Bayesian approach to estimation

- One can roughly say that there are two "estimation schools":
  1. The frequentist school
  2. The Bayesian school
- Likelihood function and related methodology belong to the frequentist school.
- Thomas Bayes (1701-1761) was an English mathematician and Presbyterian minister. Bayes never published what would eventually become his most famous accomplishment; his notes were edited and published after his death by Richard Price.(source: Wikipedia)

Frequentist vs Bayesian

- Frequentist is a person who view probability associated to an experiment which can be repeated with the same setting arbitrary many times (e.g. tossing a coin).
- Probability of an event (“heads” or “tails”) tells the long-run expected frequency of occurrence when repeating the experiment.
- Frequentist reports his analysis in form of (point) estimates and their standard errors, sampling distribution and confidence intervals.
- The Bayesian school has a subjective view on probability in which probability can be associated to (almost) any event.
- Person (subject) defines probability to an event which tells his degree of belief of occurrence of an event.

Frequentist vs Bayesian

- Bayesian thinks that an unknown parameter θ can have a fixed true value (θ_0), but he incorporates his prior knowledge and beliefs on the plausibility of different parameter values in terms of the prior probability distribution of θ.
- Thus θ is seen as a RV and the prior information about its distribution should be exploited in the estimation.
- Bayesian reports his analysis in form of posterior distribution of θ (the pdf of θ after the data has been observed) and its characteristics such as its mode, median or mean or mean.

Frequentist vs Bayesian

- Both Bayesian and frequentist approve the same (frequentist) probabality model $f(\mathbf{y}|\theta)$ or likelihood function.
- The difference is that Bayesian uses his prior knowledge/beliefs on plausibility of different parameter values in terms of prior distribution $f(\theta)$.
- Frequentist reports his analysis using tests, estimates, confidence intervals and sampling distributions ("What did this experiment tell us").
- Bayesian reports the posterior disribution $f(\theta|\mathbf{y})$ ("What did I learn of θ by observing the data"), the MAP and MMSE estimators, etc.

Frequentist vs Bayesian

Frequentist often criticices Bayesian because:

1. The prior distribution is mystically obtained. Often it is selected so that the derivation of the posterior distribution and the related MAP/MMSE estimator would be easier.
2. Often it is assumed that the joint distribution $f(\mathbf{y}, \theta)$ is Gaussian. When the Gaussian assumption can not be made, the Bayesian estimators are difficult to derive in closed form and require computationally intensive methods.
3. Often there is *NO* reliable information on the prior distribution of θ or such information may be difficult to represent quantitatively in a form of a pdf. or pmf. This can also lead to large bias especially for very small sample lengths. Morever, different prior distributions give different end results (posterior, MAP estimator, etc).

Recap of Basic probability rules

- Conditional probability:

  $\mathrm{Pr}(A|B) =$ probability of A , given B event is true

- Chain rule:

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

- Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Law of total probability:

$$P(A) = \sum_n P(A|B_n)P(B_n)$$

- Conditional independence:

$$P(A, B|C) = P(A|C)P(B|C)$$

Estimation of Random Parameters - Basic Concepts

- In this section we study Mean Square and Maximum A Posteriori estimation.
- Bayesian approach: Unknown parameter is assumed to be random and we may have some prior information about is pdf.
- Parameter is a random quantity which is statistically related to the observations.
- Parameter θ to be estimated is a realization of random variable Θ. It has a prior pdf (pdf before any data is observed)
- Prior knowledge about θ can be exploited in deriving estimator.
- The estimators obtained in this manner are optimal on the average or with respect to the assumed prior pdf of the parameter θ

Estimation of Random Parameters - Basic Concepts ...

- After a large number of observations are acquired, the prior information has less influence on the computed estimates. In particular in the cases where the parameter does not change in time.
- Static model, parameters are constant in time (vs. dynamical parameter/signal in optimum filtering).
- Estimation error is a random vector and needs to be described probabilistically.

Bayesian approach to estimating random parameters

- Prior knowledge about θ is incorporated into our estimator. θ is assumed to be a RV with a prior pdf $f(\theta)$.
- Bayes estimator combines the prior knowledge and information contained in the data. It updates the prior knowledge in an optimal manner using observed data
- The implementations are based on the Bayes' theorem:

$$f(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)f(\theta)}{f(\mathbf{y})}$$

where $f(\theta|\mathbf{y})$ is a posterior pdf (pdf after the data have been observed) and $f(\theta)$

$$f(\theta) = \int f(\mathbf{y}, \theta) d\mathbf{y}$$

may be thought as a priori pdf of θ or our prior knowledge on θ before the data are observed.

Bayesian approach to estimating random parameters ...

- The resulting estimators are optimal on the average or with respect to the assumed prior pdf of θ
- The prior knowledge becomes less and less influential as the number of observations grows.
- One needs to know the a posteriori density which can be determined using the Bayes rule.
- In order to do it a priori density has to be known. Often there is no reliable information on a priori density.
- Prior knowledge may also be difficult to represent in a form of a prior pdf.
- A bayes estimator may be highly biased if the prior statistics are not correct. This is the case in particular for small sample sizes N.
- If no a priori info is available θ is treated as an unknown deterministic constant

Bayesian approach to estimating random parameters ...

- In case we can determine the posterior pdf, it's mean, mode or median is typically used as an estimator.
- In particular, two types of estimators are studied: Mean Squared (MS, mean of the posterior pdf) and Maximum A Posteriori (MAP, mode of the posterior pdf).

Cost functions

- Any estimate based on finite number of observations is expected to contain some error
- Therefore we define estimators that minimize estimation error in some sense.
- The problem is to find an estimate $\hat{\theta}$ that provides the best guess in some sense of the true value of the parameter given the observations $\mathbf{y}$.
- We can associate a cost averaged over observations $\mathbf{y}$ for each θ

$$R_\theta = E[C(\hat{\theta}(\mathbf{y})), \theta]$$

  where C is a cost function.
- The Bayes risk is then

$$r(\hat{\theta}) = E[R_\Theta(\hat{\theta})] = E[E[C(\hat{\theta}(\mathbf{y})), \Theta|\mathbf{y}]]$$

  and the design goal is to find an estimator that minimizes $r(\hat{\theta})$.

Cost functions ...

- A reasonable cost function must be symmetric; convex (not strictly) or nondecreasing; and zero error must have zero cost.
- various cost functions: Quadratic cost. Mean Square error

$$C(\hat{\theta}) = (\hat{\theta}(\mathbf{y}) - \theta)^2$$

  The Minimum Mean Square estimate is denoted by $\hat{\theta}_{MSE}$
- Uniform cost

$$C(\hat{\theta}) = \begin{cases} 0 & \text{if } |\hat{\theta}(y) - \theta| \leq a \\ 1 & \text{if } |\hat{\theta}(y) - \theta| > a \end{cases}$$

  where threshold $a > 0$;
- Mode of the a posteriori pdf (MAP estimator $\hat{\theta}_{MAP}$) minimizes the uniform cost criterion

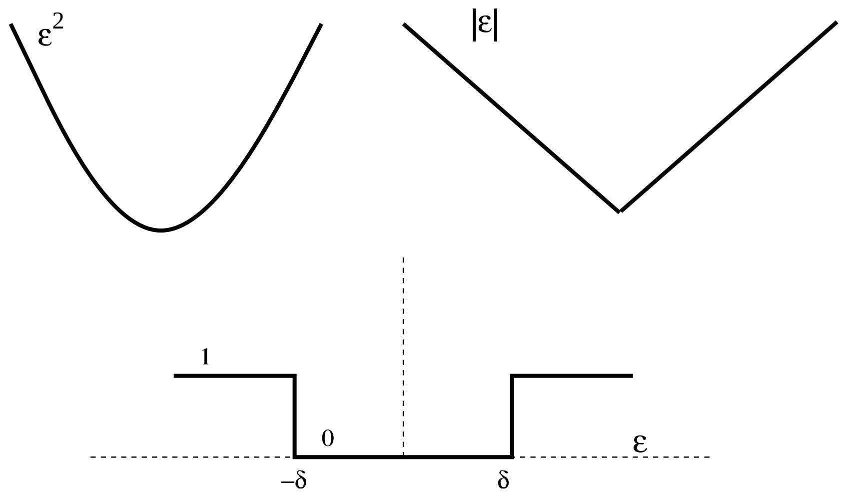
Cost functions ...

- Mean Absolute Error (median of the a posteriori density)

$$C(\hat{\theta}) = |\hat{\theta}(\mathbf{y}) - \theta|.$$

  The minimum mean absolute value estimate is denoted $\hat{\theta}_{MAE}$
- Median of the a posteriori pdf minimizes the MAE
- The MAP estimator corresponds to the mode of the posteriori density function whereas the MS estimator corresponds to the mean of the posteriori pdf.
  - minimizing Bayes risk using quadratic cost function gives the MMSE (mean of the posterior pdf).
  - minimizing Bayes risk using Hit-or-miss cost function gives the MAP estimator (mode of the posterior pdf).
  - Minimizing absolute error cost function gives an estimator which corresponds to the median of the posterior pdf.

Cost functions ...



Cost functions ...

- Cost functions for vector parameters. E.g. Euclidean distance
$$\sum_{i=1}^{m} C_i(\hat{\theta}_i, \theta_i) = \sum_{i=1}^{m} (\hat{\theta}_i - \theta_i)^2$$
Sum of absolute distances
$$\sum_{i=1}^{m} |\hat{\theta}_i - \theta_i|$$
and
$$C(\hat{\theta}) = \begin{cases} 0 & \text{if } max_{i=1,\ldots,m} \, |\hat{\theta}_i - \theta_i| \leq a \\ 1 & \text{if } max_{i=1,\ldots,m} \, |\hat{\theta}_i - \theta_i| > a \end{cases}$$
- Generalization of quadratic cost function:
$$(\hat{\theta}_i - \theta_i)^T A (\hat{\theta}_i - \theta_i).$$
The bayes risk is then $tr(AE[Cov(\Theta|\mathbf{y})])$.

Mean Squared (MS) Estimation of Random Parameters

- θ is now viewed as a vector of random unknown parameters.
- Given the measurements we should determine the parameters
- Measurements $\mathbf{y}$ that are assumed to depend on θ are available to us (no specific structural dependency is assumed, however).
- Performance index related to the square of the estimation error. Mean square cost function. Large errors have large cost.
- The risk function to be minimized is the expected value of the mean square error.

$$E[(\theta - \hat{\theta})^T(\theta - \hat{\theta})]$$

where θ is a random variable (not an unknown constant as in the case of classical MSE).

Mean Squared (MS) Estimation of Random Parameters ...

- The mean square estimate is the expected value of the conditional density i.e. mean of the a posteriori density (pdf after data has been observed)

$$\hat{\theta}_{MS} = E[\theta|\mathbf{y}] = \int \theta f(\theta|\mathbf{y}) d\theta.$$

- Given measurements $y(1), ..., y(k)$ we shall determine an estimator $\hat{\theta}_{MS}$ such that the Mean Squared Error (MSE)

$$J[\tilde{\theta}_{MS}] = E[\tilde{\theta}_{MS}^T \tilde{\theta}_{MS}]$$

is minimized. Estimation error is denoted by $\tilde{\theta}_{MS} = \theta - \hat{\theta}$.

- In case of linear estimator

$$\hat{\theta}_{MS} = \sum_{i=1}^{N} A(i) y(i)$$

Mean Squared (MS) Estimation of Random Parameters ...

- Minimizing conditional expectation $E[\tilde{\theta}_{MS}^T \tilde{\theta}_{MS} | \mathbf{y}]$ with respect to $\hat{\theta}_{MS}$ is equivalent to our original objective of minimizing the total expectation $E[\tilde{\theta}_{MS}^T \tilde{\theta}_{MS}]$
- The mean square estimation problem:
Given measurements $y(1), ..., y(N)$ we shall determine an estimator $\hat{\theta}_{MS}$ such that the conditional MS error

$$J[\tilde{\theta}_{MS}] = E[\tilde{\theta}_{MS}^T \tilde{\theta}_{MS} | y(1), ..., y(N)]$$

is minimized.

Derivation of the MS Estimator

- Theorem: The estimator that minimizes the mean squared error is

$$\hat{\theta}_{MS} = E[\theta | \mathbf{y}]$$

- Proof:

$$J[\tilde{\theta}_{MS}] = E[\tilde{\theta}_{MS}^T \tilde{\theta}_{MS} | \mathbf{y}]$$

$$= E[\theta^T \theta - \theta \hat{\theta}_{MS} - \hat{\theta}_{MS}^T \theta + \hat{\theta}_{MS}^T \hat{\theta}_{MS} | \mathbf{y}]$$

$$= E[\theta^T \theta | \mathbf{y}] - E[\theta^T | \mathbf{y}] \hat{\theta}_{MS} - \hat{\theta}_{MS}^T E[\theta | \mathbf{y}] + \hat{\theta}_{MS}^T \hat{\theta}_{MS}$$

$$= E[\theta^T \theta | \mathbf{y}] + [\hat{\theta}_{MS} - E[\theta | \mathbf{y}]]^T [\hat{\theta}_{MS} - E[\theta | \mathbf{y}]] - E[\theta^T | \mathbf{y}] E[\theta | \mathbf{y}]$$

Note that $E[\hat{\theta}_{MS} | \mathbf{y}] = \hat{\theta}_{MS}$ because $\hat{\theta}_{MS}$ is a function of $\mathbf{y}$ by definition. The first and the last terms do not depend on $\hat{\theta}_{MS}$ hence the middle term is minimized by choosing $\hat{\theta}_{MS} = E[\theta | \mathbf{y}]$. QED.

Derivation of the MS Estimator ...

- Two major cases are:
  - $\mathbf{y}$ and θ are jointly Gaussian.
  - $\mathbf{y}$ and θ are not jointly Gaussian.
    (subcases linear and nonlinear estimators)
- Corollary: When θ and $\mathbf{y}$ are jointly Gaussian the Minimum MS Estimator (MMSE) is
$$\hat{\theta}_{MS} = \mathbf{m}_\theta + P_{\theta y} P_y^{-1}[\mathbf{y} - \mathbf{m}_y]$$
  i.e., it can be evaluated using conditional mean
$$E[\mathbf{x}|\mathbf{y}] = \mathbf{m}_x + P_{xy} P_y^{-1}[\mathbf{y} - \mathbf{m}_y].$$
- In case θ and $\mathbf{y}$ are not necessarily jointly Gaussian and we know $\mathbf{m}_\theta, \mathbf{m}_y, P_y, P_{\theta y}$. In this case an estimator that is constrained to be an affine transformation of $\mathbf{y}$ and minimizes the mean square error is given by the above corollary.

Derivation of the MS Estimator ...

- Linear Minimum MS Estimator (LMMSE) and MS Estimator are the same when $\mathbf{y}$ and θ are jointly Gaussian.
- If θ and $\mathbf{y}$ are not jointly Gaussian then
$$\hat{\theta}_{MS} = E[\theta|\mathbf{y}]$$
  which is a nonlinear function of measurements $\mathbf{y}$ (nonlinear estimator).

Fourier analysis using MS estimator

- Let the data model be

$$y(n) = \alpha cos(2\pi f_0 n) + \beta sin(2\pi f_0 n) + v(n),\ n = 0, ..., N-1$$

where f_0 is a multiple of $\frac{1}{N}$ and $v(n)$ is white Gaussian noise with variance σ^2. We want to estimate $\theta = [\alpha\ \beta]^T$.

- Now α, β are assumed to be RV's with prior pdf

$$\theta \sim N(0, \sigma_\theta^2 I)$$

and θ is independent of $v(t)$.

Fourier analysis using MS estimator

- The data model in matrix form is

$$\mathbf{y} = H\theta + \mathbf{v}$$

where

$$H = \begin{bmatrix} 1 & 0 \\ cos(2\pi f_0) & sin(2\pi f_0) \\ \vdots & \vdots \\ cos(2\pi f_0(N-1)) & sin(2\pi f_0(N-1)) \end{bmatrix}$$

Fourier analysis using MS estimator

- Let $m_\theta = 0, P_\theta = \sigma_\theta^2 I,$ and $P_v = \sigma_v^2 I$. As a result we obtain

$$\hat{\theta}_{MS} = E[\theta|\mathbf{y}] = \sigma_\theta^2 H^T [H\sigma_\theta^2 H^T + \sigma_v^2 I]^{-1}\mathbf{y}$$

and

$$P_{\theta|\mathbf{y}} = \sigma_\theta^2 I - \sigma_\theta^2 H^T [H\sigma_\theta^2 H^T + \sigma_v^2 I]^{-1} H\sigma_\theta^2$$

- These can rewritten for Bayesian linear model

$$\hat{\theta}_{MS} = m_\theta + [P_\theta^{-1} + H^T P_v^{-1} H]^{-1} H^T P_v^{-1}(\mathbf{y} - Hm_\theta)$$

$$\hat{\theta}_{MS} = (\frac{1}{\sigma_\theta^2} I + H^T \frac{1}{\sigma_v^2} H)^{-1} H^T \frac{1}{\sigma_v^2}\mathbf{y}$$

and

$$P_{\theta|\mathbf{y}} = (\frac{1}{\sigma_\theta^2} I + H^T \frac{1}{\sigma_v^2} H)^{-1}$$

Fourier analysis using MS estimator

- Because the columns of H are orthogonal (due to the choice of frequencies) we get

$$H^T H = \frac{N}{2} I$$

which is substituted into the formula for $\hat{\theta}_{MS}$ and we get

$$\hat{\theta}_{MS} = (\frac{1}{\sigma_\theta^2} I + \frac{N}{2\sigma_v^2} I)^{-1} \frac{H^T\mathbf{y}}{\sigma_v^2}$$

$$= \frac{\frac{1}{\sigma_v^2}}{\frac{1}{\sigma_\theta^2} + \frac{N}{2\sigma_v^2}} H^T\mathbf{y}$$

Fourier analysis using MS estimator

- The MS estimator gives

$$\hat{\alpha} = \frac{1}{1 + \frac{2\sigma_v^2/N}{\sigma_\theta^2}}[\frac{2}{N}\sum_{n=0}^{N-1} y(n)cos(2\pi f_0 n)]$$

and

$$\hat{\beta} = \frac{1}{1 + \frac{2\sigma_v^2/N}{\sigma_\theta^2}}[\frac{2}{N}\sum_{n=0}^{N-1} y(n)sin(2\pi f_0 n)]$$

- The posterior covariance matrix is

$$P_{\theta|\mathbf{y}} = \frac{1}{\frac{1}{\sigma_\theta^2} + \frac{N}{2\sigma_v^2}} I$$

which does not depend on $\mathbf{y}$.

Fourier analysis using MS estimator

- The mean square error for each parameter are

$$MSE(\hat{\alpha}) = \frac{1}{\frac{1}{\sigma_\theta^2} + \frac{N}{2\sigma_v^2}}$$

and

$$MSE(\hat{\beta}) = \frac{1}{\frac{1}{\sigma_\theta^2} + \frac{N}{2\sigma_v^2}}$$

- If there were no prior knowledge, i.e., $P_\theta^{-1} \rightarrow 0$ we would have obtained the classical estimator

$$\hat{\theta} = [H^T P_V^{-1} H]^{-1} H^T P_V^{-1}\mathbf{y}$$

Orthogonality Principle

- Suppose $f(\mathbf{y})$ is any function of the data $\mathbf{y}$. The error in the MS estimation is orthogonal to $f(\mathbf{y})$ in the sense that

$$E[[\theta - \hat{\theta}_{MS}]f^T(\mathbf{y})] = 0 = E[\tilde{\theta}_{MS} f^T(\mathbf{y})]$$

- Proof:

$$E[[\theta - \hat{\theta}_{MS}]f^T(\mathbf{y})] = E[E[(\theta - \hat{\theta}_{MS})|\mathbf{y}]f^T(\mathbf{y})$$
$$= E[[\hat{\theta}_{MS} - \hat{\theta}_{MS}]f^T(\mathbf{y})] = 0$$

- Note: $E[\theta|\mathbf{y}] = \hat{\theta}_{MS}$ and $\hat{\theta}_{MS}$ is not random when $\mathbf{y}$ is specified.
- Frequently we deal with a situation where $f[\mathbf{y}] = \hat{\theta}_{MS}$ and we can write

$$E[\tilde{\theta}_{MS}\hat{\theta}_{MS}^T] = 0$$

Properties of MS estimators

- θ and $\mathbf{y}$ are jointly Gaussian
  – Unbiasedness: The Mean Squared estimator is unbiased
  $$E[\hat{\theta}_{MS}] = \mathbf{m}_\theta$$
  – Minimum Variance Estimator (MVE): The MS estimator is MVE
  $$\sigma^2_{\tilde{\theta}_i,MS} = E[\tilde{\theta}^2_{i,MS}]$$
  $$J[\tilde{\theta}_{MS}] = \sum_{i=1}^{n} \sigma^2_{\tilde{\theta}_i,MS}$$
  – Linearity: $\hat{\theta}_{MS}$ is a linear (affine) estimator.
  – Gaussianity: Both $\hat{\theta}_{MS}$ and $\tilde{\theta}_{MS}$ are Gaussian. $\hat{\theta}_{MS}$ is an affine transformation of $\mathbf{y}$ which are Gaussian. Estimation error is an affine transformation of jointly Gaussian vectors $\theta, \mathbf{y}$ hence $\tilde{\theta}_{MS}$ is also Gaussian.

Properties of MS estimators ...

- Generic Linear and Gaussian model

$$\mathbf{y} = H\theta + \mathbf{v}$$

where H is deterministic, $\mathbf{v}$ is white Gaussian with known covariance matrix R. θ is multivariate Gaussian with known mean $\mathbf{m}_\theta$ and covariance P_θ, i.e., $\theta \sim N(\theta; \mathbf{m}_\theta, P_\theta)$ and $\theta, \mathbf{v}$ are mutually uncorrelated (jointly Gaussian since they are Gaussian and uncorrelated)

- Theorem: for the generic linear model above in which H is deterministic, $\mathbf{v}$ is white Gaussian with known covariance matrix R, $\theta \sim N(\theta; \mathbf{m}_\theta, P_\theta)$ and V are mutually uncorrelated

$$\hat{\theta}_{MS} = \mathbf{m}_\theta + P_\theta H^T [H P_\theta H^T + R]^{-1} [\mathbf{y} - H\mathbf{m}_\theta]$$

Properties of MS estimators ...

- Let P_{MS} be the error covariance matrix associated with $\hat{\theta}_{MS}$ then

$$P_{MS} = [P_\theta^{-1} + H^T R^{-1} H]^{-1}$$

so that $\hat{\theta}_{MS}$ can be re-expressed as

$$\hat{\theta}_{MS} = \mathbf{m}_\theta + P_{MS} H^T R^{-1} [\mathbf{y} - H\mathbf{m}_\theta]$$

- Proof omitted.
- Note: θ depends on all given information $\mathbf{y}, H, \mathbf{m}_\theta, P_\theta, R$. Prior to any measurement we can estimate θ by $\mathbf{m}_\theta$. After measurement $\mathbf{m}_\theta$ is updated by the second term. $H\mathbf{m}_\theta$ represents the predicted value of $\mathbf{y}$.

Relations to BLUE

- The assumption of deterministic θ was never needed in the derivation of $\hat{\theta}_{BLUE}$.
- $\hat{\theta}_{BLUE}$ is applicable to random and deterministic parameter in our linear model. Since $\hat{\theta}_{BLUE}$ is a special case of $\hat{\theta}_{WLS}$ it also is applicable to both random and deterministic parameter in our linear model.
- Theorem: If $P_\theta^{-1} = 0$ and H deterministic then
$$\hat{\theta}_{MS} = \hat{\theta}_{BLUE}$$
By setting $P_\theta^{-1} = 0$ in previous theorem we get
$$P_{MS} = [H^T R^{-1} H]^{-1}$$
and
$$\hat{\theta}_{MS} = [H^T R^{-1} H]^{-1} H^T R^{-1} \mathbf{y}$$

Relations to BLUE ...

- Note: Suppose, for example, that θ are uncorrelated and consequently we have diagonal $P_\theta^{-1} = 0$.
- This means that all variances are large and we have no idea where θ_i is located about its mean value. In other words, dependence of $\hat{\theta}_{MS}$ on a priori statistical information on θ is removed.

MAP estimation of random parameters

- Maximum A Posteriori (MAP) estimation
- Bayes rule is used:
$$f(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)f(\theta)}{f(\mathbf{y})}$$
where $f(\theta|\mathbf{y})$ is a posteriori conditional density function and $f(\theta)$ is the prior pdf for θ.
- Both $f(\mathbf{y}|\theta)$ and $f(\theta)$ have to be specified. Then the value of θ that maximizes $f(\theta|\mathbf{y})$ have to be found (or $ln\ f(\theta|\mathbf{y})$).
- The MAP estimator is given by
$$\hat{\theta}_{MAP} = arg\ max_\theta\ f(\mathbf{y}|\theta)f(\theta)$$
This resembles the Maximum Likelihood estimator except for the prior pdf $f(\theta)$.

MAP estimation of random parameters ...

- Theorem: If $\mathbf{y}$ and θ are jointly Gaussian then
$$\hat{\theta}_{MAP} = \hat{\theta}_{MS}$$
- Proof: If $\mathbf{y}$ and θ are jointly Gaussian then
$$f(\theta|\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^n |P_{\theta|y}|}} exp \frac{-1}{2}[\theta - m]^T P_{\theta|y}^{-1}[\theta - m]$$
where $m = E[\theta|\mathbf{y}]$.
- $\hat{\theta}_{MAP}$ is found by maximizing $f(\theta|\mathbf{y})$ or by minimizing the argument of the exponential $[\theta - m]^T P_{\theta|y}^{-1}[\theta - m]$ which is obtained by setting it to zero and this occurs when $\theta - m = 0, \theta = \hat{\theta}_{MAP}$, i.e., $\hat{\theta}_{MAP} = E[\theta|\mathbf{y}]$ and we conclude that $\hat{\theta}_{MAP} = \hat{\theta}_{MS}$. QED.

MAP estimation of random parameters ...

- For linear Gaussian model $\mathbf{y} = H\theta + \mathbf{v}$, H deterministic

$$\hat{\theta}_{MAP} = \hat{\theta}_{MS} = \hat{\theta}_{BLU},$$

  i.e., all of these approaches lead to the same estimator.
- Maximum Likelihood (ML) estimator may be considered as a special case of MAP with a uniform prior.

MAP example

- Assume that

$$f(y(n)|\theta) = \begin{cases} \theta exp[-\theta y(n)] & y(n) > 0 \\ 0 & y(n) \leq 0 \end{cases}$$

  and for $y(n)$'s

$$f(\mathbf{y}|\theta) = \prod_{n=0}^{N-1} f(y(n)|\theta),$$

  i.e., conditionally i.i.d.
- The prior pdf is

$$f(\theta) == \begin{cases} \lambda exp[-\lambda\theta] & \theta > 0 \\ 0 & \theta \leq 0 \end{cases}$$

MAP example

- The MAP estimator is found by maximizing

$$\hat{\theta}_{MAP} = arg\ max_{\theta}\ f(\mathbf{y}|\theta)f(\theta)$$

or

$$\hat{\theta}_{MAP} = arg\ max_{\theta}[\ ln\ f(\mathbf{y}|\theta) + ln\ f(\theta)]$$

- Now we maximize

$$f(\theta) = ln\ f(\mathbf{y}|\theta) + ln\ f(\theta)$$

$$= ln\ [\theta^N exp\{-\theta \sum_{n=0}^{N-1} y(n)\}] + ln\ [\lambda exp\{-\lambda\theta\}]$$

$$= N ln\ \theta - N\theta\overline{y} + ln\ \lambda - \lambda\theta$$

for $\theta > 0$.

MAP example

- Now we will differentiate wrt. θ and set the derivative to 0:

$$\frac{df(\theta)}{d\theta} = \frac{N}{\theta} - N\overline{y} - \lambda = 0$$

which yields the MAP estimator

$$\hat{\theta}_{MAP} = \frac{1}{\overline{y} + \frac{\lambda}{N}}$$

MAP and MMSE example

- Estimation of signal power: The power x is at most equal to the square of magnitude of the observed signal y. The joint density function of the RV's x, y is

$$f_{yx}(y,x) = \begin{cases} 10x & 0 \leq x \leq y^2, 0 \leq y \leq 1 \\ 0 & otherwise \end{cases}$$

- The marginal density of y is obtained by integrating

$$f_y(y) = \int_0^{y^2} 10x \, dx = |_0^{y^2} 5x^2 = 5y^4$$

and the conditional density is

$$f(x|y) = \frac{10x}{5y^4} = \frac{2x}{y^4}, \;\; 0 \leq x \leq y^2$$

MAP and MMSE example

- The maximum of the conditional density occurs at y^2

$$\hat{x}_{MAP} = y^2$$

whereas the MS estimate is obtained by computing the mean of the conditional density

$$\int_{-\infty}^{\infty} x f(x|y) dx = \int_0^{y^2} x \frac{2x}{y^4} dx = |_0^{y^2} \frac{2}{3} \frac{x^3}{y^4} = \frac{2}{3} y^2$$

and

$$\hat{x}_{MS} = \frac{2}{3} y^2$$

MAP and MMSE example

- The minimum MS error is

$$E[(x-\hat{x}_{MS})^2] = \int_0^1 \int_0^{y^2} (x - \frac{2}{3}y^2)^2 10x \; dxdy = 0.0309$$

which is less that the error for $\hat{x}_{MAP}$ obtained by

$$E[(x-\hat{x}_{MAP})^2] = \int_0^1 \int_0^{y^2} (x - y^2)^2 10x \; dxdy = 0.0926$$

MAP and MMSE example

- Consider the situation in which messages are arriving to a switch at rate characterized by the following conditional density function given θ:

$$f(y|\theta) = \begin{cases} \theta e^{-\theta y}, & y \geq 0 \\ 0, & y < 0 \end{cases}$$

Suppose that our prior information on θ has the following density

$$f(\theta) = \begin{cases} \alpha e^{-\alpha \theta}, & \theta \geq 0 \\ 0, & \theta < 0 \end{cases}$$

where $\alpha > 0$ is known. Find the Mean Square and MAP estimates of θ.

MAP and MMSE example

- Remember the Bayes rule:

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)}$$

- Now we know $f(y|\theta)$ and $f(\theta)$ and we want to find out a posteriori density $f(\theta|Y)$. In order to apply Bayes rule we need to compute $f(y)$ first.
- The formula for conditional distribution is

$$f(y|\theta) = \frac{f(y,\theta)}{f(\theta)} \Leftrightarrow f(y|\theta)f(\theta) = f(y,\theta)$$

and find the marginal for $f(y)$ by integrating θ out from $f(y,\theta)$.

MAP and MMSE example

- Place this integral into denominator and A Posteriori density becomes

$$f(\theta|y) = \frac{\alpha\theta e^{-(\alpha+y)\theta}}{\int_0^\infty \alpha\theta e^{-(\alpha+y)\theta} d\theta}$$

- The integration in the denominator is done by part using $u = \theta, dv = \alpha e^{-\theta(\alpha+y)}$. The result of the integration is

$$f(y) = \frac{\alpha}{(\alpha+y)^2}$$

and A Posteriori density is as follows

$$f(\theta|y) = (\alpha+y)^2\theta e^{-\theta(\alpha+y)}$$

MAP and MMSE example

- The MS estimate is the mean of the A Posteriori density, i.e.,

$$\hat{\theta}_{MS} = \int_0^\infty \theta f(\theta|y) d\theta$$

$$= (\alpha + y)^2 \int_0^\infty \theta^2 e^{-\theta(\alpha+y)} = \frac{2}{\alpha + y}$$

- $\hat{\theta}_{MS} = \frac{2}{\alpha+y}$

MAP and MMSE example

- The MAP estimate is obtained by maximizing

$$\frac{\partial}{\partial \theta}[log\ f(y|\theta) f(\theta)]$$

$$= \frac{\partial}{\partial \theta}[log\ \theta - \theta y + log\ \alpha - \alpha\theta] = \theta^{-1} - (\alpha + y).$$

By setting this to zero we get

$$\hat{\theta}_{MAP} = \frac{1}{\alpha + y}$$

MAP and MMSE example

- Check if this is maximum

$$\frac{\partial^2}{\partial\theta}[log\ f(y,\theta)f(\theta)] = -\frac{1}{\theta^2}\ < 0$$

- So $f(\theta|y)$ has a unique maximum at

$$\hat{\theta}_{MAP} = \frac{1}{\alpha + y}$$