Estimation Tasks

Short Course on
Image Quality
Matthew A. Kupinski

Introduction

- Section 13.3 in B&M
- Keep in mind the similarities between estimation and classification
- Image-quality is a statistical concept

What we Will Cover

- Bias, variance, MSE, EMSE
- Bayesian estimation
- Maximum-likelihood estimation
- Fisher information
- Linear estimators

Basic Concept

$$\boldsymbol{g} = \mathcal{H}\boldsymbol{f} + \boldsymbol{n}$$

- Classification

  $$\boldsymbol{g} \to D_i$$

  Compare to H_i

- Estimation

  $$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{g})$$

  Compare to $\boldsymbol{\theta}$

Some PDFs

Random quantities
$$\boldsymbol{g} = \mathcal{H}\boldsymbol{f} + \boldsymbol{n}$$
$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{g})$$

$pr(\boldsymbol{\theta})$

$pr(\boldsymbol{f}|\boldsymbol{\theta})$

$pr(\boldsymbol{g}|\boldsymbol{f})$

$$pr(\boldsymbol{g}|\boldsymbol{\theta}) = \int d\boldsymbol{f}\, pr(\boldsymbol{g}|\boldsymbol{f}) pr(\boldsymbol{f}|\boldsymbol{\theta})$$

$pr(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})$

Example

Estimate tumor volume
$$\boldsymbol{g} = \mathcal{H}\boldsymbol{f} + \boldsymbol{n}$$
$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{g})$$

$pr(\boldsymbol{\theta})$ Distribution of volumes

$pr(\boldsymbol{f}|\boldsymbol{\theta})$ Object distribution given a volume

$pr(\boldsymbol{g}|\boldsymbol{f})$ Noise

$$pr(\boldsymbol{g}|\boldsymbol{\theta}) = \int d\boldsymbol{f}\, pr(\boldsymbol{g}|\boldsymbol{f}) pr(\boldsymbol{f}|\boldsymbol{\theta})$$

Data distribution given the true value of the parameter

$pr(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})$ Distribution of estimate given truth

Example

Detect a tumor

$$\boldsymbol{g} = \mathcal{H}\boldsymbol{f} + \boldsymbol{n}$$
$$T(\boldsymbol{g}) = D_i$$

$pr(H_i)$ Probability of truth states

$pr(\boldsymbol{f}|H_i)$ Object distribution given truth

$pr(\boldsymbol{g}|\boldsymbol{f})$ Noise

$$pr(\boldsymbol{g}|H_i) = \int d\boldsymbol{f}\, pr(\boldsymbol{g}|f)pr(f|H_i)$$

Data distribution given the truth state

$pr(D_i|H_j)$ Sensitivity or specificity

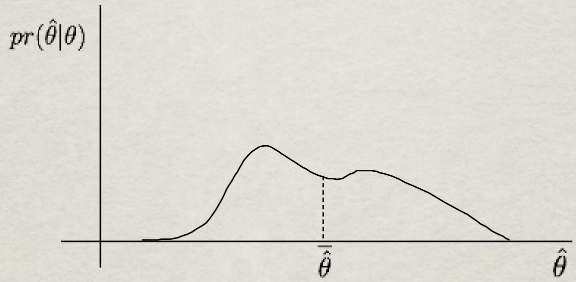
Bias

- Given a fixed value for the parameter we wish to estimate θ

$$\overline{\hat{\theta}}(\theta) = \left\langle \hat{\theta}(\boldsymbol{g}) \right\rangle_{\boldsymbol{g}|\theta} = \int d\boldsymbol{g}\, pr(\boldsymbol{g}|\theta)\hat{\theta}(\boldsymbol{g}) \qquad (13.274)$$

- Compare $\overline{\hat{\theta}}$ and θ

Bias



$$b(\theta) = \bar{\hat{\theta}} - \theta \qquad (13.276)$$

- If $b(\theta) = 0$ for all θ then $\bar{\hat{\theta}}(\boldsymbol{g})$ is unbiased

Variance

- Given a fixed value for the parameter we wish to estimate θ

$$\mathrm{Var}(\theta) = \sigma_{\hat{\theta}}^2 = \left\langle |\hat{\theta}(\boldsymbol{g}) - \bar{\hat{\theta}}|^2 \right\rangle_{\boldsymbol{g}|\theta} \qquad (13.279)$$

- Would like to have a small variance

Mean-Square Error

- Given a fixed value for the parameter we wish to estimate θ

$$\mathrm{MSE}(\theta) = \left\langle |\hat{\theta} - \theta|^2 \right\rangle_{\boldsymbol{g}|\theta} \qquad (13.280)$$

$$\mathrm{MSE}(\theta) = b(\theta)^2 + \mathrm{Var}(\theta)$$

- Would like to have a small MSE

Ensemble Forms

$$\overline{\overline{\hat{\theta}}} = \left\langle \overline{\hat{\theta}}(\theta) \right\rangle_{\theta} = \int d\theta pr(\theta) \int d\boldsymbol{g}\, pr(\boldsymbol{g}|\theta)\hat{\theta}(\boldsymbol{g}) \qquad (13.277)$$

$$\overline{\theta} = \langle \theta \rangle_{\theta}$$

- Ensemble bias

$$\overline{b} = \langle b(\theta) \rangle_{\theta} = \overline{\overline{\hat{\theta}}} - \overline{\theta} \qquad (13.277)$$

- Ensemble variance

$$\overline{\mathrm{Var}} = \langle \mathrm{Var}(\theta) \rangle_{\theta}$$

- Ensemble MSE (EMSE)

$$\mathrm{EMSE} = \left\langle \left\langle |\hat{\theta} - \theta|^2 \right\rangle_{\boldsymbol{g}|\theta} \right\rangle_{\theta} \qquad (13.281)$$

Vector Forms

- Bias

$$\boldsymbol{b}(\boldsymbol{\theta}) = \overline{\hat{\boldsymbol{\theta}}} - \boldsymbol{\theta} \qquad (13.283)$$

- Variance

$$\boldsymbol{K}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}) = \left\langle \left(\hat{\boldsymbol{\theta}} - \overline{\hat{\boldsymbol{\theta}}}\right) \left(\hat{\boldsymbol{\theta}} - \overline{\hat{\boldsymbol{\theta}}}\right)^{\dagger} \right\rangle_{\boldsymbol{g}|\boldsymbol{\theta}} \qquad (13.285)$$

- MSE

$$\mathrm{MSE}(\boldsymbol{\theta}) = \mathrm{tr}\left[\boldsymbol{K}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta})\right] + \mathrm{tr}\left[\boldsymbol{b}(\boldsymbol{\theta})\boldsymbol{b}(\boldsymbol{\theta})^{\dagger}\right] \qquad (13.287)$$

- EMSE

$$\mathrm{EMSE} = \mathrm{tr}\left[\overline{\boldsymbol{K}}_{\hat{\boldsymbol{\theta}}}\right] + \mathrm{tr}\left[\overline{\boldsymbol{b}}(\boldsymbol{\theta})\overline{\boldsymbol{b}}(\boldsymbol{\theta})^{\dagger}\right] \qquad (13.288)$$

Bayesian Estimation

- Determination of $\boldsymbol{\theta}$ though minimization of Bayes risk
- Knowledge of $pr(\boldsymbol{\theta})$ is assumed
- Cost function must be specified $C(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$
- Choose $\hat{\boldsymbol{\theta}}$ that minimizes the average cost $\left\langle C(\hat{\boldsymbol{\theta}}(\boldsymbol{g}), \boldsymbol{\theta}) \right\rangle_{\boldsymbol{g},\boldsymbol{\theta}}$

MMSE

- Let us choose

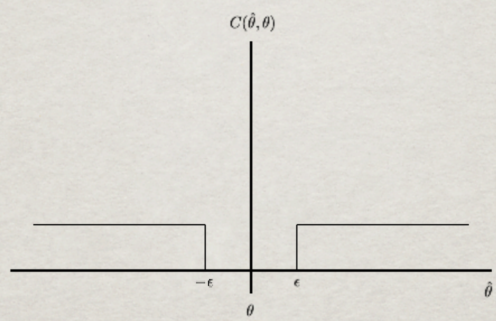
$$C(\hat{\theta}, \theta) = \text{EMSE} = \left\langle \left(\hat{\theta} - \theta\right)^2 \right\rangle_{g,\theta} \qquad (13.309)$$

- Estimator is given by the posterior mean

$$\hat{\theta}_{\text{MMSE}} = \int d\theta\, \theta\, pr(\theta|\boldsymbol{g}) = \langle \theta \rangle_{\theta|\boldsymbol{g}} \qquad (13.311b)$$

$$= \int d\theta\, \theta \left[\frac{pr(\boldsymbol{g}|\theta) pr(\theta)}{pr(\boldsymbol{g})} \right] \qquad (13.312)$$

MAP Estimation



$$C(\hat{\theta}, \theta) = C(\hat{\theta} - \theta) = 1 - \text{rect}\left(\frac{\hat{\theta} - \theta}{2\epsilon}\right) \qquad (13.322)$$

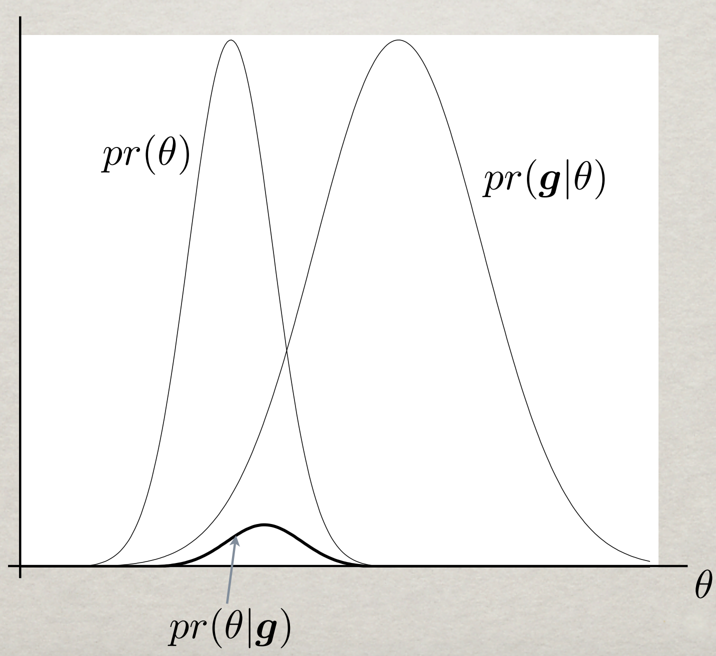
- All wrong answers beyond ϵ are equally costly

MAP Estimation

$$
\begin{aligned}
\hat{\theta}_{\mathrm{MAP}} &= \arg\max_{\theta}\ pr(\boldsymbol{g}|\theta)pr(\theta) \qquad (13.325)\\
&= \arg\max_{\theta}\ pr(\theta|\boldsymbol{g}) \qquad (13.327)\\
&= \arg\max_{\theta}\ \{\ln[pr(\boldsymbol{g}|\theta)] + \ln[pr(\theta)]\} \qquad (13.328)
\end{aligned}
$$

- MAP estimation is a type of Bayesian estimation!
- $\hat{\theta}_{\mathrm{MAP}}$ and $\hat{\theta}_{\mathrm{MMSE}}$ are equivalent under certain constraints (eqn. 13.316)

MAP Estimation



Maximum Likelihood

- ML estimation does not require us to know or assert $pr(\theta)$
- ML estimation is purely data driven, i.e., focus is on $\boldsymbol{g}$
- As we will see, ML estimator have many desirable properties

Maximum Likelihood

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} \; pr(\boldsymbol{g}|\theta) \qquad (13.348)$$

$$= \arg\max_{\theta} \; \ln\left[pr(\boldsymbol{g}|\theta)\right] \qquad (13.349)$$

- Limit of $\hat{\theta}_{\mathrm{MAP}}$ when the prior $pr(\theta)$ is sufficiently broad that it can be ignored

Fisher Information

- The score is given by

$$\boldsymbol{s} = \frac{\partial}{\partial \boldsymbol{\theta}} \ln \left[pr(\boldsymbol{g}|\boldsymbol{\theta}) \right] \qquad (13.358)$$

$$\langle \boldsymbol{s} \rangle_{\boldsymbol{g}|\boldsymbol{\theta}} = 0 \qquad (13.359)$$

- The Fisher information matrix

$$\boldsymbol{F}(\boldsymbol{\theta}) = \left\langle \boldsymbol{s}\boldsymbol{s}^{\dagger} \right\rangle_{\boldsymbol{g}|\boldsymbol{\theta}} \qquad (13.360)$$

- The Fisher information matrix measures the curvature of the average log-likelihood

Cramér-Rao Bound

$$\mathrm{Var}\left[\hat{\theta}_n\right] \geq \left[\boldsymbol{F}^{-1}\right]_{nn} \qquad (13.371)$$

- Fisher information provides us with a lower bound on all **unbiased** estimators of θ_n

- An unbiased estimator that meets this lower bound is known as an **efficient** estimator

Properties of ML Estimators

- Invariance
  $\boldsymbol{\Theta} = \tau(\boldsymbol{\theta})$
  $\hat{\boldsymbol{\Theta}}_{\mathrm{ML}} = \tau(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})$

- Efficiency
  If an efficient estimator exists $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is it

- Asymptotic properties
  ML estimators are asymptotically unbiased, efficient, and normal

Bayesian and ML Limitations

- We need
  $pr(\boldsymbol{g}|\boldsymbol{\theta})$
  $pr(\boldsymbol{\theta})$
  or both

- Consider the case of varying backgrounds

$$pr(\boldsymbol{g}|\boldsymbol{\theta}) = \int d\boldsymbol{f}\, pr(\boldsymbol{g}|\boldsymbol{f})pr(\boldsymbol{f}|\boldsymbol{\theta})$$

- How can we possibly hope to know this?!

Before we Continue...

- Random variable x

$$\sigma_x^2 = \left\langle (x - \overline{x})^2 \right\rangle_x$$

- Random vector $\boldsymbol{x}$

$$\boldsymbol{K_x} = \left\langle (\boldsymbol{x} - \overline{\boldsymbol{x}})(\boldsymbol{x} - \overline{\boldsymbol{x}})^\dagger \right\rangle_{\boldsymbol{x}} \qquad [\boldsymbol{K_x}]_{i,j} = \langle (x_i - \overline{x}_i)(x_j - \overline{x}_j) \rangle_{x_i, x_j}$$

- Two random vectors $\boldsymbol{x}$ $\boldsymbol{y}$

$$K_{\boldsymbol{x},\boldsymbol{y}} = \left\langle (\boldsymbol{x} - \overline{\boldsymbol{x}})(\boldsymbol{y} - \overline{\boldsymbol{y}})^\dagger \right\rangle_{\boldsymbol{x},\boldsymbol{y}}$$

$$[K_{\boldsymbol{x},\boldsymbol{y}}]_{i,j} = \left\langle (x_i - \overline{x}_i)(y_j - \overline{y}_j) \right\rangle_{x_i, y_j}$$

Linear Estimators

- Consider a linear estimator of the form

$$\hat{\boldsymbol{\theta}} = \boldsymbol{W}^\dagger \boldsymbol{g} + \boldsymbol{c}$$

- What is the matrix $\boldsymbol{W}$ that minimizes the EMSE?

$$\text{EMSE} = \left\langle \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^\dagger \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \right\rangle_{\boldsymbol{g},\boldsymbol{\theta}} = \left\langle \left(\boldsymbol{W}^\dagger \boldsymbol{g} - \boldsymbol{\theta}\right)^\dagger \left(\boldsymbol{W}^\dagger \boldsymbol{g} - \boldsymbol{\theta}\right) \right\rangle_{\boldsymbol{g},\boldsymbol{\theta}} \qquad (13.385)$$

- The Wiener estimator is given by

$$\hat{\boldsymbol{\theta}}_{\text{WE}} = \overline{\boldsymbol{\theta}} + \boldsymbol{W}^\dagger \left[\boldsymbol{g} - \overline{\boldsymbol{g}}\right]$$

$$\boldsymbol{W}^\dagger = \boldsymbol{K}_{\boldsymbol{\theta},\overline{\boldsymbol{g}}} \boldsymbol{K}_{\boldsymbol{g}}^{-1} \qquad (13.391)$$

Wiener Estimator

$$\boldsymbol{W}^{\dagger} = \boldsymbol{K}_{\boldsymbol{\theta},\overline{\boldsymbol{g}}}\boldsymbol{K}_{\boldsymbol{g}}^{-1} \qquad (13.391)$$

- $\boldsymbol{K}_{\boldsymbol{g}}$ is the covariance of the data averaged over all sources of randomness
- $\boldsymbol{K}_{\overline{\boldsymbol{g}},\boldsymbol{\theta}}$ is the cross-covariance of $\boldsymbol{\theta}$ and $\overline{\boldsymbol{g}} = \langle \boldsymbol{g} \rangle_{\boldsymbol{g}|\boldsymbol{\theta}}$

Wiener Estimator

- The WE is the linear observer that minimizes the EMSE
- In the case of jointly Gaussian data, the WE minimizes the EMSE of all estimators

Similarities with Detection

- Bayesian estimation
  - Minimize Bayes risk
  - Requires complete knowledge of risks, priors, and data densities

Similarities with Detection

$$\hat{\theta}_{\mathrm{MMSE}} = \int d\boldsymbol{\theta}\, \boldsymbol{\theta} \left[\frac{pr(\boldsymbol{g}|\boldsymbol{\theta})pr(\boldsymbol{\theta})}{pr(\boldsymbol{g})} \right] \qquad (13.312)$$

$$\Lambda(\boldsymbol{g}) \begin{array}{c} D_2 \\ > \\ < \\ D_1 \end{array} \frac{(C_{12} - C_{22})Pr(H_2)}{(C_{21} - C_{11})Pr(H_1)} \qquad (13.58)$$

Similarities with Detection

- MAP Estimation
  - Applies equal cost to all wrong decisions and 0 cost to correct decisions
  - Requires complete knowledge of prior and data densities

Similarities with Detection

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} \; pr(\boldsymbol{g}|\boldsymbol{\theta})pr(\boldsymbol{\theta}) \qquad (13.325)$$

$$\Lambda(\boldsymbol{g}) \underset{D_1}{\overset{D_2}{\gtrless}} \frac{Pr(H_2)}{Pr(H_1)} \qquad (13.61)$$

Similarities with Detection

- ML Estimation
  - Assumes a flat (or wide) prior distribution
  - Requires complete knowledge data densities

Similarities with Detection

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \; pr(\boldsymbol{g}|\boldsymbol{\theta}) \qquad (13.348)$$

$$\Lambda(\boldsymbol{g}) \underset{D_1}{\overset{D_2}{\gtrless}} 1 \qquad (13.71)$$

Similarities with Detection

- Ideal linear observers
  - Limited to linear manipulations of the data
  - Requires only first- and second-order statistics

Similarities with Detection

$$\hat{\boldsymbol{\theta}}_{\mathrm{WE}} = \overline{\boldsymbol{\theta}} + \boldsymbol{W}^{\dagger} \left[\boldsymbol{g} - \overline{\overline{\boldsymbol{g}}}\right]$$

$$\boldsymbol{W}^{\dagger} = \boldsymbol{K}_{\boldsymbol{\theta},\overline{\boldsymbol{g}}} \boldsymbol{K}_{\boldsymbol{g}}^{-1} \qquad (13.391)$$

$$\Delta\overline{\boldsymbol{g}}^{\dagger} \boldsymbol{K}^{-1} \boldsymbol{g} \underset{D_1}{\overset{D_2}{\gtrless}} t_c$$

To Reiterate

- Estimation and classification are very similar
- Estimation is a statistical concept and one should consider all sources of variation when dealing with estimation tasks

Final Thoughts

- Possible estimation task figures of merit
  - EMSE
  - Bayes risk
  - Fisher information matrix
- Computational methods will be discussed in a later talk