University of Michigan - NAME 568/EECS 568/ROB 530 **Winter 2022**

Note: Probability Theory and Statistics

Lecturer and Author: Maani Ghaffari

For this and all future notes, if you would like to add or extend any sections, and it is within the scope of the subject, just contact me and I will update the note.

1 Notation

Vectors are column-wise and $1{:}n$ means integers from 1 to n. The Euclidean and Frobenius norms are shown by $\|\cdot\|$ and $\|\cdot\|_{\mathrm{F}}$, respectively. $|X|$ denotes the determinant of matrix X. For the sake of compactness, random variables, such as X, and their realizations, x, are sometimes denoted interchangeably where it is evident from context. An alphabet such as $\mathcal{X}$ denotes a set, and the cardinality of the set is denoted by $|\mathcal{X}|$. $\mathrm{vec}(x_1, \dots, x_n)$ denotes a vector such as x constructed by stacking x_i, $\forall i \in \{1{:}n\}$. The n-by-n identity and zero matrices are denoted by I_n and 0_n, respectively. Finally, $\mathbb{E}[\cdot]$, $\mathbb{V}[\cdot]$, and $\mathrm{Cov}[\cdot]$ denote the expected value, variance, and covariance (for random vectors) of a random variable, respectively.

2 Probability Theory and Statistics

Let X be a random variable that maps the sample space Ω (set of all possible outcomes) to the state space $\mathcal{X}$. Let $p(X = x) \geq 0$ be the probability of the random variable X taking a specific value x. If X is a discrete random variable, then

$$\sum_{x \in \mathcal{X}} p(X = x) = 1, \tag{1}$$

and for the continuous form we can write

$$\int_{x \in \mathcal{X}} p(X = x) dx = 1. \tag{2}$$

For the sake of simplicity, it is common to use $p(x)$ instead of $p(X = x)$ and sometimes refer to x as the random variable itself. Let Y be another random variable, the joint distribution of X and Y is $p(x, y) = p(X = x \text{ and } Y = y)$ and if X and Y are independent

$$p(x, y) = p(x)p(y). \tag{3}$$

The conditional probability of X given another random variable Y is defined as

$$p(x|y) = \frac{p(x, y)}{p(y)} \qquad p(y) > 0. \tag{4}$$

Given the joint distribution of X and Y, the marginalization rule states that the marginal distribution of X can be computed by summing (integration) over Y. The law of total probability is its variant which uses the conditional probability definition and can be written as

$$p(x) = \sum_{y \in \mathcal{Y}} p(x, y) = \sum_{y \in \mathcal{Y}} p(x|y)p(y), \tag{5}$$

and for continuous random variables is

$$p(x) = \int_{y \in \mathcal{Y}} p(x, y)dy = \int_{y \in \mathcal{Y}} p(x|y)p(y)dy. \tag{6}$$

Given three random variables X, Y, and Z, Bayes rule relates the prior probability distribution, $p(x|z)$, and the likelihood function, $p(y|x, z)$, as follows.

$$p(x|y, z) = \frac{p(y|x, z)p(x|z)}{p(y|z)}. \tag{7}$$

The term $p(x|y, z)$ is called the posterior probability distribution over X.

$$p(\text{hypothesis}|\text{data}) = \frac{p(\text{data}|\text{hypothesis})p(\text{hypothesis})}{p(\text{data})},$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence (Marginal Likelihood)}}.$$

The expected value of the random variable X is

$$\mathbb{E}[X] = \int_{\Omega} X dp = \int_{\mathcal{X}} xp(x)dx \tag{8}$$

and if X is discrete

$$\mathbb{E}[X] = \sum_{\mathcal{X}} xp(x) \tag{9}$$

The expectation operator is linear which follows from linearity of integration and has the following properties:

- $\mathbb{E}[a] = a$
- $\mathbb{E}[aX] = a\mathbb{E}[X]$
- $\mathbb{E}[a + X] = a + \mathbb{E}[X]$
- $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$,

where X and Y are two arbitrary random variables and a and b are constants.

The variance of X can be calculated as

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \tag{10}$$

The covariance of a random vector $X = x$ can be calculated as

$$\text{Cov}[X] = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^{\mathsf{T}}] = \mathbb{E}[XX^{\mathsf{T}}] - \mathbb{E}[X]\mathbb{E}[X]^{\mathsf{T}}. \tag{11}$$

A covariance matrix such as Σ is symmetric, i.e., $\Sigma = \Sigma^{\mathsf{T}}$, and positive semi-definite, i.e., $x^{\mathsf{T}}\Sigma x \geq 0$ and all eigenvalues are nonnegative. A symmetric positive definite matrix ($x^{\mathsf{T}}\Sigma x > 0$) has positive eigenvalues and a unique Cholesky decomposition, i.e., $\Sigma = LL^{\mathsf{T}}$, where L is a lower triangular matrix.

The correlation coefficient is defined as

$$\rho_{XY} = \frac{\mathrm{Cov}[XY]}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}} \qquad |\rho_{XY}| \leq 1, \tag{12}$$

where the bound follows from the Cauchy-Schwarz inequality which asserts (assuming $\mathbb{E}[X^2]$ and $\mathbb{E}[Y^2]$ are finite)

$$|\mathbb{E}[XY]|^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]. \tag{13}$$

Remark 1. *In general, $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$, unless X and Y are uncorrelated. Now it is clear that if X and Y are independent, i.e., $X \perp Y$, then $\mathrm{Cov}[XY] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$ and X and Y are uncorrelated ($\rho_{XY} = 0$). However, uncorrelated random variables are not necessarily independent (e.g., Y is a non-constant function of X).*

Remark 2. *If a random vector has a multivariate normal distribution then any two or more of its components that are uncorrelated are independent. However, it is not true that two random variables that are (separately, marginally) normally distributed and uncorrelated are independent [1].*

Problem 1 (Uncorrelated but not independent)**.** *Let $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$. Show that X and Y are not independent, but are uncorrelated. Hint: $\mathbb{E}[X] = \mathbb{E}[X^3] = 0$.*

The univariate (one-dimensional) *Gaussian (or normal) distribution* with mean μ (location) and variance σ^2 (scale) has the following Probability Density Function (PDF).

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right). \tag{14}$$

We often write $x \sim \mathcal{N}(\mu, \sigma^2)$ or $\mathcal{N}(x; \mu, \sigma^2)$ to imply that x follows a Gaussian distribution with mean $\mu = \mathbb{E}[x]$ and variance $\sigma^2 = \mathbb{V}[x]$.

The *multivariate* Gaussian distribution of an n-dimensional random vector $x \sim \mathcal{N}(\mu, \Sigma)$, with mean $\mu = \mathbb{E}[x] \in \mathbb{R}^n$ and covariance $\Sigma = \mathrm{Cov}[x] = \mathbb{E}[(x-\mu)(x-\mu)^{\mathsf{T}}]$, can be written as follows.

$$p(x) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu)\right) \tag{15}$$

Lemma 1 (Marginalization and conditioning of normal distribution)**.** *Let x and y be jointly Gaussian random vectors*

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} A & C \\ C^{\mathsf{T}} & B \end{bmatrix}\right), \tag{16}$$

then the marginal distribution of x is

$$x \sim \mathcal{N}(\mu_x, A), \tag{17}$$

and the conditional distribution of x given y is

$$x|y \sim \mathcal{N}(\mu_x + CB^{-1}(y - \mu_y), A - CB^{-1}C^{\mathsf{T}}). \tag{18}$$

Problem 2. *Show that uncorrelated Gaussian random vectors are independent. Hint: In* (16)*, if x and y are uncorrelated, then $C = 0$; show $p(x, y) = p(x)p(y)$.*

Proposition 2 (Affine transformation of a multivariate normal distribution)**.** *Suppose $x \sim \mathcal{N}(\mu, \Sigma)$ and $y = Ax + b$. Then $y \sim \mathcal{N}(A\mu + b, A\Sigma A^{\mathsf{T}})$.*

Proof. Exercise. □

The canonical parametrization of a multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$ can be identified by the *information (or precision) matrix* $\Lambda = \Sigma^{-1}$, and the *information vector* $\eta = \Sigma^{-1}\mu$. Then we can write $p(x) = \mathcal{N}(\mu, \Sigma) = \mathcal{N}^{-1}(\eta, \Lambda)$.

Problem 3 (Derivation of canonical parametrization)**.** *Using direct calculation, show that the canonical parametrization of a multivariate normal distribution has the following form and find the constant γ.*

$$\mathcal{N}^{-1}(x; \eta, \Lambda) = \gamma \exp(-\frac{1}{2}x^{\mathsf{T}}\Lambda x + x^{\mathsf{T}}\eta) \tag{19}$$

Example 1 (Bayes rule [2])**.** *A diagnostic test has a probability* 0.95 *of giving a positive result when applied to a person suffering from a certain disease, and a probability* 0.10 *of giving a (false) positive when applied to a non-sufferer. It is estimated that* 0.5% *of the population are sufferers. Suppose that the test is now administered to a person about whom we have no relevant information relating to the disease (apart from the fact that he/she comes from this population). Calculate the following probabilities:*

1. *that the test result will be positive;*
2. *that, given a positive result, the person is a sufferer;*
3. *that, given a negative result, the person is a non-sufferer;*
4. *that the person will be misclassified.*

Let us first define the following random variables: T (Test positive), S (Sufferer), M (Misclassified). From given information we have $p(T|S) = 0.95$, $p(T|\neg S) = 0.1$, and $P(S) = 0.005$. Then

$$p(T) = \sum_S p(T|S)p(S) = p(T|S)p(S) + p(T|\neg S)p(\neg S) = 0.95 \times 0.005 + 0.1 \times (1 - 0.005) = 0.10425$$

$$p(S|T) = \frac{p(T|S)p(S)}{p(T)} = (0.95 \times 0.005)/0.10425 = 0.04556$$

$$p(\neg S|\neg T) = \frac{p(\neg T|\neg S)p(\neg S)}{p(\neg T)} = (1 - 0.1) \times (1 - 0.005)/(1 - 0.10425) = 0.99972$$

$$\begin{aligned} p(M) &= p(T, \neg S) + p(\neg T, S) = p(T|\neg S)p(\neg S) + p(\neg T|S)p(S) \\ &= 0.1 \times (1 - 0.005) + (1 - 0.95) \times 0.005 = 0.09975 \end{aligned}$$

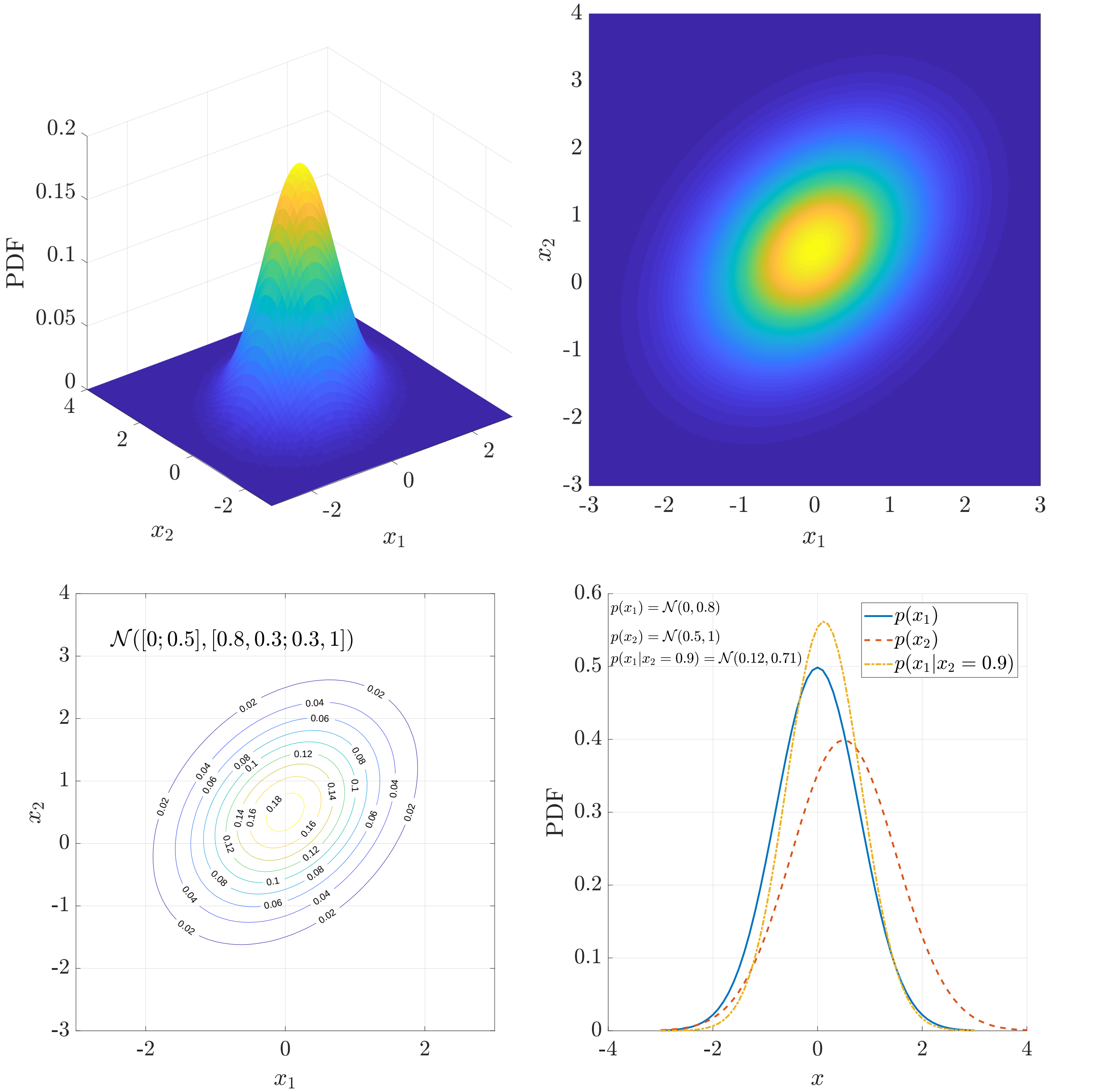

Figure 1: From top left, the plots show two-dimensional PDF, top view of the first plot, the contour plot of the PDF, and the marginals and the conditional distribution of $p(x_1|x_2 = 0.9)$.

Example 2 (Visualizing multivariate Gaussian)**.** *Let* $x = \mathrm{vec}(x_1, x_2)$ *and* $x \sim \mathcal{N}(\mu, \Sigma)$ *where*

$$\mu = \begin{bmatrix} 0.0 \\ 0.5 \end{bmatrix}, \Sigma = \begin{bmatrix} 0.8 & 0.3 \\ 0.3 & 1.0 \end{bmatrix}$$

Figure 1 shows the PDF plot of $\mathcal{N}(\mu, \Sigma)$ *as well as marginal and conditional distributions. See* `mvn_plots.m` *for more details and reproducing the plots.*

3 Sampling from Gaussian Distributions

Suppose we wish to draw samples from $Y \sim \mathcal{N}(\mu, \Sigma)$. If the covariance matrix of Y is not degenerate (is positive definite), we have $\Sigma = LL^{\mathsf{T}}$ where L is a lower triangular matrix computed using Cholesky decomposition. Now, let $X \sim \mathcal{N}(0, I_n)$ be a vector of standard normal random variables where n is the dimension (length) of Y. Define $Z := LX + \mu$. Using Proposition 2, we have $\mathbb{E}[Z] = \mu$ and $\mathrm{Cov}[Z] = LL^{\mathsf{T}}$. Therefore, using samples from $\mathcal{N}(0, 1)$, we are able to draw samples from $\mathcal{N}(\mu, \Sigma)$.

4 Sample Mean and Covariance

Sometimes we do not know the distribution of data, but instead we have access to samples or observations. Let $X = x$ be a random vector and $x_1, \dots, x_n$ be n independent samples (realization) of X. The sample or empirical mean, $\overline{x}$, and (unbiased) covariance, $\overline{\Sigma}$, can be computed as follows.

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{20}$$

$$\overline{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})^{\mathsf{T}} \tag{21}$$

Note that the sample mean is a random variable.

5 Chi-Square Distribution, Mahalanobis Distance, and Confidence Ellipsoid

Let $X \sim \mathcal{N}(\mu, \Sigma)$ be an n-dimensional Gaussian random vector. The scalar random variable, q, defined by the quadratic form

$$q = (x - \mu)^{\mathsf{T}} \Sigma^{-1} (x - \mu), \tag{22}$$

is the sum of the squares of n independent zero-mean, unity-variance Gaussian random variables. We say q has a chi-square distribution with n Degrees Of Freedom (DOF) [1], i.e., $q \sim \chi^2_n$. It can be shown that $\mathbb{E}[q] = n$ and $\mathbb{V}[q] = 2n$. If $q_1 \sim \chi^2_{n_1}$ and $q_2 \sim \chi^2_{n_2}$, then $q_3 = q_1 + q_2 \sim \chi^2_{n_1+n_2}$. Chi-square goodness-of-fit is a statistical test that determines if an observation sample comes from a specified probability distribution. In particular, we would like to test the null hypothesis that the data in x comes from a normal distribution such as $\mathcal{N}(\mu, \Sigma)$. This test is useful for data association. The distance $\sqrt{(x - \mu)^{\mathsf{T}} \Sigma^{-1} (x - \mu)}$ is known as Mahalanobis distance.

Remark 3. *The weighted sum of independent identically distributed (i.i.d.) chi-square random variables does not follow a chi-square distribution. For more details, see Bar-Shalom et al. [3, page 60].*

[1] In statistics, the number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary.

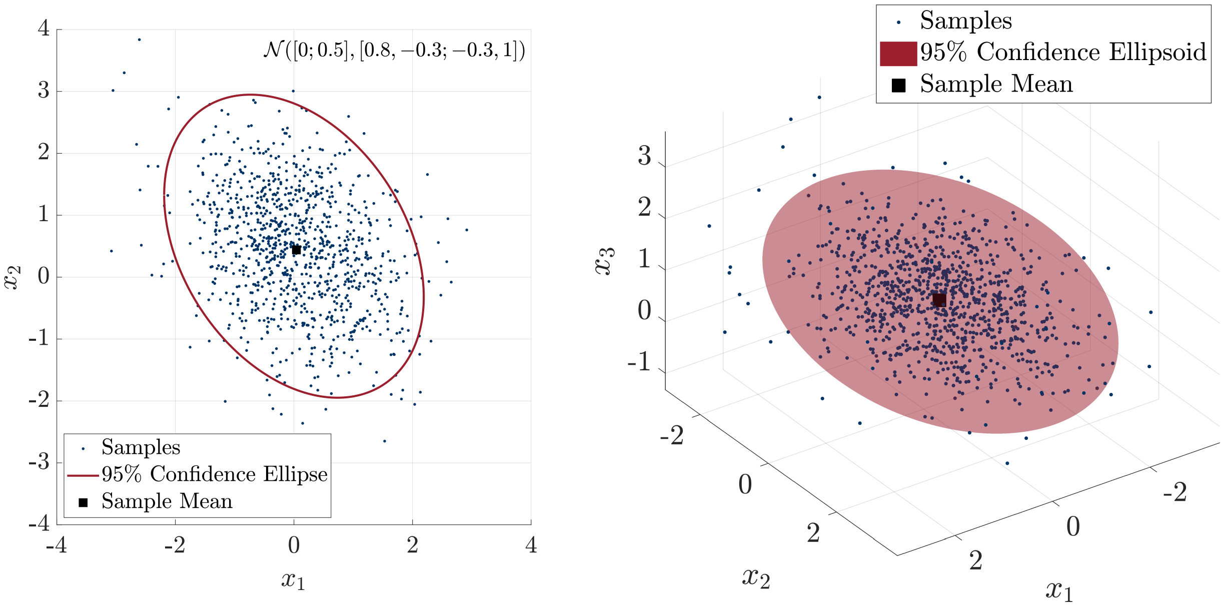

Figure 2: Illustrative examples of drawing confidence ellipsoids. See `confidence_ellipsoid_plots.m` for details and reproducing the plots.

Suppose we are a give a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, and we would like to illustrate the confidence region that the mentioned distribution covers with a certain probability. Geometrically, an equation such as $(x-b)^\mathsf{T} A(x-b) = 1$, where A is positive definite and $x, b \in \mathbb{R}^n$, corresponds to an ellipsoid in $\mathbb{R}^n$ centered at b. Since A is positive definite, we have $A = LL^\mathsf{T}$. It can be shown that L corresponds to a linear transformation that rotates and scale a sphere to the desired ellipsoid.

Now it is clear that $q = (x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu)$ also corresponds to an ellipsoid. The chi-square value, q, for a desired significance level (p-value) can be found using the pre-calculated chi-square table. Finally, to map a point from a unit sphere, x_s, to the desired ellipsoid, x_e, we can use the following formula:

$$x_e = \sqrt{q} L x_s + \mu, \tag{23}$$

where L is the Cholesky factor of the covariance matrix, i.e., $\Sigma = LL^\mathsf{T}$. Figure 2 shows two examples of the 95% confidence ellipsoid where samples are drawn from known normal distributions.

6 Resources

Here I list some of the good references for background knowledge. Of course, there are many other notable books and resources available. For probability and statistics see Papoulis and Pillai [4]. For optimal filtering including Kalman filtering, linear systems, and estimation see Anderson and Moore [5]. For a classic textbook on estimation with a focus on target tracking and navigation that also covers an introduction to probability theory see Bar-Shalom et al. [3].

References

[1] "Multivariate normal distribution," https://en.wikipedia.org/wiki/Multivariate_normal_distribution, accessed: 2022-01-01.

[2] "Examples of Bayes' theorem in practice," http://wwwf.imperial.ac.uk/~atw/Bayes.pdf, accessed: 2017-12-26.

[3] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2001.

[4] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*. McGraw-Hill Education, 2002.

[5] B. D. Anderson and J. B. Moore, *Optimal filtering*. Englewood Cliffs, 1979.