(\theta)]$$

Linear MMSE Estimation

[THm.2] Jointly Gaussian

Assume that
$$\begin{bmatrix} x \\ \theta \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_\theta \end{bmatrix}, \begin{bmatrix} C_{xx} & C_{x\theta} \\ C_{\theta x} & C_{\theta\theta} \end{bmatrix}\right)$$
Then, posterior pdf p(θ | x) is also Gaussian, with the first two moments given by
$$E[\theta|X = x] = \mu_\theta + C_{\theta x}C_{xx}^{-1}(x - \mu_x)$$
$$C_{\theta|x} = C_{\theta\theta} - C_{\theta x}C_{xx}^{-1}C_{x\theta}.$$

[THm.3] Bayesian General Model

Assume that: **x = Hθ + w**; where H is known and w ~ N (0, C_w), *independent* of θ, C_w known.
If *prior* pdf for θ is π(θ) = N(μ_θ,C_θ) [μ_θ and C_θ known], *posterior* pdf p(θ | x) is also *Gaussian*:
$$\begin{aligned}E[\theta|X=x] &= \mu_\theta + C_\theta H^T(HC_\theta H^T + C_w)^{-1}(x - H\mu_\theta) \\ &= (H^TC_w^{-1}H + C_\theta^{-1})^{-1}(H^TC_w^{-1}x + C_\theta^{-1}\mu_\theta)\end{aligned}$$
$$\begin{aligned}C_{\theta|x} &= C_\theta - C_\theta H^T(HC_\theta H^T + C_w)^{-1}HC_\theta. \\ &= (H^TC_w^{-1}H + C_\theta^{-1})^{-1}\end{aligned}$$

■ **Matrix Inversion**
$$(A+B)^{-1} = A^{-1} - A^{-1}(B^{-1} + A^{-1})^{-1}A^{-1}$$
$$(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

@ Detection Theory

$P_{\mathrm{FA}} = \mathrm{E}_{\boldsymbol{x}\,|\,\theta}[\phi(\boldsymbol{X})\,|\,\theta] = \int_{\mathcal{X}_1} p(\boldsymbol{x}\,|\,\theta)\,d\boldsymbol{x}$ for θ in Θ_0

$P_{\mathrm{M}} = \mathrm{E}_{\boldsymbol{x}\,|\,\theta}[1-\phi(\boldsymbol{X})\,|\,\theta] = 1 - \int_{\mathcal{X}_1} p(\boldsymbol{x}\,|\,\theta)\,d\boldsymbol{x}$

$= \int_{\mathcal{X}_0} p(\boldsymbol{x}\,|\,\theta)\,d\boldsymbol{x}$ for θ in Θ_1

critical region: χ_1 = {decide H_1};

Best Critical Region: critical region that attains the maximum *power*

Power of Test:

$P_{\mathrm{D}} = 1 - P_{\mathrm{M}} = \mathrm{E}_{\boldsymbol{x}\,|\,\theta}[\phi(\boldsymbol{X})\,|\,\theta] = \int_{\mathcal{X}_1} p(\boldsymbol{x}\,|\,\theta)\,d\boldsymbol{x}$ for θ in Θ_1

operating point on *ROC*: (P_{FA}, P_D)

$$P_{\mathrm{D}} = Q(Q^{-1}(P_{\mathrm{FA}}) - d).$$

Given P_{FA}, P_D *depends only on the deflection coefficient*: (Signal-to-Noise Ratio)

$$d^2 = \frac{NA^2}{\sigma^2} = \frac{\{\mathrm{E}\,[T(\boldsymbol{X})\,|\,a=A] - \mathrm{E}\,[T(\boldsymbol{X})\,|\,a=0]\}^2}{\mathrm{var}[T(\boldsymbol{X}\,|\,a=0)]}$$

Δ Remark: Performance improves with d^2. **Bayesian** and **NP tests** both hit specific points on the curve.

size(α):

$$\alpha = \max_{\theta\in\Theta_0} P[\boldsymbol{x}\in\mathcal{X}_1\,|\,\theta] = \text{max possible } P_{\mathrm{FA}}$$

p-value: used to reject or fail to reject H_0 - but not declare H_1.

$$p\ value = \inf\{\alpha\,:\,\boldsymbol{x}\in\mathcal{X}_{1,\alpha}\}.$$

• **NP-Thm** tells us how to choose X_1 if we are given $p(x\,;\theta_0)$, $p(x\,;\theta_1)$: To maximize P_D for a given P_{FA} = α. **Decide H_1 if** *Likelihood ratio*:

$$L(X) = p(x\,;\theta_1)/p(x\,;\theta_0) > \lambda$$

Determine threshold that achieve a specified P_{FA}

1. based on the pdf of x

$$\int_{\boldsymbol{x}\,:\,\Lambda(\boldsymbol{x})>\lambda} p(\boldsymbol{x}\,;\theta_0)\,d\boldsymbol{x} = P_{\mathrm{FA}} = \alpha$$

2. based on the pdf of *sufficient statistics* Λ(x)

$$\int_{\lambda}^{\infty} p_{\Lambda;\theta_0}(l\,;\theta_0)\,dl = \alpha$$

ΔRemark:

· Usually we simplify *likelihood ratio* into *a test statistic*, since we know test statistic's distribution it is easy to calculate P_D/P_{FA}.

· *Exponential Family:* (可加性)

Discrete: Ber/Bin/Poisson

Continuous: Normal/Gamma/Exp/Chi-square

• Bayesian Detection

posterior expected loss

$$\rho(\text{action}\,|\,\boldsymbol{x}) = \int_{\Theta} \mathrm{L}(\theta,\text{action})\,p(\theta\,|\,\boldsymbol{x})\,d\theta$$

Loss function: described by the quantities

L(declared | true): L(1|1) = L(0|0) = 0

L(1 | 0) quantifies loss due to a false alarm;

L(0 | 1) quantifies loss due to a miss;

$$\rho_0(\boldsymbol{x}) = \int_{\Theta_1} \mathrm{L}(0\,|\,1)\,p(\theta\,|\,\boldsymbol{x})\,d\theta + \int_{\Theta_0} \underbrace{\mathrm{L}(0\,|\,0)}_{0}\,p(\theta\,|\,\boldsymbol{x})\,d\theta$$

$$\rho_1(\boldsymbol{x}) = \int_{\Theta_0} \mathrm{L}(1\,|\,0)\,p(\theta\,|\,\boldsymbol{x})\,d\theta$$

Bayes' decision rule is to **Minimizes *Posterior Expected Loss*** : this rule corresponds to choosing data-space partitioning as follows:

$$\mathcal{X}_1 = \{\boldsymbol{x}\,:\,\rho_1(\boldsymbol{x}) \le \rho_0(\boldsymbol{x})\}$$

or, equivalently, upon applying Bayes' rule:

$$\mathcal{X}_1 = \left\{\boldsymbol{x}\,:\,\frac{\int_{\Theta_1} p(\boldsymbol{x}\,|\,\theta)\,\pi(\theta)\,d\theta}{\int_{\Theta_0} p(\boldsymbol{x}\,|\,\theta)\,\pi(\theta)\,d\theta} \ge \frac{\mathrm{L}(1\,|\,0)}{\mathrm{L}(0\,|\,1)}\right\}$$

/Special Loss/ *0-1 loss*: For L(1|0) = L(0|1) = 1

Preposterior (Bayes) Risk for *rule* φ(x) is to minimize it.

$$\mathrm{E}_{\,\boldsymbol{x},\theta}[\text{loss}] = \int_{\mathcal{X}_1}\int_{\Theta_0} \mathrm{L}(1\,|\,0)\,p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta\,d\boldsymbol{x}$$

$$+ \int_{\mathcal{X}_0}\int_{\Theta_1} \mathrm{L}(0\,|\,1)\,p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta\,d\boldsymbol{x}$$

$$= \underbrace{\text{const}}_{\text{not dependent on }\phi(\boldsymbol{x})} + \int_{\mathcal{X}_1}\Big\{\mathrm{L}(1\,|\,0)\cdot\int_{\Theta_0} p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta$$

$$-\mathrm{L}(0\,|\,1)\cdot\int_{\Theta_1} p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta\Big\}\,d\boldsymbol{x}$$

implying that X_1 should be chosen as

$$\Big\{\mathcal{X}_1 : \mathrm{L}(1\,|\,0)\cdot\int_{\Theta_0} p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta - \mathrm{L}(0\,|\,1)\cdot\int_{\Theta_1} p(\boldsymbol{x}\,|\,\theta)\pi(\theta) < 0$$

ΔRemark: Minimizes *posterior expected loss* ⇔ Minimizes *preposterior risk* for every x because of *joint density* for x and θ.

/Special Loss/ *0-1 loss*: i.e. L(1|0) = L(0|1) = 1 *preposterior (Bayes) risk* for *rule* φ(x) is

$$\mathrm{E}_{\,\boldsymbol{x},\theta}[\text{loss}] = \int_{\mathcal{X}_1}\int_{\Theta_0} p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta\,d\boldsymbol{x}$$

$$+ \int_{\mathcal{X}_0}\int_{\Theta_1} p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta\,d\boldsymbol{x}$$

which is *average error probability*, with averaging performed over *joint probability density* or *mass function* or data x and parameters θ.

Bayes' decision rule for simple hypotheses

$$\underbrace{\Lambda(\boldsymbol{x})}_{\text{likelihood ratio}} = \frac{p(\boldsymbol{x}\,|\,\theta_1)}{p(\boldsymbol{x}\,|\,\theta_0)} \overset{\mathcal{H}_1}{\gtrless} \frac{\pi_0\,\mathrm{L}(1\,|\,0)}{\pi_1\,\mathrm{L}(0\,|\,1)} \equiv \tau$$

Λ(x): sufficient statistic for detection problem

Minimum *a.v error probability* Detection

/Special Loss/ *0-1 loss case*: L(1|0) = L(0|1) = 1

$$\text{av. error probability} = \pi_0\cdot\underbrace{\int_{\mathcal{X}_1} p(\boldsymbol{x}\,|\,\theta_0)\,d\boldsymbol{x}}_{P_{\mathrm{FA}}} + \pi_1\cdot\underbrace{\int_{\mathcal{X}_0} p(\boldsymbol{x}\,|\,\theta_1)\,d\boldsymbol{x}}_{P_{\mathrm{M}}}$$

Maximum likelihood test occurs for $\pi(H_0)$ = $\pi(H_1)$ = 0.5 and L(0|1) = L(1|0), where λ = 1 and the decision is based on which hypothesis has a larger likelihood value.

Handling Nuisance Parameters φ

Integrate φ out:

$$p_{\theta\,|\,\boldsymbol{x}}(\theta\,|\,\boldsymbol{x}) = \int p_{\theta,\varphi\,|\,\boldsymbol{x}}(\theta,\varphi\,|\,\boldsymbol{x})\,d\varphi$$

Updated Decision rule:

$$\frac{\int_{\Theta_1}\int p_{\boldsymbol{x}\,|\,\theta,\varphi}(\boldsymbol{x}\,|\,\theta,\varphi)\,\pi_{\theta,\varphi}(\theta,\varphi)\,d\varphi\,d\theta}{\int_{\Theta_0}\int p_{\boldsymbol{x}\,|\,\theta,\varphi}(\boldsymbol{x}\,|\,\theta,\varphi)\,\pi_{\theta,\varphi}(\theta,\varphi)\,d\varphi\,d\theta} \overset{\mathcal{H}_1}{\gtrless} \frac{\mathrm{L}(1\,|\,0)}{\mathrm{L}(0\,|\,1)}.$$

Simple hypo & independent priors for θ & φ:

$$\underbrace{\frac{\int p_{\boldsymbol{x}\,|\,\theta,\varphi}(\boldsymbol{x}\,|\,\theta_1,\varphi)\,\pi_{\varphi}(\varphi)\,d\varphi}{\int p_{\boldsymbol{x}\,|\,\theta,\varphi}(\boldsymbol{x}\,|\,\theta_0,\varphi)\,\pi_{\varphi}(\varphi)\,d\varphi}}_{\text{integrated likelihood ratio}} = \frac{p(\boldsymbol{x}\,|\,\theta_1)}{p(\boldsymbol{x}\,|\,\theta_0)} \overset{\mathcal{H}_1}{\gtrless} \frac{\pi_0\,\mathrm{L}(1\,|\,0)}{\pi_1\,\mathrm{L}(0\,|\,1)}$$

where $\pi_0 = \pi_\theta(\theta_0)$, $\pi_1 = \pi_\theta(\theta_1) = 1 - \pi_0$.

Testing Multiple Hypotheses

$\Theta_0, \Theta_1, \ldots, \Theta_{M-1}$ that form a partition of parameter space Θ

H_0 : θ∈Θ_0 vs. H_1 : θ∈Θ_1 vs. … H_{M-1} : θ∈Θ_{M-1}

$$\rho_m(\boldsymbol{x}) = \sum_{i=0}^{M-1} \mathrm{L}(m\,|\,i)\int_{\Theta_i} p(\theta\,|\,\boldsymbol{x})\,d\theta,\quad m = 0,1,\ldots,M-1.$$

Then, *Bayes' decision rule* φ⋆ is defined via the following data-space partitioning:

$$\mathcal{X}_m^{\star} = \{\boldsymbol{x}\,:\,\rho_m(\boldsymbol{x}) = \min_{0\le l\le M-1}\rho_l(\boldsymbol{x})\},\quad m = 0,1,\ldots,M-1$$

The *preposterior (Bayes) risk* for rule φ(x) is

$$\mathrm{E}_{\,\boldsymbol{x},\theta}[\text{loss}] = \sum_{m=0}^{M-1}\int_{\mathcal{X}_m}\underbrace{\sum_{i=0}^{M-1}\mathrm{L}(m\,|\,i)\int_{\Theta_i} p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta}_{h_m(\theta)}\,d\boldsymbol{x}$$

Then, for an arbitrary $h_m(x)$,

$$\Big[\sum_{m=0}^{M-1}\int_{\mathcal{X}_m} h_m(\boldsymbol{x})\,d\boldsymbol{x}\Big] - \Big[\sum_{m=0}^{M-1}\int_{\mathcal{X}_m^{\star}} h_m(\boldsymbol{x})\,d\boldsymbol{x}\Big] \ge 0$$

which verifies that *Bayes' decision rule* φ⋆ minimizes *preposterior (Bayes) risk.*

@ Composite Hypothesis Testing

Ex1. DC Level in WGN with Unknown A

$\mathcal{H}_0:\quad x[n] = w[n],\quad n = 1,2,\ldots,N$

$\mathcal{H}_1:\quad x[n] = A + w[n],\quad n = 1,2,\ldots,N$

where $w[n]$ is zero-mean white Gaussian noise with known variance σ^2. Here is an alternative formulation: Consider this family of probability density functions (pdfs):

$$p(\boldsymbol{x}\,;a,\sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^N}}\cdot\exp\Big[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x[n]-a)^2\Big]$$

and the following (equivalent) hypotheses:

$\mathcal{H}_0:\quad \theta = 0$ (signal absent), $\Theta_0 = \{0\}$ versus

$\mathcal{H}_1:\quad \theta = A > 0$ (signal present), $\Theta_1 = (0,\infty)$

where A is unknown, except for its sign. Let us try classical **NP** approach: **decide H_1 if**

$$\Lambda(\boldsymbol{x}) = \frac{1/(2\pi\sigma^2)^{N/2}\cdot\exp[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x[n]-A)^2]}{1/(2\pi\sigma^2)^{N/2}\cdot\exp(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}x[n]^2)} > \lambda.$$

Taking log etc. leads to

$$A\sum_{n=1}^{N}x[n] > \sigma^2\log\lambda + N\,A^2/2.$$

Since A > 0,

$$T(\boldsymbol{x}) = \overline{x} = \frac{1}{N}\sum_{n=1}^{N}x[n] > \lambda'.$$

How to determine the threshold λ'?

Under H_0 : T(X) | θ = 0 ~ N(0,σ²/N) and hence

$P_{\mathrm{FA}} = Q\Big(\frac{\lambda'}{\sqrt{\sigma^2/N}}\Big)$ where $\lambda' = \sqrt{\frac{\sigma^2}{N}}\cdot Q^{-1}(P_{\mathrm{FA}})$.

which does not depend on A.

However, under H_1 : T(X) | θ = A ~ N(A,σ²/N)

$P_{\mathrm{D}} = Q\big(Q^{-1}(P_{\mathrm{FA}}) - \sqrt{d^2}\big)$ where $d^2 = NA^2/\sigma^2$

which does depend on A.

Uniformly Most Powerful Test

- **NP test** is *optimal* in terms of maximizing P_D s.t a specified P_{FA}, all other tests are poorer w.r.t this criterion.

* For a **UMP** test to exist, the parameter test must be **one-sided**.

* *Doea UMP test exist for θ_0 vs. $\theta > \theta_0$?*

The NP optimal test for level *α is the same for testing θ_0 vs. θ for each $\theta > \theta_0$*, so it follows that this NP-optimal test is indeed a UMP test. (*meaning α is fixed and independent of* θ_1.)

Ex2. DC Level in WGN with Unknown A (Continued)

$\mathcal{H}_0:\quad a = 0$ versus

$\mathcal{H}_1:\quad a = \underbrace{A}_{\text{unknown}} \neq 0.$

GLRT decides H_1 if

$$\Lambda_{\mathrm{GLR}}(\boldsymbol{x}) = \frac{\max_a p(\boldsymbol{x}\,;a)}{p(\boldsymbol{x}\,;a=0)} > \gamma$$

By **MLE**, $\widehat{A}_{\mathrm{MLE}} = \overline{X}$

$$\log\Lambda_{\mathrm{GLR}}(\boldsymbol{x}) = -\frac{1}{2\sigma^2}\Big\{\sum_{n=1}^{N}(x[n]-\overline{x})^2 - \sum_{n=1}^{N}x^2[n]\Big\} = \frac{N\,\overline{x}^2}{2\sigma^2}$$

Decide H_1 if $(\bar{X})^2 < \gamma' \Leftrightarrow |\bar{X}| < \gamma''$

Compare this detector with (unrealizable, also called clairvoyant) *NP detector* with known A. Assuming that sign of A is known, we can construct *UMP/NP/clairvoyant detector*, whose performance is described by

$P_{\mathrm{D}} = Q\big(Q^{-1}(P_{\mathrm{FA}}) - \sqrt{d^2}\big)$ where $d^2 = NA^2/\sigma^2$

ΔRemark: To make sure that the GLR test is implementable, we must be able to specify a threshold γ'' independent of A.

In this case, the *GLR test* is only slightly worse than the *clairvoyant detector(optimal)*.

Ex3. DC Level in WGN with A and variance both Unknown

$\mathcal{H}_0:\quad a = 0$ versus

$\mathcal{H}_1:\quad a = \underbrace{A}_{\text{unknown}} \neq 0.$

GLRT decides H_1 if

$$\Lambda_{\mathrm{GLR}}(\boldsymbol{x}) = \frac{\max_{\theta,\sigma^2} p(\boldsymbol{x}\,;\theta,\sigma^2)}{\max_{\sigma^2} p(\boldsymbol{x}\,;\theta=0,\sigma^2)} > \gamma$$

By **MLE**, $[\hat{A}\,,\hat{\sigma}_1{}^2]$ is **MLE** of vector parameter $\theta_1 = [A\,,\sigma^2]$ under H_1;

$\hat{\sigma}_0{}^2$ is the **MLE** of the parameter $\theta_0 = \sigma^2$.

$$\widehat{\sigma}_0^2 = \frac{1}{N}\sum_{n=1}^{N}x^2[n]\ \text{and}\ \widehat{\sigma}_1^2 = \frac{1}{N}\sum_{n=1}^{N}(x[n]-\overline{x})^2$$

*Note that we need to estimate σ^2 under both hypotheses.

i.e. the **GLR test** fits data with "best" DC-level signal $\hat{A}_{MLE} = \bar{X}$ finds the residual variance estimate $\hat{\sigma}_1{}^2$, and compares this estimate with variance estimate $\hat{\sigma}_0{}^2$ under H_0 (i.e. for θ = 0).

When signal is present, $\hat{\sigma}_1{}^2 \ll \hat{\sigma}_0{}^2 \Rightarrow \Lambda_{GLR}(x) \gg 1$

$$\widehat{\sigma}_1^2 = \frac{1}{N}\sum_{n=1}^{N}(\overline{x}-x[n])^2 = \widehat{\sigma}_0^2 - \overline{x}^2$$

Hence, $0 < (\bar{x})^2/\ \hat{\sigma}_0{}^2 < 1$

$$2\log\Lambda_{\mathrm{GLR}}(\boldsymbol{x}) = N\log\Big(\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_0^2 - \overline{x}^2}\Big) = N\log\Big(\frac{1}{1-\overline{x}^2/\widehat{\sigma}_0^2}\Big)$$

$\Rightarrow T(x) = (\bar{x})^2/\ \hat{\sigma}_0{}^2 > \lambda'$

Under H_0, pdf of T(x) doesn't depend on σ^2 ⇒ **GLR test** can be implemented, i.e. it is *CFAR*.

- A test is *Constant False Alarm Rate* if we can find a **threshold** that yields a detector with **constant** (specified) P_{FA}.

Large-data Record Performance of GLR Tests

Asymptotic assumptions:

(i) N is large; (ii) $\vartheta = [\theta, \varphi] \xrightarrow{a} N(\vartheta, I(\vartheta)^{-1})$

$1. General result (Nuisance):

Consider parametric model *p(x; ϑ)*

$$\vartheta = \begin{bmatrix}\theta\\ \varphi\end{bmatrix} = \begin{bmatrix}r\times 1\\ s\times 1\end{bmatrix}$$

Here, θ is to be tested and φ is a nuisance parameter vector. And equivalent hypotheses:

$\mathcal{H}_0\ :\ \boldsymbol{\theta} = \boldsymbol{\theta}_0,\quad \boldsymbol{\varphi}$

$\mathcal{H}_1\ :\ \boldsymbol{\theta} \neq \boldsymbol{\theta}_0,\quad \boldsymbol{\varphi}.$

GLRT decides H_1 if

$$\Lambda_{\mathrm{GLR}}(\boldsymbol{x}) = \frac{\max_{\boldsymbol{\theta},\boldsymbol{\varphi}} p(\boldsymbol{x}\,;\boldsymbol{\theta},\boldsymbol{\varphi})}{\max_{\boldsymbol{\varphi}} p(\boldsymbol{x}\,;\boldsymbol{\theta}=\boldsymbol{\theta}_0,\boldsymbol{\varphi})} > \lambda.$$

Then, as N ⟶ ∞,

$$2\ln L_G(x) \xrightarrow{a} \begin{cases} x^2{}_r\ , under\ H_0 \\ x'^2{}_r(\lambda)\ , under\ H_1\end{cases}$$

where r is *degree of freedom*, and *Non-centrality Parameter(λ)*:

$\lambda = [\theta_1 - \theta_0]^T[I_{\theta\theta}(\theta_0,\varphi) - I_{\theta\varphi}(\theta_0,\varphi)(I_{\varphi\varphi}(\theta_0,\varphi))^{-1}I_{\varphi\theta}(\theta_0,\varphi)][\theta_1 - \theta_0]$,

where $I(\theta,\varphi) = \begin{bmatrix} I_{\theta\theta} & I_{\theta\varphi}\\ I_{\varphi\theta} & I_{\varphi\varphi}\end{bmatrix} = \begin{bmatrix} r\times r & r\times s\\ s\times r & s\times s\end{bmatrix}$

$2. No nuisance parameter:

θ is an r × 1 vector and we test

$\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$

$\mathcal{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$

$2\ln L_G(x)$ are the same as above, with

$$\lambda = [\theta_1 - \theta_0]^T I(\theta_0)[\theta_1 - \theta_0]$$

Wald Test

$\mathcal{H}_0:\ \boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0}$ versus $\mathcal{H}_1:\ \boldsymbol{h}(\boldsymbol{\theta}) \neq \boldsymbol{0}$

where h is an r × 1 [r ⩽ dim(θ)] once continuously differentiable function. **Decide H_1 if**

$$T_{\mathrm{W}}(\boldsymbol{x}) = \boldsymbol{h}(\widehat{\boldsymbol{\theta}})^T\Big[H(\widehat{\boldsymbol{\theta}})\cdot\mathrm{CRB}(\widehat{\boldsymbol{\theta}})\cdot H(\widehat{\boldsymbol{\theta}})^T\Big]^{-1}\boldsymbol{h}(\widehat{\boldsymbol{\theta}}) > \lambda$$

where H(θ) = ∂h(θ)/∂θᵀ (having full rank r), CRB(θ) = I(θ)−1, and $\hat{\theta}$ is an unrestricted ML estimator of θ (under H_1). Then

$$T_W(x) \sim x^2{}_r \text{ under } H_0$$

Rao Test (Simplest)

Decide H_1 if

$$T_{\mathrm{R}}(\boldsymbol{x}) = \boldsymbol{s}(\widetilde{\boldsymbol{\theta}})^T\mathrm{CRB}(\widetilde{\boldsymbol{\theta}})\boldsymbol{s}(\widetilde{\boldsymbol{\theta}}) > \lambda$$

where s(θ) = ∂ log $p(x;\theta)/\partial\theta$, and $\tilde{\theta}$ is the restricted estimate of θ (under H_0). Then

$$T_R(x) \sim x^2{}_r \text{ under } H_0$$

Detection Performance

$P_{FA} = \Pr(T(x) > \gamma; H_0) =$

$\Pr(x > \sqrt{\gamma}; H_0) + \Pr(x < -\sqrt{\gamma}; H_0) = 2Q(\sqrt{\gamma})$

$P_D = Q\left(\sqrt{\gamma} - \sqrt{d^2}\right)$, *where* $d^2 = NA^2/\sigma^2$

\# For **Frequentist detection**, λ is a parameter chosen by setting *a probability of false alarm* α and solving: p(Λ(x) > λ|H0) = α for λ.

Classical Linear Model

Recall the *classical linear model*: **x = Hθ + w**;

where **x** is a measured N ×1 vector and **H** is a known deterministic N × p matrix, where N ⩾ p. Assume **w** ∼ N (0, σ²I) and **σ²** **known**. **θ** is deterministic unknown parameter vector, to be tested.

In general, we consider

$$H_0 : A\theta = b \text{ vs. } H_1 : A\theta \neq b$$

A is r×p(r≤p) of rank r. GLR test **decides H_1 if:**

$$T(\boldsymbol{x}) = \frac{(\boldsymbol{A}\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{b})^T[\boldsymbol{A}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{b})}{\sigma^2} > \tau$$

where $\widehat{\boldsymbol{\theta}}_1 = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{x}$ is the ML estimator of $\boldsymbol{\theta}$ under $\mathcal{H}_1$ (no restrictions).

Detection Performance is

$P_{\mathrm{FA}} = Q_{F_{r,N-p}}(\tau)$ and $P_{\mathrm{D}} = Q_{F_{r,N-p}(\lambda)}(\tau)$

where

$$\lambda = \frac{(\boldsymbol{A}\boldsymbol{\theta}_1 - \boldsymbol{b})^T[\boldsymbol{A}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\boldsymbol{\theta}_1 - \boldsymbol{b})}{\sigma^2}$$

GLRT for linear model where σ² unknown

GLR test decides H_1 if

$$T(\boldsymbol{x}) = \frac{N-p}{r}\cdot\frac{(\boldsymbol{A}\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{b})^T[\boldsymbol{A}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{b})}{\boldsymbol{x}^T(\boldsymbol{I} - \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T)\boldsymbol{x}}$$

where $\widehat{\boldsymbol{\theta}}_1 = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{x}$ is the ML estimator of $\boldsymbol{\theta}$ under $\mathcal{H}_1$ (no restrictions).

Detection Performance is:

$P_{\mathrm{FA}} = Q_{F_{r,N-p}}(\tau)$ and $P_{\mathrm{D}} = Q_{F_{r,N-p}(\lambda)}(\tau)$

where

$$\lambda = \frac{(\boldsymbol{A}\boldsymbol{\theta}_1 - \boldsymbol{b})^T[\boldsymbol{A}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\boldsymbol{\theta}_1 - \boldsymbol{b})}{\sigma^2}$$

Midterm Pr1.d MMSE for Bernoulli

From Bayesian perspective, prior distribution : π(p) = 6p(1−p), for p∈(0,1).

1. Find posterior probability distribution of p based on the N observations , {X1 , X2 , . . . , XN }, i.e., P (p|X1 , . . . , XN).

$$P(p|X_1,X_2,\ldots,X_N) = \frac{P(X_1,X_2,\ldots,X_N|p)\pi(p)}{\int_0^1 P(X_1,X_2,\ldots,X_N|p)\pi(p)dp}$$

$$= \frac{p^{1+\sum_{n=1}^N X_n}(1-p)^{N+1-\sum_{n=1}^N X_n}}{\int_0^1 p^{1+\sum_{n=1}^N X_n}(1-p)^{N+1-\sum_{n=1}^N X_n}dp}$$

$$= \frac{p^{1+\sum_{n=1}^N X_n}(1-p)^{N+1-\sum_{n=1}^N X_n}}{\mathrm{Beta}(2+\sum_{n=1}^N X_n, N+2-\sum_{n=1}^N X_n)}$$

2. Also find the Bayesian minimum mean squared error (MMSE) estimator for p which is given as E[p|X1, . . . , XN].

$$\mathbb{E}(p|X_1,X_2,\ldots,X_N) = \int_0^1 pP(p|X_1,\ldots,X_N)dp$$

$$= \frac{\int_0^1 p^{2+\sum_{n=1}^N X_n}(1-p)^{N+1-\sum_{n=1}^N X_n}dp}{\mathrm{Beta}(2+\sum_{n=1}^N X_n, N+2-\sum_{n=1}^N X_n)}$$

$$= \frac{\mathrm{Beta}(3+\sum_{n=1}^N X_n, N+2-\sum_{n=1}^N X_n)}{\mathrm{Beta}(2+\sum_{n=1}^N X_n, N+2-\sum_{n=1}^N X_n)}$$

HW7. Uniformly Most powerful test

Let x have a density: p(x; θ) = (1+θx) /2 for −1≤x≤1

a) If the hypotheses are:H_0 :θ=θ_0 H_1 :θ=θ_1 where θ_0∈ [−1, 0] and θ_1∈ [0, 1] are known. The likelihood ratio with threshold λ is:

$$1 + \theta_1 x \gtrless \lambda\ 1+\theta_0 x$$

For a level α,

$$\alpha = P_{FA} = \int_{\lambda'}^{1}\frac{1+\theta_0 x}{2}dx = \frac{1-\lambda'+\theta_0(1-\lambda'^2)/2}{2}$$

b) For a fixed θ_1, we know that the threshold value is 1 − 2α. This doesn't depend on value of θ_1 at all, so likelihood ratio test is always the most powerful test for all θ_1 > 0.

c) If hypotheses are:H_0 :θ≤0 vs. H_1 :θ>0 GLRT:

$$\frac{\max_{\theta_1\in[0,1]} 1+\theta_1 x}{\max_{\theta_0\in[-1,0]} 1+\theta_0 x} \gtrless \lambda$$

If the measurement x > 0, then the maximum value is θ1 = 1, θ0 = 0 and vice versa if x < 0.