Information Theory and Image/Video Coding

Ming Jiang

School of Mathematical Sciences
Peking University

ming-jiang@pku.edu.cn

March 12, 2012

Outline

Bayes' Theorem

- For events A and B, provided $\mathbf{Pr}(B) \neq 0$,

$$\mathbf{Pr}(A|B) = \frac{\mathbf{Pr}(A)\mathbf{Pr}(B|A)}{\mathbf{Pr}(B)}. \tag{1}$$

- For two continuous random variables, Bayes' theorem is stated with the density functions.

Bayes' Theorem at wikipedia

Bayes' Theorem

- For events A and B, provided $\mathbf{Pr}(B) \neq 0$,

$$\mathbf{Pr}(A|B) = \frac{\mathbf{Pr}(A)\mathbf{Pr}(B|A)}{\mathbf{Pr}(B)}. \tag{1}$$

- For two continuous random variables, Bayes' theorem is stated with the density functions.

Bayes' Theorem at wikipedia

Bayes' Theorem: derivation

- Bayes' theorem may be derived from the definition of conditional probability

$$\mathbf{Pr}(A|B) = \frac{\mathbf{Pr}(A \cap B)}{\mathbf{Pr}(B)} \qquad \mathbf{Pr}(B) \neq 0; \qquad (2)$$

$$\mathbf{Pr}(B|A) = \frac{\mathbf{Pr}(A \cap B)}{\mathbf{Pr}(A)} \qquad \mathbf{Pr}(A) \neq 0. \qquad (3)$$

-

$$\mathbf{Pr}(A \cap B) = \mathbf{Pr}(A|B)\mathbf{Pr}(B) = \mathbf{Pr}(B|A)\mathbf{Pr}(A). \qquad (4)$$

-

$$\mathbf{Pr}(A|B) = \frac{\mathbf{Pr}(A)\mathbf{Pr}(B|A)}{\mathbf{Pr}(B)}. \qquad (5)$$

Bayes' Theorem at wikipedia

Bayes' Theorem: derivation

- Bayes' theorem may be derived from the definition of conditional probability

$$\mathbf{Pr}(A|B) = \frac{\mathbf{Pr}(A \cap B)}{\mathbf{Pr}(B)} \qquad \mathbf{Pr}(B) \neq 0; \tag{2}$$

$$\mathbf{Pr}(B|A) = \frac{\mathbf{Pr}(A \cap B)}{\mathbf{Pr}(A)} \qquad \mathbf{Pr}(A) \neq 0. \tag{3}$$

- $$\mathbf{Pr}(A \cap B) = \mathbf{Pr}(A|B)\mathbf{Pr}(B) = \mathbf{Pr}(B|A)\mathbf{Pr}(A). \tag{4}$$

- $$\mathbf{Pr}(A|B) = \frac{\mathbf{Pr}(A)\mathbf{Pr}(B|A)}{\mathbf{Pr}(B)}. \tag{5}$$

Bayes' Theorem at wikipedia

Bayes' Theorem: derivation

- Bayes' theorem may be derived from the definition of conditional probability

$$\mathbf{Pr}(A|B) = \frac{\mathbf{Pr}(A \cap B)}{\mathbf{Pr}(B)} \qquad \mathbf{Pr}(B) \neq 0; \qquad (2)$$

$$\mathbf{Pr}(B|A) = \frac{\mathbf{Pr}(A \cap B)}{\mathbf{Pr}(A)} \qquad \mathbf{Pr}(A) \neq 0. \qquad (3)$$

- $$\mathbf{Pr}(A \cap B) = \mathbf{Pr}(A|B)\mathbf{Pr}(B) = \mathbf{Pr}(B|A)\mathbf{Pr}(A). \qquad (4)$$

- $$\mathbf{Pr}(A|B) = \frac{\mathbf{Pr}(A)\mathbf{Pr}(B|A)}{\mathbf{Pr}(B)}. \qquad (5)$$

Bayes' Theorem at wikipedia

Bayes' Formula: extended form

- Let $A_1, \cdots, A_n$ be a partition of the event space, i.e., independent events with positive probabilities $\mathbf{Pr}(A_i) > 0$ and that $\cup_{i=1}^{n} A_i$ is the whole event space.
- By the law of total probability,

$$\mathbf{Pr}(B) = \sum_{i=1}^{n} \mathbf{Pr}(B|A_i)\mathbf{Pr}(A_i). \tag{6}$$

- For $1 \leq j \leq n$,

$$\mathbf{Pr}(A_j|B) = \frac{\mathbf{Pr}(A_j)\mathbf{Pr}(B|A_j)}{\sum_{i=1}^{n} \mathbf{Pr}(A_i)\mathbf{Pr}(B|A_i)}. \tag{7}$$

Bayes' Theorem at wikipedia

Bayes' Formula: extended form

- Let $A_1, \cdots, A_n$ be a partition of the event space, i.e., independent events with positive probabilities $\mathbf{Pr}(A_i) > 0$ and that $\cup_{i=1}^{n} A_i$ is the whole event space.
- By the law of total probability,

$$\mathbf{Pr}(B) = \sum_{i=1}^{n} \mathbf{Pr}(B|A_i)\mathbf{Pr}(A_i). \tag{6}$$

- For $1 \leq j \leq n$,

$$\mathbf{Pr}(A_j|B) = \frac{\mathbf{Pr}(A_j)\mathbf{Pr}(B|A_j)}{\sum_{i=1}^{n} \mathbf{Pr}(A_i)\mathbf{Pr}(B|A_i)}. \tag{7}$$

Bayes' Theorem at wikipedia

Bayes' Formula: extended form

- Let $A_1, \cdots, A_n$ be a partition of the event space, i.e., independent events with positive probabilities $\mathbf{Pr}(A_i) > 0$ and that $\cup_{i=1}^{n} A_i$ is the whole event space.
- By the law of total probability,

$$\mathbf{Pr}(B) = \sum_{i=1}^{n} \mathbf{Pr}(B|A_i)\mathbf{Pr}(A_i). \tag{6}$$

- For $1 \leq j \leq n$,

$$\mathbf{Pr}(A_j|B) = \frac{\mathbf{Pr}(A_j)\mathbf{Pr}(B|A_j)}{\sum_{i=1}^{n} \mathbf{Pr}(A_i)\mathbf{Pr}(B|A_i)}. \tag{7}$$

Bayes' Theorem at wikipedia

Outline

Bayesian Interpretation

- Events are equivalent to propositions.
- Probability measures a degree of belief.
- Bayes' theorem then links the degree of belief in a proposition before and after accounting for evidence.
- For proposition A and evidence B,
  - $\mathbf{Pr}(A)$, the `prior`, is the initial degree of belief in A before B is observed.
  - $\mathbf{Pr}(A|B)$, the `posterior`, is the degree of belief in A after B is observed.
  - $\frac{\mathbf{Pr}(B|A)}{\mathbf{Pr}(B)}$ is a factor representing the impact of B on the degree of belief in A.
  - The numerator $\mathbf{Pr}(B|A)$ is called the `likelihood`.

Bayes' Theorem at wikipedia

Bayesian Inference at wikipedia

Bayesian Interpretation

- Events are equivalent to propositions.
- Probability measures a degree of belief.
- Bayes' theorem then links the degree of belief in a proposition before and after accounting for evidence.
- For proposition A and evidence B,
  - $\mathbf{Pr}(A)$, the `prior`, is the initial degree of belief in A before B is observed.
  - $\mathbf{Pr}(A|B)$, the `posterior`, is the degree of belief in A after B is observed.
  - $\dfrac{\mathbf{Pr}(B|A)}{\mathbf{Pr}(B)}$ is a factor representing the impact of B on the degree of belief in A.
  - The numerator $\mathbf{Pr}(B|A)$ is called the `likelihood`.

Bayes' Theorem at wikipedia

Bayesian Inference at wikipedia

Bayesian Interpretation

- Events are equivalent to propositions.
- Probability measures a degree of belief.
- Bayes' theorem then links the degree of belief in a proposition before and after accounting for evidence.
- For proposition A and evidence B,
  - $\mathbf{Pr}(A)$, the `prior`, is the initial degree of belief in A before B is observed.
  - $\mathbf{Pr}(A|B)$, the `posterior`, is the degree of belief in A after B is observed.
  - $\frac{\mathbf{Pr}(B|A)}{\mathbf{Pr}(B)}$ is a factor representing the impact of B on the degree of belief in A.
  - The numerator $\mathbf{Pr}(B|A)$ is called the `likelihood`.

Bayes' Theorem at wikipedia

Bayesian Inference at wikipedia

Bayesian Interpretation

- Events are equivalent to propositions.
- Probability measures a degree of belief.
- Bayes' theorem then links the degree of belief in a proposition before and after accounting for evidence.
- For proposition A and evidence B,
  - $\mathbf{Pr}(A)$, the `prior`, is the initial degree of belief in A before B is observed.
  - $\mathbf{Pr}(A|B)$, the `posterior`, is the degree of belief in A after B is observed.
  - $\frac{\mathbf{Pr}(B|A)}{\mathbf{Pr}(B)}$ is a factor representing the impact of B on the degree of belief in A.
  - The numerator $\mathbf{Pr}(B|A)$ is called the `likelihood`.

Bayes' Theorem at wikipedia

Bayesian Inference at wikipedia

Bayesian Interpretation

- Events are equivalent to propositions.
- Probability measures a degree of belief.
- Bayes' theorem then links the degree of belief in a proposition before and after accounting for evidence.
- For proposition A and evidence B,
  - $\mathbf{Pr}(A)$, the `prior`, is the initial degree of belief in A before B is observed.
  - $\mathbf{Pr}(A|B)$, the `posterior`, is the degree of belief in A after B is observed.
  - $\frac{\mathbf{Pr}(B|A)}{\mathbf{Pr}(B)}$ is a factor representing the impact of B on the degree of belief in A.
  - The numerator $\mathbf{Pr}(B|A)$ is called the `likelihood`.

Bayes' Theorem at wikipedia

Bayesian Inference at wikipedia

Bayesian Interpretation

- Events are equivalent to propositions.
- Probability measures a degree of belief.
- Bayes' theorem then links the degree of belief in a proposition before and after accounting for evidence.
- For proposition A and evidence B,
  - $\mathbf{Pr}(A)$, the `prior`, is the initial degree of belief in A before B is observed.
  - $\mathbf{Pr}(A|B)$, the `posterior`, is the degree of belief in A after B is observed.
  - $\frac{\mathbf{Pr}(B|A)}{\mathbf{Pr}(B)}$ is a factor representing the impact of B on the degree of belief in A.
  - The numerator $\mathbf{Pr}(B|A)$ is called the `likelihood`.

Bayes' Theorem at wikipedia

Bayesian Inference at wikipedia

Bayesian Interpretation

- Events are equivalent to propositions.
- Probability measures a degree of belief.
- Bayes' theorem then links the degree of belief in a proposition before and after accounting for evidence.
- For proposition A and evidence B,
  - $\mathbf{Pr}(A)$, the `prior`, is the initial degree of belief in A before B is observed.
  - $\mathbf{Pr}(A|B)$, the `posterior`, is the degree of belief in A after B is observed.
  - $\frac{\mathbf{Pr}(B|A)}{\mathbf{Pr}(B)}$ is a factor representing the impact of B on the degree of belief in A.
  - The numerator $\mathbf{Pr}(B|A)$ is called the `likelihood`.

Bayes' Theorem at wikipedia

Bayesian Inference at wikipedia

Bayesian Interpretation

- Events are equivalent to propositions.
- Probability measures a degree of belief.
- Bayes' theorem then links the degree of belief in a proposition before and after accounting for evidence.
- For proposition A and evidence B,
  - $\mathbf{Pr}(A)$, the `prior`, is the initial degree of belief in A before B is observed.
  - $\mathbf{Pr}(A|B)$, the `posterior`, is the degree of belief in A after B is observed.
  - $\frac{\mathbf{Pr}(B|A)}{\mathbf{Pr}(B)}$ is a factor representing the impact of B on the degree of belief in A.
  - The numerator $\mathbf{Pr}(B|A)$ is called the `likelihood`.

Bayes' Theorem at wikipedia

Bayesian Inference at wikipedia

Bayesian Inference

- In statistics, `Bayesian inference` is a method of statistical inference in which Bayes' theorem is used to calculate how the degree of belief in a proposition changes due to evidence.
- Bayes' theorem provides the rational update given the evidence.
- The initial degree of belief is called the `prior` and the updated degree of belief the `posterior`.
- Bayesian inference has applications in science, engineering, medicine and law.
- Research has suggested that the brain may employ Bayesian inference.

Bayesian Inference at wikipedia

Bayesian Inference

- In statistics, `Bayesian inference` is a method of statistical inference in which Bayes' theorem is used to calculate how the degree of belief in a proposition changes due to evidence.
- Bayes' theorem provides the rational update given the evidence.
- The initial degree of belief is called the `prior` and the updated degree of belief the `posterior`.
- Bayesian inference has applications in science, engineering, medicine and law.
- Research has suggested that the brain may employ Bayesian inference.

Bayesian Inference at wikipedia

Bayesian Inference

- In statistics, `Bayesian inference` is a method of statistical inference in which Bayes' theorem is used to calculate how the degree of belief in a proposition changes due to evidence.
- Bayes' theorem provides the rational update given the evidence.
- The initial degree of belief is called the `prior` and the updated degree of belief the `posterior`.
- Bayesian inference has applications in science, engineering, medicine and law.
- Research has suggested that the brain may employ Bayesian inference.

Bayesian Inference at wikipedia

Bayesian Inference

- In statistics, `Bayesian inference` is a method of statistical inference in which Bayes' theorem is used to calculate how the degree of belief in a proposition changes due to evidence.
- Bayes' theorem provides the rational update given the evidence.
- The initial degree of belief is called the `prior` and the updated degree of belief the `posterior`.
- Bayesian inference has applications in science, engineering, medicine and law.
- Research has suggested that the brain may employ Bayesian inference.

Bayesian Inference at wikipedia

Bayesian Inference

- In statistics, `Bayesian inference` is a method of statistical inference in which Bayes' theorem is used to calculate how the degree of belief in a proposition changes due to evidence.
- Bayes' theorem provides the rational update given the evidence.
- The initial degree of belief is called the `prior` and the updated degree of belief the `posterior`.
- Bayesian inference has applications in science, engineering, medicine and law.
- Research has suggested that the brain may employ Bayesian inference.

Bayesian Inference at wikipedia

Bayesian Inference in Data Processing

- A unknown quantity θ which is to be inferred is called the `state`.
- Θ denotes the set of all possible states and is called `state space`.
- Typically, experiments are performed to obtain information about θ.
- Experiments are designed so that the observations are distributed according to some probability distribution, which has θ as an unknown parameter.
- In such situations, θ is called the `parameter` and Θ the `parameter space`.
- References
  - i [Mumford, 1994].
  - ii [Berger, 1985].

Bayesian Inference in Data Processing

- A unknown quantity θ which is to be inferred is called the `state`.
- Θ denotes the set of all possible states and is called `state space`.

- Typically, experiments are performed to obtain information about θ.
- Experiments are designed so that the observations are distributed according to some probability distribution, which has θ as an unknown parameter.
- In such situations, θ is called the `parameter` and Θ the `parameter space`.

- References
  - i [Mumford, 1994].
  - ii [Berger, 1985].

Bayesian Inference in Data Processing

- A unknown quantity θ which is to be inferred is called the `state`.
- Θ denotes the set of all possible states and is called `state space`.
- Typically, experiments are performed to obtain information about θ.
- Experiments are designed so that the observations are distributed according to some probability distribution, which has θ as an unknown parameter.
- In such situations, θ is called the `parameter` and Θ the `parameter space`.
- References
  i [Mumford, 1994].
  ii [Berger, 1985].

Bayesian Inference in Data Processing

- A unknown quantity θ which is to be inferred is called the `state`.
- Θ denotes the set of all possible states and is called `state space`.
- Typically, experiments are performed to obtain information about θ.
- Experiments are designed so that the observations are distributed according to some probability distribution, which has θ as an unknown parameter.
- In such situations, θ is called the `parameter` and Θ the `parameter space`.
- References
  i [Mumford, 1994].
  ii [Berger, 1985].

Bayesian Inference in Data Processing

- A unknown quantity θ which is to be inferred is called the `state`.
- Θ denotes the set of all possible states and is called `state space`.
- Typically, experiments are performed to obtain information about θ.
- Experiments are designed so that the observations are distributed according to some probability distribution, which has θ as an unknown parameter.
- In such situations, θ is called the `parameter` and Θ the `parameter space`.
- References
  1. [Mumford, 1994].
  2. [Berger, 1985].

Bayesian Inference in Data Processing

- A unknown quantity θ which is to be inferred is called the `state`.
- Θ denotes the set of all possible states and is called `state space`.
- Typically, experiments are performed to obtain information about θ.
- Experiments are designed so that the observations are distributed according to some probability distribution, which has θ as an unknown parameter.
- In such situations, θ is called the `parameter` and Θ the `parameter space`.
- References
  i [Mumford, 1994].
  ii [Berger, 1985].

Bayesian Inference in Data Processing

- A unknown quantity θ which is to be inferred is called the `state`.
- Θ denotes the set of all possible states and is called `state space`.
- Typically, experiments are performed to obtain information about θ.
- Experiments are designed so that the observations are distributed according to some probability distribution, which has θ as an unknown parameter.
- In such situations, θ is called the `parameter` and Θ the `parameter space`.
- References
  - i [Mumford, 1994].
  - ii [Berger, 1985].

Bayesian Inference in Data Processing

- A unknown quantity θ which is to be inferred is called the `state`.
- Θ denotes the set of all possible states and is called `state space`.
- Typically, experiments are performed to obtain information about θ.
- Experiments are designed so that the observations are distributed according to some probability distribution, which has θ as an unknown parameter.
- In such situations, θ is called the `parameter` and Θ the `parameter space`.
- References
  i [Mumford, 1994].
  ii [Berger, 1985].

Bayesian Analysis

- When a statistical investigation is performed to obtain information about θ, the outcome (a random variable) will be denoted by X.
- Often X will be a vector, $X = (X_1, \cdots, X_n)$.
- A particular realization will be denoted by x.
- The set of possible outcomes is the `sample space`, and will be denoted by $\mathfrak{X}$.
- X is either a continuous or discrete random variable, with the conditional density $\mathbf{Pr}(x|\theta)$.
- $\mathbf{Pr}(x|\theta)$ is called the `data model`.
- References
  - i [Mumford, 1994].
  - ii [Berger, 1985].

Bayesian Analysis

- When a statistical investigation is performed to obtain information about θ, the outcome (a random variable) will be denoted by X.
- Often X will be a vector, $X = (X_1, \cdots, X_n)$.
- A particular realization will be denoted by x.
- The set of possible outcomes is the `sample space`, and will be denoted by $\mathfrak{X}$.
- X is either a continuous or discrete random variable, with the conditional density $\mathbf{Pr}(x|\theta)$.
- $\mathbf{Pr}(x|\theta)$ is called the `data model`.
- References
  - i [Mumford, 1994].
  - ii [Berger, 1985].

Bayesian Analysis

- When a statistical investigation is performed to obtain information about θ, the outcome (a random variable) will be denoted by X.
- Often X will be a vector, $X = (X_1, \cdots, X_n)$.
- A particular realization will be denoted by x.

- The set of possible outcomes is the `sample space`, and will be denoted by $\mathfrak{X}$.
- X is either a continuous or discrete random variable, with the conditional density $\mathbf{Pr}(x|\theta)$.
- $\mathbf{Pr}(x|\theta)$ is called the `data model`.

- References
  - i [Mumford, 1994].
  - ii [Berger, 1985].

Bayesian Analysis

- When a statistical investigation is performed to obtain information about θ, the outcome (a random variable) will be denoted by X.
- Often X will be a vector, $X = (X_1, \cdots, X_n)$.
- A particular realization will be denoted by x.
- The set of possible outcomes is the `sample space`, and will be denoted by $\mathfrak{X}$.
- X is either a continuous or discrete random variable, with the conditional density $\mathbf{Pr}(x|\theta)$.
- $\mathbf{Pr}(x|\theta)$ is called the `data model`.
- References
  i [Mumford, 1994].
  ii [Berger, 1985].

Bayesian Analysis

- When a statistical investigation is performed to obtain information about θ, the outcome (a random variable) will be denoted by X.
- Often X will be a vector, $X = (X_1, \cdots, X_n)$.
- A particular realization will be denoted by x.
- The set of possible outcomes is the `sample space`, and will be denoted by $\mathfrak{X}$.
- X is either a continuous or discrete random variable, with the conditional density $\mathbf{Pr}(x|\theta)$.
- $\mathbf{Pr}(x|\theta)$ is called the `data model`.
- References
  - i [Mumford, 1994].
  - ii [Berger, 1985].

Bayesian Analysis

- When a statistical investigation is performed to obtain information about θ, the outcome (a random variable) will be denoted by X.
- Often X will be a vector, $X = (X_1, \cdots, X_n)$.
- A particular realization will be denoted by x.
- The set of possible outcomes is the `sample space`, and will be denoted by $\mathfrak{X}$.
- X is either a continuous or discrete random variable, with the conditional density $\mathbf{Pr}(x|\theta)$.
- $\mathbf{Pr}(x|\theta)$ is called the `data model`.
- References
  - i [Mumford, 1994].
  - ii [Berger, 1985].

Bayesian Analysis

- When a statistical investigation is performed to obtain information about θ, the outcome (a random variable) will be denoted by X.
- Often X will be a vector, $X = (X_1, \cdots, X_n)$.
- A particular realization will be denoted by x.

- The set of possible outcomes is the `sample space`, and will be denoted by $\mathfrak{X}$.
- X is either a continuous or discrete random variable, with the conditional density $\mathbf{Pr}(x|\theta)$.
- $\mathbf{Pr}(x|\theta)$ is called the `data model`.

- References
  i [Mumford, 1994].
  ii [Berger, 1985].

Bayesian Analysis

- When a statistical investigation is performed to obtain information about θ, the outcome (a random variable) will be denoted by X.
- Often X will be a vector, $X = (X_1, \cdots, X_n)$.
- A particular realization will be denoted by x.
- The set of possible outcomes is the `sample space`, and will be denoted by $\mathfrak{X}$.
- X is either a continuous or discrete random variable, with the conditional density $\mathbf{Pr}(x|\theta)$.
- $\mathbf{Pr}(x|\theta)$ is called the `data model`.
- References
  - i [Mumford, 1994].
  - ii [Berger, 1985].

Bayesian Analysis

- When a statistical investigation is performed to obtain information about θ, the outcome (a random variable) will be denoted by X.
- Often X will be a vector, $X = (X_1, \cdots, X_n)$.
- A particular realization will be denoted by x.
- The set of possible outcomes is the `sample space`, and will be denoted by $\mathfrak{X}$.
- X is either a continuous or discrete random variable, with the conditional density $\mathbf{Pr}(x|\theta)$.
- $\mathbf{Pr}(x|\theta)$ is called the `data model`.
- References
  - i [Mumford, 1994].
  - ii [Berger, 1985].

Bayesian Analysis

- Prior information about θ is seldom very precise.
- The symbol $\mathbf{Pr}(\theta)$ will be used to represent a prior density of θ.
- The `posterior` distribution of θ given x is $\mathbf{Pr}(\theta|x)$, the conditional distribution of θ given the sample observation x.
- Bayesian analysis is conducted by combing the prior information and the sample information into the `posterior` distribution, from which all decision and inference are made.
- References
  i [Mumford, 1994].
  ii [Berger, 1985].

Bayesian Analysis

- Prior information about θ is seldom very precise.
- The symbol $\mathbf{Pr}(\theta)$ will be used to represent a prior density of θ.
- The posterior distribution of θ given x is $\mathbf{Pr}(\theta|x)$, the conditional distribution of θ given the sample observation x.
- Bayesian analysis is conducted by combing the prior information and the sample information into the posterior distribution, from which all decision and inference are made.
- References
  1. [Mumford, 1994].
  2. [Berger, 1985].

Bayesian Analysis

- Prior information about θ is seldom very precise.
- The symbol $\mathbf{Pr}(\theta)$ will be used to represent a prior density of θ.
- The `posterior` distribution of θ given x is $\mathbf{Pr}(\theta|x)$, the conditional distribution of θ given the sample observation x.

- Bayesian analysis is conducted by combing the prior information and the sample information into the `posterior` distribution, from which all decision and inference are made.

- References
  i [Mumford, 1994].
  ii [Berger, 1985].

Bayesian Analysis

- Prior information about θ is seldom very precise.
- The symbol **Pr**(θ) will be used to represent a prior density of θ.
- The `posterior` distribution of θ given x is **Pr**$(\theta|x)$, the conditional distribution of θ given the sample observation x.

- Bayesian analysis is conducted by combing the prior information and the sample information into the `posterior` distribution, from which all decision and inference are made.

- References
  i [Mumford, 1994].
  ii [Berger, 1985].

Bayesian Analysis

- Prior information about θ is seldom very precise.
- The symbol $\mathbf{Pr}(\theta)$ will be used to represent a prior density of θ.
- The `posterior` distribution of θ given x is $\mathbf{Pr}(\theta|x)$, the conditional distribution of θ given the sample observation x.
- Bayesian analysis is conducted by combing the prior information and the sample information into the `posterior` distribution, from which all decision and inference are made.
- References
  i [Mumford, 1994].
  ii [Berger, 1985].

Bayesian Analysis

- Prior information about θ is seldom very precise.
- The symbol $\mathbf{Pr}(\theta)$ will be used to represent a prior density of θ.
- The `posterior` distribution of θ given x is $\mathbf{Pr}(\theta|x)$, the conditional distribution of θ given the sample observation x.
- Bayesian analysis is conducted by combing the prior information and the sample information into the `posterior` distribution, from which all decision and inference are made.
- References
  - i [Mumford, 1994].
  - ii [Berger, 1985].

Bayesian Analysis

- Prior information about θ is seldom very precise.
- The symbol $\mathbf{Pr}(\theta)$ will be used to represent a prior density of θ.
- The `posterior` distribution of θ given x is $\mathbf{Pr}(\theta|x)$, the conditional distribution of θ given the sample observation x.
- Bayesian analysis is conducted by combing the prior information and the sample information into the `posterior` distribution, from which all decision and inference are made.
- References
  i [Mumford, 1994].
  ii [Berger, 1985].

Joint and Marginal Densities

- θ and X have joint density

$$h(x,\theta) = \mathbf{Pr}(\theta, x) = \mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta), \tag{8}$$

and that X has the (unconditional) marginal density

$$m(x) = \int_{\Theta} \mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta)\, d\theta. \tag{9}$$

- Providing $m(x) \neq 0$, by Bayes' Theorem,

$$\mathbf{Pr}(\theta|x) = \frac{h(x,\theta)}{m(x)} = \frac{\mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta)}{m(x)}. \tag{10}$$

- References
  - [Berger, 1985].

Joint and Marginal Densities

- θ and X have `joint density`

$$h(x,\theta) = \mathbf{Pr}(\theta, x) = \mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta), \tag{8}$$

and that X has the (unconditional) `marginal density`

$$m(x) = \int_{\Theta} \mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta)\, d\theta. \tag{9}$$

- Providing $m(x) \neq 0$, by Bayes' Theorem,

$$\mathbf{Pr}(\theta|x) = \frac{h(x,\theta)}{m(x)} = \frac{\mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta)}{m(x)}. \tag{10}$$

- References
  - [Berger, 1985].

Joint and Marginal Densities

- θ and X have `joint density`

$$h(x,\theta) = \mathbf{Pr}(\theta, x) = \mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta), \tag{8}$$

  and that X has the (unconditional) `marginal density`

$$m(x) = \int_{\Theta} \mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta)\, d\theta. \tag{9}$$

- Providing $m(x) \neq 0$, by Bayes' Theorem,

$$\mathbf{Pr}(\theta|x) = \frac{h(x,\theta)}{m(x)} = \frac{\mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta)}{m(x)}. \tag{10}$$

- References
  - [Berger, 1985].

Joint and Marginal Densities

- θ and X have `joint density`

$$h(x, \theta) = \mathbf{Pr}(\theta, x) = \mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta), \tag{8}$$

and that X has the (unconditional) `marginal density`

$$m(x) = \int_{\Theta} \mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta)\, d\theta. \tag{9}$$

- Providing $m(x) \neq 0$, by Bayes' Theorem,

$$\mathbf{Pr}(\theta|x) = \frac{h(x, \theta)}{m(x)} = \frac{\mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta)}{m(x)}. \tag{10}$$

- References
  - [Berger, 1985].

Bayesian Decision Rules: MAP

- To estimate θ, a number of classical techniques can be applied to the posterior distribution.

- **Definition**

  The `maximum a posterior` *(MAP) estimate of θ is the largest mode of $\mathbf{Pr}(\theta|x)$ (i.e., the value θ which maximizes $\mathbf{Pr}(\theta|x)$)*

$$\theta_{MAP} = \arg\max_{\theta} \mathbf{Pr}(\theta|x) = \arg\max_{\theta} \mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta). \quad (11)$$

- References
  - i [Berger, 1985].
  - ii [Wrinkler, 1995].

Bayesian Decision Rules: MAP

- To estimate θ, a number of classical techniques can be applied to the posterior distribution.

- **Definition**

  The maximum a posterior *(MAP) estimate of θ is the largest mode of $\Pr(\theta|x)$ (i.e., the value θ which maximizes $\Pr(\theta|x)$)*

$$\theta_{MAP} = \arg\max_{\theta} \Pr(\theta|x) = \arg\max_{\theta} \Pr(\theta)\Pr(x|\theta). \quad (11)$$

- References
  i [Berger, 1985].
  ii [Wrinkler, 1995].

Bayesian Decision Rules: MAP

- To estimate θ, a number of classical techniques can be applied to the posterior distribution.

- Definition

The maximum a posterior *(MAP) estimate of θ is the largest mode of $\mathbf{Pr}(\theta|x)$ (i.e., the value θ which maximizes $\mathbf{Pr}(\theta|x)$)*

$$\theta_{MAP} = \arg\max_{\theta} \mathbf{Pr}(\theta|x) = \arg\max_{\theta} \mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta). \quad (11)$$

- References

i [Berger, 1985].
ii [Wrinkler, 1995].

Bayesian Decision Rules: MAP

- To estimate θ, a number of classical techniques can be applied to the posterior distribution.

- Definition

The `maximum a posterior` *(MAP) estimate of θ is the largest mode of $\mathbf{Pr}(\theta|x)$ (i.e., the value θ which maximizes $\mathbf{Pr}(\theta|x)$)*

$$\theta_{MAP} = \arg\max_{\theta} \mathbf{Pr}(\theta|x) = \arg\max_{\theta} \mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta). \quad (11)$$

- References

i [Berger, 1985].
ii [Wrinkler, 1995].

Bayesian Decision Rules: MAP

- To estimate θ, a number of classical techniques can be applied to the posterior distribution.

- Definition

The `maximum a posterior` *(MAP) estimate of θ is the largest mode of $\mathbf{Pr}(\theta|x)$ (i.e., the value θ which maximizes $\mathbf{Pr}(\theta|x)$)*

$$\theta_{MAP} = \arg\max_{\theta} \mathbf{Pr}(\theta|x) = \arg\max_{\theta} \mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta). \quad (11)$$

- References
  i [Berger, 1985].
  ii [Wrinkler, 1995].

Bayesian Decision Rules: MAP

- To estimate θ, a number of classical techniques can be applied to the posterior distribution.

- Definition

The `maximum a posterior` *(MAP) estimate of θ is the largest mode of* $\mathbf{Pr}(\theta|x)$ *(i.e., the value θ which maximizes* $\mathbf{Pr}(\theta|x)$*)*

$$\theta_{MAP} = \arg\max_{\theta} \mathbf{Pr}(\theta|x) = \arg\max_{\theta} \mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta). \quad (11)$$

- References

  i [Berger, 1985].

  ii [Wrinkler, 1995].

Maximum Likelihood Principle

- The maximum likelihood (ML) estimate is the estimate of θ, which maximizes the likelihood function $\mathbf{Pr}(x|\theta)$:

$$\theta_{\text{ML}} = \arg\max_{\theta} \mathbf{Pr}(x|\theta). \quad (12)$$

- The maximum likelihood principle is implicitly assumed in the MAP, when there is no prior information about θ other than contained in $\mathbf{Pr}(x|\theta)$ (for the given x).

- References
  i [Berger, 1985].
  ii [Mumford, 1994].

Maximum Likelihood Principle

- The `maximum likelihood` (ML) estimate is the estimate of θ, which maximizes the likelihood function $\mathbf{Pr}(x|\theta)$:

$$\theta_{\text{ML}} = \arg\max_{\theta} \mathbf{Pr}(x|\theta). \tag{12}$$

- The maximum likelihood principle is implicitly assumed in the MAP, when there is no prior information about θ other than contained in $\mathbf{Pr}(x|\theta)$ (for the given x).

- References
  i [Berger, 1985].
  ii [Mumford, 1994].

Maximum Likelihood Principle

- The `maximum likelihood` (ML) estimate is the estimate of θ, which maximizes the likelihood function $\mathbf{Pr}(x|\theta)$:

$$\theta_{\mathrm{ML}} = \mathop{\mathbf{arg\,max}}_{\theta} \mathbf{Pr}(x|\theta). \tag{12}$$

- The maximum likelihood principle is implicitly assumed in the MAP, when there is no prior information about θ other than contained in $\mathbf{Pr}(x|\theta)$ (for the given x).

- References
  - i [Berger, 1985].
  - ii [Mumford, 1994].

Maximum Likelihood Principle

- The `maximum likelihood` (ML) estimate is the estimate of θ, which maximizes the likelihood function $\mathbf{Pr}(x|\theta)$:

$$\theta_{\mathrm{ML}} = \mathop{\mathbf{arg\,max}}_{\theta} \mathbf{Pr}(x|\theta). \tag{12}$$

- The maximum likelihood principle is implicitly assumed in the MAP, when there is no prior information about θ other than contained in $\mathbf{Pr}(x|\theta)$ (for the given x).

- References
  i [Berger, 1985].
  ii [Mumford, 1994].

Maximum Likelihood Principle

- The `maximum likelihood` (ML) estimate is the estimate of θ, which maximizes the likelihood function $\mathbf{Pr}(x|\theta)$:

$$\theta_{\mathrm{ML}} = \mathop{\mathbf{arg\,max}}_{\theta} \mathbf{Pr}(x|\theta). \tag{12}$$

- The maximum likelihood principle is implicitly assumed in the MAP, when there is no prior information about θ other than contained in $\mathbf{Pr}(x|\theta)$ (for the given x).

- References
  - i [Berger, 1985].
  - ii [Mumford, 1994].

Bayesian Decision Rules: MMSE

- Another reasonable estimate is the mean value of the posterior distribution.

Definition

The `minimum mean squares estimate` *(MMSE) of θ is the mean value of $\mathbf{Pr}(\theta|x)$:*

$$\theta_{MMSE} = \int_{\theta \in \Theta} \theta \mathbf{Pr}(\theta|x)\, d\theta = \int_{\theta \in \Theta} \theta \mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta)\, d\theta. \quad (13)$$

- References
  i [Berger, 1985].
  ii [Wrinkler, 1995].

Bayesian Decision Rules: MMSE

- Another reasonable estimate is the mean value of the posterior distribution.

Definition

The minimum mean squares estimate *(MMSE) of θ is the mean value of $\mathbf{Pr}(\theta|x)$:*

$$\theta_{MMSE} = \int_{\theta \in \Theta} \theta \mathbf{Pr}(\theta|x)\, d\theta = \int_{\theta \in \Theta} \theta \mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta)\, d\theta. \quad (13)$$

- References
  - i [Berger, 1985].
  - ii [Wrinkler, 1995].

Bayesian Decision Rules: MMSE

- Another reasonable estimate is the mean value of the posterior distribution.

Definition

The `minimum mean squares estimate` *(MMSE) of* θ *is the mean value of* $\mathbf{Pr}(\theta|x)$*:*

$$\theta_{MMSE} = \int_{\theta \in \Theta} \theta \mathbf{Pr}(\theta|x)\, d\theta = \int_{\theta \in \Theta} \theta \mathbf{Pr}(\theta) \mathbf{Pr}(x|\theta)\, d\theta. \quad (13)$$

- References
  i [Berger, 1985].
  ii [Wrinkler, 1995].

Bayesian Decision Rules: MMSE

- Another reasonable estimate is the mean value of the posterior distribution.

Definition

The `minimum mean squares estimate` *(MMSE) of* θ *is the mean value of* $\mathbf{Pr}(\theta|x)$*:*

$$\theta_{MMSE} = \int_{\theta\in\Theta} \theta\mathbf{Pr}(\theta|x)\, d\theta = \int_{\theta\in\Theta} \theta\mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta)\, d\theta. \quad (13)$$

- References
  i [Berger, 1985].
  ii [Wrinkler, 1995].

Bayesian Decision Rules: MMSE

- Another reasonable estimate is the mean value of the posterior distribution.

Definition

The `minimum mean squares estimate` *(MMSE) of* θ *is the mean value of* $\mathbf{Pr}(\theta|x)$*:*

$$\theta_{MMSE} = \int_{\theta\in\Theta} \theta \mathbf{Pr}(\theta|x)\, d\theta = \int_{\theta\in\Theta} \theta \mathbf{Pr}(\theta)\mathbf{Pr}(x|\theta)\, d\theta. \quad (13)$$

- References
  - i [Berger, 1985].
  - ii [Wrinkler, 1995].

Bayesian Risk

- The performance of estimators are studied in terms of loss functions.
- The loss between a true θ and its estimate $\hat{\theta}(x)$ is measured by a `loss function` such that $L : \Theta \times \Theta \to \mathbf{R}_+$ and
  - $L(\theta, \hat{\theta}) \geq 0$;
  - $L(\theta, \theta) = 0$.
- The `Bayesian risk` of the estimate is the mean loss

$$\hat{R} = \int_{\Theta \times \mathfrak{X}} L(\theta, \hat{\theta}(x)) \mathbf{Pr}(\theta, x). \tag{14}$$

- References
  - [Wrinkler, 1995].

Bayesian Risk

- The performance of estimators are studied in terms of loss functions.
- The loss between a true θ and its estimate $\hat{\theta}(x)$ is measured by a `loss function` such that $L : \Theta \times \Theta \to \mathbf{R}_+$ and
  - $L(\theta, \hat{\theta}) \geq 0$;
  - $L(\theta, \theta) = 0$.
- The `Bayesian risk` of the estimate is the mean loss

$$\hat{R} = \int_{\Theta \times \mathfrak{X}} L(\theta, \hat{\theta}(x)) \mathbf{Pr}(\theta, x). \tag{14}$$

- References
  - [Wrinkler, 1995].

Bayesian Risk

- The performance of estimators are studied in terms of loss functions.
- The loss between a true θ and its estimate $\hat{\theta}(x)$ is measured by a `loss function` such that $L : \Theta \times \Theta \to \mathbf{R}_+$ and
  - $L(\theta, \hat{\theta}) \geq 0$;
  - $L(\theta, \theta) = 0$.
- The `Bayesian risk` of the estimate is the mean loss

$$\hat{R} = \int_{\Theta \times \mathfrak{X}} L(\theta, \hat{\theta}(x)) \mathbf{Pr}(\theta, x). \tag{14}$$

- References
  - [Wrinkler, 1995].

Bayesian Risk

- The performance of estimators are studied in terms of loss functions.
- The loss between a true θ and its estimate $\hat{\theta}(x)$ is measured by a `loss function` such that $L : \Theta \times \Theta \to \mathbf{R}_+$ and
  - $L(\theta, \hat{\theta}) \geq 0$;
  - $L(\theta, \theta) = 0$.
- The `Bayesian risk` of the estimate is the mean loss

$$\hat{R} = \int_{\Theta \times \mathfrak{X}} L(\theta, \hat{\theta}(x)) \mathbf{Pr}(\theta, x). \tag{14}$$

- References
  - [Wrinkler, 1995].

Bayesian Risk

- The performance of estimators are studied in terms of loss functions.
- The loss between a true θ and its estimate $\hat{\theta}(x)$ is measured by a `loss function` such that $L : \Theta \times \Theta \rightarrow \mathbf{R}_+$ and
  - $L(\theta, \hat{\theta}) \geq 0$;
  - $L(\theta, \theta) = 0$.
- The `Bayesian risk` of the estimate is the mean loss

$$\hat{R} = \int_{\Theta \times \mathfrak{X}} L(\theta, \hat{\theta}(x)) \mathbf{Pr}(\theta, x). \tag{14}$$

- References
  - [Wrinkler, 1995].

Bayesian Risk

- The performance of estimators are studied in terms of loss functions.
- The loss between a true θ and its estimate $\hat{\theta}(x)$ is measured by a `loss function` such that $L : \Theta \times \Theta \to \mathbf{R}_+$ and
  - $L(\theta, \hat{\theta}) \geq 0$;
  - $L(\theta, \theta) = 0$.
- The `Bayesian risk` of the estimate is the mean loss

$$\hat{R} = \int_{\Theta \times \mathfrak{X}} L(\theta, \hat{\theta}(x)) \mathbf{Pr}(\theta, x). \quad (14)$$

- References
  - [Wrinkler, 1995].

Bayesian Risk

- The performance of estimators are studied in terms of loss functions.
- The loss between a true θ and its estimate $\hat{\theta}(x)$ is measured by a `loss function` such that $L : \Theta \times \Theta \to \mathbf{R}_+$ and
  - $L(\theta, \hat{\theta}) \geq 0$;
  - $L(\theta, \theta) = 0$.
- The `Bayesian risk` of the estimate is the mean loss

$$\hat{R} = \int_{\Theta \times \mathfrak{X}} L(\theta, \hat{\theta}(x)) \mathbf{Pr}(\theta, x). \tag{14}$$

- References
  - i [Wrinkler, 1995].

Bayesian Estimator

- An estimator minimizing this risk is called a `Bayesian estimator`.
- The quality of an estimator depends on both the prior model and the loss function.
- Prior information about θ is seldom very precise.
- The choice of L is problem specific.
- The MAP, ML and MMSE estimators are Bayes estimators for certain loss functions.
- One of the reasons why the above estimators were introduced is that they can be computed (or at least approximated).
- References
  - [Wrinkler, 1995].

Bayesian Estimator

- An estimator minimizing this risk is called a `Bayesian estimator`.
- The quality of an estimator depends on both the prior model and the loss function.
- Prior information about θ is seldom very precise.
- The choice of L is problem specific.
- The MAP, ML and MMSE estimators are Bayes estimators for certain loss functions.
- One of the reasons why the above estimators were introduced is that they can be computed (or at least approximated).
- References
  - [Wrinkler, 1995].

Bayesian Estimator

- An estimator minimizing this risk is called a `Bayesian estimator`.
- The quality of an estimator depends on both the prior model and the loss function.
- Prior information about θ is seldom very precise.
- The choice of L is problem specific.
- The MAP, ML and MMSE estimators are Bayes estimators for certain loss functions.
- One of the reasons why the above estimators were introduced is that they can be computed (or at least approximated).
- References
  - i [Wrinkler, 1995].

Bayesian Estimator

- An estimator minimizing this risk is called a `Bayesian estimator`.
- The quality of an estimator depends on both the prior model and the loss function.

- Prior information about θ is seldom very precise.
- The choice of L is problem specific.

- The MAP, ML and MMSE estimators are Bayes estimators for certain loss functions.
- One of the reasons why the above estimators were introduced is that they can be computed (or at least approximated).
- References
  - [Wrinkler, 1995].

Bayesian Estimator

- An estimator minimizing this risk is called a `Bayesian estimator`.
- The quality of an estimator depends on both the prior model and the loss function.
- Prior information about θ is seldom very precise.
- The choice of L is problem specific.
- The MAP, ML and MMSE estimators are Bayes estimators for certain loss functions.
- One of the reasons why the above estimators were introduced is that they can be computed (or at least approximated).
- References
  - [Wrinkler, 1995].

Bayesian Estimator

- An estimator minimizing this risk is called a `Bayesian estimator`.
- The quality of an estimator depends on both the prior model and the loss function.
- Prior information about θ is seldom very precise.
- The choice of L is problem specific.
- The MAP, ML and MMSE estimators are Bayes estimators for certain loss functions.
- One of the reasons why the above estimators were introduced is that they can be computed (or at least approximated).
- References
  - [Wrinkler, 1995].

Bayesian Estimator

- An estimator minimizing this risk is called a `Bayesian estimator`.
- The quality of an estimator depends on both the prior model and the loss function.
- Prior information about θ is seldom very precise.
- The choice of L is problem specific.
- The MAP, ML and MMSE estimators are Bayes estimators for certain loss functions.
- One of the reasons why the above estimators were introduced is that they can be computed (or at least approximated).
- References
  - [Wrinkler, 1995].

Bayesian Estimator

- An estimator minimizing this risk is called a `Bayesian estimator`.
- The quality of an estimator depends on both the prior model and the loss function.

- Prior information about θ is seldom very precise.
- The choice of L is problem specific.

- The MAP, ML and MMSE estimators are Bayes estimators for certain loss functions.
- One of the reasons why the above estimators were introduced is that they can be computed (or at least approximated).
- References
  - i [Wrinkler, 1995].

0-1 Loss and MAP

- 0-1 loss function

$$L(\theta, \hat{\theta}) = \begin{cases} 0, & \text{if } \theta = \hat{\theta}; \\ 1, & \text{if } \theta \neq \hat{\theta}. \end{cases} \tag{15}$$

- The Bayesian risk

$$\hat{R} = \int_{\Theta \times \mathfrak{X}} L(\theta, \hat{\theta}(x)) \mathbf{Pr}(\theta, x) \tag{16}$$

$$= \int_{\mathfrak{X}} \int_{\Theta} L(\theta, \hat{\theta}(x)) \mathbf{Pr}(\theta, x) \tag{17}$$

$$= \int_{\mathfrak{X}} \left\{ m(x) - \mathbf{Pr}(\hat{\theta}(x), x) \right\} \tag{18}$$

$$= 1 - \int_{\mathfrak{X}} \mathbf{Pr}(\hat{\theta}(x), x) \tag{19}$$

  is minimized when $\mathbf{Pr}(\hat{\theta}(x), x)$ is maximized, by (10).

- References
  - [Wrinkler, 1995].

0-1 Loss and MAP

- 0-1 loss function

$$L(\theta, \hat{\theta}) = \begin{cases} 0, & \text{if } \theta = \hat{\theta}; \\ 1, & \text{if } \theta \neq \hat{\theta}. \end{cases} \quad (15)$$

- The Bayesian risk

$$\hat{R} = \int_{\Theta \times \mathfrak{X}} L(\theta, \hat{\theta}(x)) \mathbf{Pr}(\theta, x) \quad (16)$$

$$= \int_{\mathfrak{X}} \int_{\Theta} L(\theta, \hat{\theta}(x)) \mathbf{Pr}(\theta, x) \quad (17)$$

$$= \int_{\mathfrak{X}} \left\{ m(x) - \mathbf{Pr}(\hat{\theta}(x), x) \right\} \quad (18)$$

$$= 1 - \int_{\mathfrak{X}} \mathbf{Pr}(\hat{\theta}(x), x) \quad (19)$$

is minimized when $\mathbf{Pr}(\hat{\theta}(x), x)$ is maximized, by (10).

- References
  - [Wrinkler, 1995].

0-1 Loss and MAP

- 0-1 loss function

$$L(\theta, \hat{\theta}) = \begin{cases} 0, & \text{if } \theta = \hat{\theta}; \\ 1, & \text{if } \theta \neq \hat{\theta}. \end{cases} \tag{15}$$

- The Bayesian risk

$$\hat{R} = \int_{\Theta \times \mathfrak{X}} L(\theta, \hat{\theta}(x))\mathbf{Pr}(\theta, x) \tag{16}$$

$$= \int_{\mathfrak{X}} \int_{\Theta} L(\theta, \hat{\theta}(x))\mathbf{Pr}(\theta, x) \tag{17}$$

$$= \int_{\mathfrak{X}} \left\{ m(x) - \mathbf{Pr}(\hat{\theta}(x), x) \right\} \tag{18}$$

$$= 1 - \int_{\mathfrak{X}} \mathbf{Pr}(\hat{\theta}(x), x) \tag{19}$$

  is minimized when $\mathbf{Pr}(\hat{\theta}(x), x)$ is maximized, by (10).
- References
  - [Wrinkler, 1995].

0-1 Loss and MAP

- 0-1 loss function

$$L(\theta, \hat{\theta}) = \begin{cases} 0, & \text{if } \theta = \hat{\theta}; \\ 1, & \text{if } \theta \neq \hat{\theta}. \end{cases} \tag{15}$$

- The Bayesian risk

$$\hat{R} = \int_{\Theta \times \mathfrak{X}} L(\theta, \hat{\theta}(x)) \mathbf{Pr}(\theta, x) \tag{16}$$

$$= \int_{\mathfrak{X}} \int_{\Theta} L(\theta, \hat{\theta}(x)) \mathbf{Pr}(\theta, x) \tag{17}$$

$$= \int_{\mathfrak{X}} \left\{ m(x) - \mathbf{Pr}(\hat{\theta}(x), x) \right\} \tag{18}$$

$$= 1 - \int_{\mathfrak{X}} \mathbf{Pr}(\hat{\theta}(x), x) \tag{19}$$

  is minimized when $\mathbf{Pr}(\hat{\theta}(x), x)$ is maximized, by (10).
- References
  i [Wrinkler, 1995].

Squared-error Loss and MMSE

- Squared-error loss function

$$L(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2 = \sum_s |\theta_s - \hat{\theta}_s|^2. \qquad (20)$$

- The Bayesian risk

$$\hat{R} = \int_{\mathfrak{X}} \int_{\Theta} \|\theta - \hat{\theta}(x)\|^2 \mathbf{Pr}(\theta, x) \qquad (21)$$

$$= \int_{\mathfrak{X}} \int_{\Theta} \left\{ \|\theta\|^2 - 2\left\langle \theta, \hat{\theta}(x) \right\rangle + \|\hat{\theta}(x)\|^2 \right\} \mathbf{Pr}(\theta, x) \qquad (22)$$

$$= \int_{\mathfrak{X}} \int_{\Theta} \|\theta\|^2 \mathbf{Pr}(\theta, x) - \int_{\mathfrak{X}} \left\{ 2\left\langle \theta_{MMSE}(x), \hat{\theta}(x) \right\rangle - \|\hat{\theta}(x)\|^2 \right\} m(x) \qquad (23)$$

$$= \int_{\mathfrak{X}} \int_{\Theta} \|\theta\|^2 \mathbf{Pr}(\theta, x) - \int_{\mathfrak{X}} \|\theta_{MMSE}(x)\|^2 m(x) + \int_{\mathfrak{X}} \left\{ \|\theta_{MMSE}(x) - \hat{\theta}(x)\|^2 \right\} m(x) \qquad (24)$$

is minimized when $\hat{\theta}(x) = \theta_{MMSE}(x)$.

- References
  - [Wrinkler, 1995].

Squared-error Loss and MMSE

- Squared-error loss function

$$L(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2 = \sum_s |\theta_s - \hat{\theta}_s|^2. \quad (20)$$

- The Bayesian risk

$$\hat{R} = \int_{\mathfrak{X}} \int_{\Theta} \|\theta - \hat{\theta}(x)\|^2 \mathbf{Pr}(\theta, x) \quad (21)$$

$$= \int_{\mathfrak{X}} \int_{\Theta} \left\{ \|\theta\|^2 - 2\left\langle \theta, \hat{\theta}(x) \right\rangle + \|\hat{\theta}(x)\|^2 \right\} \mathbf{Pr}(\theta, x) \quad (22)$$

$$= \int_{\mathfrak{X}} \int_{\Theta} \|\theta\|^2 \mathbf{Pr}(\theta, x) - \int_{\mathfrak{X}} \left\{ 2\left\langle \theta_{MMSE}(x), \hat{\theta}(x) \right\rangle - \|\hat{\theta}(x)\|^2 \right\} m(x) \quad (23)$$

$$= \int_{\mathfrak{X}} \int_{\Theta} \|\theta\|^2 \mathbf{Pr}(\theta, x) - \int_{\mathfrak{X}} \|\theta_{MMSE}(x)\|^2 m(x) + \int_{\mathfrak{X}} \left\{ \|\theta_{MMSE}(x) - \hat{\theta}(x)\|^2 \right\} m(x) \quad (24)$$

is minimized when $\hat{\theta}(x) = \theta_{MMSE}(x)$.

- References
  - [Wrinkler, 1995].

Squared-error Loss and MMSE

- Squared-error loss function

$$L(\theta,\hat{\theta}) = \|\theta - \hat{\theta}\|^2 = \sum_s |\theta_s - \hat{\theta}_s|^2. \qquad (20)$$

- The Bayesian risk

$$\hat{R} = \int_{\mathfrak{X}}\int_{\Theta} \|\theta - \hat{\theta}(x)\|^2 \mathbf{Pr}(\theta, x) \qquad (21)$$

$$= \int_{\mathfrak{X}}\int_{\Theta} \left\{ \|\theta\|^2 - 2\left\langle \theta, \hat{\theta}(x)\right\rangle + \|\hat{\theta}(x)\|^2 \right\} \mathbf{Pr}(\theta, x) \qquad (22)$$

$$= \int_{\mathfrak{X}}\int_{\Theta} \|\theta\|^2 \mathbf{Pr}(\theta, x) - \int_{\mathfrak{X}} \left\{ 2\left\langle \theta_{MMSE}(x), \hat{\theta}(x)\right\rangle - \|\hat{\theta}(x)\|^2 \right\} m(x) \qquad (23)$$

$$= \int_{\mathfrak{X}}\int_{\Theta} \|\theta\|^2 \mathbf{Pr}(\theta, x) - \int_{\mathfrak{X}} \|\theta_{MMSE}(x)\|^2 m(x) + \int_{\mathfrak{X}} \left\{ \|\theta_{MMSE}(x) - \hat{\theta}(x)\|^2 \right\} m(x) \qquad (24)$$

is minimized when $\hat{\theta}(x) = \theta_{MMSE}(x)$.

- References

[Wrinkler, 1995].

Squared-error Loss and MMSE

- Squared-error loss function

$$L(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2 = \sum_s |\theta_s - \hat{\theta}_s|^2. \tag{20}$$

- The Bayesian risk

$$\hat{R} = \int_{\mathfrak{X}} \int_{\Theta} \|\theta - \hat{\theta}(x)\|^2 \mathbf{Pr}(\theta, x) \tag{21}$$

$$= \int_{\mathfrak{X}} \int_{\Theta} \left\{ \|\theta\|^2 - 2\left\langle \theta, \hat{\theta}(x) \right\rangle + \|\hat{\theta}(x)\|^2 \right\} \mathbf{Pr}(\theta, x) \tag{22}$$

$$= \int_{\mathfrak{X}} \int_{\Theta} \|\theta\|^2 \mathbf{Pr}(\theta, x) - \int_{\mathfrak{X}} \left\{ 2\left\langle \theta_{MMSE}(x), \hat{\theta}(x) \right\rangle - \|\hat{\theta}(x)\|^2 \right\} m(x) \tag{23}$$

$$= \int_{\mathfrak{X}} \int_{\Theta} \|\theta\|^2 \mathbf{Pr}(\theta, x) - \int_{\mathfrak{X}} \|\theta_{MMSE}(x)\|^2 m(x) + \int_{\mathfrak{X}} \left\{ \|\theta_{MMSE}(x) - \hat{\theta}(x)\|^2 \right\} m(x) \tag{24}$$

  is minimized when $\hat{\theta}(x) = \theta_{MMSE}(x)$.
- References
  - [Wrinkler, 1995].

Other Estimators and Loss Functions

- Marginal posterior mode estimate (MPME) minimizes the Bayesian risk for the Hamming distance

$$L(\theta, \hat{\theta}) = \frac{1}{|S|} |\{s \in S : \theta_s \neq \hat{\theta}_s\}|. \tag{25}$$

- Posterior median minimizes the Bayesian risk for the absolute-value loss function:

$$L(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_1 = \sum_s |\theta_s - \hat{\theta}_s|. \tag{26}$$

- References
  i [Wrinkler, 1995].
  ii Point estimation at wikipedia.

Other Estimators and Loss Functions

- Marginal posterior mode estimate (MPME) minimizes the Bayesian risk for the Hamming distance

$$L(\theta, \hat{\theta}) = \frac{1}{|S|}|\{s \in S : \theta_s \neq \hat{\theta}_s\}|. \tag{25}$$

- Posterior median minimizes the Bayesian risk for the absolute-value loss function:

$$L(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_1 = \sum_s |\theta_s - \hat{\theta}_s|. \tag{26}$$

- References
  i [Wrinkler, 1995].
  ii Point estimation at wikipedia.

Other Estimators and Loss Functions

- `Marginal posterior mode estimate` (MPME) minimizes the Bayesian risk for the Hamming distance

$$L(\theta, \hat{\theta}) = \frac{1}{|S|}|\{s \in S : \theta_s \neq \hat{\theta}_s\}|. \quad (25)$$

- `Posterior median` minimizes the Bayesian risk for the absolute-value loss function:

$$L(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_1 = \sum_s |\theta_s - \hat{\theta}_s|. \quad (26)$$

- References
  - i [Wrinkler, 1995].
  - ii Point estimation at wikipedia.

Other Estimators and Loss Functions

- `Marginal posterior mode estimate` (MPME) minimizes the Bayesian risk for the Hamming distance

$$L(\theta, \hat{\theta}) = \frac{1}{|S|} |\{s \in S : \theta_s \neq \hat{\theta}_s\}|. \tag{25}$$

- `Posterior median` minimizes the Bayesian risk for the absolute-value loss function:

$$L(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_1 = \sum_s |\theta_s - \hat{\theta}_s|. \tag{26}$$

- References
  - i [Wrinkler, 1995].
  - ii Point estimation at wikipedia.

Other Estimators and Loss Functions

- `Marginal posterior mode estimate` (MPME) minimizes the Bayesian risk for the Hamming distance

$$L(\theta, \hat{\theta}) = \frac{1}{|S|}|\{s \in S : \theta_s \neq \hat{\theta}_s\}|. \tag{25}$$

- `Posterior median` minimizes the Bayesian risk for the absolute-value loss function:

$$L(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_1 = \sum_s |\theta_s - \hat{\theta}_s|. \tag{26}$$

- References
  - i [Wrinkler, 1995].
  - ii Point estimation at wikipedia.

Outline

Bayesian Inference

- Bayes' Theorem
- Bayesian Inference
- Prior Information

Bayesian Approach

- The Bayesian approach entails constructing the prior distribution **Pr**(θ) and finding algorithm to compute the Bayesian reconstruction.
- This consists of identifying the prior and specifying the data model.
- The following are several approaches to assign the prior distribution for θ.

- References
  - [Berger, 1985].

Bayesian Approach

- The Bayesian approach entails constructing the prior distribution $\mathbf{Pr}(\theta)$ and finding algorithm to compute the Bayesian reconstruction.
- This consists of identifying the prior and specifying the data model.
- The following are several approaches to assign the prior distribution for θ.
- References
  - [Berger, 1985].

Bayesian Approach

- The Bayesian approach entails constructing the prior distribution $\mathbf{Pr}(\theta)$ and finding algorithm to compute the Bayesian reconstruction.
- This consists of identifying the prior and specifying the data model.
- The following are several approaches to assign the prior distribution for θ.

- References
  - [Berger, 1985].

Bayesian Approach

- The Bayesian approach entails constructing the prior distribution $\mathbf{Pr}(\theta)$ and finding algorithm to compute the Bayesian reconstruction.
- This consists of identifying the prior and specifying the data model.
- The following are several approaches to assign the prior distribution for θ.
- References
  - [Berger, 1985].

Bayesian Approach

- The Bayesian approach entails constructing the prior distribution **Pr**(θ) and finding algorithm to compute the Bayesian reconstruction.
- This consists of identifying the prior and specifying the data model.
- The following are several approaches to assign the prior distribution for θ.
- References
  - i [Berger, 1985].

Non-informative Priors

- There have been attempts to use the Bayesian approach even when no (or minimal) prior information is available.
- What is needed is a `non-informative prior`, by which is meant a prior which contains no information about θ.
- The simplest situation to consider is when Θ is a finite set, consisting of n elements.
- The obvious prior is to then give each of Θ probability $1/n$.
- For infinite set, the uniform non-informative prior $\mathbf{Pr}(\theta) = c$ is proposed, where c is a constant.
- Comments: lack of invariance under transformation and improper probability distribution.
- References
  - [Berger, 1985].

Non-informative Priors

- There have been attempts to use the Bayesian approach even when no (or minimal) prior information is available.
- What is needed is a `non-informative prior`, by which is meant a prior which contains no information about θ.
- The simplest situation to consider is when Θ is a finite set, consisting of n elements.
- The obvious prior is to then give each of Θ probability $1/n$.
- For infinite set, the uniform non-informative prior $\mathbf{Pr}(\theta) = c$ is proposed, where c is a constant.
- Comments: lack of invariance under transformation and improper probability distribution.
- References
  - [Berger, 1985].

Non-informative Priors

- There have been attempts to use the Bayesian approach even when no (or minimal) prior information is available.
- What is needed is a `non-informative prior`, by which is meant a prior which contains no information about θ.
- The simplest situation to consider is when Θ is a finite set, consisting of n elements.
- The obvious prior is to then give each of Θ probability $1/n$.
- For infinite set, the uniform non-informative prior $\Pr(\theta) = c$ is proposed, where c is a constant.
- Comments: lack of invariance under transformation and improper probability distribution.
- References
  - [Berger, 1985].

Non-informative Priors

- There have been attempts to use the Bayesian approach even when no (or minimal) prior information is available.
- What is needed is a `non-informative prior`, by which is meant a prior which contains no information about θ.
- The simplest situation to consider is when Θ is a finite set, consisting of n elements.
- The obvious prior is to then give each of Θ probability $1/n$.
- For infinite set, the uniform non-informative prior $\Pr(\theta) = c$ is proposed, where c is a constant.
- Comments: lack of invariance under transformation and improper probability distribution.
- References
    i [Berger, 1985].

Non-informative Priors

- There have been attempts to use the Bayesian approach even when no (or minimal) prior information is available.
- What is needed is a `non-informative prior`, by which is meant a prior which contains no information about θ.
- The simplest situation to consider is when Θ is a finite set, consisting of n elements.
- The obvious prior is to then give each of Θ probability $1/n$.
- For infinite set, the uniform non-informative prior $\mathbf{Pr}(\theta) = c$ is proposed, where c is a constant.
- Comments: lack of invariance under transformation and improper probability distribution.
- References
  - [Berger, 1985].

Non-informative Priors

- There have been attempts to use the Bayesian approach even when no (or minimal) prior information is available.
- What is needed is a `non-informative prior`, by which is meant a prior which contains no information about θ.

- The simplest situation to consider is when Θ is a finite set, consisting of n elements.
- The obvious prior is to then give each of Θ probability $1/n$.

- For infinite set, the uniform non-informative prior $\mathbf{Pr}(\theta) = c$ is proposed, where c is a constant.
- Comments: lack of invariance under transformation and improper probability distribution.
- References
  - [Berger, 1985].

Non-informative Priors

- There have been attempts to use the Bayesian approach even when no (or minimal) prior information is available.
- What is needed is a `non-informative prior`, by which is meant a prior which contains no information about θ.
- The simplest situation to consider is when Θ is a finite set, consisting of n elements.
- The obvious prior is to then give each of Θ probability $1/n$.
- For infinite set, the uniform non-informative prior $\mathbf{Pr}(\theta) = c$ is proposed, where c is a constant.
- Comments: lack of invariance under transformation and improper probability distribution.
- References
  - [Berger, 1985].

Non-informative Priors

- There have been attempts to use the Bayesian approach even when no (or minimal) prior information is available.
- What is needed is a `non-informative prior`, by which is meant a prior which contains no information about θ.
- The simplest situation to consider is when Θ is a finite set, consisting of n elements.
- The obvious prior is to then give each of Θ probability $1/n$.
- For infinite set, the uniform non-informative prior $\mathbf{Pr}(\theta) = c$ is proposed, where c is a constant.
- Comments: lack of invariance under transformation and improper probability distribution.
- References
  - [Berger, 1985].

Jeffreys' Rule

Given RULE for assigning prior distribution for θ. Assume that g is a function of θ. One can also assign a prior distribution for the random variable $\xi = g(\theta)$ by this RULE. Then it should hold that

$$\mathbf{Pr}(\theta) = \mathbf{Pr}(\xi)|\det(\nabla g(\theta))| = \mathbf{Pr}(g(\theta))|\det(\nabla g(\theta))| \quad (27)$$

- Since
$$\int_{g(A)} \mathbf{Pr}(\xi)d\xi = \int_A \mathbf{Pr}(\theta)d\theta,$$
Jeffreys' Rule requires that the prior distribution is invariant under transformation.
- References
  i [Berger, 1985].
  ii [Zhang and Cheng, 1994].

Jeffreys' Rule

Given RULE for assigning prior distribution for θ. Assume that g is a function of θ. One can also assign a prior distribution for the random variable $\xi = g(\theta)$ by this RULE. Then it should hold that

$$\mathbf{Pr}(\theta) = \mathbf{Pr}(\xi)|\det(\nabla g(\theta))| = \mathbf{Pr}(g(\theta))|\det(\nabla g(\theta))| \quad (27)$$

- Since
$$\int_{g(A)} \mathbf{Pr}(\xi) d\xi = \int_A \mathbf{Pr}(\theta) d\theta,$$
Jeffreys' Rule requires that the prior distribution is invariant under transformation.

- References
  i [Berger, 1985].
  ii [Zhang and Cheng, 1994].

Jeffreys' Rule

Given RULE for assigning prior distribution for θ. Assume that g is a function of θ. One can also assign a prior distribution for the random variable $\xi = g(\theta)$ by this RULE. Then it should hold that

$$\textbf{Pr}(\theta) = \textbf{Pr}(\xi)|\det(\nabla g(\theta))| = \textbf{Pr}(g(\theta))|\det(\nabla g(\theta))| \quad (27)$$

- Since
$$\int_{g(A)} \textbf{Pr}(\xi)d\xi = \int_A \textbf{Pr}(\theta)d\theta,$$
Jeffreys' Rule requires that the prior distribution is invariant under transformation.
- References
  i [Berger, 1985].
  ii [Zhang and Cheng, 1994].

Jeffreys' Rule

Given RULE for assigning prior distribution for θ. Assume that g is a function of θ. One can also assign a prior distribution for the random variable $\xi = g(\theta)$ by this RULE. Then it should hold that

$$\mathbf{Pr}(\theta) = \mathbf{Pr}(\xi)|\det(\nabla g(\theta))| = \mathbf{Pr}(g(\theta))|\det(\nabla g(\theta))| \quad (27)$$

- Since
$$\int_{g(A)} \mathbf{Pr}(\xi) d\xi = \int_A \mathbf{Pr}(\theta) d\theta,$$
Jeffreys' Rule requires that the prior distribution is invariant under transformation.
- References
  - i [Berger, 1985].
  - ii [Zhang and Cheng, 1994].

Fisher Information matrix

- Jeffrey showed that if

$$I(\theta) = E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\theta_i}\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\theta_j}\right)\right] \\ = E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\theta}\right)\cdot\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\theta}\right)^{\mathrm{tr}}\right], \tag{28}$$

  then

$$\mathbf{Pr}(\theta) = |\det I(\theta)|^{1/2} \tag{29}$$

  is a prior satisfying (27).

- References
  - i [Berger, 1985].
  - ii [Zhang and Cheng, 1994].

Fisher Information matrix

- Jeffrey showed that if

$$I(\theta) = E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta_i} \frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta_j}\right)\right]$$
$$= E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right)^{\mathrm{tr}}\right], \quad (28)$$

  then

$$\mathbf{Pr}(\theta) = |\det I(\theta)|^{1/2} \quad (29)$$

  is a prior satisfying (27).

- References
  i [Berger, 1985].
  ii [Zhang and Cheng, 1994].

Fisher Information matrix

- Jeffrey showed that if

$$I(\theta) = E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta_i}\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta_j}\right)\right] = E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right)\cdot\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right)^{tr}\right], \quad (28)$$

then

$$\mathbf{Pr}(\theta) = |\det I(\theta)|^{1/2} \quad (29)$$

is a prior satisfying (27).

- References
  i [Berger, 1985].
  ii [Zhang and Cheng, 1994].

Fisher Information matrix

- Jeffrey showed that if

$$I(\theta) = E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta_i} \frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta_j}\right)\right] = E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right)^{\mathrm{tr}}\right], \quad (28)$$

then

$$\mathbf{Pr}(\theta) = |\det I(\theta)|^{1/2} \quad (29)$$

is a prior satisfying (27).

- References
  - i [Berger, 1985].
  - ii [Zhang and Cheng, 1994].

Proof

- We assume that $g(\cdot)$ is a smooth homeomorphism.
- Since

$$\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right) = \left(\frac{\partial g(\theta)}{\partial \theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right),$$

- we have

$$\begin{aligned} I(\theta) &= E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right)^{\mathrm{tr}}\right] \\ &= E\left[\left(\frac{\partial g(\theta)}{\partial \theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right)^{\mathrm{tr}} \cdot \left(\frac{\partial g(\theta)}{\partial \theta}\right)^{\mathrm{tr}}\right] \\ &= \left(\frac{\partial g(\theta)}{\partial \theta}\right) \cdot E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right)^{\mathrm{tr}}\right] \cdot \left(\frac{\partial g(\theta)}{\partial \theta}\right)^{\mathrm{tr}}. \end{aligned}$$

- Therefore

$$\det I(\theta) = \det I(\xi) \cdot \left|\det \left(\frac{\partial g(\theta)}{\partial \theta}\right)\right|^2.$$

  So (27) holds.
- References
  - [Zhang and Cheng, 1994].

Proof

- We assume that $g(\cdot)$ is a smooth homeomorphism.
- Since

$$\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right) = \left(\frac{\partial g(\theta)}{\partial \theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right),$$

- we have

$$\begin{aligned} I(\theta) &= E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right)^{\mathrm{tr}}\right] \\ &= E\left[\left(\frac{\partial g(\theta)}{\partial \theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right)^{\mathrm{tr}} \cdot \left(\frac{\partial g(\theta)}{\partial \theta}\right)^{\mathrm{tr}}\right] \\ &= \left(\frac{\partial g(\theta)}{\partial \theta}\right) \cdot E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right)^{\mathrm{tr}}\right] \cdot \left(\frac{\partial g(\theta)}{\partial \theta}\right)^{\mathrm{tr}}. \end{aligned}$$

- Therefore

$$\det I(\theta) = \det I(\xi) \cdot \left|\det\left(\frac{\partial g(\theta)}{\partial \theta}\right)\right|^2.$$

  So (27) holds.
- References
  - [Zhang and Cheng, 1994].

Proof

- We assume that $g(\cdot)$ is a smooth homeomorphism.
- Since

$$\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\theta}\right) = \left(\frac{\partial g(\theta)}{\partial\theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\xi}\right),$$

- we have

$$\begin{aligned} I(\theta) &= E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\theta}\right)^{\mathrm{tr}}\right] \\ &= E\left[\left(\frac{\partial g(\theta)}{\partial\theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\xi}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\xi}\right)^{\mathrm{tr}} \cdot \left(\frac{\partial g(\theta)}{\partial\theta}\right)^{\mathrm{tr}}\right] \\ &= \left(\frac{\partial g(\theta)}{\partial\theta}\right) \cdot E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\xi}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\xi}\right)^{\mathrm{tr}}\right] \cdot \left(\frac{\partial g(\theta)}{\partial\theta}\right)^{\mathrm{tr}}. \end{aligned}$$

- Therefore

$$\det I(\theta) = \det I(\xi) \cdot \left|\det\left(\frac{\partial g(\theta)}{\partial\theta}\right)\right|^2.$$

  So (27) holds.
- References
  - [Zhang and Cheng, 1994].

Proof

- We assume that $g(\cdot)$ is a smooth homeomorphism.
- Since

$$\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right) = \left(\frac{\partial g(\theta)}{\partial \theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right),$$

- we have

$$\begin{aligned} I(\theta) &= E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right)^{\text{tr}}\right] \\ &= E\left[\left(\frac{\partial g(\theta)}{\partial \theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right)^{\text{tr}} \cdot \left(\frac{\partial g(\theta)}{\partial \theta}\right)^{\text{tr}}\right] \\ &= \left(\frac{\partial g(\theta)}{\partial \theta}\right) \cdot E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right)^{\text{tr}}\right] \cdot \left(\frac{\partial g(\theta)}{\partial \theta}\right)^{\text{tr}}. \end{aligned}$$

- Therefore

$$\det I(\theta) = \det I(\xi) \cdot \left|\det \left(\frac{\partial g(\theta)}{\partial \theta}\right)\right|^2.$$

So (27) holds.

- References
  - [Zhang and Cheng, 1994].

Proof

- We assume that $g(\cdot)$ is a smooth homeomorphism.
- Since

$$\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right) = \left(\frac{\partial g(\theta)}{\partial \theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right),$$

- we have

$$\begin{aligned} I(\theta) &= E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \theta}\right)^{\mathrm{tr}}\right] \\ &= E\left[\left(\frac{\partial g(\theta)}{\partial \theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right)^{\mathrm{tr}} \cdot \left(\frac{\partial g(\theta)}{\partial \theta}\right)^{\mathrm{tr}}\right] \\ &= \left(\frac{\partial g(\theta)}{\partial \theta}\right) \cdot E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial \xi}\right)^{\mathrm{tr}}\right] \cdot \left(\frac{\partial g(\theta)}{\partial \theta}\right)^{\mathrm{tr}}. \end{aligned}$$

- Therefore

$$\det I(\theta) = \det I(\xi) \cdot \left|\det \left(\frac{\partial g(\theta)}{\partial \theta}\right)\right|^2.$$

So (27) holds.
- References
  - [Zhang and Cheng, 1994].

Proof

- We assume that $g(\cdot)$ is a smooth homeomorphism.
- Since

$$\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\theta}\right) = \left(\frac{\partial g(\theta)}{\partial\theta}\right) \cdot \left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\xi}\right),$$

- we have

$$\begin{aligned} I(\theta) &= E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\theta}\right)\cdot\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\theta}\right)^{\mathrm{tr}}\right] \\ &= E\left[\left(\frac{\partial g(\theta)}{\partial\theta}\right)\cdot\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\xi}\right)\cdot\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\xi}\right)^{\mathrm{tr}}\cdot\left(\frac{\partial g(\theta)}{\partial\theta}\right)^{\mathrm{tr}}\right] \\ &= \left(\frac{\partial g(\theta)}{\partial\theta}\right)\cdot E\left[\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\xi}\right)\cdot\left(\frac{\partial \ln \mathbf{Pr}(X|\theta)}{\partial\xi}\right)^{\mathrm{tr}}\right]\cdot\left(\frac{\partial g(\theta)}{\partial\theta}\right)^{\mathrm{tr}}. \end{aligned}$$

- Therefore

$$\det I(\theta) = \det I(\xi) \cdot \left|\det\left(\frac{\partial g(\theta)}{\partial\theta}\right)\right|^2.$$

  So (27) holds.
- References
  - i [Zhang and Cheng, 1994].

Maximum Entropy Principle

- When partial prior information is available, it is desired to use a prior that is as non-informative as possible.
- E.g., suppose the prior mean is specified. Among prior distributions with this mean the most non-informative distribution is sought.
- A useful method of dealing with this problem is through the concept of entropy.

- References
  i [Berger, 1985].
  ii [Zhang and Cheng, 1994].

Maximum Entropy Principle

- When partial prior information is available, it is desired to use a prior that is as non-informative as possible.
- E.g., suppose the prior mean is specified. Among prior distributions with this mean the most non-informative distribution is sought.
- A useful method of dealing with this problem is through the concept of entropy.

- References
  1. [Berger, 1985].
  2. [Zhang and Cheng, 1994].

Maximum Entropy Principle

- When partial prior information is available, it is desired to use a prior that is as non-informative as possible.
- E.g., suppose the prior mean is specified. Among prior distributions with this mean the most non-informative distribution is sought.
- A useful method of dealing with this problem is through the concept of entropy.

- References
  1. [Berger, 1985].
  2. [Zhang and Cheng, 1994].

Maximum Entropy Principle

- When partial prior information is available, it is desired to use a prior that is as non-informative as possible.
- E.g., suppose the prior mean is specified. Among prior distributions with this mean the most non-informative distribution is sought.
- A useful method of dealing with this problem is through the concept of entropy.

- References
  i [Berger, 1985].
  ii [Zhang and Cheng, 1994].

Maximum Entropy Principle

- When partial prior information is available, it is desired to use a prior that is as non-informative as possible.
- E.g., suppose the prior mean is specified. Among prior distributions with this mean the most non-informative distribution is sought.
- A useful method of dealing with this problem is through the concept of entropy.

- References
  i [Berger, 1985].
  ii [Zhang and Cheng, 1994].

Maximum Entropy Principle

- When partial prior information is available, it is desired to use a prior that is as non-informative as possible.
- E.g., suppose the prior mean is specified. Among prior distributions with this mean the most non-informative distribution is sought.
- A useful method of dealing with this problem is through the concept of entropy.

- References
  i [Berger, 1985].
  ii [Zhang and Cheng, 1994].

Entropy

- Entropy is most easily understood for discrete distributions.

- **Definition**

  Assume Θ is discrete and let $\mathbf{Pr}(\cdot)$ be a probability density on Θ. The entropy of $\mathbf{Pr}(\cdot)$, denoted by $\mathfrak{E}(\mathbf{Pr}(\cdot))$, is defined as

$$\mathfrak{E}(\mathbf{Pr}(\cdot)) = -\sum_{\Theta} \mathbf{Pr}(\theta_i) \log \mathbf{Pr}(\theta_i) \tag{30}$$

  If Θ is continuous,

$$\mathfrak{E}(\mathbf{Pr}(\cdot)) = -\int_{\Theta} \mathbf{Pr}(\theta) \log \frac{\mathbf{Pr}(\theta)}{\pi_0(\theta)} d\theta \tag{31}$$

  where π_0 is a natural "invariant" non-informative prior for the problem.

  - [Berger, 1985].

Entropy

- Entropy is most easily understood for discrete distributions.

- **Definition**
  Assume Θ is discrete and let $\mathbf{Pr}(\cdot)$ be a probability density on Θ. The entropy of $\mathbf{Pr}(\cdot)$, denoted by $\mathfrak{E}(\mathbf{Pr}(\cdot))$, is defined as

$$\mathfrak{E}(\mathbf{Pr}(\cdot)) = -\sum_{\Theta} \mathbf{Pr}(\theta_i) \log \mathbf{Pr}(\theta_i) \tag{30}$$

If Θ is continuous,

$$\mathfrak{E}(\mathbf{Pr}(\cdot)) = -\int_{\Theta} \mathbf{Pr}(\theta) \log \frac{\mathbf{Pr}(\theta)}{\pi_0(\theta)} d\theta \tag{31}$$

where π_0 is a natural “invariant” non-informative prior for the problem.

  i [Berger, 1985].

Entropy

- Entropy is most easily understood for discrete distributions.

- Definition
  Assume Θ is discrete and let $\mathbf{Pr}(\cdot)$ be a probability density on Θ. The entropy of $\mathbf{Pr}(\cdot)$, denoted by $\mathfrak{E}(\mathbf{Pr}(\cdot))$, is defined as

$$\mathfrak{E}(\mathbf{Pr}(\cdot)) = -\sum_{\Theta} \mathbf{Pr}(\theta_i) \log \mathbf{Pr}(\theta_i) \tag{30}$$

If Θ is continuous,

$$\mathfrak{E}(\mathbf{Pr}(\cdot)) = -\int_{\Theta} \mathbf{Pr}(\theta) \log \frac{\mathbf{Pr}(\theta)}{\pi_0(\theta)} d\theta \tag{31}$$

where π_0 is a natural "invariant" non-informative prior for the problem.

i [Berger, 1985].

Entropy

- Entropy is most easily understood for discrete distributions.

- **Definition**

  Assume Θ is discrete and let $\mathbf{Pr}(\cdot)$ be a probability density on Θ. The entropy of $\mathbf{Pr}(\cdot)$, denoted by $\mathfrak{E}(\mathbf{Pr}(\cdot))$, is defined as

$$\mathfrak{E}(\mathbf{Pr}(\cdot)) = -\sum_{\Theta} \mathbf{Pr}(\theta_i) \log \mathbf{Pr}(\theta_i) \tag{30}$$

  If Θ is continuous,

$$\mathfrak{E}(\mathbf{Pr}(\cdot)) = -\int_{\Theta} \mathbf{Pr}(\theta) \log \frac{\mathbf{Pr}(\theta)}{\pi_0(\theta)} d\theta \tag{31}$$

  where π_0 is a natural "invariant" non-informative prior for the problem.

  i [Berger, 1985].

Entropy Maximization

- Entropy has a direct relationship to information theory.
- It is a measure of the amount of uncertainty inherent in the probability distribution.
- The principle is to seek the prior distribution which maximizes entropy among all those distributions which satisfy the given set of restrictions.
- Entropy maximization was first proposed as a general inference procedure by Jaynes [Jaynes, 1957a, Jaynes, 1957b].
- It has historical roots in physics [Elsasser, 1937].
- It has been applied successfully in a remarkable variety of fields.

i [Berger, 1985].

Entropy Maximization

- Entropy has a direct relationship to information theory.
- It is a measure of the amount of uncertainty inherent in the probability distribution.
- The principle is to seek the prior distribution which maximizes entropy among all those distributions which satisfy the given set of restrictions.
- Entropy maximization was first proposed as a general inference procedure by Jaynes [Jaynes, 1957a, Jaynes, 1957b].
- It has historical roots in physics [Elsasser, 1937].
- It has been applied successfully in a remarkable variety of fields.

i [Berger, 1985].

Entropy Maximization

- Entropy has a direct relationship to information theory.
- It is a measure of the amount of uncertainty inherent in the probability distribution.
- The principle is to seek the prior distribution which maximizes entropy among all those distributions which satisfy the given set of restrictions.
- Entropy maximization was first proposed as a general inference procedure by Jaynes [Jaynes, 1957a, Jaynes, 1957b].
- It has historical roots in physics [Elsasser, 1937].
- It has been applied successfully in a remarkable variety of fields.

i [Berger, 1985].

Entropy Maximization

- Entropy has a direct relationship to information theory.
- It is a measure of the amount of uncertainty inherent in the probability distribution.
- The principle is to seek the prior distribution which maximizes entropy among all those distributions which satisfy the given set of restrictions.
- Entropy maximization was first proposed as a general inference procedure by Jaynes [Jaynes, 1957a, Jaynes, 1957b].
- It has historical roots in physics [Elsasser, 1937].
- It has been applied successfully in a remarkable variety of fields.

i [Berger, 1985].

Entropy Maximization

- Entropy has a direct relationship to information theory.
- It is a measure of the amount of uncertainty inherent in the probability distribution.
- The principle is to seek the prior distribution which maximizes entropy among all those distributions which satisfy the given set of restrictions.
- Entropy maximization was first proposed as a general inference procedure by Jaynes [Jaynes, 1957a, Jaynes, 1957b].
- It has historical roots in physics [Elsasser, 1937].
- It has been applied successfully in a remarkable variety of fields.

i [Berger, 1985].

Entropy Maximization

- Entropy has a direct relationship to information theory.
- It is a measure of the amount of uncertainty inherent in the probability distribution.
- The principle is to seek the prior distribution which maximizes entropy among all those distributions which satisfy the given set of restrictions.
- Entropy maximization was first proposed as a general inference procedure by Jaynes [Jaynes, 1957a, Jaynes, 1957b].
- It has historical roots in physics [Elsasser, 1937].
- It has been applied successfully in a remarkable variety of fields.

  i [Berger, 1985].

Entropy Maximization

- Entropy has a direct relationship to information theory.
- It is a measure of the amount of uncertainty inherent in the probability distribution.
- The principle is to seek the prior distribution which maximizes entropy among all those distributions which satisfy the given set of restrictions.
- Entropy maximization was first proposed as a general inference procedure by Jaynes [Jaynes, 1957a, Jaynes, 1957b].
- It has historical roots in physics [Elsasser, 1937].
- It has been applied successfully in a remarkable variety of fields.

  i [Berger, 1985].

Controversies

- The foundations of the principle is the entropy's unique properties as an uncertainty measure.
- To some, entropy's unique properties make it obvious that entropy maximization is the correct way to account for constraint information.
- To others, such an informal and intuitive justification yields plausibility but not proof — why maximize entropy; why not some other function?
- A more serious problem is that the maximizer may not exist.

i [Berger, 1985].

Controversies

- The foundations of the principle is the entropy's unique properties as an uncertainty measure.
- To some, entropy's unique properties make it obvious that entropy maximization is the correct way to account for constraint information.
- To others, such an informal and intuitive justification yields plausibility but not proof — why maximize entropy; why not some other function?
- A more serious problem is that the maximizer may not exist.

i [Berger, 1985].

Controversies

- The foundations of the principle is the entropy's unique properties as an uncertainty measure.
- To some, entropy's unique properties make it obvious that entropy maximization is the correct way to account for constraint information.
- To others, such an informal and intuitive justification yields plausibility but not proof — why maximize entropy; **why not some other function**?
- A more serious problem is that the maximizer may not exist.

i [Berger, 1985].

Controversies

- The foundations of the principle is the entropy's unique properties as an uncertainty measure.
- To some, entropy's unique properties make it obvious that entropy maximization is the correct way to account for constraint information.
- To others, such an informal and intuitive justification yields plausibility but not proof — why maximize entropy; **why not some other function**?
- A more serious problem is that the maximizer may not exist.

i [Berger, 1985].

Controversies

- The foundations of the principle is the entropy's unique properties as an uncertainty measure.
- To some, entropy's unique properties make it obvious that entropy maximization is the correct way to account for constraint information.
- To others, such an informal and intuitive justification yields plausibility but not proof — why maximize entropy; **why not some other function**?
- A more serious problem is that the maximizer may not exist.

i [Berger, 1985].

Edwin Thompson Jaynes

- The maximum entropy distribution "*is uniquely determined as the one which is maximally noncommittal with regard to missing information*"([Jaynes, 1957a, p. 623])
- It "*agrees with what is known, but expresses 'maximum uncertainty' with regard to all other matters, and thus leaves a maximum possible freedom for our final decision to be influenced by the subsequent sample data*"([Jaynes, 1968, p. 231]).
- Jaynes demonstrated that the maximum entropy distribution is equal to the frequency distribution that can be realized in the great number of ways.
- In [Shore and Johonson, 1980]: maximizing any function but entropy will lead to inconsistencies unless that function and entropy have identical maxima.

Edwin Thompson Jaynes

- The maximum entropy distribution "*is uniquely determined as the one which is maximally noncommittal with regard to missing information*"([Jaynes, 1957a, p. 623])
- It "*agrees with what is known, but expresses 'maximum uncertainty' with regard to all other matters, and thus leaves a maximum possible freedom for our final decision to be influenced by the subsequent sample data*"([Jaynes, 1968, p. 231]).
- Jaynes demonstrated that the maximum entropy distribution is equal to the frequency distribution that can be realized in the great number of ways.
- In [Shore and Johonson, 1980]: maximizing any function but entropy will lead to inconsistencies unless that function and entropy have identical maxima.

Edwin Thompson Jaynes

- The maximum entropy distribution "*is uniquely determined as the one which is maximally noncommittal with regard to missing information*"([Jaynes, 1957a, p. 623])
- It "*agrees with what is known, but expresses 'maximum uncertainty' with regard to all other matters, and thus leaves a maximum possible freedom for our final decision to be influenced by the subsequent sample data*"([Jaynes, 1968, p. 231]).
- Jaynes demonstrated that the maximum entropy distribution is equal to the frequency distribution that can be realized in the great number of ways.
- In [Shore and Johonson, 1980]: maximizing any function but entropy will lead to inconsistencies unless that function and entropy have identical maxima.

Edwin Thompson Jaynes

- The maximum entropy distribution "*is uniquely determined as the one which is maximally noncommittal with regard to missing information*"([Jaynes, 1957a, p. 623])
- It "*agrees with what is known, but expresses 'maximum uncertainty' with regard to all other matters, and thus leaves a maximum possible freedom for our final decision to be influenced by the subsequent sample data*"([Jaynes, 1968, p. 231]).
- Jaynes demonstrated that the maximum entropy distribution is equal to the frequency distribution that can be realized in the great number of ways.
- In [Shore and Johonson, 1980]: maximizing any function but entropy will lead to inconsistencies unless that function and entropy have identical maxima.

Kullback's Theorem

- Assume that the distribution density **Pr**(θ) satisfies

$$E[g_k(\theta)] = \mu_k, \quad i = 1, \cdots, m \tag{32}$$

where $g_k(\cdot)$ and μ_k are known functions and constants.

- Theorem

(Kullback's Theorem) If the maximum entropy distribution density $\hat{\pi}$ of θ subject to the constraints (32) exists, then

$$\hat{\pi}(\theta) = \frac{\pi_0(\theta)\mathbf{e}^{\sum_{k=1}^{m} \lambda_k g_k(\theta)}}{\int_\Theta \pi_0(\theta)\mathbf{e}^{\sum_{k=1}^{m} \lambda_k g_k(\theta)}d\theta}. \tag{33}$$

where λ_k are constants to be determined from the constraints in (32).

i [Berger, 1985].

Kullback's Theorem

- Assume that the distribution density $\mathbf{Pr}(\theta)$ satisfies

$$E[g_k(\theta)] = \mu_k, \quad i = 1, \cdots, m \tag{32}$$

where $g_k(\cdot)$ and μ_k are known functions and constants.

- Theorem

(Kullback's Theorem) If the maximum entropy distribution density $\hat{\pi}$ of θ subject to the constraints (32) exists, then

$$\hat{\pi}(\theta) = \frac{\pi_0(\theta)\mathbf{e}^{\sum_{k=1}^{m} \lambda_k g_k(\theta)}}{\int_\Theta \pi_0(\theta)\mathbf{e}^{\sum_{k=1}^{m} \lambda_k g_k(\theta)} d\theta}. \tag{33}$$

where λ_k are constants to be determined from the constraints in (32).

i [Berger, 1985].

Kullback's Theorem

- Assume that the distribution density $\mathbf{Pr}(\theta)$ satisfies

$$E[g_k(\theta)] = \mu_k, \quad i = 1, \cdots, m \tag{32}$$

  where $g_k(\cdot)$ and μ_k are known functions and constants.

- Theorem

  (Kullback's Theorem) If the maximum entropy distribution density $\hat{\pi}$ of θ subject to the constraints (32) exists, then

$$\hat{\pi}(\theta) = \frac{\pi_0(\theta)\mathbf{e}^{\sum_{k=1}^{m}\lambda_k g_k(\theta)}}{\int_\Theta \pi_0(\theta)\mathbf{e}^{\sum_{k=1}^{m}\lambda_k g_k(\theta)}d\theta}. \tag{33}$$

  where λ_k are constants to be determined from the constraints in (32).

i [Berger, 1985].

Kullback's Theorem

- Assume that the distribution density $\mathbf{Pr}(\theta)$ satisfies

$$E[g_k(\theta)] = \mu_k, \quad i = 1, \cdots, m \tag{32}$$

where $g_k(\cdot)$ and μ_k are known functions and constants.

- Theorem
(Kullback's Theorem) If the maximum entropy distribution density $\hat{\pi}$ of θ subject to the constraints (32) exists, then

$$\hat{\pi}(\theta) = \frac{\pi_0(\theta)\mathbf{e}^{\sum_{k=1}^{m}\lambda_k g_k(\theta)}}{\int_\Theta \pi_0(\theta)\mathbf{e}^{\sum_{k=1}^{m}\lambda_k g_k(\theta)}d\theta}. \tag{33}$$

where λ_k are constants to be determined from the constraints in (32).

i [Berger, 1985].

Proof

- By Lagrange's multiplier method, let

$$G(\pi) = -\int_\Theta \mathbf{Pr}(\theta)\log\frac{\mathbf{Pr}(\theta)}{\pi_0(\theta)}d\theta + \sum_{k=1}^m \lambda_k\left[E[g_k(\theta)] - \mu_k\right] + \mu\left[\int_\Theta \mathbf{Pr}(\theta)d\theta - 1\right].$$

- if the maximum entropy distribution density $\hat{\pi}$ of θ subject to the constraints (32) exists, we have

$$0 =< G'(\hat{\pi}), \varphi >= \int_\Theta \left[-\log\frac{\hat{\pi}(\theta)}{\pi_0(\theta)} - 1 + \sum_{k=1}^m \lambda_k g_k(\theta) + \mu\right]\cdot\varphi d\theta.$$

- Then

$$-\log\frac{\hat{\pi}(\theta)}{\pi_0(\theta)} - 1 + \sum_{k=1}^m \lambda_k g_k(\theta) + \mu = 0.$$

- Therefore

$$\hat{\pi}(\theta) = \pi_0(\theta)\cdot \mathbf{e}^{-1+\mu+\sum_{k=1}^m \lambda_k g_k(\theta)}$$

- Because $\int_\Theta \hat{\pi}(\theta)d\theta = 1$, it follows that

$$\mathbf{e}^{-1+\mu} = \frac{1}{\int_\Theta \pi_0(\theta)\mathbf{e}^{\sum_{k=1}^m \lambda_k g_k(\theta)}}.$$

Therefore, $\hat{\pi}$ is given by (33).

Proof

- By Lagrange's multiplier method, let

$$G(\pi) = -\int_\Theta \mathbf{Pr}(\theta)\log\frac{\mathbf{Pr}(\theta)}{\pi_0(\theta)}d\theta + \sum_{k=1}^{m}\lambda_k\left[E[g_k(\theta)] - \mu_k\right] + \mu\left[\int_\Theta \mathbf{Pr}(\theta)d\theta - 1\right].$$

- if the maximum entropy distribution density $\hat{\pi}$ of θ subject to the constraints (32) exists, we have

$$0 =< G'(\hat{\pi}), \varphi >= \int_\Theta \left[-\log\frac{\hat{\pi}(\theta)}{\pi_0(\theta)} - 1 + \sum_{k=1}^{m}\lambda_k g_k(\theta) + \mu\right]\cdot\varphi d\theta.$$

- Then

$$-\log\frac{\hat{\pi}(\theta)}{\pi_0(\theta)} - 1 + \sum_{k=1}^{m}\lambda_k g_k(\theta) + \mu = 0.$$

- Therefore

$$\hat{\pi}(\theta) = \pi_0(\theta)\cdot \mathbf{e}^{-1+\mu+\sum_{k=1}^{m}\lambda_k g_k(\theta)}$$

- Because $\int_\Theta \hat{\pi}(\theta)d\theta = 1$, it follows that

$$\mathbf{e}^{-1+\mu} = \frac{1}{\int_\Theta \pi_0(\theta)\mathbf{e}^{\sum_{k=1}^{m}\lambda_k g_k(\theta)}}.$$

Therefore, $\hat{\pi}$ is given by (33).

Proof

- By Lagrange's multiplier method, let

$$G(\pi) = -\int_\Theta \mathbf{Pr}(\theta)\log\frac{\mathbf{Pr}(\theta)}{\pi_0(\theta)}d\theta+\sum_{k=1}^m \lambda_k\left[E[g_k(\theta)]-\mu_k\right]+\mu[\int_\Theta \mathbf{Pr}(\theta)d\theta-1].$$

- if the maximum entropy distribution density $\hat{\pi}$ of θ subject to the constraints (32) exists, we have

$$0 =< G'(\hat{\pi}),\varphi >= \int_\Theta \left[-\log\frac{\hat{\pi}(\theta)}{\pi_0(\theta)} - 1 + \sum_{k=1}^m \lambda_k g_k(\theta) + \mu\right]\cdot\varphi d\theta.$$

- Then

$$-\log\frac{\hat{\pi}(\theta)}{\pi_0(\theta)} - 1 + \sum_{k=1}^m \lambda_k g_k(\theta) + \mu = 0.$$

- Therefore

$$\hat{\pi}(\theta) = \pi_0(\theta)\cdot \mathbf{e}^{-1+\mu+\sum_{k=1}^m \lambda_k g_k(\theta)}$$

- Because $\int_\Theta \hat{\pi}(\theta)d\theta = 1$, it follows that

$$\mathbf{e}^{-1+\mu} = \frac{1}{\int_\Theta \pi_0(\theta)\mathbf{e}^{\sum_{k=1}^m \lambda_k g_k(\theta)}}.$$

Therefore, $\hat{\pi}$ is given by (33).

Proof

- By Lagrange's multiplier method, let

$$G(\pi) = -\int_\Theta \mathbf{Pr}(\theta)\log\frac{\mathbf{Pr}(\theta)}{\pi_0(\theta)}d\theta+\sum_{k=1}^{m}\lambda_k\left[E[g_k(\theta)]-\mu_k\right]+\mu[\int_\Theta \mathbf{Pr}(\theta)d\theta-1].$$

- if the maximum entropy distribution density $\hat{\pi}$ of θ subject to the constraints (32) exists, we have

$$0 =< G'(\hat{\pi}),\varphi >= \int_\Theta\left[-\log\frac{\hat{\pi}(\theta)}{\pi_0(\theta)}-1+\sum_{k=1}^{m}\lambda_k g_k(\theta)+\mu\right]\cdot\varphi d\theta.$$

- Then

$$-\log\frac{\hat{\pi}(\theta)}{\pi_0(\theta)}-1+\sum_{k=1}^{m}\lambda_k g_k(\theta)+\mu = 0.$$

- Therefore

$$\hat{\pi}(\theta) = \pi_0(\theta)\cdot\mathbf{e}^{-1+\mu+\sum_{k=1}^{m}\lambda_k g_k(\theta)}$$

- Because $\int_\Theta \hat{\pi}(\theta)d\theta = 1$, it follows that

$$\mathbf{e}^{-1+\mu} = \frac{1}{\int_\Theta \pi_0(\theta)\mathbf{e}^{\sum_{k=1}^{m}\lambda_k g_k(\theta)}}.$$

Therefore, $\hat{\pi}$ is given by (33).

Proof

- By Lagrange's multiplier method, let

$$G(\pi) = -\int_\Theta \mathbf{Pr}(\theta)\log\frac{\mathbf{Pr}(\theta)}{\pi_0(\theta)}d\theta + \sum_{k=1}^m \lambda_k\left[E[g_k(\theta)] - \mu_k\right] + \mu\left[\int_\Theta \mathbf{Pr}(\theta)d\theta - 1\right].$$

- if the maximum entropy distribution density $\hat{\pi}$ of θ subject to the constraints (32) exists, we have

$$0 =< G'(\hat{\pi}), \varphi >= \int_\Theta \left[-\log\frac{\hat{\pi}(\theta)}{\pi_0(\theta)} - 1 + \sum_{k=1}^m \lambda_k g_k(\theta) + \mu\right] \cdot \varphi d\theta.$$

- Then

$$-\log\frac{\hat{\pi}(\theta)}{\pi_0(\theta)} - 1 + \sum_{k=1}^m \lambda_k g_k(\theta) + \mu = 0.$$

- Therefore

$$\hat{\pi}(\theta) = \pi_0(\theta)\cdot \mathbf{e}^{-1+\mu+\sum_{k=1}^m \lambda_k g_k(\theta)}$$

- Because $\int_\Theta \hat{\pi}(\theta)d\theta = 1$, it follows that

$$\mathbf{e}^{-1+\mu} = \frac{1}{\int_\Theta \pi_0(\theta)\mathbf{e}^{\sum_{k=1}^m \lambda_k g_k(\theta)}}.$$

Therefore, $\hat{\pi}$ is given by (33).

References I

- Berger, J. O. (1985).
  Statistical Decision Theory and Bayesian Analysis.
  Springer – Verlag, New York INC., New York, 2nd edition.
- Elsasser, W. M. (1937).
  On quantun measurements and the role of the uncertainty relations in statistical mechanics.
  Physical Review, 52:987 – 999.
- Jaynes, E. T. (1957a).
  Information theory and statistical mechanics i.
  Physical Review, 106:620 – 630.
- Jaynes, E. T. (1957b).
  Information theory and statistical mechanics ii.
  Physical Review, 108:171 – 190.

References II

- Jaynes, E. T. (1968).
Prior probabilities.
IEEE Transactions on Systems, Man, and Cybernetics, SSC-4:227 – 241.
- Mumford, D. (1994).
Pattern theory: a unifying perspective.
In Anthony, J. et al., editors, *First European Congress of Mathematics*, pages 187 – 224. Springer–Verlag.
- Shore, J. E. and Johonson, R. W. (1980).
Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy.
IEEE Transactions on Information Theory, IT-26(1):26 – 37.

References III

- Wrinkler, G. (1995).
  Image Analysis, Random Fields and Dynamic Monte Carlo Methods.
  Springer, Berlin–Heideberg.
- Zhang, X. T. and Cheng, H. F. (1994).
  Bayesian Statistical Inference.
  Science Press, Beijing.
  in Chinese.