Contents

Chapter 1

Introduction

1. Overview of model based image processing
   (a) Direct approach to image processing
      i. One step approach ⇒ easy
   (b) Model based approach
      i. Model of both unknown and system ⇒ probability and statistics
      ii. Estimation of unknown quantity from model ⇒ inverse methods
      iii. Two step approach ⇒ more difficult
   (c) Image modeling
   (d) Properties of images that make them challenging to model
      i. not Gaussian
      ii. 2-D ⇒ no natural causality
   (e) Issues to cover
      i. Only consider discrete-parameter (time and space) models
      ii. Random variables being modeled can be continuous Gaussian, continuous non-Gaussian, and discrete.
      iii. Indexed field can be 1-D and 2 or more-D
      iv. Dependencies can be causal or non-causal
   (f) Order of presentation
      i. continuous Gaussian, 1-D, causal models (1D AR)
      ii. continuous Gaussian, 1-D, non-causal models (1D GMRFs)
      iii. continuous Gaussian, 2-D, causal models (2D AR Processes)

iv. continuous Gaussian, 2-D, non-causal models (2D GMRFs)
v. continuous non-Gaussian, 1 and 2-D, non-causal models (continuously valued MRFs)
vi. discrete, 1-D, causal models (Markov Chains)
vii. discrete, 1-D and 2-D, non-causal models (discrete valued MRFs)

Conventions:

1. Examples should be numbered 1 to N within a **Chapter**.

```
\bigskip
\noindent
{\em Example~\ref{chapter:NameOfChapter}.1:}
This is the example.
\bigskip
```

2. Properties should be numbered 1 to N within a **Chapter**.

```
\noindent
{\bf Property \ref{chapter:NameOfChapter}.1}:
{\em Name of Property} -
This is the property
```

Chapter 2

Probability, Estimation, and Random Processes

This chapter provides a quick review of probability, estimation, and random processes. It also introduces many of the basic tools that will be needed for modeling physical data.

Two important themes throughout this book will be the use of both frequentist and Bayesian methods for the estimation of unknown quantities. We will illustrate the frequentist approach using the ubiquitous **maximum likelihood** (ML) estimator, and we will illustrate the Bayesian approach using the **minimum mean squared error** (MMSE) estimator. Our goal is to show that each approach, frequentist or Bayesian, has its potential advantages, and the best choice tends to depend on the particular situation. Generally, we will see that the ML estimator tends to be a good choice when the ratio of the amount of data to the number of unknowns is $>> 1$, whereas the MMSE estimator tends to be a good choice when this ratio approaches or is less than 1.

2.1 Random Variables and Expectation

Let X be a real valued random variable with **cumulative distribution function** (CDF) given by

$$F(t) \triangleq P\{X \leq t\}$$

where $P\{X \leq t\}$ is the probability of the event that X is less than or equal to t. If $F(t)$ is an absolutely continuous function, then it will have an associated

probability density function (PDF), $p(t)$, such that

$$F(t) = \int_{-\infty}^{t} p(\tau)d\tau \ .$$

For most physical problems, it is reasonable to assume that such a density function exists, and when it does then it is given by

$$p(t) = \frac{dF(t)}{dt} \ .$$

Any function[1] of a random variable is itself a random variable. So for example, if we define the new random variable $Y = g(X)$ where $g : \Re \to \Re$, then Y will also be a random variable.

Armed with the random variable X and its distribution, we may define the expectation as

$$\mathrm{E}[X] \stackrel{\triangle}{=} \int_{-\infty}^{\infty} \tau dF(\tau) = \int_{-\infty}^{\infty} \tau p(\tau)d\tau \ .$$

The first integral form is known as a Lebesgue-Stieltjes integral and is defined even when the probability density does not exist. However, the second integral containing the density is perhaps more commonly used.

The expectation is a very basic and powerful tool which exists under very general conditions.[2] An important property of expectation, which directly results from it definition as an integral, is linearity.

Property 2.1: *Linearity of expectation* - For all random variables X and Y,

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y] \ .$$

Of course, it is also possible to specify the distribution of groups of random variables. So let $X_1, \cdots, X_n$ be n random variables. Then we may specify the joint distribution of these n random variable via the n-dimensional CDF given by

$$F(t_1, \cdots, t_n) = P\{X_1 \leq t_1, \cdots, X_n \leq t_n\} \ .$$

[1] Technically, this can only be a Lebesgue-measurable function; but in practice, measurability is a reasonable assumption in any physically meaningful situation.

[2] In fact, for any positive random variable, X, the $E[X]$ takes on a well-defined value on the extended real line of $(-\infty, \infty]$. In addition, whenever $E[|X|] < \infty$, then $E[X]$ is real valued and well-defined. So we will generally assume that $E[|X|] < \infty$ for all random variables we consider.

In this case, there is typically an associated n-dimensional PDF, $p(t_1, \cdots, t_n)$, so that

$$F(t_1, \cdots, t_n) = \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_n} p(\tau_1, \cdots, \tau_n) d\tau_1 \cdots d\tau_n \ .$$

Again, any function of the vector $X = (X_1, \cdots, X_n)$ is then a new random variable. So for example, if $Y = g(X)$ where $g : \Re^n \rightarrow \Re$, then Y is a random variable, and we may compute its expectation as

$$\mathrm{E}[Y] = \int_{\Re^n} g(\tau_1, \cdots, \tau_n) p(\tau_1, \cdots, \tau_n) d\tau_1 \cdots d\tau_n \ .$$

If we have a finite set of random variables, $X_1, \cdots, X_n$, then we say the random variables are **jointly independent** if we can factor the CDF (or equivalently the PDF if it exists) into a product form,

$$F(t_1, \cdots, t_n) = F_{X_1}(t_1) \cdots F_{X_n}(t_n) \ ,$$

where $F_{X_k}(t_k)$ denotes the CDF for the random variable X_k. This leads to another important property of expectation.

Property 2.2: *Expectation of independent random variables* - If $X_1, \cdots, X_n$ are a set of jointly independent random variables, then we have that

$$\mathrm{E}\left[\prod_{k=1}^{n} X_k\right] = \prod_{k=1}^{n} \mathrm{E}[X_k] \ .$$

So when random variables are jointly independent, then expectation and multiplication can be interchanged. Perhaps surprisingly, pair wise independence of random variables does not imply joint independence.

One of the most subtile and important concepts in probability is conditional probability and conditional expectation. Let X and Y be two random variables with a joint CDF given by $F(x, y)$ and conditional CDF given by $F_y(y) = \lim_{x \rightarrow \infty} F(x, y)$. The conditional CDF of X given Y is then any function, $F(x|y)$, which solves the equation

$$F(x, y) = \int_{-\infty}^{y} F(x|t) dF_y(t) dt \ .$$

At least one solution to this equation is guaranteed to exist by an application of the famous Radon-Nikodym theorem.[3] So the conditional CDF, $F(x|y)$, is

[3]Interesting, the solution is generally not unique, so there can be many such functions $F(x|y)$. However, it can be shown that the resulting conditional expactations resulting from any two such functions are almost surely equal.

guaranteed to exist. However, this definition of the conditional CDF is somewhat unsatisfying because it does not specify how to construction $F(x|y)$.

Fortunately, the conditional CDF can be calculated in most practical situations as

$$F(x|y) = \frac{dF(x,y)}{dy}\left(\frac{dF(\infty,y)}{dy}\right)^{-1} \quad .$$

More typically, we will just work with probability density functions so that

$$\begin{aligned} p(x,y) &= \frac{d^2F(x,y)}{dxdy} \\ p_y(y) &= \frac{dF(\infty,y)}{dy} \ . \end{aligned}$$

Then the PDF of X given $Y = y$ is given by

$$p(x|y) = \frac{p(x,y)}{p_y(y)} \ .$$

We may now ask what is the conditional expectation of X given Y? However, the answer to this question has a subtile twist because what we are given, Y, is itself a random variable. So the conditional expectation of X given Y is given by

$$\mathrm{E}[X|Y] = \int_{-\infty}^{\infty} x dF(x|Y) \ ,$$

or alternatively using the conditional CDF, it is given by

$$\mathrm{E}[X|Y] = \int_{-\infty}^{\infty} x p(x|Y) dx \ .$$

Notice that in both cases, the integral expression for $E[X|Y]$ is a function of the random variable Y. This means that the conditional expectation is itself a random variable!

The fact that the conditional expectation is a random variable is very useful. For example, consider the so-called indicator function denoted by

$$I_A(X) \stackrel{\triangle}{=} \begin{cases} 1 & \text{if } X \in A \\ 0 & \text{otherwise} \end{cases} \quad .$$

Using the indicator function, we may define the conditional probability of an event, $\{X \in A\}$, given Y as

$$P\{X \in A|Y\} = \mathrm{E}[I_A(X)|Y] \ .$$

Conditional expectation also has many useful properties. We list some below.

Property 2.3: *Filtration property of conditional expectation* - For all random variables X, Y, and Z

$$\mathrm{E}[\,\mathrm{E}[X|Y,Z]\,|Y] = \mathrm{E}[X|Y] \ .$$

A special case of the filtration property is that $\mathrm{E}[X] = \mathrm{E}[\,\mathrm{E}[X|Y]]$. This can be a useful relationship because sometimes it is much easier to evaluate an expectation in two steps, by first computing $\mathrm{E}[X|Z]$ and then taking the expectation over Z.

Property 2.4: *Conditional expectation of known quantities* - For all random variables X, Y, and Z, and for all functions $f(\cdot)$

$$\mathrm{E}[f(X)Z|X,Y] = f(X)\mathrm{E}[Z|X,Y] \ ,$$

which implies that $\mathrm{E}[X|X,Y] = X$.

This property states the obvious fact that if we are given knowledge of X, then the value of $f(X)$ is known. For example, if we are told that the height of a building is 100 ft, then its expected height is simply 100 ft. Notice that since $\mathrm{E}[X|Y] = f(Y)$ for some function $f(\cdot)$, Property 2.4 may be used to show that

$$\begin{aligned} \mathrm{E}[\mathrm{E}[X|Y]\,|Y,Z] &= \mathrm{E}[f(Y)|Y,Z] \\ &= f(Y) \\ &= \mathrm{E}[X|Y] \ , \end{aligned}$$

Therefore, we know that for all random variables, X, Y, and Z

$$\mathrm{E}[\,\mathrm{E}[X|Y,Z]\,|Y] = \mathrm{E}[X|Y] = \mathrm{E}[\,\mathrm{E}[X|Y]\,|Y,Z] \ .$$

Finally, we should introduce some common distributions and their associated notation. A widely used distribution for a random variable, X, is the **Gaussian distribution** denoted by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

where μ and σ^2 are the mean and variance; so that

$$\begin{aligned} \mathrm{E}[X] &= \mu \\ \mathrm{E}\big[(X-\mu)^2\big] &= \sigma^2 \end{aligned}$$

We use the notation $X \sim N(\mu, \sigma^2)$ to indicate that X is a Gaussian random variable with mean μ and variance σ^2.

A generalization of the Gaussian distribution is the **multivariate Gaussian distribution**. We can use vector-matrix notation to compactly represent the distributions of multivariate Gaussian random vectors. Let $X \in \Re^p$ denote a p-dimensional random column vector, and X^t denote its transpose. Then the PDF of X is given by

$$p(x) = \frac{1}{(2\pi)^{p/2}} |R|^{-1/2} \exp\left\{ -\frac{1}{2}(x-\mu)^t R^{-1} (x-\mu) \right\} ,$$

where $\mu \in \Re^p$ and $R \in \Re^{p \times p}$ are the mean vector and covariance matrix; so that

$$\begin{aligned} \mathrm{E}[X] &= \mu \\ \mathrm{E}\big[(X-\mu)(X-\mu)^t\big] &= R \ . \end{aligned}$$

Generalizing our previous notation, we write $X \sim N(\mu, R)$ to indicate that X is a multivariate Gaussian distribution with mean μ and covariance R.

Jointly Gaussian random variables have many useful properties, one of which is listed below.

Property 2.5: *Linearity of conditional expectation for Gaussian random variables* - If $X \in \Re^n$ and $Y \in \Re^p$ are jointly Gaussian random vectors, then

$$\mathrm{E}[X|Y] = AY + b \ ,$$

where $A \in \Re^{n \times p}$ is a matrix and $b \in \Re$ is a scalar constant. Furthermore, if X and Y are zero mean, then $b = 0$ and the conditional expectation is a linear function of Y.

The **Bernoulli distribution** is an example of a discrete distribution which we will find very useful in image and signal modeling applications. If X has a Bernoulli distribution, then

$$\begin{aligned} P\{X = 1\} &= \theta \\ P\{X = 0\} &= 1 - \theta \ . \end{aligned}$$

Since X takes on discrete values, its CDF has discontinuous jumps. The easiest solution to this problem is to represent the distribution with a **probability mass function (PMF)** rather than a probability density function.

The PMF of a random variable or random vector specifies the actual probability of the random quantity taking on a specific value. If X has a Bernoulli distribution, then its PMF is given by

$$p(x) = (1-\theta)^{1-x}\theta^x \ .$$

In some cases, it is more convenient to use the following different, but equivalent form of the PMF given by

$$p(x) = (1-\theta)\delta(x-0) + \theta\delta(x-1) \ ,$$

where the function $\delta(x)$ is 1 when its argument is 0, and 0 otherwise.

It is often the case that a series of n experiments are perform, each of which results in a measurement with the same distribution. In this case, we might have a set of jointly independent random variables, $X_1, \cdots, X_n$, which are **independent and identically distributed (i.i.d.)**. If the random variables are i.i.d. with distribution $N(\mu, \sigma^2)$, then their joint distribution is given by

$$p(x) = p(x_1, \cdots, x_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2\right\} \ .$$

Notice that for notational simplicity, x is used to denote the entire vector of measurements.

Example 2.1: Let $X_1, \cdots, X_n$ be n i.i.d. Gaussian random vectors each with multivariate Gaussian distribution given by $N(\mu, R)$, where $\mu \in \Re^p$ and $R \in \Re^p \times \Re^p$. We would like to know the density function for the entire observation. In order to simplify notation, we will represent the data as a matrix

$$X = [X_1, \cdots, X_n] \ .$$

Since the n observations, $X_1, \cdots, X_n$, are i.i.d., we know we can represent the density of X as the product of densities for each vector X_k.

$$\begin{aligned} p(x) &= \prod_{k=1}^{n} \frac{1}{(2\pi)^{p/2}}|R|^{-1/2} \exp\left\{-\frac{1}{2}(x_k-\mu)^t R^{-1}(x_k-\mu)\right\} \\ &= \frac{1}{(2\pi)^{np/2}}|R|^{-n/2} \exp\left\{-\frac{1}{2}\sum_{k=1}^{n}(x_k-\mu)^t R^{-1}(x_k-\mu)\right\} \end{aligned}$$

We may further simplify this expression, by defining two new sample statistics.

$$\begin{aligned} b &= \sum_{k=1}^{n} x_k \\ S &= \sum_{k=1}^{n} x_k x_k^t \end{aligned}$$

A **statistic** is any function of the observed data, but typically, statistics summarize the data in some useful way. Using these two statistics, the PDF of X may be written as

$$p(x) = \frac{1}{(2\pi)^{np/2}} |R|^{-n/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}\left\{ S R^{-1} \right\} + b^t R^{-1} \mu - \frac{n}{2} \mu^t R^{-1} \mu \right\} \ , \quad (2.1)$$

where $\mathrm{tr}\{A\}$ denotes the trace of the matrix A.

Example 2.2: Let $X_1, \cdots, X_n$ be n i.i.d. Bernoulli random vectors, each with parameter θ. We would like to know the probability mass function for the entire observation. Again, since the n observations are i.i.d., we know that the PMF is given by

$$\begin{aligned} p(x) &= \prod_{k=1}^{n} \left\{ (1-\theta)^{1-x_k} \theta^{x_k} \right\} \\ &= (1-\theta)^{n-N_1} \theta^{N_1} \end{aligned}$$

where $N_1 = \sum_{k=1}^{n} x_k$ is a statistic which counts the number of random variables X_k which are equal to 1.

2.2 Frequentist Estimation and the ML Estimator

Once we have decided on a model for our data, our next step is usually to estimate some information related to the data. There are two general frameworks for estimating unknown information from data. We will refer to these two general frameworks as the frequentist and Bayesian approaches. We will see that each approach is valuable, and the key is understanding what combination of approaches is best for a particular application.

In the frequentist approach, one treats the unknown quantity as a deterministic, but unknown parameter vector, $\theta \in \Omega$. So for example, after we observe the random vector $Y \in \Re^n$, then our objective is to use Y to estimate the unknown scalar or vector θ. In order to formulate this problem, we will assume that the vector Y has a PDF given by $p_\theta(y)$ where θ parameterizes a family of density functions for Y. We may then use this family of distributions to determine a function, $T : \Re^n \rightarrow \Omega$, that can be used to compute an estimate of the unknown parameter as

$$\hat{\theta} = T(Y) \ .$$

Notice, that since $T(Y)$ is a function of the random vector Y, the estimate, $\hat{\theta}$, is a random vector. The mean of the estimator, $\bar{\theta}$, can be computed as

$$\bar{\theta} = \mathrm{E}_\theta\Big[\hat{\theta}\Big] = \mathrm{E}_\theta[T(Y)] = \int_{\Re^n} T(y) p_\theta(y) dy \ .$$

The difference between the mean of the estimator and the value of the parameter is known as the **bias** and is given by

$$\mathrm{bias}_\theta = \bar{\theta} - \theta \ .$$

Similarly, the variance of the estimator is given by

$$\mathrm{var}_\theta = \mathrm{E}_\theta\left[\left(\hat{\theta} - \bar{\theta}\right)^2\right] \ ,$$

and it is easily shown that the **mean squared error (MSE)** of the estimate is then given by

$$\mathrm{MSE}_\theta = \mathrm{E}_\theta\left[\left(\hat{\theta} - \theta\right)^2\right] = \mathrm{var}_\theta + (\mathrm{bias}_\theta)^2 \ .$$

Since the bias, variance, and MSE of the estimator will depend on the specific value of θ, it is often unclear precisely how to compare the accuracy of different estimators. Even estimators that seem quite poor may produce small or zero error for certain values of θ. For example, consider the estimator which is fixed to the value $\hat{\theta} = 1$, independent of the data. This would seem to be a very poor estimator, but it has an MSE of 0 when $\theta = 1$.

An estimator is said to be **consistent** if for all $\theta \in \Omega$, the MSE of the estimator goes to zero as the number of independent data samples, n, goes to

infinity. If an estimator is not consistent, this means that even with arbitrarily large quantities of data, the estimate will not approach the true value of the parameter. Consistency would seem to be the least we would expect of an estimator, but we will later see that even some very intuitive estimators are not always consistent.

From this we can see that the trick is to select an estimator which has uniformly low bias and/or variance for all values of θ. This is not always possible, but when it is we have names for such estimators. For example, $\hat{\theta}$ is said to be an **unbiased estimator** if for all values of θ the bias is zero, i.e. $\theta = \bar{\theta}$. If in addition, for all values of θ, the variance of an estimator is less than all other unbiased estimators, then we say that the estimator is a **uniformly minimum variance unbiased (UMVU)** estimator.

There are many excellent estimators that have been proposed through the years for many different types of problems. However, one very widely used Frequentist estimator is known as the **maximum likelihood (ML)** estimator given by

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta \in \Omega} p_\theta(Y) \\ &= \arg\max_{\theta \in \Omega} \log p_\theta(Y) \ . \end{aligned}$$

where the notation "arg max" denotes the value of the argument that achieves the global maximum of the function. Notice that these formulas for the ML estimate actually use the random variable Y as an argument to the density function $p_\theta(y)$. This implies that $\hat{\theta}$ is a function of Y, which in turn means that $\hat{\theta}$ is a random variable.

When the density function, $p_\theta(y)$, is a continuously differentiable function of θ, then a necessary condition when computing the ML estimate is that the gradient of the likelihood is zero.

$$\nabla_\theta \, p_\theta(Y)|_{\theta=\hat{\theta}} = 0 \ .$$

While the ML estimate is generally not unbiased, it does have a number of desirable properties. It can be shown that under realitively mild technical conditions, the ML estimate is both consistent and **asymptotically efficient**. This means that the ML estimate attains the **Cramér-Rao bound** asymptotically as the number of independent data samples, n, goes to infinity. Since this Cramér-Rao bound is a lower bound on the variance of an

unbiased estimator, this means that the ML estimate is about as good as any unbiased estimator can be. So if one has no idea what values of θ may occur, and needs to guarantee good performance for all cases, then the ML estimator is usually a good choice.

Example 2.3: Let $\{Y_k\}_{k=1}^n$ be i.i.d. random variables with Gaussian distribution $N(\theta, 1)$ and unknown mean parameter θ. For this case, the logarithm of the density function is given by

$$\log p_\theta(Y) = -\frac{1}{2\sigma^2}\sum_{k=1}^{n}(Y_k - \theta)^2 - \frac{n}{2}\log\left(2\pi\sigma^2\right) \ .$$

Differentiating the log likelihood results in the following expression.

$$\left.\frac{d\log p_\theta(Y)}{d\theta}\right|_{\theta=\hat{\theta}} = \frac{1}{\sigma^2}\sum_{k=1}^{n}(Y_k - \hat{\theta}) = 0$$

From this we obtain the ML estimate for θ.

$$\hat{\theta} = \frac{1}{n}\sum_{k=1}^{n} Y_k$$

Notice that, by the following argument, this particular ML estimate is unbiased.

$$\mathrm{E}_\theta\left[\hat{\theta}\right] = \mathrm{E}_\theta\left[\frac{1}{n}\sum_{k=1}^{n} Y_k\right] = \frac{1}{n}\sum_{k=1}^{n}\mathrm{E}_\theta[Y_k] = \frac{1}{n}\sum_{k=1}^{n}\theta = \theta$$

Moreover, for this special case, it can be shown that the ML estimator is also the UMVU estimator [1].

Example 2.4: Again, let $X_1, \cdots, X_n$ be n i.i.d. Gaussian random vectors each with multivariate Gaussian distribution given by $N(\mu, R)$, where $\mu \in \Re^p$. We would like to compute the ML estimate of the parameter vector $\theta = [\mu, R]$. Using the statistics b and S from Example 2.1, we can define the sample mean and covariance given by

$$\begin{aligned}\hat{\mu} &= b/n \\ \hat{R} &= (S/n) - \hat{\mu}\hat{\mu}^t \ ,\end{aligned}$$

then by manipulation of equation (2.1) the probability distribution of X can be written in the form

$$p(x|\theta) = \frac{1}{(2\pi)^{np/2}}|R|^{-n/2}\exp\left\{-\frac{n}{2}\mathrm{tr}\left\{\hat{R}R^{-1}\right\} - \frac{n}{2}(\hat{\mu}-\mu)^t R^{-1}(\hat{\mu}-\mu)\right\} \quad (2.2)$$

Maximizing this expression with respect to θ then results in the ML estimate of the mean and covariance given by

$$\arg \max_{[\mu,R]} p(x|\mu, R) = [\hat{\mu}, \hat{R}] \ .$$

So from this we see that the ML estimate of the mean and covariance of a multivariate Gaussian distribution is simply given by the sample mean and covariance of the observations.

2.3 Bayesian Estimation and the MMSE Estimator

Since the ML estimator is asymptotically efficient, it is guaranteed to do well for all values of the unknown parameter θ, as the amount of data grows towards infinity. So it might seem that nothing more can be done.

However, Bayesian methods attempt to exceed the accuracy of Frequentist estimators, such as the ML estimate, by making assumptions about values of the parameters that are most likely to occur. In practice, we will see that Bayesian estimation methods are most useful when the amount of data is relatively small compared to the dimension of the unknown parameter.

In the Bayesian approach, we model the unknown quantity as a random, rather than deterministic, vector. In order to emphasize this distinction, we use the random variable X to denote the unknown quantity in Bayesian estimation, as opposed to the unknown parameter, θ, used in frequentist approach. As before, the estimator is then a function of the observations with the form

$$\hat{X} = T(Y) \ ,$$

and once again the estimate, $\hat{X}$, is a random vector.

The good news with Bayesian estimators is that once we make the assumption that X is random, then the design of estimators is conceptually straight forward, and the MSE of our estimator can be reduced by focusing on the estimation values of X that are most likely to occur. However, the bad news is that when we assume that X is a random vector, we must select some distribution for it. In some cases, the distribution for X may be easy to select, but in many cases, it may be very difficult to select a reasonable distribution or model for X. And in some cases, a poor choice of this prior distribution can severely degrade the accuracy of the estimator.

In any case, let us assume that the joint distribution of both Y and X is known, and let $C(x, \hat{x}) \geq 0$ be the **cost** of choosing $\hat{x}$ as our estimate if x is the correct value of the unknown. If our estimate is always perfect, then $C(x, x) = 0$, and the cost is zero. Of course, in most cases our estimates will not be perfect, so our objective is to select an estimator that minimizes the expected cost. For a given value of $X = x$, the expected cost is given by

$$\begin{aligned} \bar{C}(x) &= \mathrm{E}[C(x, T(Y))|X = x] \\ &= \int_{\Re^n} C(x, T(y)) p_{y|x}(y|x) dy \end{aligned}$$

where $p_{y|x}(y|x)$ is the conditional distribution of the data, Y, given the unknown, X. Here again, the cost is a function of the unknown, X, but we can remove this dependency by taking the expectation to form what is known as the **Bayes' risk**.

$$\begin{aligned} \text{Risk} &= \mathrm{E}[C(X, T(Y))] \\ &= \mathrm{E}[\,\mathrm{E}[C(X, T(Y))|X]] \\ &= \mathrm{E}\big[\bar{C}(X)\big] \\ &= \int_{\Omega} \bar{C}(x) p_x(x) dx \end{aligned}$$

Notice, that the risk is now no longer a function of X. This stands in contrast to the Frequentist case, where the MSE of the estimator was a function of θ. However, the price we pay for removing this dependency is that the risk now depends critically on the choice of the density function $p_x(x)$, which specifies the distribution on the unknown X. The distribution of X is known as the **prior distribution** since it is the distribution that X is assumed to have prior to the observation of any data.

Of course, the choice of a specific cost function will determine the form of the resulting best estimator. For example, it can be shown that when the cost is squared error, so that $C(x, \hat{x}) = |x - \hat{x}|^2$, then the optimal estimator is given by the conditional expectation

$$\hat{X}_{MMSE} = \mathrm{E}[X|Y] \ ,$$

and when the cost is absolute error, so that $C(x, \hat{x}) = |x - \hat{x}|$, then the optimal estimator is given by the conditional median.

$$\int_{-\infty}^{\hat{X}_{median}} p_{x|y}(x|y) dx = \frac{1}{2}$$

Another important case, that will be used extensively in this book, is when the cost is given by

$$C(x, \hat{x}) = \delta\left(x - \hat{x}\right) \ .$$

Here, the cost is only zero when $x = \hat{x}$, and otherwise the cost is fixed at 1. This is a very pessimistic cost function since it assigns the same value whether the error is very small or very large. The optimal solution for this cost function is given by the so-called **maximum a posteriori (MAP)** estimate,

$$\hat{X}_{MAP} \ = \ \arg\max_{x \in \Omega} p_{x|y}(x|y) \ ,$$

where Ω is the set of feasible values for X, and the conditional distribution $p_{x|y}(x|y)$ is known as the **posterior distribution** because it specifies the distribution of the unknown X given the observation Y.

Example 2.5: Let X be a Gaussian random variable with distribution $N(0, \sigma_x^2)$, and let W be a Gaussian variables with distribution $N(0, \sigma_w^2)$. Furthermore, let $Y = X + W$. Our task is then to estimate X from the observations of Y.

Using the definition of condition probability, we may express the joint distribution of X and Y as

$$p_{y,x}(y, x) = p_{x|y}(x|y)p_y(y) = p_{y|x}(y|x)p_x(x) \ .$$

Reorganizing terms results in **Bayes' rule**, which is the source of the terminology Bayesian estimation.

$$p_{x|y}(x|y) = \frac{p_{y|x}(y|x)p_x(x)}{p_y(y)}$$

One approach to computing the posterior distribution is to directly evaluate Bayes' rule. However, this is somewhat cumbersome in this case, and for some problems, it will actually be impossible. An alternative approach is to consider the posterior as

$$p_{x|y}(x|y) = \frac{1}{z_y} p_{y,x}(y, x) \tag{2.3}$$

where z_y is a normalizing constant or **partition function** for our problem. Since the posterior distribution is a probability density in x, the right side

of equation (2.3) must integrate to 1 as a function of x. This means that z_y is uniquely specified as the normalizing constant of the distribution. Notice, that the partition function may depend on y, but it may not depend on x.

Using this approach, we first evaluate the conditional data distribution, $p_{y|x}(y|x)$ and prior distribution, $p_x(x)$, for our problem. The prior is assumed to be

$$p_x(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left\{-\frac{1}{2\sigma_x^2}x^2\right\} \ ,$$

and the conditional density of the data is given by

$$p_{y|x}(y|x) = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left\{-\frac{1}{2\sigma_w^2}(y-x)^2\right\} \ .$$

This means the joint distribution of X and Y can be written in the form

$$\begin{aligned} p_{x|y}(x|y) &= \frac{1}{z_y}\, p_{y|x}(y|x)p_x(x) \\ &= \frac{1}{z'_y}\, \exp\left\{-\frac{1}{2\sigma_w^2}(y-x)^2 - \frac{1}{2\sigma_x^2}x^2\right\} \\ &= \frac{1}{z''_y}\, \exp\left\{a(x-b)^2\right\} \ , \end{aligned} \tag{2.4}$$

where a and b are two appropriately chosen parameters. As with the partition function z_y, the parameters a and b can be functions of y, but they can not be functions of x. Also notice that in general z_y, z'_y, and z''_y are not equal, but in each case, the partition function collects together all the multiplicative factors in the expression that do not depend on x.

We can solve for the values of the parameters a and b by taking the first two derivatives of the logarithms of the expressions and setting the derivatives to be equal. For the first derivative, this results in the following calculation.

$$\begin{aligned} \frac{d}{dx}\log\left(\frac{1}{z'_y}\, \exp\left\{-\frac{1}{2\sigma_w^2}(y-x)^2 - \frac{1}{2\sigma_x^2}x^2\right\}\right) &= \frac{d}{dx}\log\left(\frac{1}{z''_y}\, \exp\left\{a(x-b)^2\right\}\right) \\ \frac{d}{dx}\left(-\frac{1}{2\sigma_w^2}(y-x)^2 - \frac{1}{2\sigma_x^2}x^2 - \log z'_y\right) &= \frac{d}{dx}\left(a(x-b)^2 - \log z''_y\right) \\ -\frac{1}{\sigma_w^2}(x-y) - \frac{1}{\sigma_x^2}x &= 2a(x-b) \end{aligned}$$

Differentiating this expression a second time yields the result that

$$a = -\frac{1}{2}\left(\frac{1}{\sigma_w^2} + \frac{1}{\sigma_x^2}\right) \ ,$$

and solving for b yields

$$b = \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_w^2}\right) y \ .$$

In fact, looking at the expression of equation (2.4), one can see that the parameters b and a specify the posterior mean and variance, respectively. More specifically, if we define $\mu_{x|y}$ to be the mean of the posterior density and $\sigma_{x|y}^2$ to be its variance, then we have the relationships that

$$\begin{aligned} -\frac{1}{2\sigma_{x|y}^2} &= a = -\frac{1}{2}\left(\frac{1}{\sigma_w^2} + \frac{1}{\sigma_x^2}\right) = -\frac{1}{2}\left(\frac{\sigma_w^2 + \sigma_x^2}{\sigma_w^2 \sigma_x^2}\right) \\ \mu_{x|y} &= b = \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_w^2}\right) y \ . \end{aligned}$$

Solving for the conditional mean and variance, we can then write an explicit expression for the conditional distribution of x given y as

$$p_{x|y}(x|y) = \frac{1}{\sqrt{2\pi}\sigma_{x|y}} \exp\left\{-\frac{1}{2\sigma_{x|y}^2}(x - \mu_{x|y})^2\right\} \ ,$$

where

$$\begin{aligned} \mu_{x|y} &= \frac{\sigma_x^2}{\sigma_x^2 + \sigma_w^2} y \\ \sigma_{x|y}^2 &= \frac{\sigma_x^2 \sigma_w^2}{\sigma_x^2 + \sigma_w^2} \ . \end{aligned}$$

At this point, there are two interesting observations to make, which are both particular to the case of jointly Gaussian random variables. First, the posterior mean is a linear function of the data y; and second, the posterior variance is not a function of y. These two properties will not hold for more general distributions.

From this posterior distribution, it is now easy to compute various Bayesian estimators. The MMSE estimator for this case is the conditional expectation

which is given by

$$\hat{X}_{MMSE} = E\left[X|Y\right] = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_w^2} Y \ .$$

Because the Gaussian posterior distribution is both a symmetric and unimodal function of x, the conditional mean, conditional median, and MAP estimates will be the same in this case.

$$\hat{X}_{median} = \hat{X}_{MMSE} = \hat{X}_{MAP}$$

However, for more general distributions, these estimators will differ.

2.4 Discrete-Time Random Processes

NEED TO WORK ON THIS SECTION

Stationarity is closely related to homogeneity in that it captures another aspect of time invariance for random processes. There are two specific forms of stationary which we define below.

> *Definition:* A 1-D discrete-time random process X_n is said to be **strict sense stationary** if all the finite dimensional distributions of X_n and $Y_n = X_{n-k}$ are the same for any constant k.
>
> *Definition:* A 1-D discrete-time random process X_n is said to be **second order** if for all n and k, $\mathrm{E}[|X_n X_k|] < \infty$ i.e. all second order moments are finite.
>
> *Definition:* A 1-D second order discrete-time random process X_n is said to be **wide sense stationary** if for all n and k, $\mathrm{E}[X_n] = \mu$ and $\mathrm{E}[(X_n - \mu)(X_k - \mu)] = R(n-k)$ where μ is a scalar constant and $R(n)$ is a scalar function of n.

Reversibility is another important invariance property which holds when a random process, X_n, and its time-reverse, X_{-n}, have the same characteristics.

> *Definition:* A 1-D discrete-time random process X_n is said to be **reversible** if all the finite dimensional distributions of the X_n and $Y_n = X_{-n}$ are the same.

Gaussian random processes have some properties which make them particularly easy to analyze. In particular, the time autocorrelation for Gaussian random processes provides all the information required to specify the density function of any finite dimension subset the random process samples. In practice, this means that the distribution of any Gaussian random process is fully specified by its mean and autocorrelation, as stated in the following property.

Property 2.6: *Autocorrelation specification of Gaussian random processes* - The distribution of a discrete-time Gaussian random process, X_n, is uniquely specified by its mean and time correlation.

$$\begin{aligned} \mathrm{E}[X_n] &= \mu_n \\ \mathrm{E}[X_n X_k] &= R(n,k) \ . \end{aligned}$$

If a Gaussian random process is also stationary, then the mean must be constant, and the power spectral density is the DTFT of the autocorrelation. The next property summarizes this important result.

Property 2.7: *Power spectral density specification of Gaussian random processes* - The distribution of a wide-sense stationary discrete-time Gaussian random process, X_n, is uniquely specified by its mean and power spectral density, $S_X\left(e^{j\omega}\right)$.

Example 2.6: Consider the P^{th} order zero mean AR random process $\cdots, X_{-1}, X_0, X_1, \cdots$ with prediction variance of σ_C^2 and prediction filter h_n.

Our first task in this example is to determine if X_n is a stationary random process. By equation (**??**), we know that the power spectrum of the random process is given by

$$S_X(e^{j\omega}) = \frac{\sigma_C^2}{|1 - H(e^{j\omega})|^2} \ ,$$

where $H(e^{j\omega})$ is the DTFT of h_n. Let $Y_n = X_{n-k}$ be a version of X_n that is delayed by k samples. Since the MMSE predictor for X_n is time-invariant, the MMSE causal predictor for Y_n must also be h_n with corresponding prediction variance σ_C^2. This means that the power spectrum for Y_n is also given by

$$S_Y(e^{j\omega}) = \frac{\sigma_C^2}{|1 - H(e^{j\omega})|^2} \ .$$

Since the power spectrum for Y_n and X_n are the same, we know by Property **??**.5 that the two random processes must have the same distribution. Since the time shifted versions of the second order random process X_n has the same distribution, it must be both wide and strict sense stationary.

The next task is to determine if X_n is a reversible random process. For this purpose, redefine $Y_n = X_{-n}$. Then by the definition of time autocorrelation, we know that

$$\begin{aligned} R_Y(n) &= \mathrm{E}\{Y_k Y_{k+n}\} \\ &= \mathrm{E}\{X_{-k} X_{-k-n}\} \\ &= \mathrm{E}\{X_{l+n} X_l\} \\ &= R_X(n) \end{aligned}$$

where $l = -k-n$. From this, we know that X_n and its time reverse, Y_n, have the same autocorrelation. By Property **??**.4, this means that X_n and Y_n have the same distribution. The conclusion is that any stationary AR process is reversible.

Interestingly, our analysis did not explicitly depend on the assumption that the process was AR. In fact, this analysis is valid for any stationary Gaussian random process, which leads to the following important property.

Property 2.8: *Reversibility of stationary Gaussian random processes* - Any wide-sense stationary Gaussian random process is also strict-sense stationary and reversible.

In this section also include:

- **Definitions of DTFT and DSFT**
- **Definition of DFT**
- **Definition of 1-D and 2-D Z-transform**

Chapter 2 Problems

1. Let Y be a random variable which is uniformly distributed on the interval $[0, 1]$. Compute the cumulative density function and probability density function of both Y and $Z = Y^2$.

2. Given a random variable, X, which is uniformly distribute on the interval $[0, 1]$, find a function $Y = f(X)$ which will generate a new random variable, Y, with a specified CDF, $F_y(t)$.

3. Give an example of three random variables, X_1, X_2, and X_3 such that for $k = 1, 2, 3$, $P\{X_k = 1\} = P\{X_k = -1\} = \frac{1}{2}$, and such X_k and X_j are independent for all k and j, but such that X_1, X_2, and X_3 are not jointly independent.

4. Given an example of a set of three random variables X_1, X_2, and X_3, such that

   i. Each random variable has a Gaussian $N(0, 1)$ distribution

   ii. But the vector $X = [X_1, X_2, X_3]^t$ is not a Gaussian random vector!

5. For each of the following cost functions, find expressions for the minimum risk Bayesian estimator, and show that it minimizes the risk over all estimators.

   a) $C(x, \hat{x}) = |x - \hat{x}|$

   b) $C(x, \hat{x}) = |x - \hat{x}|^2$.

6. Let $\{Y_i\}_{i=1}^n$ be i.i.d. RV's with distribution

$$\begin{aligned} P(y_i = 1) &= \theta \\ P(y_i = 0) &= 1 - \theta \end{aligned}$$

   Compute the ML estimate of θ.

7. Let X be a p-dimensional zero mean Gaussian random vector. Show that if for all i and j, X_i and X_j are uncorrelated, then the components are X are jointly independent.

8. Let $\{Y_i\}_{i=1}^n$ be i.i.d. random variables with distribution

$$P(x_i = k) \;=\; \pi_k$$

where $\sum_{k=1}^m \pi_k = 1$. Compute the ML estimate of the parameter vector $\theta = [\pi_1, \cdots, \pi_m]$.

9. Let $Y_1, \cdots, Y_n$ be i.i.d.$\sim N(\mu, \sigma^2)$ random variables. Calculate the ML estimate of the parameter vector (μ, σ^2).

10. Let $X_1, \cdots, X_n$ be i.i.d. Gaussian random vectors with distribution $N(\mu, R)$ where $\mu \in \Re^p$ and $R \in \Re^{p \times p}$ is the positive definite matrix. Let $\theta = [\mu, R]$ denote the parameter vector for the distribution.

    a) Derive the expressions for the probability density of $p(x|\theta)$ with the forms given in equations (2.1) and (2.2).

    b) Compute the joint ML estimate of μ and R.

11. Let X, N, and Y be Gaussian random vectors such that $X \sim N(0, R_x)$ and $W \sim N(0, R_w)$, and let θ be a deterministic vector.

    a) First assume that $Y = \theta + W$, and calculate the ML estimate of θ given Y .

    b) For the next parts, assume that $Y = X + W$, and calculate an expression for $p_{x|y}(x|y)$, the conditional density of X given Y.

    c) Calculate the MMSE estimate of X when $Y = X + W$.

    d) Calculate an expression for the conditional variance of X given Y.

12. Show that if X and Y are jointly Gaussian random vectors, then the conditional distribution of X given Y is also Gaussian.

13. Prove that if X and Y are zero-mean jointly Gaussian random vectors, the $\mathrm{E}[X|Y] = AY$ i.e. Property 2.5.

14. Prove that two zero-mean discrete-time Gaussian random processes have the same distribution, if and only if they have the same time autocorrelation function.

15. Prove all Gaussian wide sense stationary random processes are:
    a) strict sense stationary
    b) reversible

16. Construct an example of a strict sense stationary random process that is not reversible.

17. Consider to zero-mean Gaussian discrete-time random processes, X_n and Y_n related by

$$Y_n = h_n * X_n \ ,$$

where $*$ denotes discrete-time convolution and h_n is an impulse response with $\sum_n |h_n| < \infty$. Show that

$$R_y(n) = R_x(n) * h_k * h_{-k} \ .$$

Bibliography

[1] S. Silvey. *Statistical Inference*. Chapman and Hall, London, 1975.