(A Quick) Probability Review

Reading:

- Go over handouts 2–5 in EE 420x notes.

Basic probability rules:

(1) $P[\Omega] = 1$, $P[\emptyset] = 0$, $0 \leq P[A] \leq 1$;
$P[\cup_{i=1}^{\infty} A_i] = \sum_{i=1}^{\infty} P[A_i]$ if $\underbrace{A_i \cap A_j}_{A_i \text{ and } A_j \text{ disjoint}} = \emptyset$ for all $i \neq j$;

(2) $P[A \cup B] = P[A] + P[B] - P[A \cap B]$, $P[A^c] = 1 - P[A]$;

(3) If $A \perp\!\!\!\perp B$, then $P[A \cap B] = P[A] \cdot P[B]$;

(4)

$$P[A \,|\, B] = \frac{P[A \cap B]}{P[B]} \quad \text{(conditional probability)}$$

or

$$P[A \cap B] = P[A \,|\, B] \cdot P[B] \quad \text{(chain rule)};$$

(5)

$$P[A] = P[A \mid B_1]\, P[B_1] + \cdots P[A \mid B_n]\, P[B_n]$$

if $B_1, B_2, \ldots, B_n$ form a *partition* of Ω;

(6) Bayes rule:

$$P[A \mid B] = \frac{P[B \mid A]\, P[A]}{P[B]};$$

(7)

$$\begin{aligned} \mathrm{E}\,[aX + bY + c] &= a \cdot \mathrm{E}\,[X] + b \cdot \mathrm{E}\,[Y] + c \\ \mathrm{var}(aX + bY + c) &= a^2\, \mathrm{var}(X) + b^2\, \mathrm{var}(Y) \\ &\quad + 2\, a\, b \cdot \mathrm{cov}(X, Y) \end{aligned}$$

where $a, b,$ and c are constants and X and Y are random variables.

(7') A vector/matrix version of (7):

$$\begin{aligned} \mathrm{E}\,[A\,\boldsymbol{X} + B\,\boldsymbol{Y} + \boldsymbol{c}] &= A\,\mathrm{E}\,[\boldsymbol{X}] + B\,\mathrm{E}\,[\boldsymbol{Y}] + \boldsymbol{c} \\ \mathrm{cov}(A\,\boldsymbol{X} + B\,\boldsymbol{Y} + \boldsymbol{c}) &= A\,\mathrm{cov}(\boldsymbol{X})\,A^T + B\,\mathrm{cov}(\boldsymbol{Y})\,B^T \\ &\quad + A\,\mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y})\,B^T + B\,\mathrm{cov}(\boldsymbol{Y}, \boldsymbol{X})\,A^T \end{aligned}$$

where "T" denotes a transpose and

$$\mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{E}\,\{(\boldsymbol{X} - \mathrm{E}\,[\boldsymbol{X}])\,(\boldsymbol{Y} - \mathrm{E}\,[\boldsymbol{Y}])^T\}.$$

(To refresh memory about covariance and its properties, see p. 12 of handout 5 in EE 420x notes. For random vectors, see handout 7 in EE 420x notes, particularly pp. 1–15.)

Some useful theorems:

(1) (handout 5 in EE 420x notes)

$$\begin{aligned} \mathrm{E}\,[X] &= \mathrm{E}_Y[\mathrm{E}_{X|Y}[X \,|\, Y]] \\ \mathrm{E}\,[g(X) \cdot h(Y) \,|\, Y = y] &= h(y) \cdot \mathrm{E}\,[g(X) \,|\, Y = y] \\ \mathrm{E}\,[g(X) \cdot h(Y)] &= \mathrm{E}\,[h(Y) \cdot \mathrm{E}\,[g(X) \,|\, Y]]; \end{aligned}$$

The vector version of (1) is the same — just put bold letters.

(2)

$$\mathrm{var}(X) = \mathrm{E}\,[\mathrm{var}(X \,|\, Y)] + \mathrm{var}(\mathrm{E}\,[X|Y]);$$

The vector/matrix version of (2) is:

(2')

$$\underbrace{\mathrm{cov}(\boldsymbol{X})}_{\substack{\text{variance/covariance} \\ \text{matrix of } \boldsymbol{X}}} = \mathrm{E}\,[\mathrm{cov}(\boldsymbol{X} \,|\, \boldsymbol{Y})] + \mathrm{cov}(\mathrm{E}\,[\boldsymbol{X}|\boldsymbol{Y}]);$$

(3)

$$\operatorname{cov}(X, Y) = \mathrm{E}\left[\operatorname{cov}(X, Y \mid Z)\right] + \operatorname{cov}(\mathrm{E}\left[X|Z\right], \mathrm{E}\left[Y|Z\right]).$$

(4) Transformation:

$$\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X}) \quad \Longleftrightarrow \quad \begin{array}{c} Y_1 = g_1(X_1, \ldots, X_n) \\ \vdots \\ Y_n = g_n(X_1, \ldots, X_n) \end{array}$$

then

$$p_Y(\boldsymbol{y}) = p_X(h_1(y_1), \ldots, h_n(y_n)) \cdot |J|$$

where $h(\cdot)$ is the unique inverse of $g(\cdot)$ and

$$J = \left|\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}^T}\right| = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}$$

Print and read the handout "Probability distributions" from the Supplementary material section on WebCT. Bring it with you to the midterm exam.

Estimator Performance

We now continue with the DC-level estimation example from handout # 0.

This handout describes the "classical world" and, therefore, arguments are made in this context. Bayesian arguments will be somewhat different.

Consider the following two estimators:

$$\begin{aligned} \widehat{A}_1 &= \frac{1}{N}\sum_{n=0}^{N-1} x[n] \\ \widehat{A}_2 &= x[0]. \end{aligned}$$

Note that $\widehat{A}_1$ is the ML estimate of A — it maximizes the likelihood function, see handout # 0. (Interestingly, the ML estimate of A is the same regardless of whether σ^2 is known or not.) It is also intuitively appealing — A is the *average* level of $x[n]$ (since $w[n]$ is zero mean).

Which estimator is better?

For a given realization, it is possible that either $\widehat{A}_1$ or $\widehat{A}_2$ is closer to A. Hence, we need statistical analysis to answer this question.

Estimator Performance (cont.)

Substitute the measurement model to perform statistical analysis. We have

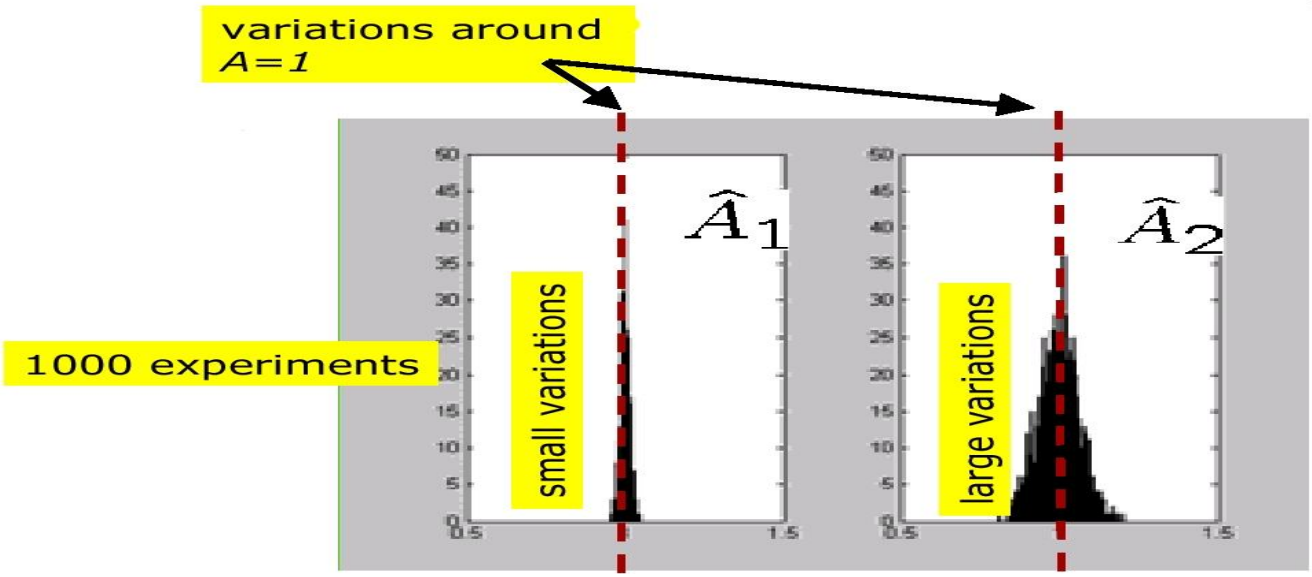
$$\begin{aligned}\widehat{A}_1 &= \frac{1}{N}\sum_{n=0}^{N-1}\{A + w[n]\} \\ \widehat{A}_2 &= A + w[0].\end{aligned}$$

Take expectation:

$$\begin{aligned}\mathrm{E}_x[\widehat{A}_1] &= \frac{1}{N}\sum_{n=0}^{N-1}(A + \overbrace{\mathrm{E}\{w[n]\}}^{0}) = A \\ \mathrm{E}_x[\widehat{A}_2] &= A + \overbrace{\mathrm{E}\{w[n]\}}^{0} = A.\end{aligned}$$

On average, both estimators are around the correct value (i.e. they are unbiased).

Estimator Performance (cont.)



Histograms of $\widehat{A}_1$ and $\widehat{A}_2$.

But $\widehat{A}_1$ is better than $\widehat{A}_2$ because its pdf is more concentrated around the true value.

"On average, $\widehat{A}_1$ is closer to $A = 1$."

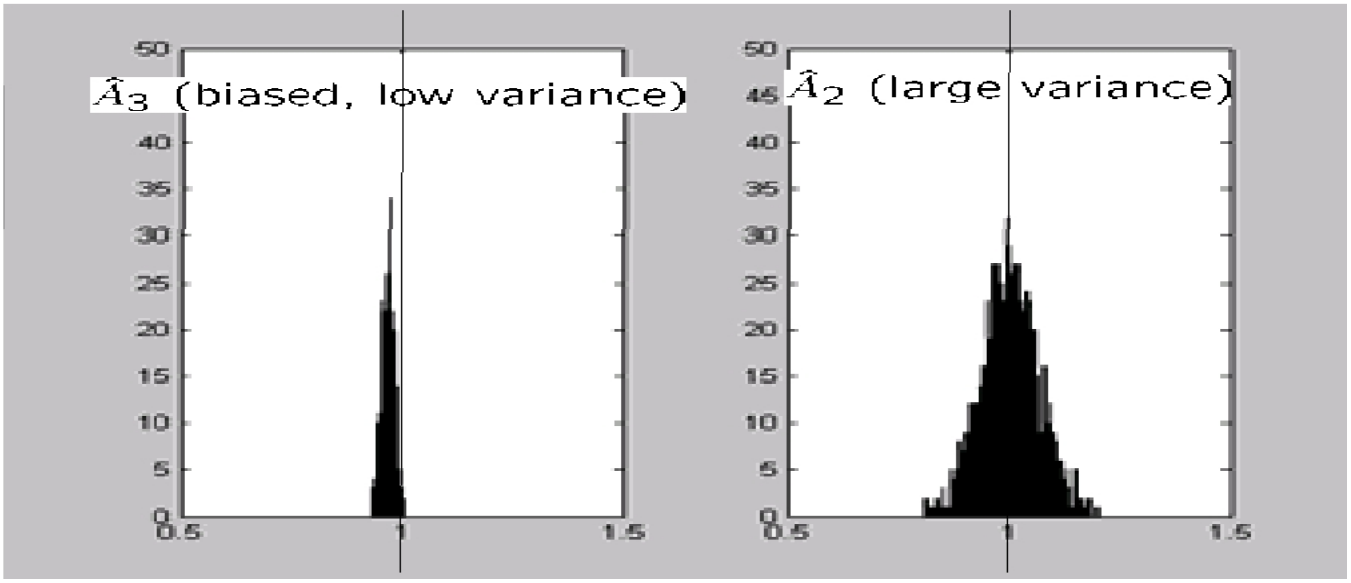
Proof.

$$\begin{aligned}
\mathrm{E}_x[\widehat{A}_1] &= \mathrm{E}[\widehat{A}_2] = A \\
\mathrm{var}_x(\widehat{A}_1) &= \mathrm{E}[(\widehat{A}_1 - \overbrace{\mathrm{E}[\widehat{A}_1]}^{A})^2] = \frac{1}{N^2}\sum_{n=0}^{N-1}\mathrm{var}_x(w[n]) = \frac{\sigma^2}{N} \\
\mathrm{var}_x(\widehat{A}_2) &= \sigma^2.
\end{aligned}$$

□

But, what is the justification for taking these expectations? We implicitly assume that we could repeat this experiment many times and plot the histogram (say) of the resulting estimates. This is what Bayesians criticize saying that, in the classical approach, "data that have never been observed are used for inference." (As you can guess, Bayesians do not need this virtual-data argument.) Suppose we are fine with the classical argument and let us continue.

Which Estimator is the Best?



Notation. Bias and mean-square error (MSE) of estimator $\widehat{\theta}$:

$$\begin{aligned} b(\theta) &= \mathrm{E}_x[\widehat{\theta}] - \theta \\ \mathrm{MSE}(\widehat{\theta}) &= \mathrm{E}_x[(\widehat{\theta} - \theta)^2] = \mathrm{var}(\widehat{\theta}) + b(\theta)^2. \end{aligned}$$

We wish to minimize the MSE, which leads to an MMSE estimator. MSE expression:

$$\begin{aligned} \mathrm{MSE}(\widehat{\theta}) &= \mathrm{E}_x[(\widehat{\theta} - \mathrm{E}_x[\widehat{\theta}] + \mathrm{E}_x[\widehat{\theta}] - \theta)^2] \\ &= \underbrace{\mathrm{E}_x[(\widehat{\theta} - \mathrm{E}_x[\widehat{\theta}])^2]}_{\mathrm{var}(\widehat{\theta})} + \underbrace{(\mathrm{E}_x[\widehat{\theta}] - \theta)^2}_{b(\theta)^2} \\ &\quad \underbrace{+2\,\mathrm{E}_x[(\widehat{\theta} - \mathrm{E}_x[\widehat{\theta}]) \cdot \underbrace{(\mathrm{E}_x[\widehat{\theta}] - \theta)}_{\text{const}}]}_{0} \end{aligned}$$

Note that the above $\mathrm{MSE}(\widehat{\theta})$ is different from Bayesian MSE. Since in Bayesian inference we assign a prior distribution to θ, Bayesian MSE would be obtained by taking the expectation of $\mathrm{MSE}(\widehat{\theta})$ with respect to θ.

In the classical setup, minimizing the MSE is our key objective, but the use of this criterion leads to unrealizable estimators $\Longrightarrow \mathrm{MSE}(\widehat{\theta}) = \mathrm{var}(\widehat{\theta}) + b(\theta)^2$ is a strong function of θ and minimizing it over some family of estimators will usually produce an "optimal" estimator $\widehat{\theta}$ that depends on θ.

Example 1. DC level in **A**dditive **W**hite **G**aussian **N**oise (AWGN), see also Section 2.4 in Kay-I:

$$\begin{aligned} x[n] &= A + \underbrace{w[n]}_{\text{AWGN}} \\ w[n] &\sim \mathcal{N}(0, \sigma^2), \quad n = 0, 1, \ldots, N-1. \end{aligned}$$

Consider the following family of estimators of A:

$$\check{A} = a\,\overline{x}$$

where

$$\overline{x} = (1/N) \sum_{n=0}^{N-1} x[n] \quad \text{(sample mean)}.$$

Here

$$\mathrm{E}\,[\check{A}] = aA, \quad \mathrm{var}(\check{A}) = a^2\sigma^2/N.$$

Find the best a (that minimizes the MSE for the given family). In other words, can we improve upon the sample mean?

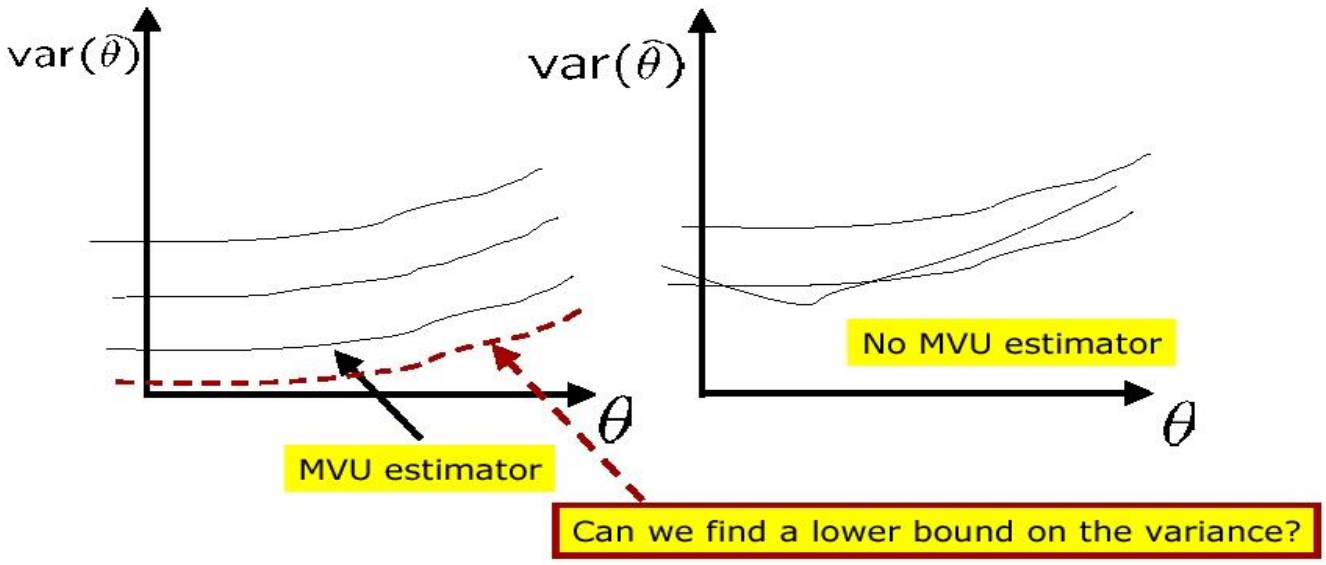
$$\begin{aligned}
\mathrm{MSE}(\check{A}) &= a^2\sigma^2/N + \overbrace{(aA - A)^2}^{\text{depends on } A} \\
\frac{d\mathrm{MSE}}{da} &= 2a\sigma^2/N + 2(aA - A)A = 0 \\
a_{\mathrm{opt}} &= \frac{A^2}{A^2 + \sigma^2/N}
\end{aligned}$$

depends on the unknown parameter A. Hence not useful (at least not directly). Observe the "shrinkage" form of the above "estimator."

Minimum Variance Unbiased (MVU) Estimation

How do we construct a "realizable" estimator?

An idea: Constrain the bias to be zero and then minimize the estimator variance (which is equal to MSE in this scenario since the bias is zero) for all values of $\theta \Longrightarrow$ MVU estimator.



MVU estimator does not always exist, as $\widehat{\theta}$ must have smallest variance for all values of θ.[1]

[1] To emphasize the fact that the MVU estimator must have the smallest variance for all values of θ, B & D refer to it as *uniformly minimum variance unbiased* (UMVU).

Comments:

- Even if it exists for a particular problem, MVU estimator is not optimal in terms of minimizing the MSE and we may be able to do better.

- Unbiasedness is nice, but not the most important $\Longrightarrow$ we can relax this condition and consider biased estimators as well, e.g. by making them *asymptotically unbiased*. By relaxing the unbiasedness condition, it is possible to outperform the MVU estimators in terms of MSE, as shown in the following example. What we really care about is minimizing the MSE!

Example 2[2]**.** Consider now estimating the variance σ^2 of independent, identically distributed (i.i.d.) zero-mean Gaussian observations, using the following estimator:

$$\widehat{\sigma}^2 = a \cdot \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] \tag{1}$$

where $a > 0$ is variable. If we choose $a = 1$, $\widehat{\sigma}^2\big|_{a=1}$ will be unbiased[3] with

$$\widehat{\sigma}^2\big|_{a=1} = \widehat{\sigma}^2_{\text{MVU}} = \frac{1}{N} \sum_{n=0}^{N-1} x^2[n]. \tag{2}$$

[2] See also P. Stoica and R. Moses, "On biased estimators and the unbiased Cramér-Rao lower bound," *Signal Processing,* vol. 21, pp. 349–350, 1991.

[3] We will show later that this choice yields an MVU estimate.

Now, in general,

$$\mathrm{E}\left[\widehat{\sigma}^2\right] = a\,\sigma^2$$

and

$$\begin{aligned}
\mathrm{MSE}(\widehat{\sigma}^2) &= \mathrm{E}\left[(\widehat{\sigma}^2 - \sigma^2)^2\right] \\
&= \mathrm{E}\left[\widehat{\sigma}^4\right] + \sigma^4 - 2\sigma^2\,\mathrm{E}\left[\widehat{\sigma}^2\right] \\
&= \mathrm{E}\left[\widehat{\sigma}^4\right] + \sigma^4\,(1-2a) \\
&= \frac{a^2}{N^2}\sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1}\mathrm{E}\left\{x^2[n_1]x^2[n_2]\right\} \\
&\quad +\sigma^4\,(1-2a) \\
&= \frac{a^2}{N^2}\left[(N^2-N)\,\sigma^4 + N\cdot\underbrace{\mathrm{E}\left\{x^4[n]\right\}}_{3\sigma^4}\right) + \sigma^4\,(1-2a) \\
&= \sigma^4\cdot\left[a^2(1+\frac{2}{N}) + (1-2a)\right]. \qquad (3)
\end{aligned}$$

To evaluate the above expression, we have used the following facts:

- For $n_1 \neq n_2$, $\mathrm{E}\left\{x^2[n_1]x^2[n_2]\right\} = \mathrm{E}\left\{x^2[n_1]\right\}\cdot\mathrm{E}\left\{x^2[n_2]\right\} = \sigma^2\cdot\sigma^2 = \sigma^4$.

- For $n_1 = n_2$, $\mathrm{E}\left\{x^2[n_1]x^2[n_2]\right\} = \mathrm{E}\left\{x^4[n_1]\right\} = 3\sigma^4$ (which is the fourth-order moment of a Gaussian distribution).

It can be easily shown that (3) is minimized for

$$a_{\rm OPT} = \frac{N}{N+2}$$

yielding the estimator

$$\widehat{\sigma}^2_\star = a_{\rm OPT} \cdot \frac{1}{N} \sum_{n=0}^{N-1} x^2[n]$$

whose MSE

$$\mathrm{MSE}_{\rm MIN} = \frac{2\,\sigma^4}{N+2}.$$

is minimum for the family of estimators in (1).

Comments:

- $\widehat{\sigma}^2_\star$ is *biased* and has *smaller MSE* than the MVU estimator in (2):

$$\mathrm{MSE}_{\rm MIN} < \mathrm{MSE}(\widehat{\sigma}^2)\Big|_{a=1} = \frac{2\,\sigma^4}{N}.$$

- Note that we are able to construct an realizable estimator in this case — compare with Example 1 in this handout.

- For large N, $\widehat{\sigma}^2_\star$ and $\widehat{\sigma}^2_{\rm MVU}$ are approximately the same since $N/(N+2) \to 1$ as $N \to \infty$. This also implies that $\widehat{\sigma}^2_\star$ is *asymptotically unbiased*.

Note: I do not wish to completely dismiss bias considerations. For example, we may have two estimators $\widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_2$ with

$$[\mathrm{bias}(\widehat{\boldsymbol{\theta}}_1)]^2 \ll \mathrm{var}(\widehat{\boldsymbol{\theta}}_1) \quad \text{and} \quad [\mathrm{bias}(\widehat{\boldsymbol{\theta}}_2)]^2 \ll \mathrm{var}(\widehat{\boldsymbol{\theta}}_2)$$

and

$$\mathrm{MSE}(\widehat{\boldsymbol{\theta}}_1) \approx \mathrm{MSE}(\widehat{\boldsymbol{\theta}}_2).$$

So, these two estimators are "equally good" as far as MSE is concerned. But, we may have

$$|\mathrm{bias}(\widehat{\boldsymbol{\theta}}_1)| \ll |\mathrm{bias}(\widehat{\boldsymbol{\theta}}_2)|$$

making $\widehat{\boldsymbol{\theta}}_1$ "more desirable" than $\widehat{\boldsymbol{\theta}}_2$. *Bias correction* methods have been developed for constructing estimators that have small bias. Hence, having small bias is typically a second-tier concern (compared with minimizing the MSE), but a valid one, particularly in the scenario outlined in this comment.

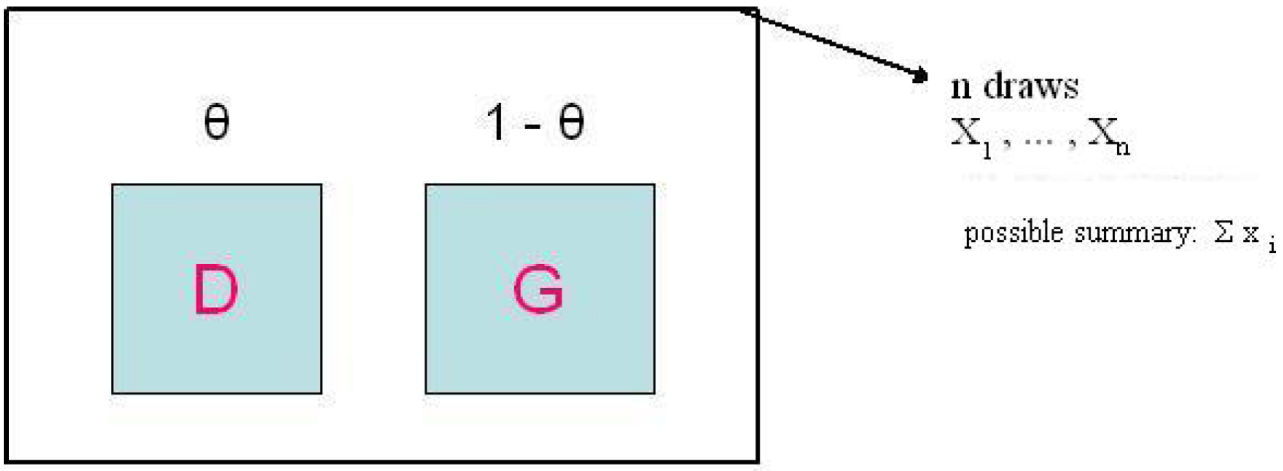
Sufficiency

Reading:

- Kay-I, chs. 5.3–5.4.

A function $T(\boldsymbol{x})$ of the observations $\boldsymbol{x}$ *only* is called a *statistic*.

Example: A machine produces n items in succession with probability θ of producing a defective product. Suppose that there is no dependence in quality of the produced items.



Then, our statistical model is

$$p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}.$$

Is there any loss of information by keeping and recording only

$\sum_{i=1}^{n} x_i$? Answer:

$$\begin{cases} \text{Yes, we are dropping a lot of information. But} \\ \text{No, in terms of inference about } \theta \\ \text{(provided that the model is correct, of course).} \end{cases}$$

We typically wish to separate out any aspects of the data that are irrelevant in the context of our model. In other words, we would like to reduce the data and deal only with the statistics "whose use involves no loss of information." For example, we could save memory and store only the reduced data. What we mean by "no loss of information" is quantified in the following definition.

Definition. *$T = T(\boldsymbol{x})$ is a sufficient statistic for θ if the conditional distribution of $\boldsymbol{X}$ given $T(\boldsymbol{X}) = T$ does not involve θ:*

$$p(\boldsymbol{x} \,|\, T(\boldsymbol{x}) = t; \theta) = p(\boldsymbol{x} \,|\, T(\boldsymbol{x}) = t).$$

Think of sufficient statistics as not throwing away any useful information about θ.

Note: if X is a random variable (RV), then $T = T(X)$ is a RV.

Trivial example: $T(\boldsymbol{x}) = \boldsymbol{x} \Longrightarrow$ full data is always sufficient.

Example: Let us continue with the previous example. Here, $\mathbf{X} = [X_1, \ldots, X_n]^T$ is the record of n Bernoulli trials with

probability θ, which can be written as

$$P\{X_i = x_i\} = \theta^{x_i}(1-\theta)^{1-x_i}.$$

where x_i is 1 (defective) or 0 (not defective). Thus

$$P\{\mathbf{X} = \boldsymbol{x}\} = P\{X_1 = x_1, \ldots, X_n = x_n\} = \theta^t(1-\theta)^{n-t}$$

where $t = \sum_{t=1}^n x_i$. Now, $T = \sum_{t=1}^n X_i$ has a binomial distribution $\mathrm{Bin}(n, \theta)$ and

$$\begin{aligned} p(\boldsymbol{x}|T(\boldsymbol{x}) = t; \theta) &= P\{\boldsymbol{X} = \boldsymbol{x} \,|\, T(\boldsymbol{X}) = t\} \\ &= \frac{P\{\mathbf{X} = \boldsymbol{x}, T(\boldsymbol{X}) = t\}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \begin{cases} 0, & \text{if } \overbrace{\sum_{t=1}^{n} x_i}^{T(\boldsymbol{x})} \neq t \\ \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}, & \text{otherwise} \end{cases} \end{aligned}$$

which is clearly not a function of θ. Here, we have used the (general) fact that

$$\{\boldsymbol{X} = \boldsymbol{x}\} \subset \{T(\boldsymbol{X}) = T(\boldsymbol{x})\}. \tag{4}$$

Thus, $T(\boldsymbol{x}) = \sum_{t=1}^n x_i$ is a sufficient statistic for θ.

In general, directly checking sufficiency is difficult because we need to compute the conditional distribution. Fortunately, we have the following theorem whose conditions are easy to verify.

Theorem. **(Factorization Theorem)** *A statistic $T(\mathbf{X})$ is sufficient for θ if and only if there exists a function $g(t,\theta)$ and a function $h(\boldsymbol{x})$ such that*

$$p(\boldsymbol{x};\theta) = \underbrace{g(T(\boldsymbol{x}),\theta)}_{\substack{\text{parameters} \\ \text{coupled with} \\ \text{suff. stat.}}} \cdot h(\boldsymbol{x}).$$

Note: $T(\boldsymbol{x})$ must be a statistic, a function of data $\boldsymbol{x}$ *only*.

Proof. To illustrate the idea of the proof and for simplicity, we concentrate on the discrete case. Suppose that $T(\boldsymbol{X})$ is sufficient. Then

$$\begin{aligned} p(\boldsymbol{x}\,;\,\theta) &= \overbrace{P\{\boldsymbol{X}=\boldsymbol{x}, T(\boldsymbol{X})=T(\boldsymbol{x})\}}^{P\{\boldsymbol{X}=\boldsymbol{x}\},\ \text{see (4)}} \\ &= P\{T(\boldsymbol{X})=T(\boldsymbol{x})\}\ \underbrace{P\{\boldsymbol{X}=\boldsymbol{x}\,|\,T(\boldsymbol{X})=T(\boldsymbol{x})\}}_{h(\boldsymbol{x}),\quad \text{by sufficiency}} \\ &= g(T(\boldsymbol{x}),\theta)\,h(\boldsymbol{x}). \end{aligned}$$

Conversely,

$$P\{\boldsymbol{X}=\boldsymbol{x}\,|\,T(\boldsymbol{X})=T(\boldsymbol{x})\}=\frac{P\{\boldsymbol{X}=\boldsymbol{x},T(\boldsymbol{X})=T(\boldsymbol{x})\}}{P\{T(\boldsymbol{X})=T(\boldsymbol{x})\}}$$

$$=\frac{\overbrace{P\{\boldsymbol{X}=\boldsymbol{x}\}}^{p(\boldsymbol{x}\,;\,\theta)}}{\underbrace{P\{T(\boldsymbol{X})=T(\boldsymbol{x})\}}_{\sum_{\boldsymbol{y}:T(\boldsymbol{y})=T(\boldsymbol{x})}\,p(\boldsymbol{y};\theta)}}$$

$$=\frac{\overbrace{g(T(\boldsymbol{x}),\theta)\,h(\boldsymbol{x})}^{\text{by the assumption}}}{\sum_{\boldsymbol{y}:T(\boldsymbol{y})=T(\boldsymbol{x})}\;\underbrace{g(T(\boldsymbol{y}),\theta)\,h(\boldsymbol{y})}_{\text{by the assumption}}}$$

$$=\frac{g(T(\boldsymbol{x}),\theta)\,h(\boldsymbol{x})}{g(T(\boldsymbol{x}),\theta)\sum_{\boldsymbol{y}:T(\boldsymbol{y})=T(\boldsymbol{x})}h(\boldsymbol{y})}$$

$$=\frac{h(\boldsymbol{x})}{\sum_{\boldsymbol{y}:T(\boldsymbol{y})=T(\boldsymbol{x})}h(\boldsymbol{y})}$$

which is *not* a function of θ. □

Example: Suppose $x_1, x_2 \ldots, x_n,$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$. Let

$\boldsymbol{\theta} = [\mu, \sigma^2]^T$. Then

$$p(x_1, \ldots, x_n; \boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \Big\}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\{-\frac{n\mu^2}{2\sigma^2}\} \exp\Big\{ -\frac{1}{2\sigma^2} \Big(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i \Big) \Big\}.$$

Clearly, $p(x_1, \ldots, x_n; \boldsymbol{\theta})$ is itself a function of $\sum_{i=1}^n x_i$, $\sum_{i=1}^n x_i^2$, and $\boldsymbol{\theta}$ only and, upon applying the factorization theorem, we conclude that

$$\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}(x_1, x_2, \ldots, x_n) = \begin{bmatrix} \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i^2 \end{bmatrix}$$

is sufficient for $\boldsymbol{\theta}$. Here, $h(\boldsymbol{x})$ is trivial: $h(\boldsymbol{x}) = 1$.

An equivalent (frequently used) sufficient statistic is

$$\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}(x_1, x_2, \ldots, x_n) = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n x_i \\ \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \end{bmatrix}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. This sufficient statistics can be obtained by suitably arranging the terms in the expression

for $p(x_1, \ldots, x_n; \boldsymbol{\theta})$:

$$p(x_1, \ldots, x_n; \boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \cdot \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} [(x_i - \bar{x}) + (\bar{x} - \mu)]^2 \right\}$$

and expanding the squares in the exponent.

Example. Suppose $x_1, x_2, \ldots, x_n,$ are i.i.d. $\mathcal{N}(\theta, 1)$. Then

$$p(\boldsymbol{x}; \theta) = \exp\{n\theta(\bar{x} - \tfrac{1}{2}\theta)\} \cdot \underbrace{(2\pi)^{-\frac{1}{2}n} \exp\{-\tfrac{1}{2} \sum_{i=1}^{n} x_i^2\}}_{h(\boldsymbol{x})}.$$

So, $\bar{x}$ is sufficient by the factorization theorem.

A Side Note on Multivariate Gaussian Pdf

An example (similar to the one in handout # 0). Consider $w[n]$ white Gaussian noise with unit variance:

$$w[n] \sim \underbrace{\mathcal{N}(0,1)}_{\text{univariate standard normal pdf}}, \quad n = 1, \ldots, d$$

implying

$$p(w[1], \ldots, w[d]) = \prod_{n=1}^{d} p(w[n])$$

$$= \frac{1}{(\sqrt{2\pi})^d} \cdot \exp\Big(-\tfrac{1}{2} \cdot \sum_{n=1}^{d} w[n]^2 \Big).$$

This expression can be succinctly written as

$$p(\boldsymbol{w}) = \underbrace{\frac{1}{(\sqrt{2\pi})^d} \exp\Big(-\tfrac{1}{2}\, \boldsymbol{w}^T \boldsymbol{w} \Big)}_{\text{multivariate standard normal pdf}} \triangleq \mathcal{N}(\mathbf{0}, I)$$

where

$$\boldsymbol{w} = \begin{bmatrix} w[1] \\ w[2] \\ \vdots \\ w[d] \end{bmatrix}.$$

We can generalize:

Definition. *A random $d \times 1$ vector $\boldsymbol{X}$ has a multivariate Gaussian pdf, denoted by*

$$\boldsymbol{X} \sim \mathcal{N}(\underbrace{\boldsymbol{\mu}}_{\mathrm{E}[\boldsymbol{X}]}, \underbrace{\Sigma}_{\mathrm{cov}(\boldsymbol{X})})$$

if its pdf is of the form:

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \, |\Sigma|}} \exp\left[-\tfrac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right]$$

where $|\Sigma| \equiv$ determinant of Σ, $\boldsymbol{\mu}$ is a $d \times 1$ vector, and Σ is a $d \times d$ symmetric positive definite matrix.

Since Σ is symmetric and positive definite, it can be shown that there exists a matrix $\Sigma^{1/2}$ called the square root of Σ with the following properties: (i) $\Sigma^{1/2}$ is symmetric, (ii) $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$, and (iii) $\Sigma^{1/2}\Sigma^{-1/2} = I$, where $\Sigma^{-1/2} = (\Sigma^{1/2})^{-1}$.

Theorem. *If $\boldsymbol{W} \sim \mathcal{N}(\boldsymbol{0}, I)$ and*

$$\boldsymbol{X} = \boldsymbol{\mu} + \Sigma^{1/2} \boldsymbol{W}$$

then

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma).$$

Conversely, if $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$*, then*

$$\Sigma^{-1/2}\,(\boldsymbol{X} - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}, I).$$

Suppose that we partition a random $d \times 1$ vector $\boldsymbol{X}$ as

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_a \\ \boldsymbol{X}_b \end{bmatrix}.$$

We can partition $\boldsymbol{\mu}$ and Σ accordingly:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}.$$

Theorem. *If* $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ *then*

(i) *The marginal distribution of* $\boldsymbol{X}_a$ *is*

$$\boldsymbol{X}_a \sim \mathcal{N}(\boldsymbol{\mu}_a, \Sigma_{aa}).$$

(ii) *The conditional distribution of* $\boldsymbol{X}_b$ *given* $\boldsymbol{X}_a = \boldsymbol{x}_a$ *is*

$$\boldsymbol{X}_b \,|\, \boldsymbol{X}_a = \boldsymbol{x}_a \sim \mathcal{N}\Big(\boldsymbol{\mu}_b + \Sigma_{ba}\Sigma_{aa}^{-1}(\boldsymbol{x}_a - \boldsymbol{\mu}_a),\, \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}\Big).$$

(iii) *If $\boldsymbol{a}$ is a constant $d \times 1$ vector, then*

$$\boldsymbol{a}^T \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{a}^T \boldsymbol{\mu}, \boldsymbol{a}^T \Sigma \boldsymbol{a}).$$

(iv)

$$(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \sim \chi^2_d \quad \textit{(Chi-square in your distr. table)}.$$

Example: Two Jointly Gaussian RVs

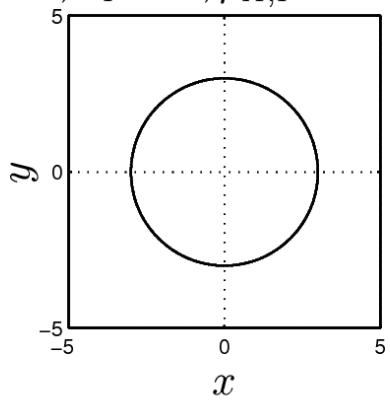
We plot contours of the joint pdf $p_{X,Y}(x,y)$ for zero-mean jointly Gaussian RVs for various values of σ_X, σ_Y, and $\rho_{x,y}$, where we have parametrized $p_{X,Y}(x,y)$ as

$$p_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2_{X,Y}}}$$
$$\cdot \exp\left\{ -\frac{1}{2(1-\rho^2_{X,Y})} \cdot \left[\frac{(x-\mu_X)^2}{\sigma^2_X} + \frac{(y-\mu_Y)^2}{\sigma^2_Y} \right.\right.$$
$$\left.\left. -2\,\rho_{X,Y} \frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} \right] \right\}$$

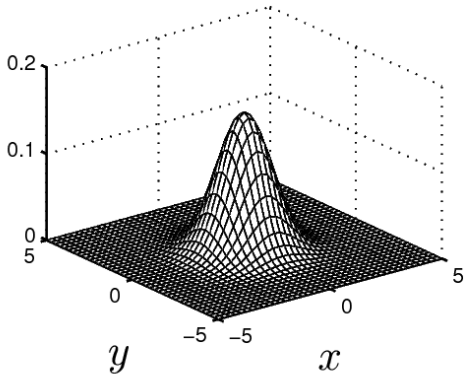
with

$$\underbrace{\rho_{X,Y}}_{\text{correlation coefficient}} = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{\text{cov}(X,Y)}{\sigma_X\,\sigma_Y}.$$

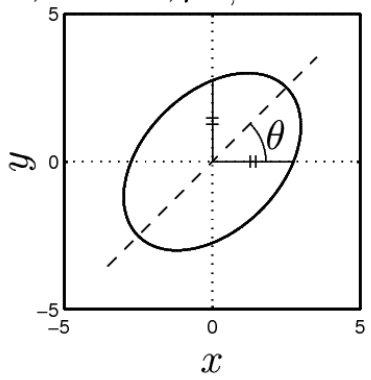
$\sigma_X = 1, \sigma_Y = 1, \rho_{X,Y} = 0$: $\theta = 0°$



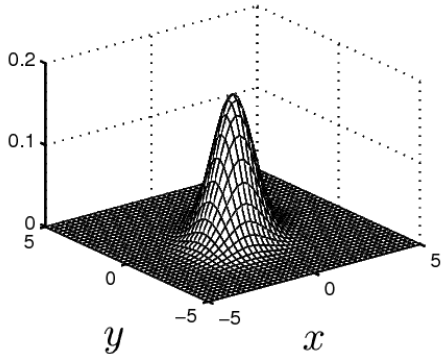
$p(x,y)$



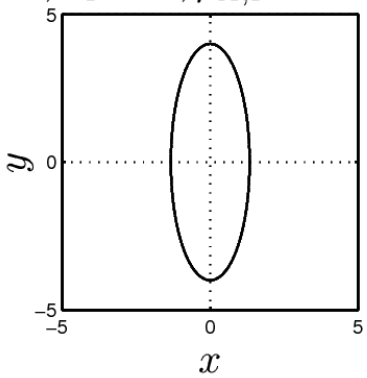
$\sigma_X = 1, \sigma_Y = 1, \rho_{X,Y} = 0.4$: $\theta = 45°$



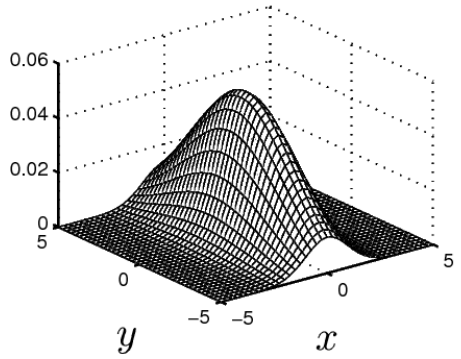
$p(x,y)$



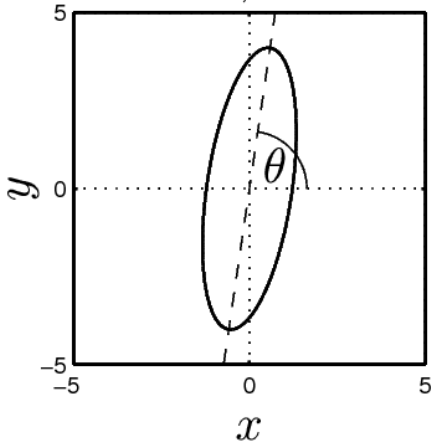
$\sigma_X = 1, \sigma_Y = 3, \rho_{X,Y} = 0$: $\theta = 90°$



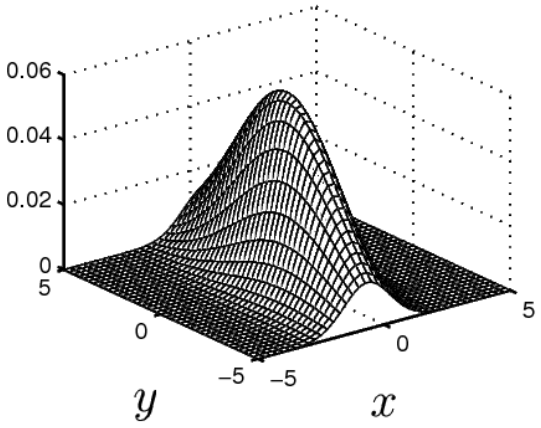
$p(x,y)$



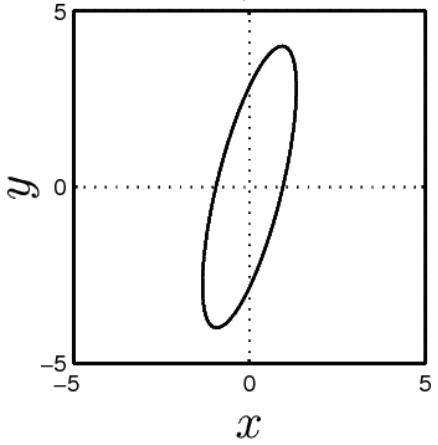
$\sigma_X = 1, \sigma_Y = 3, \rho_{X,Y} = 0.4$: $\theta = 81.65°$



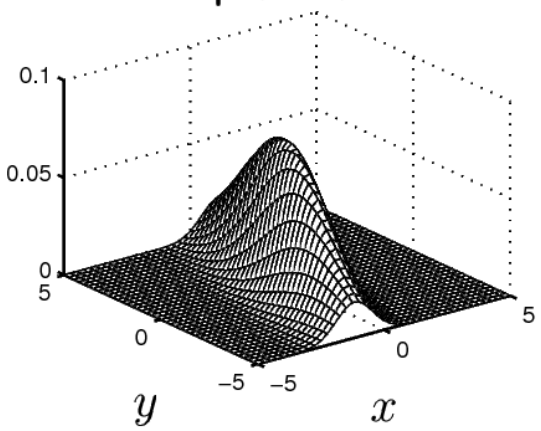
$p(x, y)$



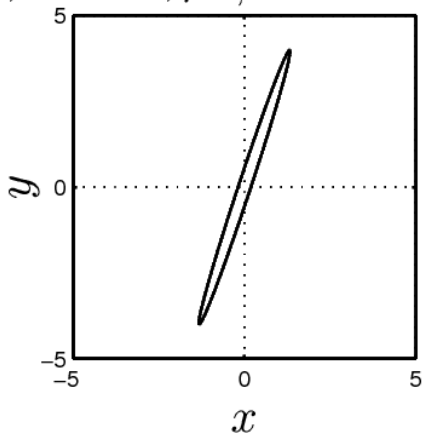
$\sigma_X = 1, \sigma_Y = 3, \rho_{X,Y} = 0.7$: $\theta = 76.15°$



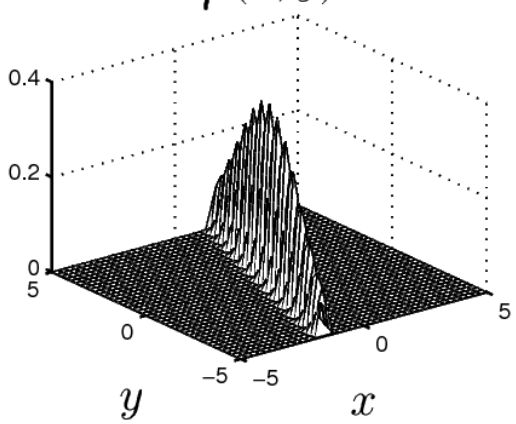
$p(x, y)$



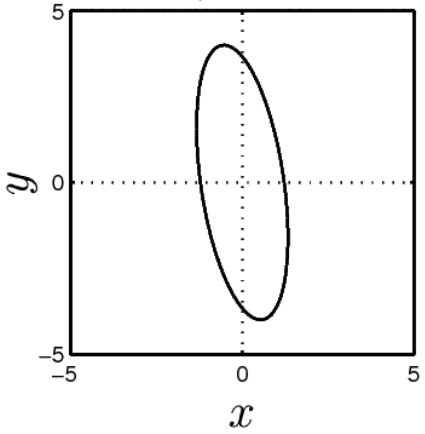
$\sigma_X = 1, \sigma_Y = 3, \rho_{X,Y} = 0.99$: $\theta = 71.7°$



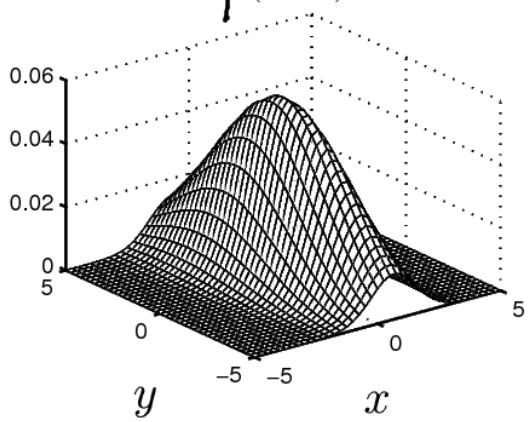
$p(x, y)$



$\sigma_X = 1, \sigma_Y = 3, \rho_{X,Y} = -0.4$: $\theta = -81.65°$



$p(x, y)$



Example (Digital Communications)

$$x(t) = \underbrace{s(t)}_{\text{signal}} + \underbrace{w(t)}_{\text{noise}}$$

where the signal $s(t)$ is usually represented using orthonormal basis functions $\varphi_k(t)$:

$$s(t) = \sum_{k=1}^{K} \alpha_k \, \varphi_k(t).$$

Note: The signal $s(t)$ is unknown, but it has known structure, incorporated in this basis-function expansion. **We wish to use this structure for data reduction.**

If $\varphi_k(t)$ are orthonormal, it is easy to show that the coefficients α_k can be computed as (for the case of real data):

$$\alpha_k = \int s(t) \, \varphi_k(t) \, dt.$$

Here, our goal at the receiver is to decide which $s(t)$ (α_k's) has been transmitted.

What is typically done in communication receivers is the following: the received data $x(t)$ are matched to the basis

functions, i.e.

$$\widehat{\alpha}_k = \int x(t)\,\varphi_k(t)\,dt, \quad k = 1, 2, \ldots, K \tag{5}$$

are computed and utilized for demodulation.

Question: Are the $\widehat{\alpha}_k$s sufficient statistics for inference about $s(t)$ (or, more precisely, for inference on the α_ks)?

Note: In some applications, sampled data $x[n], n = 0, 1, \ldots, N-1$ are available and

$$\widehat{\alpha}_k = \sum_{n=0}^{N-1} x[n]\,\varphi_k[n], \quad k = 1, 2, \ldots, K$$

are used to approximate the integrals in (5) (up to a scaling factor). We focus on this scenario, having in mind that we can easily switch from sums to integrals by letting the sampling interval go to zero — then N will go to infinity. Clearly, N is much larger than the number of basis functions K, i.e.

$$K \ll N.$$

In the sampled-data case, our model is

$$x[n] = \underbrace{s[n]}_{\text{signal}} + \underbrace{w[n]}_{\text{noise}}$$

where

$$s[n] = \sum_{k=1}^{K} \alpha_k \, \varphi_k[n].$$

Define

$$\boldsymbol{x} = \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}, \; \boldsymbol{w} = \begin{bmatrix} w[0] \\ w[1] \\ \vdots \\ w[N-1] \end{bmatrix}, \; \boldsymbol{\mu}(\boldsymbol{\alpha}) = \begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[N-1] \end{bmatrix}$$

and

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_K \end{bmatrix}$$

implying

$$\boldsymbol{x} = \boldsymbol{\mu}(\boldsymbol{\alpha}) + \boldsymbol{w}.$$

If the noise is additive zero-mean Gaussian with covariance matrix $\mathrm{E}\left[\boldsymbol{w}\boldsymbol{w}^T\right] = \boldsymbol{C}$, then

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\alpha}), \boldsymbol{C})$$

which is a multivariate Gaussian pdf:

$$p(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{C}|}} \exp\left\{-\tfrac{1}{2}\left[\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\alpha})\right]^T \boldsymbol{C}^{-1} \left[\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\alpha})\right]\right\} \tag{6}$$

and

$$\boldsymbol{\mu}(\boldsymbol{\alpha}) = \begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[N-1] \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{K} \alpha_k \, \varphi_k[0] \\ \sum_{k=1}^{K} \alpha_k \, \varphi_k[1] \\ \vdots \\ \sum_{k=1}^{K} \alpha_k \, \varphi_k[N-1] \end{bmatrix} = \boldsymbol{F}\boldsymbol{\alpha}$$

where

$$\boldsymbol{F} = \begin{bmatrix} \varphi_1(0) & \varphi_2(0) & \cdots & \varphi_K(0) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(N-1) & \varphi_2(N-1) & \cdots & \varphi_K(N-1) \end{bmatrix}$$

is an $N \times K$ matrix. So

$$p(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{C}|}} \cdot \exp\left\{ -\tfrac{1}{2} (\boldsymbol{x} - \boldsymbol{F}\boldsymbol{\alpha})^T \boldsymbol{C}^{-1} (\boldsymbol{x} - \boldsymbol{F}\boldsymbol{\alpha}) \right\}.$$

What are the sufficient statistics for inference on $\boldsymbol{\alpha}$? It depends on our knowledge about $\boldsymbol{C}$. If $\boldsymbol{C}$ is *unknown*, we clearly cannot separate out any non-trivial sufficient statistics for both $\boldsymbol{\alpha}$ and $\boldsymbol{C}$. If $\boldsymbol{C}$ is *known*, then the vector of sufficient statistics for $\boldsymbol{\alpha}$ is

$$\boldsymbol{F}^T \boldsymbol{C}^{-1} \boldsymbol{x} \tag{7}$$

which is a $K \times 1$ vector. Since, $K \ll N$, (7) achieves dimensionality reduction compared with the raw data $\boldsymbol{x}$. Note

that, if $\boldsymbol{C}$ is unknown, we cannot compute (7) $\Longrightarrow$ not realizable.

If $\boldsymbol{C} = \sigma^2 \boldsymbol{I}$ (i.e. the noise is white) and the noise variance σ^2 is *known*, then (7) simplifies to (up to a known proportionality factor):

$$\boldsymbol{F}^T \boldsymbol{x} = \begin{bmatrix} \sum_{n=0}^{N-1} \varphi_1[n] x[n] \\ \sum_{n=0}^{N-1} \varphi_2[n] x[n] \\ \vdots \\ \sum_{n=0}^{N-1} \varphi_K[n] x[n] \end{bmatrix} = \begin{bmatrix} \widehat{\alpha}_1 \\ \widehat{\alpha}_2 \\ \vdots \\ \widehat{\alpha}_K \end{bmatrix}.$$

If σ^2 is *unknown*, then

$$p(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \cdot \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{x} - \boldsymbol{F}\boldsymbol{\alpha})^T (\boldsymbol{x} - \boldsymbol{F}\boldsymbol{\alpha}) \right\}.$$

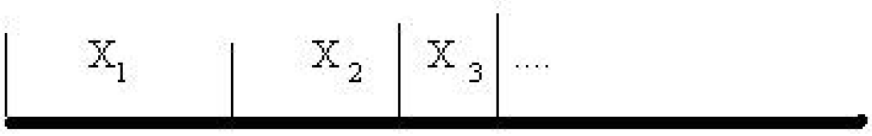
where σ^2 and $\boldsymbol{\alpha} \equiv$ parameters and $\boldsymbol{x} \equiv$ data. Now

$$\boldsymbol{x}^T \boldsymbol{x} = \sum_{n=0}^{N-1} x^2[n] \quad \text{and} \quad \boldsymbol{F}^T \boldsymbol{x} = \begin{bmatrix} \widehat{\alpha}_1 \\ \widehat{\alpha}_2 \\ \vdots \\ \widehat{\alpha}_K \end{bmatrix}$$

are jointly sufficient for $\boldsymbol{\alpha}$ and σ^2.

Examples: Computing the Sufficient Statistics

Example: Suppose that elements of $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$ are i.i.d. inter-arrival times of packets arriving at a node in a communication network.



We assume that x_i, $i = 1, 2, \ldots, n$ come from an exponential $\text{Expon}(\theta)$ distribution, implying

$$p(\boldsymbol{x}\,;\,\theta) = \prod_{i=1}^{n} \theta \exp(-\theta x_i) = \theta^n \exp\Big(-\theta \underbrace{\sum_{i=1}^{n} x_i}_{T(\boldsymbol{x})}\Big), \quad \forall x_i \geq 0.$$

Example: Elements of $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$ are i.i.d. $\text{uniform}(0, \theta)$:

$$p(\boldsymbol{x};\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} i_{[0,\theta]}(x_i) = \underbrace{\frac{1}{\theta^n}\, i_{[-\infty,\theta]}(\overbrace{\max x_i}^{T(\boldsymbol{x})})}_{g(T(\boldsymbol{x}),\theta)} \cdot \underbrace{i_{[0,\infty]}(\min x_i)}_{h(\boldsymbol{x})}$$

where $i_A(x)$ denotes the indicator function:

$$i_A(x) = \begin{cases} 1, & x \in A, \\ 0, & \text{otherwise} \end{cases}$$

Here, we have used the facts that

$$x_1, x_2, \ldots, x_N < \theta \quad \Longleftrightarrow \quad \max x_i < \theta$$

and

$$x_1, x_2, \ldots, x_N > 0 \quad \Longleftrightarrow \quad \min x_i > 0.$$

Example. Detection problem: $\theta \in \{0, 1\}$ and

$$\begin{aligned} p(x;\theta) &= \theta\, p(x;1) + (1-\theta)\, p(x;0) \\ &= \underbrace{\Big[\theta \overbrace{\frac{p(x;1)}{p(x;0)}}^{T(x)} + (1-\theta)\Big]}_{g(T(x),\theta)} \underbrace{p(x;0)}_{h(x)} \end{aligned}$$

$T(x)$ is the likelihood ratio! It is a very useful sufficient statistics because it is one-dimensional regardless of the nature of $p(x;\theta)$. See also Poor, Example IV.C.1.

Definition. *The statistic $T(\boldsymbol{x})$ is minimally sufficient if it is sufficient and provides a greater reduction of the data than any other sufficient statistic $S(\boldsymbol{x})$.*

(Multiparameter) Exponential Family of Distributions

$$p(x;\boldsymbol{\theta}) = h(x)\,\exp\underbrace{\left[\sum_{i=1}^{k}\eta_i(\boldsymbol{\theta})\,T_i(x) - B(\boldsymbol{\theta})\right]}_{\text{coupling between parameters and data has a very specific form}}$$

By the factorization theorem, $\boldsymbol{T}(X) = [T_1(X), \ldots, T_k(X)]^T$ is sufficient. It is the *natural sufficient statistic* of the family. The exponential family is important:

- It covers quite a few useful distributions, including some that are fairly complex; e.g. Markov random fields, used in image analysis, are virtually all in the exponential-family form;
- It is popular in graphical models as well (Markov random fields again);
- Many methods greatly simplify in the case of exponential family, e.g. the EM algorithm (to be discussed later).

If the support of $p(x;\theta)$ depends on θ, then $p(x;\theta)$ cannot belong to the exponential family. For example, $\mathrm{uniform}(0,\theta)$ is not a member of the exponential family.

Multiple i.i.d. measurements making $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$. Then,

$$
\begin{aligned}
p(\boldsymbol{x}; \theta) &= \prod_{l=1}^{n} \left\{ h(x_l) \exp\left[\sum_{i=1}^{k} \eta_i(\boldsymbol{\theta})\, T_i(x_l) - B(\boldsymbol{\theta}) \right] \right\} \\
&= \underbrace{\left[\prod_{l=1}^{n} h(x_l) \right] \cdot \exp\left[\sum_{i=1}^{k} \eta_i(\boldsymbol{\theta}) \sum_{l=1}^{n} T_i(x_l) - nB(\boldsymbol{\theta}) \right]}_{\text{again the exponential family}}
\end{aligned}
$$

and hence the vector of natural sufficient statistics is

$$
\boldsymbol{T}(X) = \left[\sum_{l=1}^{n} T_1(X_l), \ldots, \sum_{l=1}^{n} T_k(X_l) \right]^T .
$$

For more about exponential families, see B & D, Chapter 1.6.

A Side Note on One-Parameter Canonical Exponential Family

Here is a simple special sub-family of the exponential family — *the one-parameter canonical exponential family*:

$$p(\boldsymbol{x};\eta) = h(\boldsymbol{x})\,\exp\Big[\underbrace{\eta}_{\substack{\text{scalar}\\ \text{canonical}\\ \text{parameter}}}\,T(\boldsymbol{x}) - A(\eta)\Big]$$

where

$$A(\eta) = \log \int \cdots \int h(\boldsymbol{\chi})\exp[\eta\,T(\boldsymbol{\chi})]\,d\boldsymbol{\chi} \quad \text{for a pdf } p(\boldsymbol{x};\eta)$$

$$A(\eta) = \log \sum_{\boldsymbol{\chi}} h(\boldsymbol{\chi})\exp[\eta\,T(\boldsymbol{\chi})] \quad \text{for a pmf } p(\boldsymbol{x};\eta).$$

If we can compute the normalizing term $A(\eta)$ in a simple form, then it is easy to find the mean and variance of $T(X)$:

$$\mathrm{E}_{\,p(x;\eta)}[T(\boldsymbol{X})] = \frac{dA(\eta)}{d\eta}, \quad \mathrm{var}_{p(x;\eta)}[T(\boldsymbol{X})] = \frac{d^2A(\eta)}{d\eta^2}.$$

Why is this useful? Here is an example. Suppose

$x_1, x_2, \ldots, x_n$ are i.i.d. from

$$p(x;\theta) = \underbrace{\frac{x}{\theta^2} \cdot \exp\Big(-\frac{x^2}{2\theta^2}\Big), \quad x > 0, \theta > 0}_{\text{Rayleigh pdf}}.$$

Define $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$ and write the pdf of $\boldsymbol{x}$:

$$\begin{aligned} p(\boldsymbol{x};\theta) &= \prod_{i=1}^{n} \Big[\frac{x_i}{\theta^2} \cdot \exp\Big(-\frac{x_i^2}{2\theta^2}\Big)\Big] \\ &= \Big(\prod_{i=1}^{n} x_i\Big) \cdot \exp\Big[\underbrace{\frac{-1}{2\theta^2}}_{\eta} \cdot \underbrace{\Big(\sum_{i=1}^{n} x_i^2\Big)}_{T(\boldsymbol{x})} - n\log\theta^2\Big] \end{aligned}$$

implying

$$\theta^2 = -\frac{1}{2\eta}$$

and, consequently,

$$A(\eta) = -n\log(-2\eta).$$

Therefore, the natural sufficient statistic $T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i^2$ has mean

$$\mathrm{E}\Big[\sum_{i=1}^{n} X_i^2\Big] = \frac{dA(\eta)}{d\eta} = -\frac{n}{\eta} = 2n\,\theta^2$$

and variance

$$\mathrm{var}\Big[\sum_{i=1}^{n} X_i^2\Big] = \frac{d^2 A(\eta)}{d\eta^2} = \frac{n}{\eta^2} = 4n\,\theta^4.$$

Direct computation of these moments would be more complicated.