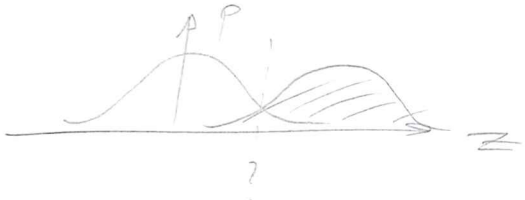
§9

Decision theory

States	measurements
s_1	$P(z \mid s_1)$ - [illegible] "healthy"
s_2	$P(z \mid s_2)$ - [illegible] "ill"



Task decide d_1 (guessed s_1) or d_2 (s_2) based on z.

$s_1 \rightarrow d_1$

$s_2 \rightarrow d_2$

$$\Lambda(z) = \frac{P(z \mid s_2)}{P(z \mid s_1)} \equiv \text{likelihood ratio}$$

Maximum likelihood test

d_1 if $P(z \mid s_1) > P(z \mid s_2)$
d_2 else

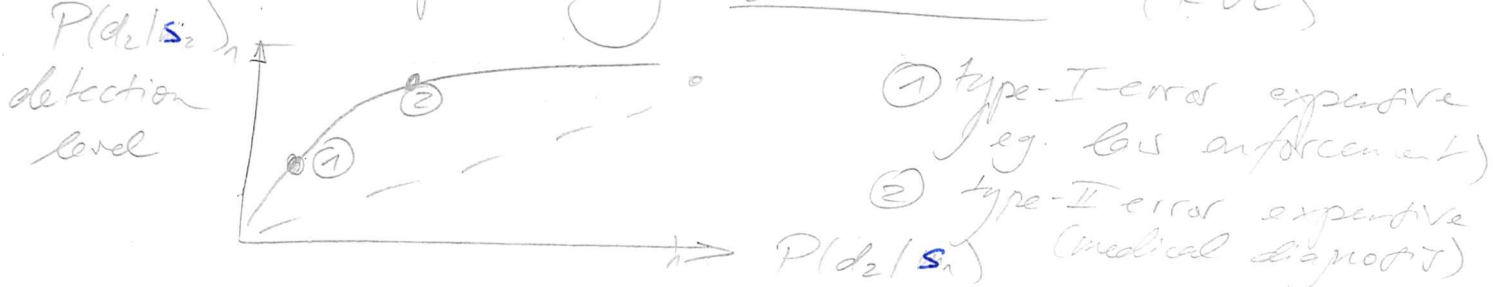
$$\Lambda(z) \underset{d_1}{\overset{d_2}{\gtrless}} 1$$

Types of error

$P(d_2 \mid s_1) \equiv$ Type-I-error ("false positive", "false alarm")

$P(d_1 \mid s_2) \equiv$ II ("false negative")

Receiver-operating-characteristic (ROC)



① type-I-error expensive (e.g. bus enforcement)

② type-II-error expensive (medical diagnosis)

Bayes risk criterion

- Cost matrix

  C_{11} = cost of d_1 if S_1 true

  C_{12} = — " — d_1 if S_2 true

- base rates $P(S_1)$, $P(S_2)$

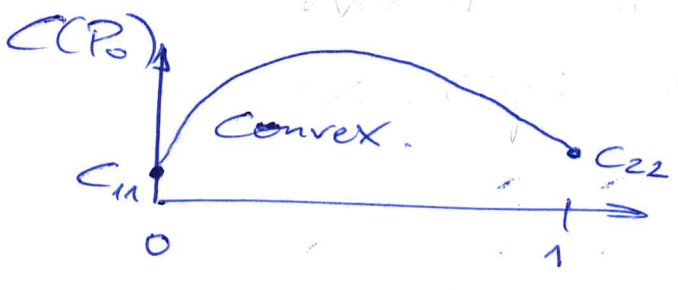
- expected cost

$$\langle C \rangle = C_{11} P(d_1 | S_1) \cdot P(S_1) + C_{21} P(d_2 | S_1) P(S_1) + \dots$$

- optimal test for likelihood ratio

$$\frac{P(z|S_2)}{P(z|S_1)} = \Lambda(z) \underset{d_1}{\overset{d_2}{\gtrless}} \frac{(C_{21} - C_{11}) P(S_1)}{(C_{12} - C_{22}) P(S_2)}$$

- expected cost for optimal test



Generalizations

- vector observations
- multiple decisions. $d_1, d_2, d_3, \dots$ for $S_1, S_2, S_3, \dots$

- Q_1, Q_2: (≡ critical base rates)

$$c_{12}\, Q_1 = C(Q_1) + \beta$$

$$c_{21}(1-Q_2) = C(Q_2) + \beta.$$

expected cost for...
- no observation
- one observation

$$T_1 = \frac{Q_1}{1-Q_1} \qquad T_2 = \frac{Q_2}{1-Q_2}$$

(≡ corresponding likelihood ratios)

$$Z_1 = \{ \Lambda_0 < T_1 \}$$

$$Z_2 = \{ \Lambda_0 > T_2 \}.$$

- Next step.

$$\Lambda_1(z) = \frac{P_1}{1-P_1}, \quad P_1 = P(S_2 | z_1)$$

$$= \frac{P(S_2 | z_1)}{P(S_1 | z_1)}$$

Assumptions: observations indep. and identically distributed

⇒ same statistics at each step, but prior updated.

$$Z_1 = \{ z_k : \Lambda_k(z_k) < T_1 \}.$$

$$Z_2 = \{ z_k : \Lambda_k(z_k) > T_2 \}.$$

Decision theory

$n \geq 2$ ~~fixed~~ alternate measurement models

↓

decide which one more likely

Statistical test

- working hypothesis. H_1
  (e.g. smoking promotes cancer)
  → no measurement model.

- null hypothesis $H_0 = \neg H_1$
  measurement model
  $p(z)$, $z \equiv$ measurement variable
  if H_0 true.

(1)

Statistical test = decision problem

d_1 : accept H_0 :

d_2 : reject H_0 :

decision rule

$$P(z \geq z_0 \mid H_0) \begin{matrix} > d_1 \\ < d_2 \end{matrix} \alpha$$

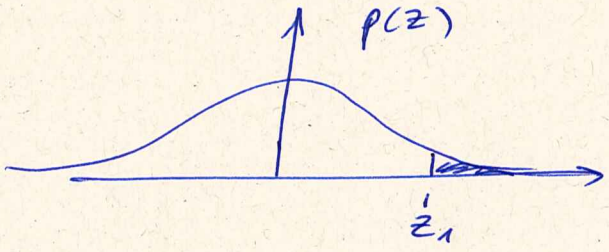
observation z_0

α = significance level

- "reject H_0 if z_0 too unlikely"
- α = probability of false positives.
- "look elsewhere fallacy"

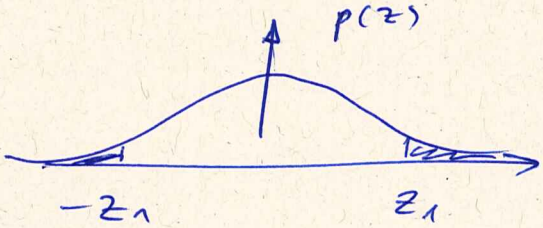
Simultaneous testing of n null hypotheses: $H_1, \ldots, H_n$ with fixed α:
even if H_j, $j = 1, \ldots, n$ true,
on average $\alpha \cdot n$ hypotheses will be rejected.

- One-sided / two-sided tests.



one-sided.

$$\int_{z_1}^{\infty} p(z)\,dz \lessgtr \alpha$$



two sided

$$\int_{\{|z|>z_1\}} p(z)\,dz \gtrless \alpha.$$

Trick depends on H_1!

Example (fair coin)

H_0: $p_{tail} = p_{head} = \frac{1}{2}$

measurement

k-times 'tail' after forming n-times

measurement model

$$p(k) = \frac{\binom{n}{k}}{2^k} \equiv \text{Binomial distribution}$$

one-sided test

$$\sum_{\tilde{k}=k_1}^{n} p(k) < \alpha,$$

H_1: bias towards tail,

two sided test

$$\sum_{k=k_1}^{n} + \sum_{k=1}^{n-k_1} p(k) < \alpha$$

H_1: unknown bias.

Student's t-test

1908: William Gosset at "Guiness"
(pen name "a student")

H_0: $\langle X \rangle = \mu$.

measurement model:

$$z = X \sim \mathcal{N}(\mu, \sigma^2).$$

with σ^2 unknown.

measurement

$z_1, \ldots, z_n$.

test

$$t = \frac{\bar{z} - \mu}{\hat{\sigma}/\sqrt{n}}$$

μ known (by H_0) assumed to be

$\hat{\sigma}$ = estimated variance

- ~~largely~~ independent of μ, σ^2.
- strategy one: Stochastic simulations.
- — " — two: analytic formula (with simplifying assumptions)
  - numerator ~ $\mathcal{N}$
  - denominator ~ $\chi^2(n-1)$

Generalization

- two populations: H_0: $\mu_1 = \mu_2$
  unpaired: e.g. group of males, group of females
  paired: e.g. group of patients before/after treatment

Def: χ^2- distribution

$y_1, \dots, y_n \sim N(0,1)$.

$\Rightarrow$

$$S = y_1^2 + \dots + y_n^2 \sim \chi^2(n)$$

mean: n

variance: $2n$.

Application to $\hat{\sigma}^2$:

$z_1, \dots, z_n \sim N(\mu, \sigma^2)$

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{j=1}^{n} (z_j - \bar{z})^2 \sim \sigma^2 \cdot \underbrace{\frac{\chi^2(n-1)}{n-1}}_{\text{mean: } 1}.$$

Proof: $(z_1, \dots, z_n) \in \mathbb{R}^n \equiv$ spherically distributed multivariate normal distribution

↓ linear proj. operator.

$(z_1 - \bar{z}, \dots, z_n - \bar{z}) \equiv$ — " — on $(n-1)$-dim'l subspace.

↑ $N(0, \frac{n-1}{n}\sigma^2)$

mean: 0

variance: $\left(\frac{n-1}{n}\right)^2 \sigma^2 + (n-1)\left(\frac{1}{n}\right)^2 \sigma^2 = \frac{n-1}{n}\sigma^2$.

Note: $\sum_{j=1}^{n} (X_j - \mu)^2 = \sum_{j=1}^{n} (X_j - \bar{X})^2 + n(\bar{X} - \mu)^2$ easy proof

Fisher's exact test:
testing categorical data for correlation

- Categorical properties
  - A, e.g. smoking
  - B, e.g. cancer

- Contingency table (absolute frequencies)

	A	$\bar{A}$	
B	N_{AB}	$N_{\bar{A}B}$	N_B ← margin totals
$\bar{B}$	$N_{A\bar{B}}$	$N_{\bar{A}\bar{B}}$	$N_{\bar{B}}$
	N_A	$N_{\bar{A}}$	N

- H_0: $p(B|A) = p(B)$

- Simplification: Suppose N_A, N_B, N fixed.

⇒ Sigle degree of freedom, e.g. N_{AB}.

$$p(N_{AB}) = \frac{\binom{N_B}{N_{AB}}\binom{N_{\bar{B}}}{N_A - N_{AB}}}{\binom{N}{N_A}}$$

≡ hyper-geometric distribution

(given N elements in total out of which K are positive; what is prob. to have k positive in selection of n)

Estimation theory

- Measurement model

  $p(z|x)$, z = measurement

  x = unknown parameter

- Max. likelihood estimate

  $\hat{x}$ that maximizes $p(z|x)$

- Later: Bayesian prior $p(x)$.

Warm-up example

How to combine noisy measurements?

- Methods 1 and 2 to measure x

  $P(z_1|x) = N_{z_1}(x, \sigma_1^2)$ — variance σ_1^2

  $P(z_2|x) = N_{z_2}(x, \sigma_2^2)$ — — — σ_2^2

- Max. likelihood estimate for x

  $P(z_1, z_2|\hat{x}) = P(z_1|\hat{x}) \cdot P(z_2|\hat{x}) \longrightarrow \max$

  $\Rightarrow$

  $$\hat{x} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} z_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} z_2$$

  = weighted mean

NB: $z_1, \ldots, z_n$ with $\sigma_1^2 = \ldots = \sigma_n^2$

$\Rightarrow$ arithmetic mean

Next: A-priori knowledge on $p(x)$

= Bayesian prior

Bayes formula

$$P(x|z) = \frac{P(z|x)\,P(x)}{P(z)}$$

$P(x)$ = Bayesian prior

$P(z|x)$ = measurement model

$P(z) = \int dx\; P(z|x)\,p(x)$

$P(x|z)$ = a-posteriori prob. distrib.

Example

$$P(x) = N(0, s^2)$$

$$P(z|x) = N(x, \theta^2) \Rightarrow P(z) = N(0, s^2 + \theta^2)$$

$$\Rightarrow P(x|z) = N(\hat{x}, p^2)$$

$$\lambda = \frac{1}{1 + \theta^2/s^2} = z, \qquad p^2 = \frac{1}{\frac{1}{s^2} + \frac{1}{\theta^2}}$$

$\hat{x}$ = weighted mean:

$$\hat{x} = \frac{\theta^2}{s^2 + \theta^2} \cdot 0 + \frac{s^2}{s^2 + \theta^2} \cdot z$$

Sequential measurements

Use previous a-posteriori distribution as new prior.

Example: Infotaxis, Kalman filter.

Curve fitting = estimation problem

- $x_i = f_i(\alpha), \quad i = 1, \dots k$

data points; physical model

α = unknown parameters

- Noisy measurements

$z_i = x_i + \xi_i \sim N(x_i, \sigma_i^2) \quad \sigma_i^2 = \sigma^2$

- Likelihood for observation

$P(\{z_i\}/\alpha) \longrightarrow \max$ for $\alpha = \hat{\alpha}$

$$P(\{z_i\}/\alpha) = \prod_{i=1}^{k} N_{z_i}(f_i(\alpha), \sigma^2)$$

$$\sim \exp\left[- \frac{\sum_{i=1}^{k} (f_i(\alpha) - z_i)^2}{2\sigma^2} \right]$$

$\Rightarrow \chi^2$ maximum $\sum_{i=1}^{k} (f_i(\alpha) - z_i)^2$

$\Rightarrow$ least-square fit.

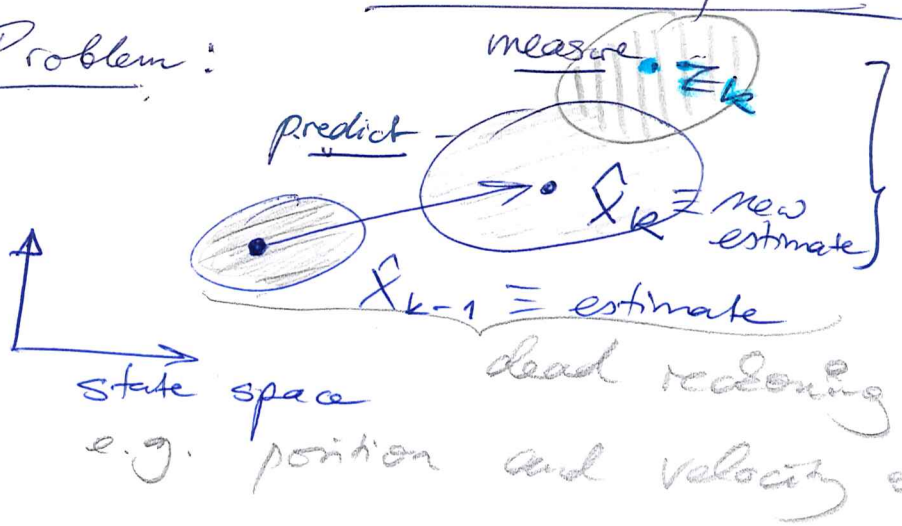
Generalizations:

- different $\sigma^2 \Rightarrow$ weighted fit
- non-Gaussian noise

Maximize $P(z_i/\alpha)$

Kalman filter

Problem:



how to combine optimally?

Application:

- Apollo.
- Missile guidance (Tomahawk)
- GPS.

Kalman filter

- Notation

$\hat{x}_{k-1|k-1}$ ≡ estimate of state at time t_{k-1}

$P_{k-1|k-1}$ ≡ covariance matrix of estimate

↓ predict

$\hat{x}_{k|k-1}$ ≡ a-priori estimate for time t_k.

$P_{k|k-1}$ ≡ a-priori-covariance

↓ measure

z_k

update (weighted average)

$\hat{x}_{k|k}$ ≡ a-posteriori estimate

$P_{k|k}$ ≡ a-posteriori covariance

Predict

$$\hat{x}_{k|k-1} = F_k x_{k|k-1} \quad = \text{a-priori}$$

$$P_{k|k-1} = F_k P_{k-1|k-1} F_k^T + Q_k$$

Measure

$$z_k$$

$$y_k = z_k - H_k \hat{x}_{k|k-1} \equiv \text{pre-fit residual (innovation)}$$

$$S_k = R + H_k P_{k|k-1} H_k^T \equiv \text{covariance of } y_k$$

Update:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k y_k \equiv \text{weighted average}$$

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} (I - K_k H_k)^T + K_k R_k K_k^T$$

$$\equiv \text{Covariance}$$

$\longrightarrow$ a-posteriori.

Optimal Kalman gain.

$$K_k = P_{k|k-1} H_k^T S_k^{-1}$$

$n \times m$ $\quad n \times n \quad n \times m \quad m \times m$

$\approx$ "ratio of covariances"

Theorem

The optimal Kalman gain

Minimizes $\quad E \|x_k - \hat{x}_{k|k}\|^2$.

N.B. For the optimal gain $P_{k|k} = (I - K_k H_k) P_{k|k-1}$.

N.B. Only update variables that have been measured (or are correlated to a measurement).

Process model

$$x_k = F_k x_{k-1} + w_k.$$

F_k = time propagator

w_k = process noise

$w_k \sim N(0, Q_k)$, i.e. mean 0 covariance Q_k.

Measurement model

$$z_k = H_k x_k + v_k.$$

H_k = measurement matrix.

N.B w.l.o.g. $H_{k}_{m \times n} = (1 \; \ldots \; 1 \; : \; 0)$

i.e. only $x_{k,1}, \ldots, x_{k,m}$, $m \leq n$ are measured

v_k = measurement noise

$v_k \sim N(0, R_k)$.

Proof

$$E = \|x_k - \hat{x}_{k|k}\|^2 = \mathrm{tr}\, P_{k|k}.$$

$$(*) \quad 0 \doteq \frac{\partial \mathrm{tr}\, P_{k|k}}{\partial K_k}.$$

We need rules of matrix gradients:

- $\frac{\partial \mathrm{tr}\, AB}{\partial A} = B^T$

because

$$\left(\frac{\partial \mathrm{tr}\, AB}{\partial A}\right)_{ij} = \frac{\partial \mathrm{tr}\, AB}{\partial A_{ij}} = \frac{\partial \sum_k (AB)_{kk}}{\partial A_{ij}} = \frac{\partial \sum_{kl} A_{kl} B_{lk}}{\partial A_{ij}} = B_{ji}$$

- $\frac{\partial \mathrm{tr}\, BAT}{\partial A} = B$

Now

$$\mathrm{tr}\, P_{k|k} = \mathrm{tr}\, P_{k|k-1} - \mathrm{tr}\, K_k H_k P_{k|k-1} - \mathrm{tr}\, P_{k|k-1}(K_k H_k)^T + \mathrm{tr}\, K_k H_k P_{k|k-1}(K_k H_k)^T + \mathrm{tr}\, K_k R_k K_k^T$$

$$\mathrm{tr}\, K_k S_k K_k^T$$

$$\frac{\partial \mathrm{tr}\, P_{k|k}}{\partial K_k} = -P_{k|k-1} H_k^T - P_{k|k-1} H_k^T + K_k S_k^T + K_k S_k$$

$$= -2 P_{k|k-1} H_k^T + 2 K_k S_k$$

Form (*): $K_k = P_{k|k-1} H_k^T S_k^{-1}$.

q.e.d.

Kalman filter example

Velocity

$$T\dot{v} = -v + \xi$$

position

$$\dot{x} = v$$

} discrete time

$$v_k = v_{k-1} - v_{k-1} \cdot \frac{\Delta t}{T} + \underbrace{\sqrt{\frac{2D_v \Delta t}{T^2}}\, N_k}_{w_k}$$

$$x_k = x_{k-1} + v_{k-1} \cdot \Delta t .$$

Measurement

$$z_k = x_k + \Theta_k, \quad x_k = x(t_k)$$

→ Matrix formulation

$$X_k = \begin{pmatrix} v_k \\ x_k \end{pmatrix}$$

$$X_k = F_k X_{k-1} + W_k .$$

$$F_k = \begin{pmatrix} 1 - \frac{\Delta t}{T} & 0 \\ \Delta t & 0 \end{pmatrix}, \quad W_k \sim N(0, Q_k)$$

$$Q_k = \begin{pmatrix} 2D_v \Delta t / T^2 & \\ & 0 \end{pmatrix}$$

$$Z_k = H_k X_k + V_k .$$

$$H_k = (0 \;\; 1)$$

$$V_k \sim N(0, R_k)$$

$$R_k = D_z .$$

Exercise

- Simulate X_k, Z_k, plot.
- Compute $\hat{x}_{k|k-1}, y_k, S_k, \hat{x}_{k|k}, P_{k|k}, K_k$.
- plot $\hat{x}_{k|k}$: effect of D_z?
- measure also velocity