# Statistical Signal Processing

*Lecture 1*

- course arrangements
- course overview
- chapter 1: parameter estimation
  stochastic parameters
  - the multivariate Gaussian distribution; Gauss-Markov theorem
  - the parameter estimation problem
  - Bayes estimation: the MMSE, absolute value and uniform cost functions
  - examples: Gaussian mean in Gaussian noise, Poisson process

# Course Arrangements

- course assistants:

  Francesco Negro, Siouar Bensaid

- theoretical part: exam (counts for 14/20), open notes and handouts

- practical part:
  - problem sessions (TDs): 2 not graded
  - computer sessions (TPs): 1 (Matlab) graded (counts for 3/20)
    $\Rightarrow$ attendance compulsory
  - homework (HW): 1 graded (counts for 3/20)

- emphasis (in exam) on problem solving, knowing how to apply the theory
  $\Rightarrow$ TDs important

- course notes and copies of viewgraphs will appear throughout the term, also on:
  web: http://teaching.eurecom.fr/courses/SSP or less updated:
http://intranet.eurecom.fr/academicAffairs/cursus_d/courses_content/techcourses/ssp_d.htm

# Course Overview

- Chapter 0: Background Material
- Chapter 1: Parameter Estimation
- Chapter 2: Spectrum Estimation
- Chapter 3: Optimal Filtering
- Chapter 4: Adaptive Filtering
- Chapter 5: Estimation of Sinusoids in Noise

# Chapter 0: Background Material

See also the MathEng (Mathematical Methods for Engineers) and Optim (Fundamentals of Optimization) courses and a

Linear Algebra review at http://teaching.eurecom.fr/courses/SSP/

- probability
- linear algebra
- multivariate Gaussian distribution, Gauss-Markov theorem
- complex Gaussian variables, circularity
- vector spaces, inner products, norms
- optimization

# Chapter 1: Parameter Estimation

- **random parameters**, Bayesian estimation, minimum mean squared error (MMSE) estimation, maximum a posteriori (MAP) estimation, equivalent estimators
- MMSE and MAP estimation for vector parameters, Cramer-Rao bound, linear MMSE estimators, orthogonality principle, the linear model
- **deterministic parameters**, uniformly minimum variance unbiased estimators (UMVUE), maximum likelihood (ML) estimation, properties of estimators (bias, efficiency, consistency), Cramer-Rao bound, Estimation-Maximization (EM) algorithm
- best linear unbiased estimator (BLUE), least-squares techniques, method of moments, the linear model
- Advanced: model order selection, case of complex measurements and/or parameters, ML optimization techniques (steepest-descent, Gauss-Newton, alternating opt., scoring), Estimation-Maximization (EM), Variational Bayesian techniques (VB), Compressed Sensing

# Chapter 2: Spectrum Estimation

- spectrum estimation = parameter estimation when parameters = spectrum
- non-parametric techniques, periodogram, windowing, spectral leakage, averaged periodogram, smoothed periodogram
- parametric random process models, autoregressive (AR) processes, moving average (MA) processes, autoregressive moving average (ARMA) processes
- parametric techniques, linear prediction, autoregressive modeling, maximum entropy, Levinson & Schur algorithms (relation to triangular covariance matrix factorization), lattice filters
- time and frequency domain localization, short-time Fourier transform, QMF, subbands, perfect reconstruction filter banks, wavelet transform, hierarchical signal representation/approximation

# Chapter 3: Optimal Filtering

- optimal filtering = Bayesian parameter estimation when parameters = signal
- **Wiener filtering**: unrealizable (non-causal), causal and FIR; lattice/ladder filters
- application to linear and decision-feedback equalization
- some elements from optimization theory, steepest descent algorithm
- linear state-space models
- **Kalman filtering**
  developed for space applications, relation to Levinson/Schur algorithms, Chandrasekhar equations
- applications (channel estimation)

# Chapter 4: Adaptive Filtering

- adaptive filtering = optimal filtering in absence of statistical knowledge
- adaptive FIR filtering
- least-mean-square (LMS) algorithm
- recursive least-squares (RLS) algorithm
- tracking of time-varying parameters, performance analysis
- applications

# Chapter 5: Estimation of Sinusoids in Noise

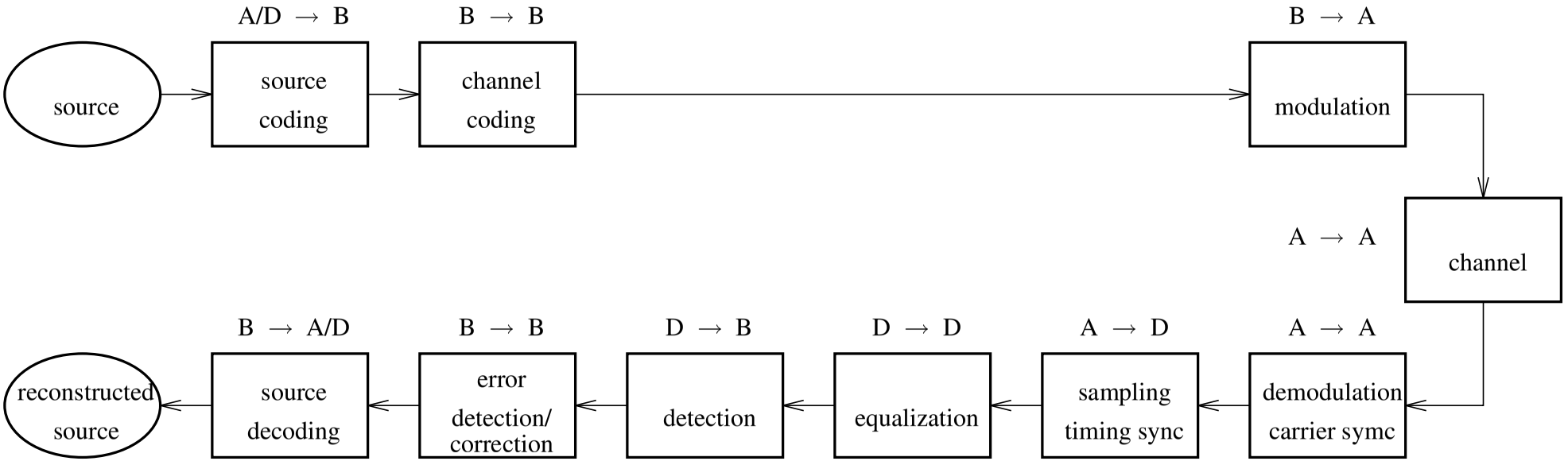
- special focus on the estimation of the sinusoid frequencies
- application of deterministic parameter estimation techniques:
  - maximum likelihood: IQML, DIQML, PQML
  - subspace fitting (method of moments)
  - linear prediction: Prony's method, Pisarenko's method
  - linear constrained minimum variance (LCMV): Capon's method
- application of adaptive filtering:
  - adaptive notch filtering

# DSP Relevance to Eurecom

- leading thread in DSP course: *statistical signal processing*. Signals as realizations of stochastic processes.

- key building block in telecommunications: *modem*
  - key issue: bit detection: Digital Communications course
  - carrier synchronization: recovery of the carrier frequency and phase
  - timing synchronization: recovery of baud rate and phase
  - channel estimation (or its inverse) for equalization, echo or interference cancellation

# MODEM Building Blocks




- modem: modulation-demodulation: digital transmission over analog medium
- signals appear in analog (A), digital (D) or binary/coded (B) form
- DigCom: channel (de)coding, (de)modulation, sampling, detection, equalization
- SSP: source (de)coding, equalization/channel estimation, parameter estimation for timing and carrier synchronization

# Applications in Corporate Communications

- source compression: (e.g. Unix commands compress and gzip), appetite for information steadily increasing (multimedia)

- hierarchical image compression: nested structure: information at a higher definition level consists of the information at a lower level plus complementary information.
  - result: info split up into streams corresponding to the details at various levels of definition
  - interaction with the network: send different streams separately, with different levels of transmission quality (success of arrival, delay,...)
  - consequences for security: not all streams need to be encrypted (to the same extent)

- communication network performance analysis: queuing theory. parameterized statistical models (e.g. Poisson processes for arrival of packets) need an estimation of their parameter values.

# Applications in Multimedia Communications

- lossy source coding for audio and video as opposed to lossless compression of data files
- videoconferencing systems
  - adaptive filters for acoustic echo cancellation (marketing studies: audio more important than video)
  - parametric face models (low bitrate TX via facial cloning), face recognition (biometry), gesture recognition, etc.
- speech recognition for voice-controlled operation: hampered by background noise (other speakers etc.). solution: speech enhancement and directive acoustic microphone arrays
- synchronizaton of multiple information streams (audio & video & ...)
- transmission of continuous sources (speech, video) over packet-mode networks (internet)
- xDSL (Digital Subscriber Loop): high-speed connectivity over the last mile

# Applications in Mobile Communications

- digital mobile communications more reliable and flexible than analog systems: requires voice coding; very limited bandwidth $\Rightarrow$ sophisticated low bit rate coding techniques required; changing environment $\Rightarrow$ adaptive multi-rate coding.
- National regulations prohibiting simultaneous driving and calling $\Rightarrow$ handsfree telephone systems $\Rightarrow$ audio conferencing systems and acoustic echos problem. Handsfree talking and dialing: speech recognition techniques in a noisy environment
- multipath propagation in a mobile environment: fast channel equalization techniques needed, possibly of limited complexity, carrier and symbol rate synchronization, Doppler and fading tracking
- user localization (911): position estimation on basis of delay (ToA) and/or direction (DoA) estimates (Time/Direction of Arrival)
- cellular communications and frequency reuse: communications limited, not by noise, but by other users (interferers). Sophisticated antenna array processing permits Spatial Division Multiple Access (SDMA): simultaneous interference reduction and multipath reduction, DoA estimation; spatio-temporal processing; interference reduction in Code DMA (CDMA) systems.

# The Multivariate Gaussian Distribution

- $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_m]^T$ : $m$ jointly Gaussian random variables: $\mathbf{X} \sim \mathcal{N}(m_X, C_{XX})$ with mean

$$m_X = E\mathbf{X} = \int dx_1 \cdots \int dx_m f_{\mathbf{X}}(X)\, X$$

$$= \begin{bmatrix} \int dx_1 \cdots \int dx_m f_{\mathbf{X}}(X)\, x_1 \\ \vdots \\ \int dx_1 \cdots \int dx_m f_{\mathbf{X}}(X)\, x_m \end{bmatrix} = \begin{bmatrix} \int dx_1\, x_1 \overbrace{\int dx_2 \cdots \int dx_m f_{\mathbf{X}}(X)}^{f_{\mathbf{x}_1}(x_1)} \\ \vdots \\ \int dx_m\, x_m \underbrace{\int dx_1 \cdots \int dx_{m-1} f_{\mathbf{X}}(X)}_{f_{\mathbf{x}_m}(x_m)} \end{bmatrix} = \begin{bmatrix} m_{x_1} \\ \vdots \\ m_{x_m} \end{bmatrix}$$

and covariance matrix

$$C_{XX} = E(\mathbf{X}-m_X)(\mathbf{X}-m_X)^T = \int dx_1 \cdots \int dx_m f_{\mathbf{X}}(X) \underbrace{(X-m_X)(X-m_X)^T}_{\geq 0\,,\ \text{rank } 1}$$

where $C_{x_i x_j} = E(\mathbf{x}_i - m_{x_i})\left(\mathbf{x}_j - m_{x_j}\right)$. As weighted average of positive semi-definite matrices: $C_{XX} = C_{XX}^T \geq 0$ symmetric and positive semidefinite: $\forall U \in \mathcal{R}^m : U^T C_{XX} U = U^T E(\mathbf{X}-m_X)(\mathbf{X}-m_X)^T U = E\left(U^T(\mathbf{X}-m_X)\right)^2 \geq 0$

- joint Gaussian probability density function (pdf):

$$f_{\mathbf{X}}(X) = (2\pi)^{-\frac{m}{2}} (\det C_{XX})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}[X-m_X]^T C_{XX}^{-1}[X-m_X]\right)$$

# Multivariate Gaussian Derivation

- goal: derive multivariate Gaussian distribution from univariate Gaussian distribution and two postulates
- $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_m]^T$ real random vector with specified mean and covariance matrix: with mean $m_X = E\mathbf{X} = [m_{x_1} \cdots m_{x_m}]^T$ and covariance matrix

$$\begin{aligned} C_{XX} &= E\left(\mathbf{X}-m_X\right)\left(\mathbf{X}-m_X\right)^T = \left[E\left(\mathbf{x}_i-m_{x_i}\right)\left(\mathbf{x}_j-m_{x_j}\right)\right]_{i,j=1}^{m} \\ &= E\left(\mathbf{X}\mathbf{X}^T-\mathbf{X}m_X^T-m_X\mathbf{X}^T+m_Xm_X^T\right) \\ &= \left(E\mathbf{X}\mathbf{X}^T\right)-\left(E\mathbf{X}\right)m_X^T-m_X\left(E\mathbf{X}\right)^T+m_Xm_X^T \\ &= R_{XX}-m_Xm_X^T-m_Xm_X^T+m_Xm_X^T = R_{XX}-m_Xm_X^T \end{aligned}$$

  $R_{XX} = E\mathbf{X}\mathbf{X}^T$ = correlation matrix,
  linearity of expectation $E$ exploited (linear operations commute)
- by definition of expectation

  $$C_{XX} = E\left(\mathbf{X}-m_X\right)\left(\mathbf{X}-m_X\right)^T = \int dx_1 \cdots \int dx_m f_{\mathbf{X}}(X)\left(X-m_X\right)\left(X-m_X\right)^T .$$

  $\Rightarrow \quad C_{XX} = C_{XX}^T$ symmetric

# Eigendecomposition Covariance Matrix

- *eigenvalues* $\lambda_i$ and corresponding *eigenvectors* $V_i$ of $C_{XX}$ : $C_{XX}\,V_i \;=\; \lambda_i\,V_i$
  fix *norm* $\|V_i\| = 1, \quad \|V_i\|^2 = V_i^T V_i$

- $(C_{XX} - \lambda_i\,I_m)\,V_i = 0 \;\Rightarrow\; (C_{XX} - \lambda_i\,I_m)$ *singular*
  $\lambda_i$ solution of $\det(C_{XX} - \lambda\,I_m) = 0$ : *characteristic equation*

- $C_{XX} = C_{XX}^T \;\Rightarrow\; \lambda_i \in \mathcal{R}\,,\; V_i^T V_j = \delta_{ij}\,,\; i,j = 1,\ldots,n$

  *Kronecker delta* : $\delta_{ij} = \begin{cases} 1 \;, & i = j \\ 0 \;, & i \neq j \end{cases}$

- matrix $V = [V_1 \cdots V_m]\;(m \times m)$ :
  $\left[V^T V\right]_{ij} = V_i^T V_j = \delta_{ij} \;\Rightarrow\; V^T V = I_m \quad V$ = *orthogonal matrix*

- $C_{XX}$ real, $\lambda_i$ real $\;\Rightarrow\;$ $V_i$ can be chosen to be real

- $I_m = V^T V \;\Rightarrow\; 1 = \det(I_m) = \det\left(V^T V\right) = \det\left(V^T\right)\,\det\left(V\right) = \left(\det V\right)^2$
  $\Rightarrow\; \det V = \pm 1$. We can choose the signs of the $V_i$ such that $\det V = 1$.

# Eigendecomposition Covariance Matrix (2)

- $C_{XX} \geq 0$ positive semidefinite:

$$\forall U \in \mathcal{R}^m : U^T C_{XX} U = E\left(U^T(\mathbf{X} - m_X)\right)^2 \geq 0$$

  (positive definite would be: $\forall U \in \mathcal{R}^m \setminus \{0\} : U^T C_{XX} U > 0$)

- $U = V_i : U^T C_{XX} U = V_i^T C_{XX} V_i = \lambda_i V_i^T V_i = \lambda_i \geq 0$
- order the $\lambda_i : \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$
- If $\lambda_m = 0$, $C_{XX}$ is singular $\Rightarrow$ $V_m^T C_{XX} V_m = E\left(V_m^T(\mathbf{X} - m_X)\right)^2 = 0$ mean and variance of $V_m^T(\mathbf{X} - m_X)$ are zero $\Rightarrow$ $V_m^T(\mathbf{X} - m_X) = 0$ in mean square. This means that at least one variable $\mathbf{x}_i$ is a linear combination of the other variables and 1. We shall in general exclude this possibility $\Rightarrow$ $C_{XX} > 0$, $\lambda_i > 0$, $i = 1, \ldots, m$
- $C_{XX} V_i - V_i \lambda_i = 0$ are the columns of the matrix $C_{XX} V - V\Lambda = 0$ where $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_m\}$. Using $V^{-1} = V^T$, we find

$$C_{XX} = V\Lambda V^T = [V_1 \cdots V_m] \,\text{diag}\{\lambda_1, \ldots, \lambda_m\} [V_1 \cdots V_m]^T = \sum_{i=1}^{m} \lambda_i V_i V_i^T$$

  Can show: $C_{XX}^{\alpha} = V\Lambda^{\alpha} V^T$ for any $\alpha$ (e.g. $\alpha = 2, -1$)

# Multivariate Gaussian Derivation (2)

- Consider now a linear transformation of variables: $\mathbf{Z} = V^T(\mathbf{X} - m_X)$.
  Then we find for the first two moments

$$\begin{aligned} m_Z &= E\,V^T(\mathbf{X} - m_X) = V^T(E\mathbf{X} - m_X) = V^T(m_X - m_X) = 0 \\ C_{ZZ} &= E\,(\mathbf{Z} - m_Z)\,(\mathbf{Z} - m_Z)^T = E\mathbf{Z}\mathbf{Z}^T = EV^T\,(\mathbf{X} - m_X)\,(\mathbf{X} - m_X)^T\,V \\ &= V^T\left(E\,(\mathbf{X} - m_X)\,(\mathbf{X} - m_X)^T\right)V = V^T C_{XX} V = V^T V\,\Lambda\,V^T V = \Lambda \end{aligned}$$

  Or hence $E\mathbf{z}_i\mathbf{z}_j = \lambda_i\delta_{ij}$: the $\mathbf{z}_i$ are zero mean and uncorrelated.

- At this point: only specified the first two moments of the $\mathbf{z}_i$. Now specify the rest of their distribution by stating that the $\mathbf{z}_i$ are jointly Gaussian. We furthermore *postulate* that zero mean uncorrelated Gaussian random variables are independent.

Note that in general $\left.\begin{matrix}\text{independent}\\ \text{zero mean}\end{matrix}\right\} \begin{matrix}\Longrightarrow\\ \not\Longleftarrow\end{matrix} \left\{\begin{matrix}\text{uncorrelated}\\ \text{zero mean}\end{matrix}\right.$

# Multivariate Gaussian Derivation (3)

- joint distribution of the independent Gaussian r.v.'s $\mathbf{z}_i$ :

$$f_{\mathbf{Z}}(Z) = \prod_{i=1}^{m} f_{\mathbf{z}_i}(z_i) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\,\lambda_i}} \exp\left(-\frac{z_i^2}{2\lambda_i}\right) = (2\pi)^{-\frac{m}{2}} (\det \Lambda)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} Z^T \Lambda^{-1} Z\right)$$

- furthermore *postulate* that a linear transformation of jointly Gaussian random variables produces again jointly Gaussian random variables.

- Since the Jacobian $(\det V^T)$ of the linear transformation between $\mathbf{X}$ and $\mathbf{Z}$ equals one, we get for the joint distribution of the $\mathbf{x}_i$ $\qquad Z = V^T(X - m_X)$

$$\begin{aligned} f_{\mathbf{X}}(X) &= f_{\mathbf{Z}}\left(V^T(X - m_X)\right) \\ &= (2\pi)^{-\frac{m}{2}} \left(\det(V^T C_{XX} V)\right)^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}\left[V^T(X - m_X)\right]^T \Lambda^{-1} \left[V^T(X - m_X)\right]\right) \\ &= (2\pi)^{-\frac{m}{2}} \left((\det V^T)(\det C_{XX})(\det V)\right)^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}\left[X - m_X\right]^T V \Lambda^{-1} V^T \left[X - m_X\right]\right) \\ &= (2\pi)^{-\frac{m}{2}} (\det C_{XX})^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}\left[X - m_X\right]^T C_{XX}^{-1} \left[X - m_X\right]\right) \end{aligned}$$

This is the general expression for a multivariate Gaussian probability density function (pdf). Notation: $\mathbf{X} \sim \mathcal{N}(m_X, C_{XX})$ : completely specified in terms of the first and second-order moments.

# The Multivariate Gaussian Distribution (2)

- $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_m]^T$ and $\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_n]^T$ : $m{+}n$ jointly Gaussian random variables:

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} m_X \\ m_Y \end{bmatrix}, \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix} \right)$$

with mean and covariance matrix

$$\begin{bmatrix} m_X \\ m_Y \end{bmatrix} = E\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \;,\; C = \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix} = E\begin{bmatrix} \mathbf{X}{-}m_X \\ \mathbf{Y}{-}m_Y \end{bmatrix}\begin{bmatrix} \mathbf{X}{-}m_X \\ \mathbf{Y}{-}m_Y \end{bmatrix}^T$$

- joint Gaussian probability density function (pdf): $f_{\mathbf{X},\mathbf{Y}}(X, Y)$

$$= (2\pi)^{-\frac{m+n}{2}} (\det C)^{-\frac{1}{2}} \exp\left( -\frac{1}{2} \begin{bmatrix} X - m_X \\ Y - m_Y \end{bmatrix}^T \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix}^{-1} \begin{bmatrix} X - m_X \\ Y - m_Y \end{bmatrix} \right)$$

- meaning of the joint pdf:

$$\Pr[\mathbf{X} \in (X, X + dX), \mathbf{Y} \in (Y, Y + dY)] = f_{\mathbf{X},\mathbf{Y}}(X, Y) dX\, dY$$

where first $dX = [dx_1 \cdots dx_m]^T$ and then $dX = dx_1\, dx_2 \cdots dx_m$

# The Conditional Gaussian Distribution

- the conditional pdf $f_{\mathbf{X}|\mathbf{Y}}(X|Y)$ means

$$\Pr\left[\mathbf{X} \in (X, X+dX)|\mathbf{Y} = Y\right] \;=\; f_{\mathbf{X}|\mathbf{Y}}(X|Y)dX$$

- the conditional pdf is defined by Bayes' rule:

$$f_{\mathbf{X}|\mathbf{Y}}(X|Y) \;=\; \frac{f_{\mathbf{X},\mathbf{Y}}(X,Y)}{f_{\mathbf{Y}}(Y)} \;\Leftrightarrow\; f_{\mathbf{X},\mathbf{Y}}(X,Y) \;=\; f_{\mathbf{Y}}(Y)\, f_{\mathbf{X}|\mathbf{Y}}(X|Y)$$

get the marginal pdf $f_{\mathbf{Y}}(Y)$ from the joint pdf by integrating out $X$:

$$\begin{aligned} f_{\mathbf{Y}}(Y) \;&=\; \textstyle\int f_{\mathbf{X},\mathbf{Y}}(X,Y)dX \\ &=\; \textstyle\int\cdots\int f_{\mathbf{x}_1,\ldots,\mathbf{x}_m,\mathbf{y}_1,\ldots,\mathbf{y}_n}(x_1,\ldots,x_m,y_1,\ldots,y_n)dx_1\cdots dx_m \\ &=\; (2\pi)^{-\frac{n}{2}}\left(\det C_{YY}\right)^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}\left[Y-m_Y\right]^T C_{YY}^{-1}\left[Y-m_Y\right]\right) \end{aligned}$$

- consider the block Upper Diagonal Lower (UDL) triangular factorization of $C$:

$$C \;=\; \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix} \;=\; \begin{bmatrix} I & K \\ 0 & I \end{bmatrix}\begin{bmatrix} P & 0 \\ 0 & C_{YY} \end{bmatrix}\begin{bmatrix} I & 0 \\ K^T & I \end{bmatrix}$$

with $K \;=\; C_{XY}C_{YY}^{-1}\,,\;\; P \;=\; C_{XX} - C_{XY}C_{YY}^{-1}C_{YX}$ (*Schur complement*)

# The Conditional Gaussian Distribution (2)

- from the UDL factorization of $C$, we can obtain the LDU factorization of $C^{-1}$

$$\begin{aligned} C^{-1} &= \begin{bmatrix} I & 0 \\ K^T & I \end{bmatrix}^{-1} \begin{bmatrix} P & 0 \\ 0 & C_{YY} \end{bmatrix}^{-1} \begin{bmatrix} I & K \\ 0 & I \end{bmatrix}^{-1} \\ &= \begin{bmatrix} I & 0 \\ -K^T & I \end{bmatrix} \begin{bmatrix} P^{-1} & 0 \\ 0 & C_{YY}^{-1} \end{bmatrix} \begin{bmatrix} I & -K \\ 0 & I \end{bmatrix} \end{aligned}$$

- we can rewrite the exponent of the joint distribution as

$$\begin{bmatrix} X - m_X \\ Y - m_Y \end{bmatrix}^T \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix}^{-1} \begin{bmatrix} X - m_X \\ Y - m_Y \end{bmatrix}$$

$$= \begin{bmatrix} X - m_X \\ Y - m_Y \end{bmatrix}^T \begin{bmatrix} I & 0 \\ -K^T & I \end{bmatrix} \begin{bmatrix} P^{-1} & 0 \\ 0 & C_{YY}^{-1} \end{bmatrix} \begin{bmatrix} I & -K \\ 0 & I \end{bmatrix} \begin{bmatrix} X - m_X \\ Y - m_Y \end{bmatrix}$$

$$= \begin{bmatrix} X - KY - (m_X - Km_Y) \\ Y - m_Y \end{bmatrix}^T \begin{bmatrix} P^{-1} & 0 \\ 0 & C_{YY}^{-1} \end{bmatrix} \begin{bmatrix} X - KY - (m_X - Km_Y) \\ Y - m_Y \end{bmatrix}$$

$$\begin{aligned} &= [X - KY - (m_X - Km_Y)]^T P^{-1} [X - KY - (m_X - Km_Y)] \\ &\quad + [Y - m_Y]^T C_{YY}^{-1} [Y - m_Y] \end{aligned}$$

# The Gauss-Markov Theorem

- By taking determinants, we also obtain

$$\begin{aligned}\det C &= \det\begin{bmatrix} I & K \\ 0 & I \end{bmatrix} \det\begin{bmatrix} P & 0 \\ 0 & C_{YY} \end{bmatrix} \det\begin{bmatrix} I & 0 \\ K^T & I \end{bmatrix} \\ &= 1 \;.\; \det\begin{bmatrix} P & 0 \\ 0 & C_{YY} \end{bmatrix} \;.\; 1 = \det P \;\; \det C_{YY}\end{aligned}$$

- **Theorem 0.1 (Gauss-Markov)** *If* $\mathbf{X}$ *and* $\mathbf{Y}$ *have the joint Gaussian distribution indicated before, then the conditional distribution is*

$$\begin{aligned}f_{\mathbf{X}|\mathbf{Y}}(X|Y) &= (2\pi)^{-\frac{m}{2}} (\det P)^{-\frac{1}{2}} \\ &\exp\left(-\tfrac{1}{2}\left[X-KY-(m_X-Km_Y)\right]^T P^{-1}\left[X-KY-(m_X-Km_Y)\right]\right)\end{aligned}$$

*The conditional distribution is again Gaussian with conditional mean*

$$E_{\mathbf{X}|\mathbf{Y}}\mathbf{X} = E\left[\mathbf{X}|\mathbf{Y}=Y\right] = m_X + \underbrace{C_{XY}C_{YY}^{-1}}_{=K}(Y-m_Y)$$

*and conditional covariance matrix*

$$E_{\mathbf{X}|\mathbf{Y}}\left[\mathbf{X}-E_{\mathbf{X}|\mathbf{Y}}\mathbf{X}\right]\left[\mathbf{X}-E_{\mathbf{X}|\mathbf{Y}}\mathbf{X}\right]^T = P = C_{XX}-C_{XY}C_{YY}^{-1}C_{YX}$$

# Example 1.1 Two correlated Gaussian r.v.'s

- $m = n = 1$, zero means $m_X = m_Y = 0$, rename $C_{XX} = \sigma_x^2$, $C_{YY} = \sigma_y^2$.
- Cauchy-Schwarz inequality : $|\langle \mathbf{x}, \mathbf{y} \rangle|^2 \leq \langle \mathbf{x}, \mathbf{x} \rangle \langle \mathbf{y}, \mathbf{y} \rangle$
- $\langle \mathbf{x}, \mathbf{y} \rangle = E\mathbf{xy} \quad \Rightarrow \quad C_{XY}^2 \leq C_{XX} C_{YY}$
- isomorphism: $\mathbf{x} = \underline{x}^T \underline{\mathbf{e}}$, $\mathbf{y} = \underline{y}^T \underline{\mathbf{e}}$, $R_{\underline{\mathbf{ee}}} = I$, $\Rightarrow$ $E\mathbf{xy} = \langle \mathbf{x}, \mathbf{y} \rangle = \langle \underline{x}, \underline{y} \rangle = \underline{x}^T \underline{y}$
- $E(\mathbf{x} + \lambda \mathbf{y})^2 = \lambda^2 r_{yy} + 2\lambda r_{xy} + r_{xx} \overset{\lambda = -r_{xy}/r_{yy}}{\geq} r_{xx} - \dfrac{r_{xy}^2}{r_{yy}} \geq 0 \;\Rightarrow\; r_{xy}^2 \leq r_{xx} r_{yy}$
- introduce normalized correlation coefficient $\rho = \dfrac{C_{XY}}{\sqrt{C_{XX} C_{YY}}} \in [-1, 1]$ can rewrite $C_{XY}$ as $C_{XY} = \rho \sigma_x \sigma_y$.
- The joint Gaussian distribution of $\mathbf{x}$ and $\mathbf{y}$ can be written as

$$f_{\mathbf{x},\mathbf{y}}(x, y) \;\leftrightarrow\; \mathcal{N}\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \right)$$

- The conditional pdf of $\mathbf{x}$ given $\mathbf{y}$ can be written as

$$f_{\mathbf{x}|\mathbf{y}}(x|y) \;\leftrightarrow\; \mathcal{N}\left( \rho \frac{\sigma_x}{\sigma_y} y, \sigma_x^2 (1 - \rho^2) \right)$$

# Example 1.1 Two correlated Gaussian r.v.'s (2)

- if $\mathbf{x}$, $\mathbf{y}$ uncorrelated ($\rho = 0$), then $f_{\mathbf{x}|\mathbf{y}}(x|y) = f_{\mathbf{x}}(x)$ : $\mathbf{x}$, $\mathbf{y}$ independent

- as $|\rho| \to 1$, $|E_{\mathbf{x}|\mathbf{y}}\mathbf{x}| \nearrow$ , conditional variance $\to 0$ :
  when $\mathbf{y} = y$ is known, it gives us some information about $\mathbf{x}$ and the residual randomness in $\mathbf{x}$ decreases as $|\rho| \to 1$.

- Extreme cases:

$$\begin{aligned} \rho = 1 \ : \quad & \mathbf{x} = \tfrac{\sigma_x}{\sigma_y}\mathbf{y} \quad && \mathbf{x} \text{ and } \mathbf{y} \text{ are perfectly correlated,} \\ \rho = -1 \ : \quad & \mathbf{x} = -\tfrac{\sigma_x}{\sigma_y}\mathbf{y} && \mathbf{x} \text{ and } \mathbf{y} \text{ are perfectly anticorrelated.} \end{aligned}$$

- *concentration ellipse*:

$$f_{\mathbf{x},\mathbf{y}}(x,y) = c' \iff \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}^{-1} \begin{bmatrix} x \\ y \end{bmatrix} = c$$

$$\iff \frac{\left(x - \rho\frac{\sigma_x}{\sigma_y}y\right)^2}{\sigma_x^2(1-\rho^2)} + \frac{y^2}{\sigma_y^2} = c$$

# Example 1.1 Two correlated Gaussian r.v.'s (3)

- the ellipse for $c = 1$

$$\frac{\left(x - \rho\frac{\sigma_x}{\sigma_y}y\right)^2}{\sigma_x^2(1-\rho^2)} + \frac{y^2}{\sigma_y^2} \leq 1$$

  contains a significant portion of the probability mass.

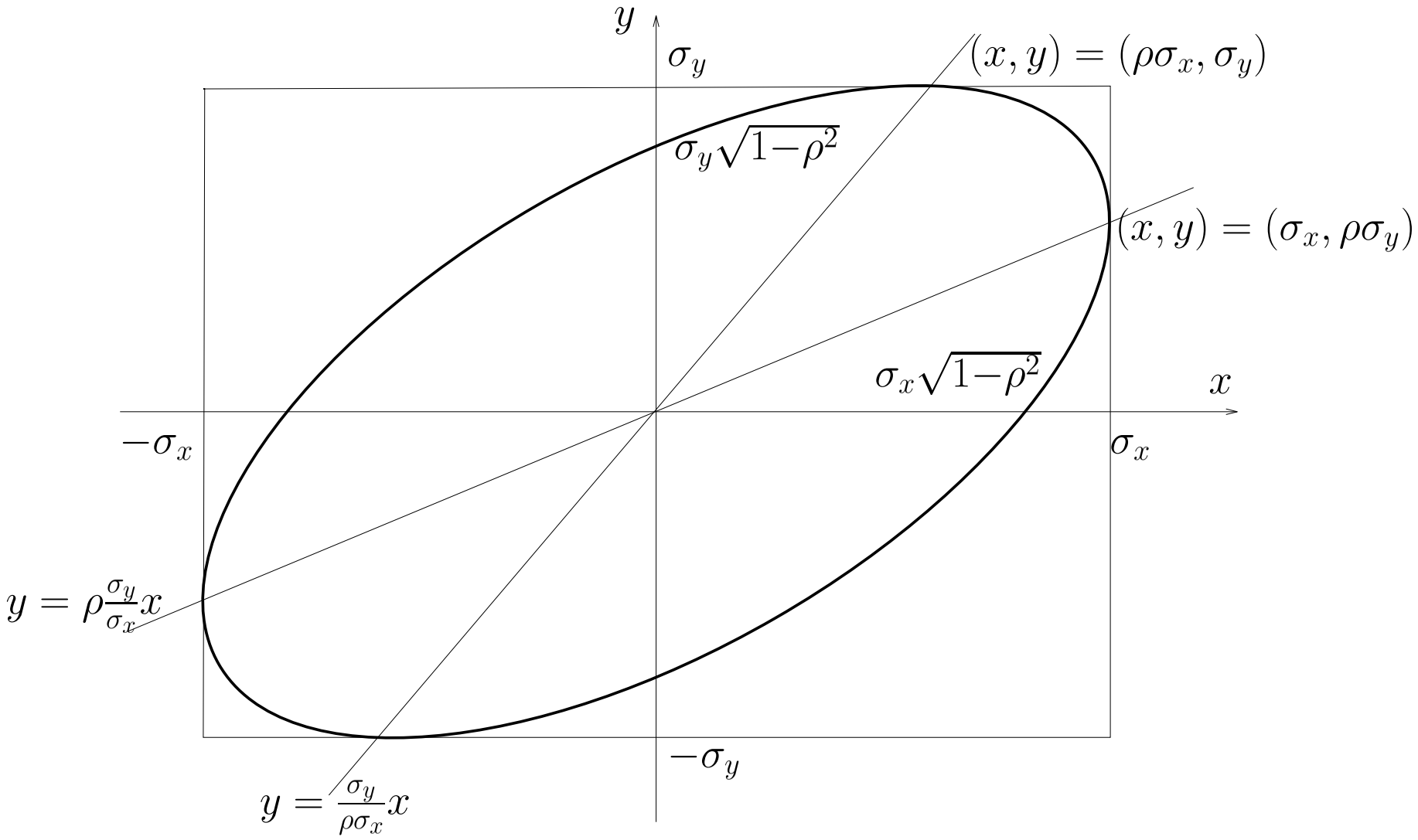
- its volume is

$$V \;=\; \pi\left(\det C\right)^{\frac{1}{2}} \;=\; \pi\sigma_x\sigma_y\sqrt{1-\rho^2} \quad \in \; [0, \pi\sigma_x\sigma_y]$$

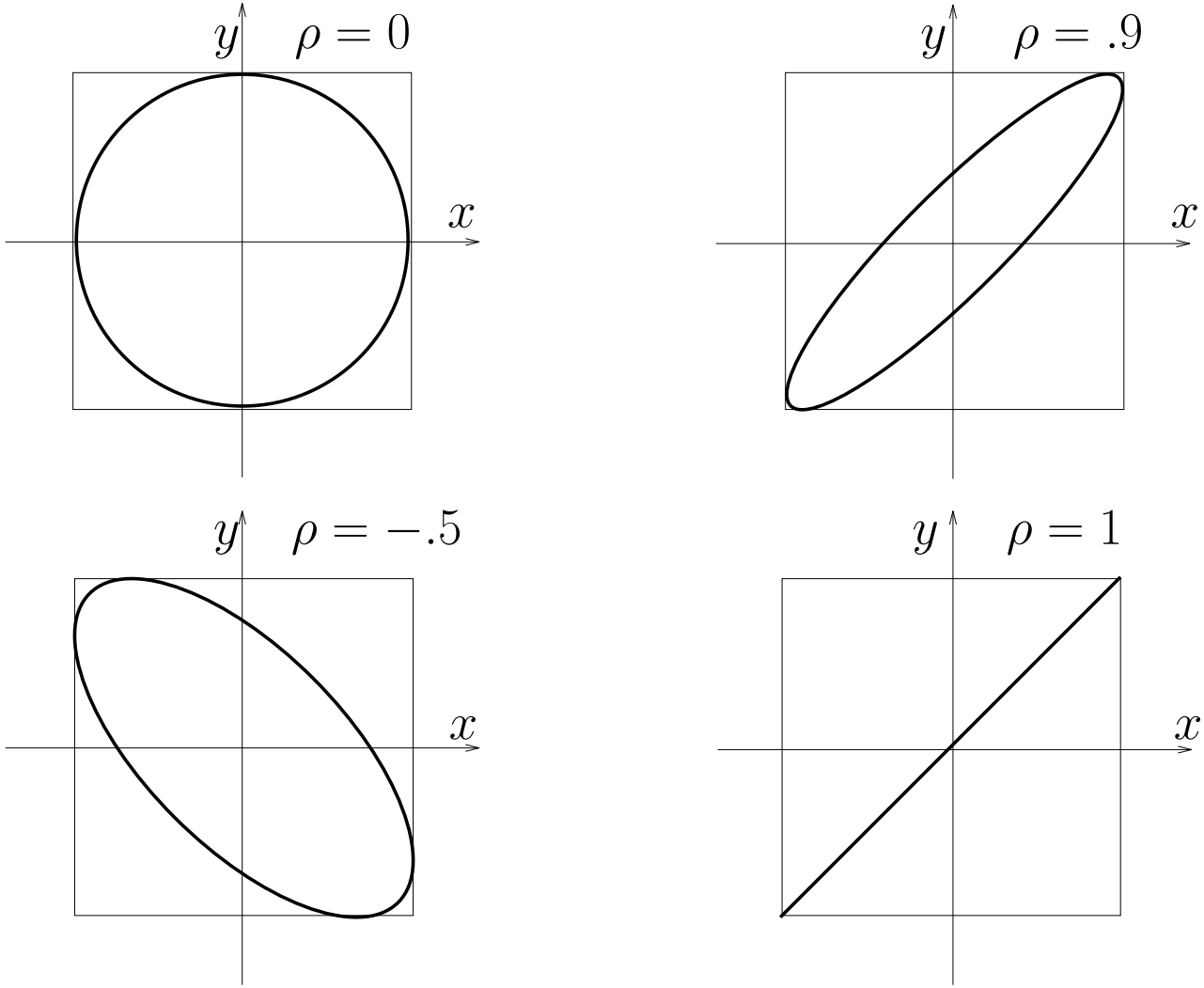
- for strongly correlated variables, the pair $(\mathbf{x}, \mathbf{y})$ takes with high probability values in a fairly small area. Hence, in some sense the randomness (or entropy) of the pair $(\mathbf{x}, \mathbf{y})$ decreases as $|\rho| \to 1$. (later: Gaussian r.v.'s: entropy $\sim \det C$)

# Example 1.1 Two correlated Gaussian r.v.'s (4)

concentration ellipses



# Example 1.1 Two correlated Gaussian r.v.'s (5)



# Gaussian r.v.'s and Linear Models

- by interchanging $X$ and $Y$, we find the conditional pdf of $\mathbf{Y}$ given $\mathbf{X}$

$$f_{\mathbf{Y}|\mathbf{X}}(Y|X) \leftrightarrow \mathcal{N}\left(m_Y + C_{YX}C_{XX}^{-1}(X - m_X),\, C_{YY} - C_{YX}C_{XX}^{-1}C_{XY}\right)$$

- introduce the Gaussian r. vector $\mathbf{V}$ of dimension $n$ also with distribution

$$f_{\mathbf{V}}(V) \leftrightarrow \mathcal{N}(m_V,\, C_{VV}) = \mathcal{N}(\underbrace{m_Y - C_{YX}C_{XX}^{-1}m_X}_{=m_V},\, \underbrace{C_{YY} - C_{YX}C_{XX}^{-1}C_{XY}}_{=C_{VV}})$$

  and $\mathbf{V}$ independent of $\mathbf{X}$,

- then $\mathbf{X}$ and $\mathbf{Y}$ generated as follows

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} I & 0 \\ C_{YX}C_{XX}^{-1} & I \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ \mathbf{V} \end{bmatrix} \Rightarrow \begin{bmatrix} I & 0 \\ C_{YX}C_{XX}^{-1} & I \end{bmatrix} \begin{bmatrix} m_X \\ m_V \end{bmatrix} = \begin{bmatrix} m_X \\ m_Y \end{bmatrix}$$

$$\begin{bmatrix} I & 0 \\ C_{YX}C_{XX}^{-1} & I \end{bmatrix} \begin{bmatrix} C_{XX} & \overbrace{C_{XV}}^{=0} \\ \underbrace{C_{VX}}_{=0} & \underbrace{C_{VV}}_{=C_{YY}-C_{YX}C_{XX}^{-1}C_{XY}} \end{bmatrix} \begin{bmatrix} I & C_{XX}^{-1}C_{XY} \\ 0 & I \end{bmatrix} = \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix}$$

  (block Lower-Diagonal-Upper (LDU) triangular factorization of $C$)
  are jointly Gaussian and have the correct mean and variance, hence have the correct pdf $f_{\mathbf{X},\mathbf{Y}}(X, Y)$.

# Gaussian r.v.'s and Linear Models (2)

- we can write explicitly

$$\mathbf{Y} \;=\; C_{YX}C_{XX}^{-1}\mathbf{X} + \mathbf{V}\ .$$

  This means that for two sets of correlated Gaussian random variables, one ($\mathbf{Y}$) can be thought of as being generated from the other ($\mathbf{X}$) through a linear model ($C_{YX}C_{XX}^{-1}$) and being corrupted by independent Gaussian "measurement" noise ($\mathbf{V}$).

- Going back to the scalar example ($m = n = 1$), the variance of the linear model part ($C_{YX}C_{XX}^{-1}\mathbf{X}$) is $\rho^2\sigma_y^2$ while the variance of the measurement noise ($\mathbf{V}$) is $(1-\rho^2)\sigma_y^2$. The two are complementary parts of the total variance $\sigma_y^2$ of $\mathbf{Y}$.
  We could define a signal to noise ratio (SNR) as the ratio of the two parts which is $\frac{\rho^2}{1-\rho^2}$. The SNR increases from $0$ to $\infty$ as $|\rho|$ increases from $0$ to $1$, as the correlation between $\mathbf{X}$ and $\mathbf{Y}$ increases.

# The Parameter Estimation Problem

description of stochastic processes known up to some parameters

- tone detector: a push button telephone emits sinusoids of which the frequencies are distinct for different buttons. The unknown parameter of interest is the sinusoid frequency. This frequency determination problem may in fact be better approached as a detection problem as we shall see shortly
- the carrier phase and timing instants in linear digital modulation schemes
- the impulse response of the transmission channel. We may take a prameterized model (e.g. FIR model) for this impulse response and then the channel identification problem becomes a problem of estimating the parameters in its model
- similar impulse response identification problems occur in the problems of the cancellation of electrical echos caused by hybrids connecting 2-wire and 4-wire sections of telephone line, and acoustic echos in teleconferencing systems

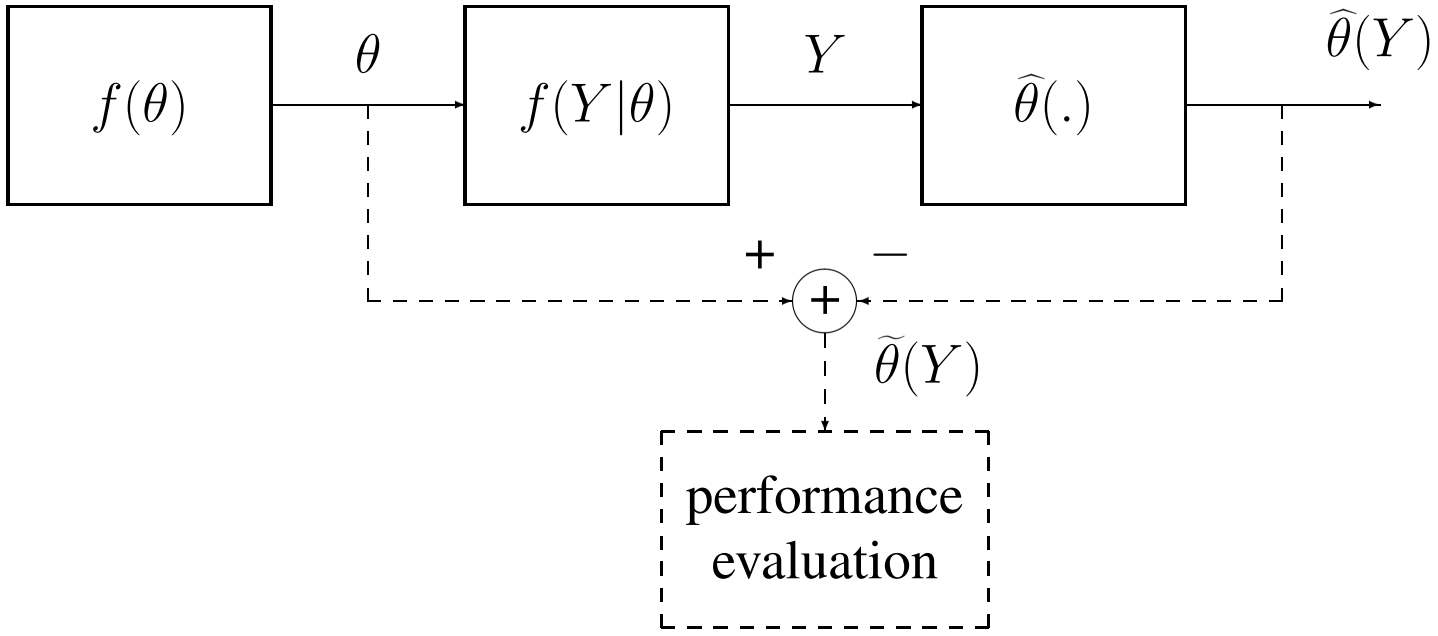
# The Parameter Estimation Problem (2)

description of stochastic processes known up to some parameters

- the average rate in a Poisson process occurring in queuing theory and network performance analysis
- the pixel value in the noisy capture of an image by a satellite
- the position and orientation of an object in an image
- security in mobile communications: the (parameterized) distribution of features characterizing the behavior of users. The deviation from typical habits may be a reason for alarm.

# Bayesian Framework

- Bayesian approach: consider parameters $\boldsymbol{\theta}$ to be random variables
- the parameters have some *a priori* distribution $f_{\boldsymbol{\theta}}(\theta)$. This distribution is called a priori because it summarizes the knowledge we have about $\theta$ before making any measurement.
- Next we make a *measurement* $Y$. The measurement is not a deterministic function of the unknown parameters. The random aspect of this relation is captured by the conditional distribution $f_{\mathbf{Y}|\boldsymbol{\theta}}(Y|\theta)$.



# Ex 1.2: Additive indep. measurement noise

- $f_{\mathbf{y}|\boldsymbol{\theta}}(y|\theta) = \dfrac{f_{\mathbf{y},\boldsymbol{\theta}}(y,\theta)}{f_{\boldsymbol{\theta}}(\theta)} = \dfrac{f_{\mathbf{v},\boldsymbol{\theta}}(y-\theta,\theta)}{f_{\boldsymbol{\theta}}(\theta)} = \dfrac{f_{\mathbf{v}}(y-\theta)\, f_{\boldsymbol{\theta}}(\theta)}{f_{\boldsymbol{\theta}}(\theta)} = f_{\mathbf{v}}(y-\theta)$

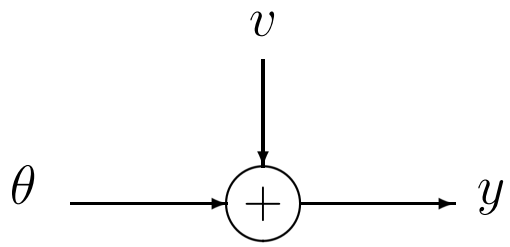
- first identity: Bayes' rule

- second identity: transformation of variables

$$\begin{bmatrix} v \\ \theta \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y \\ \theta \end{bmatrix}$$

  linear transformation $\Rightarrow$ Jacobian = determinant of the transformation matrix, which equals 1.

- third identity: independence of $v$ and $\theta$



# Estimators

- *estimator* : a function $\widehat{\theta}(.)$ to be applied to the measurement $\mathbf{Y}$
- *estimate* : $\widehat{\theta}(Y)$ is the estimator $\widehat{\theta}(\mathbf{Y})$ evaluated at $\mathbf{Y} = Y$.
- we consider here a *point estimator*, i.e. it delivers one value $\widehat{\theta} = \widehat{\theta}(Y) = t(y_1, \ldots, y_n)$ that we hope to be close to $\theta$.
- The function $\widehat{\theta}(\mathbf{Y}) = t(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ of the measurement data is called a *statistic*.
- Another type of estimator would be an *interval estimator*: two statistics $t_1(y_1, \ldots, y_n)$ and $t_2(y_1, \ldots, y_n)$ are used to define an interval such that

$$\Pr\{\theta \in (t_1(y_1, \ldots, y_n),\, t_2(y_1, \ldots, y_n))\} \;=\; \gamma$$

  can be determined. $\gamma$ is called the *confidence coefficient*.

# Ex: Estimators for the mean of Gaussian r.v.'s

- Let the $\mathbf{y}_i \sim \mathcal{N}(\theta, 5)$ be i.i.d. (independent and identically distributed) normal random variables with unknown mean $\theta$ and variance equal to 5.

- The arithmetic mean is a *point estimator* of $\theta$ :

$$\widehat{\theta} \;=\; \overline{y} \;=\; \frac{1}{n}\sum_{i=1}^{n} y_i \;\;\sim\;\; \mathcal{N}(\theta, \frac{5}{n}) \;.$$

- Now, since $\overline{y} \sim \mathcal{N}(\theta, \frac{5}{n})$ or $\overline{y} = \theta + v$ with $v \sim \mathcal{N}(0, \frac{5}{n})$, we can also state

$$\Pr\left\{\theta \in \left(\overline{y} - 2\sqrt{\frac{5}{n}}, \overline{y} + 2\sqrt{\frac{5}{n}}\right)\right\} \;=\; \Pr\left\{\overline{y} \in \left(\theta - 2\sqrt{\frac{5}{n}}, \theta + 2\sqrt{\frac{5}{n}}\right)\right\} \;=\; 0.95$$

since

$$\overline{y} - 2\sqrt{\frac{5}{n}} < \theta < \overline{y} + 2\sqrt{\frac{5}{n}} \quad \Leftrightarrow \quad \theta - 2\sqrt{\frac{5}{n}} < \overline{y} < \theta + 2\sqrt{\frac{5}{n}}$$

so that $\left(\overline{y} - 2\sqrt{\frac{5}{n}}, \overline{y} + 2\sqrt{\frac{5}{n}}\right)$ is an *interval estimator* for $\theta$ with confidence coefficient $0.95$.

# Estimation Considerations

- The design of interval estimators is the result of the following compromise. On the one hand, one wants a small interval so that $\theta$ can be known fairly accurately. On the other hand, a small interval necessarily leads to a small confidence coefficient since in general we can only state with a low probability level that $\theta$ is contained in a small interval. But we would like the confidence coefficient to be high so that we would be able to say with high probability where $\theta$ is located. This again would lead to a large interval etc.
- We will restrict the further discussion to point estimators. In general, the estimators we shall consider will asymptotically (for many measurements, large $n$) have a Gaussian distribution so that we can easily construct an interval estimator from the point estimator as in the example above.
- As a final remark, we shall assume that the parameters (and their estimators) can take on a continuous range of values so that their distribution is described by a probability density function. In case the parameters can take on only a *discrete set of values* and the task is to decide which one of the values the parameters actually take, then the estimation problem is called a *detection* or *decision* problem (see DIGICOM course).

# Bayes Estimation

- nonnegative cost function $\mathcal{C}(\theta, \widehat{\theta})$

- since $\theta, \widehat{\theta}(Y)$ are random variables, we can never hope to make this cost function small for every outcome of $\theta$ and $Y$. All we can hope for is to minimize the expected value of the cost function, also called the risk

$$\mathcal{R}(\widehat{\theta}(.)) \;=\; E\,\mathcal{C}(\theta, \widehat{\theta}(Y)) \;=\; E_{\theta,Y}\mathcal{C}(\theta, \widehat{\theta}(Y))\;.$$

  So the estimator is that function $\theta(.)$ that minimizes the Bayes risk

$$\widehat{\theta}(.) \;=\; \arg\min_{\widehat{\theta}(.)} \mathcal{R}(\widehat{\theta}(.))$$

  It is common practice to limit the choice of the cost function to a nonnegative cost function of the parameter estimation error

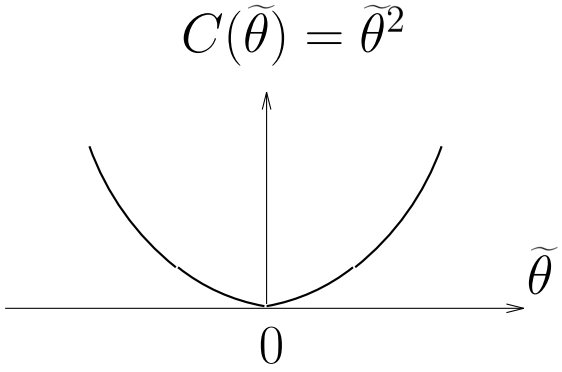
$$\widetilde{\theta} \;=\; \theta - \widehat{\theta}(Y)$$

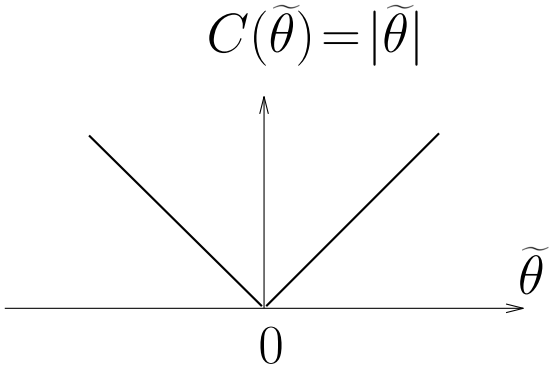
  only, in which case the estimation problem becomes

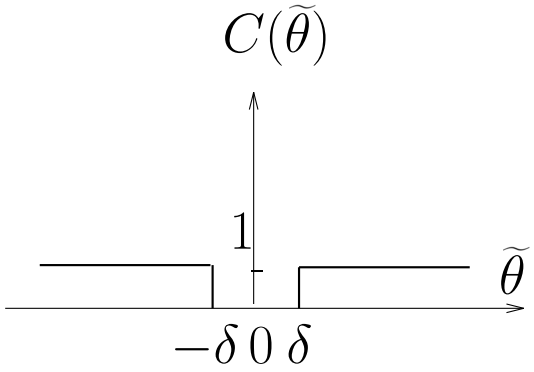
$$\widehat{\theta}(.) \;=\; \arg\min_{\widehat{\theta}(.)} E\,\mathcal{C}(\widetilde{\theta})$$

# Bayes Cost Functions

- Three popular Bayes cost functions: squared parameter deviation, absolute parameter deviation, and the uniform cost function. All three choices assign no cost when there is no error.



# Bayes Risk Minimization

- manipulate the Bayes optimization criterion

$$\min_{\widehat{\theta}(.)} \mathcal{R}(\widehat{\theta}(.)) \;=\; \min_{\widehat{\theta}(.)} E\,\mathcal{C}(\widetilde{\theta}) \;=\; \min_{\widehat{\theta}(.)} \int\int f(Y,\theta)\,\mathcal{C}(\theta - \widehat{\theta}(Y))\,dY\,d\theta$$

$$= \min_{\widehat{\theta}(.)} \int \underbrace{f(Y)}_{\geq 0}\, dY \underbrace{\int f(\theta|Y)\,\mathcal{C}(\theta - \widehat{\theta}(Y))\,d\theta}_{\mathcal{R}(\widehat{\theta}(.)|Y)} = \min_{\widehat{\theta}(.)} E_Y \mathcal{R}(\widehat{\theta}(.)|Y) \;=\; E_Y \min_{\widehat{\theta}(Y)} \mathcal{R}(\widehat{\theta}(Y)|Y)$$

  $\mathcal{R}(\widehat{\theta}(.)|Y) = \mathcal{R}(\widehat{\theta}(Y)|Y)$ is a function of $Y$.

- Since the contributions of the conditional risk $\mathcal{R}(\widehat{\theta}(Y)|Y)$ for every $Y$ are combined via the nonnegative weighting factor $f(Y)$ to obtain the global risk $\mathcal{R}(\widehat{\theta}(.))$, we can minimize $\mathcal{R}(\widehat{\theta}(.))$ by minimizing $\mathcal{R}(\widehat{\theta}(Y)|Y)$ for every particular $Y$.

- While the minimization of the global risk $\mathcal{R}(\widehat{\theta}(.))$ is with respect to (*w.r.t.*) the estimator function $\widehat{\theta}(.)$, the minimization of the conditional risk $\mathcal{R}(\widehat{\theta}(Y)|Y)$ is w.r.t. the particular estimate $\widehat{\theta}(Y)$, which is simply a number (the estimator function $\widehat{\theta}(.)$ evaluated at $\mathbf{Y} = Y$).

# Bayes Risk Minimization (2)

- This minimization of the conditional risk $\mathcal{R}(\widehat{\theta}(Y)|Y)$ requires the *a posteriori* distribution $f(\theta|Y)$ of $\theta$ given the measurement $Y$.

- The a posteriori distribution describes the randomness left in $\theta$ after we have made a measurement of (the related quantity) $Y$.

- The a posteriori distribution $f(\theta|Y)$ can be determined from the conditional distribution $f(Y|\theta)$ and the a priori distribution $f(\theta)$, which are normally given in the problem description, as follows (using Bayes' rule):

$$f(\theta|Y) = \frac{f(Y,\theta)}{f(Y)} = \frac{f(Y|\theta)f(\theta)}{\int_\Theta f(Y,\theta)d\theta} = \frac{f(Y|\theta)f(\theta)}{\int_\Theta f(Y|\theta)f(\theta)d\theta}$$

  where $\Theta$ is the region of support for $\theta$.

# The MMSE Criterion

- using the quadratic cost function $\mathcal{C}_{MMSE}(\widetilde{\theta}) = |\widetilde{\theta}|^2$, minimizing the conditional Bayes risk yields

$$\min_{\widehat{\theta}(Y)} \mathcal{R}_{MMSE}(\widehat{\theta}(Y)|Y) \;=\; \min_{\widehat{\theta}(Y)} \int_{-\infty}^{\infty} f(\theta|Y)\,(\theta - \widehat{\theta}(Y))^2 d\theta$$

- to take the derivative w.r.t. $\widehat{\theta}$, recall Leibnitz's rule:

$$\frac{\partial}{\partial u} \int_{g_1(u)}^{g_2(u)} h(u,v)\,dv \;=\; \int_{g_1(u)}^{g_2(u)} \frac{\partial\, h(u,v)}{\partial u}\,dv + \frac{dg_2(u)}{du} h(u, g_2(u)) - \frac{dg_1(u)}{du} h(u, g_1(u))$$

- Using Leibnitz's rule, we obtain

$$\frac{\partial}{\partial \widehat{\theta}} \mathcal{R}_{MMSE}(\widehat{\theta}|Y) \;=\; 2 \int_{-\infty}^{\infty} f(\theta|Y)\,(\widehat{\theta} - \theta) d\theta \quad = 0$$

  where we put the derivative equal to zero in order to find an extremum.

# The MMSE Criterion (2)

- We can rewrite as

$$\widehat{\theta}(Y) \underbrace{\int_{-\infty}^{\infty} f(\theta|Y)\, d\theta}_{=1} = \int_{-\infty}^{\infty} \theta\, f(\theta|Y)\, d\theta \;\Rightarrow\; \widehat{\theta}_{MMSE}(Y) = E(\theta|Y)$$

  which is the *mean* of the a posteriori distribution of $\theta$ given $Y$.

- To know whether this extremum is a local minimum, we can apply Leibnitz's rule a second time to obtain

$$\frac{\partial^2}{\partial \widehat{\theta}^2} \mathcal{R}_{MMSE}(\widehat{\theta}|Y) = 2 \int_{-\infty}^{\infty} f(\theta|Y)\, d\theta = 2 > 0$$

  Hence the extremum at $\widehat{\theta}(Y) = E(\theta|Y)$ is a local minimum and it is furthermore the global minimum since it is the unique local minimum. Note that $E(\theta|Y)$ is indeed a function of $Y$.

# The absolute value cost function

- For the absolute value cost function $\mathcal{C}_{ABS}(\widetilde{\theta}) = |\widetilde{\theta}|$, the minimization problem of the conditional Bayes risk becomes

$$\begin{aligned}\min_{\widehat{\theta}(Y)} \mathcal{R}_{ABS}(\widehat{\theta}(Y)|Y) &= \min_{\widehat{\theta}(Y)} \int_{-\infty}^{\infty} f(\theta|Y)\,|\theta - \widehat{\theta}(Y)|\,d\theta \\ &= \min_{\widehat{\theta}(Y)} \left[\int_{-\infty}^{\widehat{\theta}} f(\theta|Y)\,(\widehat{\theta}-\theta)d\theta \;+\; \int_{\widehat{\theta}}^{\infty} f(\theta|Y)\,(\theta-\widehat{\theta})d\theta\right] .\end{aligned}$$

- We shall again use Leibnitz's rule to take the derivative. For the first integral, we let $h(\widehat{\theta},\theta) = f(\theta|Y)\,(\widehat{\theta}-\theta)$ and we get

$$h(\widehat{\theta}, g_2(\widehat{\theta})) = h(\widehat{\theta},\widehat{\theta}) = f(\widehat{\theta}|Y)\,(\widehat{\theta}-\widehat{\theta}) = 0$$

  and $\frac{d\,g_1(\widehat{\theta})}{d\widehat{\theta}} = 0$ since the lower limit does not depend on $\widehat{\theta}$. The corresponding terms for the second integral are similarly equal to zero.

- So the only terms in the derivative remaining are those obtained by differentiating the integrands directly:

$$\frac{\partial}{\partial\widehat{\theta}}\mathcal{R}_{ABS}(\widehat{\theta}|Y) \;=\; \int_{-\infty}^{\widehat{\theta}} f(\theta|Y)\,d\theta \;-\; \int_{\widehat{\theta}}^{\infty} f(\theta|Y)\,d\theta \;=\; 0$$

  where again we put the derivative equal to zero in order to find an extremum.

# The absolute value cost function (2)

- So the condition for an extremum becomes

$$F(\widehat{\theta}|Y) \;=\; \int_{-\infty}^{\hat{\theta}} f(\theta|Y)\,d\theta \;=\; \int_{\hat{\theta}}^{\infty} f(\theta|Y)\,d\theta \;=\; 1 - F(\widehat{\theta}|Y)$$

  where $F(\theta|Y)$ is the a posteriori cumulative distribution function (cdf) of $\theta$ given $Y$.

- We can rewrite this also as

$$F(\widehat{\theta}_{ABS}(Y)|Y) \;=\; \frac{1}{2}$$

  which means that $\widehat{\theta}_{ABS}(Y)$ is the *median* of the a posteriori distribution of $\theta$ given $Y$.

# The absolute value cost function (3)

- Again, to know whether this extremum is a local minimum, we can apply Leibnitz's rule a second time to obtain

$$\frac{\partial^2}{\partial \widehat{\theta}^2} \mathcal{R}_{ABS}(\widehat{\theta}|Y)\Big|_{\widehat{\theta}=\widehat{\theta}_{ABS}} = 2\, f(\widehat{\theta}_{ABS}|Y) \geq 0 \,.$$

If $f(\widehat{\theta}_{ABS}|Y) = 0$ then, since $f(\theta|Y) \geq 0$, the first nonzero derivative w.r.t. $\theta$ of $f(\theta|Y)$ will be of even order and positive (this reasoning can be extended to the case where $f(\theta|Y)$ would not be sufficiently differentiable). Hence the extremum at $\widehat{\theta}_{ABS}(Y)$ is a local minimum and it is furthermore the global minimum since it is the unique local minimum. Note again that $\widehat{\theta}_{ABS}(Y)$ is indeed a function of $Y$.

# The uniform cost function

- For the uniform cost function we get

$$\begin{aligned}\min_{\hat{\theta}(Y)} \mathcal{R}_{UNIF}(\widehat{\theta}(Y)|Y) &= \min_{\hat{\theta}(Y)} \left[\left(\int_{-\infty}^{\hat{\theta}-\delta} + \int_{\hat{\theta}+\delta}^{\infty}\right) f(\theta|Y)\, d\theta\right] \\ &= \min_{\hat{\theta}(Y)} \Big[\underbrace{\int_{-\infty}^{\infty} f(\theta|Y)\, d\theta}_{=1} - \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} f(\theta|Y)\, d\theta\Big] \ .\end{aligned}$$

- Since we take $\delta$ to be arbitrarily small, the optimization problem becomes

$$\max_{\hat{\theta}(Y)} \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} f(\theta|Y)\, d\theta \approx \max_{\hat{\theta}(Y)} 2\delta\, f(\widehat{\theta}|Y) = 2\delta \max_{\hat{\theta}(Y)} f(\widehat{\theta}|Y)$$

  Hence, for $\delta$ arbitrarily small, $\mathcal{R}_{UNIF}(\widehat{\theta}(Y)|Y)$ is minimized by choosing $\widehat{\theta}$ to be the *mode* (location of the maximum) of the a posteriori distribution of $\theta$ given $Y$.

- For this reason, the estimator corresponding to a uniform cost function is normally called the *Maximum A Posteriori* (MAP) estimator:

$$\widehat{\theta}_{MAP}(Y) = \arg\max_{\theta} f(\theta|Y)$$

  which is again a function of $Y$.

# MAP Estimator: Remarks

- We may note that the same estimator is obtained by choosing the cost function $\mathcal{C}(\widetilde{\theta}) = 1 - \delta(\widetilde{\theta})$. While this cost function is cleaner in that it involves no limiting operation with $\delta$, it might be considered non-intuitive since it is not nonnegative. However, one should keep in mind that adding or subtracting a constant to a cost function does not influence its minimizing argument.

- Instead of maximizing $f(\theta|Y)$, any strictly increasing function of it may be maximized. Since $f(\theta|Y)$ is often given in factored form and often contains exponential distributions, a convenient choice is to maximize

$$\ln f(\theta|Y) \;=\; \ln f(Y|\theta) + \ln f(\theta) - \ln f(Y)\ .$$

- Often $f(\theta|Y)$ satisfies certain regularity conditions so that $\widehat{\theta}_{MAP}$ is a solution of

$$\frac{\partial}{\partial\theta} \ln f(\theta|Y) \;=\; 0 \;=\; \frac{\partial}{\partial\theta} \ln f(Y|\theta) + \frac{\partial}{\partial\theta} \ln f(\theta) \quad .$$

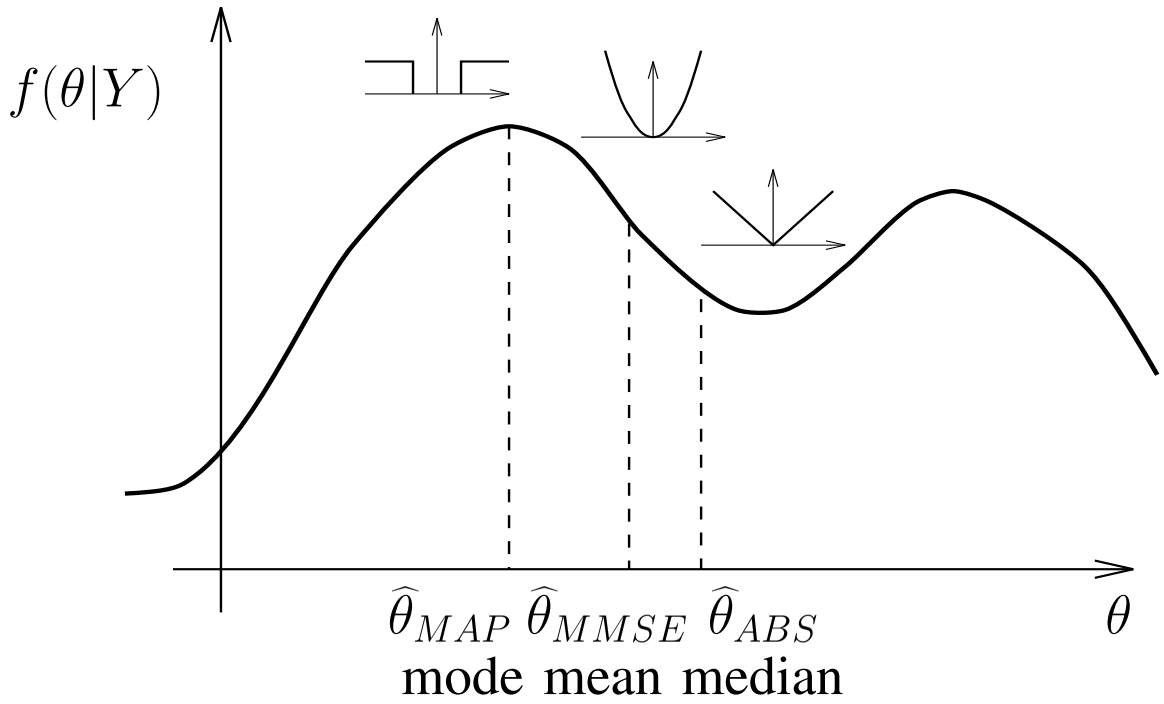
  Note that $\frac{\partial \ln f(\theta|Y)}{\partial\theta} = \frac{1}{f(\theta|Y)} \frac{\partial f(\theta|Y)}{\partial\theta}$. The conditions for a maximum (rather than another form of extremum) need to be verified of course.

# MAP Estimator: Remarks (2)

- The previous equation indicates that, starting from the description of the problem which includes $f(Y|\theta)$ and $f(\theta)$, the calculation of $\widehat{\theta}_{MAP}$ should be relatively straightforward.
- The MAP estimator is given by the *global* maximum of $f(\theta|Y)$. If there are several local maxima, all of them need to be examined and compared to find the global maximum.
- Even if $f(\theta|Y)$ satisfies regularity conditions, the maximum may occur at the boundary of the parameter space $\Theta$ (which may not necessarily be $(-\infty, \infty)$). In that case, the maximum is not a local extremum.

# 3 Bayes estimators

- in general, the three Bayes estimators may differ



# Ex: Gaussian mean in Gaussian noise

- estimating an unknown dc level in additive Gaussian noise with zero mean: e.g. satellite transmission of a digital image (accumulate many noisy images)

- for any given pixel, the measurement problem can be modeled as

$$y_i \;=\; \theta + v_i \;, \quad i = 1, \ldots, n$$

  where $\theta$ is the true grey value of the pixel, $y_i$ are the consecutive noisy measured grey values, the $v_i \sim \mathcal{N}(0, \sigma_v^2)$ are i.i.d. (Central Limit Theorem).

- Even though the images vary very slowly, the image of one group of $n = 100$ shots will not differ very much from the image of the previous 100 shots. Therefore, we can consider the image estimate (with value $m_\theta$ at the pixel considered) from the previous group to be prior information for the problem of estimating the image of the current group (with value $\theta$ at the same pixel).

# Ex: Gaussian mean in Gaussian noise (2)

- This prior information is not perfect due to the estimation variance associated with the processing of the previous group and also the variation from one image (group) to the next. Therefore we model the prior information as $\theta \sim \mathcal{N}(m_\theta, \sigma_\theta^2)$ where $\sigma_\theta^2$ reflects the joint effect of estimation variability and variablity due to change in time. Also, $\theta$ is assumed to be independent of the $v_i$.

- We can write the measurements in vector form

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \theta \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \theta \mathbf{1} + V$$

With $X = \theta$, we can consider the following invertible transformation

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \mathbf{1} & I_n \end{bmatrix} \begin{bmatrix} \theta \\ V \end{bmatrix}.$$

Now, since $\theta$ and $V$ are independent and both Gaussian, they are jointly Gaussian. A linear transformation of jointly Gaussian random variables leads again to Gaussian random variables. Hence, $X = \theta$ and $Y$ are jointly Gaussian.

# Ex: Gaussian mean in Gaussian noise (3)

- to determine a Bayes estimator, we need the a posteriori pdf $f(\theta|Y) = f(X|Y)$. Since $X$ and $Y$ are jointly Gaussian, the Gauss-Markov theorem tells us that $f(X|Y)$ is also Gaussian and it also tells us how to compute it from the first and second-order moments of $X$ and $Y$. So we shall compute these moments.

- First-order moments:
$$\begin{aligned} m_X &= E\,X = m_\theta \\ m_Y &= E\,Y = E\,(\theta \mathbf{1} + V) = (E\theta)\,\mathbf{1} + E\,V = m_\theta \mathbf{1} \end{aligned}$$

- $C_{VV} = EVV^T = \sigma_v^2\, I_n$ since $(EVV^T)_{i,j} = Ev_iv_j = \sigma_v^2\,\delta_{i,j} = (\sigma_v^2\, I_n)_{i,j}$

- The joint covariance matrix is

$$C = E\begin{bmatrix} X-m_X \\ Y-m_Y \end{bmatrix}\begin{bmatrix} X-m_X \\ Y-m_Y \end{bmatrix}^T = E\begin{bmatrix} \theta-m_\theta \\ (\theta-m_\theta)\,\mathbf{1}+V \end{bmatrix}\begin{bmatrix} \theta-m_\theta \\ (\theta-m_\theta)\,\mathbf{1}+V \end{bmatrix}^T$$

$$= \begin{bmatrix} \sigma_\theta^2 & \sigma_\theta^2 \mathbf{1}^T \\ \sigma_\theta^2 \mathbf{1} & \sigma_\theta^2 \mathbf{1}\mathbf{1}^T + \sigma_v^2 I \end{bmatrix} = \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix}$$

where we used $E\{(\theta-m_\theta)\,V\} = E\{(\theta-m_\theta)\}\;E\{V\} = 0\,.\,0 = 0$ ,

$E\,[(\theta-m_\theta)\,\mathbf{1}+V]\,[(\theta-m_\theta)\,\mathbf{1}+V]^T = E(\theta-m_\theta)^2\,\mathbf{1}\mathbf{1}^T + E\,VV^T + \mathbf{1}E\,V^T(\theta-m_\theta)$
$+E\,V(\theta-m_\theta)\,\mathbf{1}^T = \sigma_\theta^2 \mathbf{1}\mathbf{1}^T + \sigma_v^2 I + 0 + 0$

# Ex: Gaussian mean in Gaussian noise (4)

- A key quantity that appears in the expression for $f(X|Y)$ is $C_{XY}C_{YY}^{-1}$. We get

$$\begin{aligned} C_{XY}C_{YY}^{-1} &= \sigma_\theta^2 \mathbf{1}^T \left[\sigma_v^2 I + \sigma_\theta^2 \mathbf{1}\,\mathbf{1}^T\right]^{-1} \\ &= \frac{\sigma_\theta^2}{\sigma_v^2}\mathbf{1}^T \left[I + \frac{\sigma_\theta^2}{\sigma_v^2}\mathbf{1}\,\mathbf{1}^T\right]^{-1} . \end{aligned}$$

  In order to compute the inverse of the matrix in brackets, we shall use the following identity.

- **Lemma 0.1 (Matrix Inversion Lemma)** *If $A$ and $C$ are respectively $n \times n$ and $m \times m$ invertible matrices and $B$ and $D$ are respectively $n \times m$ and $m \times n$ matrices, then*

$$[A + BCD]^{-1} = A^{-1} - A^{-1}B\left[DA^{-1}B + C^{-1}\right]^{-1} DA^{-1}$$

  *if the inverses exists.*

- This lemma is very useful when we have to compute $[A + BCD]^{-1}$, we know $A^{-1}$, and $m$ is (much) smaller than $n$. We shall apply this lemma with $A = I$, $B = \mathbf{1}$, $C = \frac{\sigma_\theta^2}{\sigma_v^2}$ and $D = \mathbf{1}^T$. Hence $m = 1$.

# Ex: Gaussian mean in Gaussian noise (5)

- We obtain

$$C_{XY}C_{YY}^{-1} = \frac{\sigma_\theta^2}{\sigma_v^2}\mathbf{1}^T[I - \mathbf{1}(\underbrace{\mathbf{1}^T\mathbf{1} + \frac{\sigma_v^2}{\sigma_\theta^2}}_{=n+\frac{\sigma_v^2}{\sigma_\theta^2}})^{-1}\mathbf{1}^T] = \frac{\sigma_\theta^2}{\sigma_v^2}\left(1 - \frac{n}{n+\frac{\sigma_v^2}{\sigma_\theta^2}}\right)\mathbf{1}^T = \frac{1}{n+\frac{\sigma_v^2}{\sigma_\theta^2}}\mathbf{1}^T$$

- From the Gauss-Markov theorem, we can now compute the a posteriori mean

$$E[\theta|Y] = m_\theta + C_{XY}C_{YY}^{-1}(Y - m_Y) = m_\theta + \frac{1}{n+\frac{\sigma_v^2}{\sigma_\theta^2}}\mathbf{1}^T(Y - m_\theta\mathbf{1})$$

$$= \left(\frac{1}{\sigma_\theta^2}m_\theta + \frac{n}{\sigma_v^2}\overline{y}\right) / \left(\frac{1}{\sigma_\theta^2} + \frac{n}{\sigma_v^2}\right) = \frac{\frac{1}{\sigma_\theta^2}}{\left(\frac{1}{\sigma_\theta^2} + \frac{n}{\sigma_v^2}\right)}m_\theta + \frac{\frac{n}{\sigma_v^2}}{\left(\frac{1}{\sigma_\theta^2} + \frac{n}{\sigma_v^2}\right)}\overline{y}$$

  where we introduced the sample mean $\overline{y} = \frac{1}{n}\mathbf{1}^TY = \frac{1}{n}\sum_{i=1}^{n} y_i$.

- Note that $E[\theta|Y]$ is of the form $\alpha\, m_\theta + (1-\alpha)\overline{y}$, which is a convex combination between the a priori mean and the sample mean obtained from the data. The respective weighting factors are proportional to the inverse of the variance associated with each of the two components.

# Ex: Gaussian mean in Gaussian noise (6)

- We shall call the inverse of the variance the amount of information available:
  - $\frac{1}{\sigma_\theta^2}$ = information in the prior distribution $f(\theta)$,
  - $\frac{n}{\sigma_v^2}$ = information in the $n$ independent measurements (conditional distribution $f(y_i|\theta) = f_{\mathbf{v}_i}(y_i - \theta)$).

- We can consider two extreme cases:
  - $\sigma_\theta^2 \ll \frac{\sigma_v^2}{n}$ : $E[\theta|Y] \approx m_\theta$. In this case, the measurements are so noisy that the reliable a priori information dominates.
  - $\sigma_\theta^2 \gg \frac{\sigma_v^2}{n}$ : $E[\theta|Y] \approx \overline{y}$. In this case, the accurate measurements dominate the unreliable a priori information. Note that this second case will eventually occur when $n \to \infty$.

# Ex: Gaussian mean in Gaussian noise (7)

- From the Gauss-Markov theorem, we can also determine the a posteriori variance, which we shall denote by $\sigma^2_{\hat{\theta}}$:

$$\sigma^2_{\hat{\theta}} \;=\; Var\left[\theta|Y\right] \;=\; C_{XX}-C_{XY}C_{YY}^{-1}C_{YX} \;=\; \sigma^2_\theta-\frac{1}{n+\frac{\sigma^2_v}{\sigma^2_\theta}}\mathbf{1}^T\mathbf{1}\sigma^2_\theta \;=\; \cdots \;=\; \left(\frac{1}{\sigma^2_\theta}+\frac{n}{\sigma^2_v}\right)^{-1}$$

  or hence

$$\frac{1}{\sigma^2_{\hat{\theta}}} \;=\; \frac{1}{\sigma^2_\theta}+\frac{n}{\sigma^2_v}$$

  which means that the a priori information and the information in the $n$ independent measurements add up to form the total information in the posterior distribution.

- This allows us to rewrite the a posteriori mean as

$$\frac{1}{\sigma^2_{\hat{\theta}}}\,E\left[\theta|Y\right] \;=\; \frac{1}{\sigma^2_\theta}m_\theta+\frac{n}{\sigma^2_v}\overline{y}\;.$$

# Ex: Gaussian mean in Gaussian noise (8)

- Using the Gauss-Markov theorem, we are finally ready to write the a posteriori distribution as

$$f(\theta|Y) \leftrightarrow \mathcal{N}\left(E\left[\theta|Y\right],\ \sigma^2_{\hat{\theta}}\right)$$

  where $E\left[\theta|Y\right]$ and $\sigma^2_{\hat{\theta}}$ are given above.

- Since the posterior distribution is Gaussian, its mean, mode and median coincide. Hence we get

$$\widehat{\theta}_{MMSE} = \widehat{\theta}_{MAP} = \widehat{\theta}_{ABS} = E\left[\theta|Y\right] .$$

- Note that the posterior distribution $f(\theta, Y)$ depends on $Y$ only through $\overline{y}$. In this case we say that the statistic $\overline{y}$ (a function of the data $Y$) is a *sufficient statistic*, meaning that, for the purpose of estimating $\theta$, the only thing we need to know about $y_1, \ldots, y_n$ is $\overline{y}$.

# Equivalences of Bayes Estimators

One characteristic of Bayes estimation is that a Bayes estimator depends on the cost function it is associated with. In the following, we shall examine three cases in which the Bayes estimator coincides with $\widehat{\theta}_{MMSE}$ meaning that the estimator minimizing the risk based on some non-quadratic cost function coincides with $\widehat{\theta}_{MMSE}$, the estimator minimizing a quadratic cost function.

1. Consider a cost function $C(\widetilde{\theta})$ with the following properties:
   - symmetric: $C(\widetilde{\theta}) = C(-\widetilde{\theta})$
   - convex : $C(\alpha\widetilde{\theta}_1+(1-\alpha)\widetilde{\theta}_2) \leq \alpha\, C(\widetilde{\theta}_1)+(1-\alpha)\, C(\widetilde{\theta}_2)\,,\ \alpha \in [0,1]$

   and let $f(\theta|Y)$ be symmetric about $E[\theta|Y]$. Then $\widehat{\theta}_C = \widehat{\theta}_{MMSE}$.

# Equivalences of Bayes Estimators (2)

2. Consider a cost function $C(\widetilde{\theta})$ with the following properties:
   - symmetric: $C(\widetilde{\theta}) = C'(|\widetilde{\theta}|)$
   - $C'(|\widetilde{\theta}|)$ is a non-decreasing function of $|\widetilde{\theta}|$ (this condition is weaker than $C(\widetilde{\theta})$ being convex)

   and let $f(\theta|Y)$ have the following properties
   - symmetric about $E[\theta|Y]$
   - unimodal (only one local maximum)

   and finally let $C(.)$ and $f(.|Y)$ be such that $\lim_{\theta\to\infty} C(\theta)\, f(\theta|Y) = 0\,,\ \forall Y$, then again $\widehat{\theta}_C = \widehat{\theta}_{MMSE}$.
3. If $f(\theta|Y)$ is Gaussian, then $\widehat{\theta}_{MMSE} = \widehat{\theta}_{MAP} = \widehat{\theta}_{ABS}$.

These equivalences show the relative importance of $\widehat{\theta}_{MMSE}$. In general however, $\widehat{\theta}_{MAP}$ is the easiest to compute, as is illustrated in the following example.

# Ex: Poisson Process

- Consider a communication network node where messages are passing. Let the observation $\mathbf{N}$ be the number of messages that pass during a certain observation period of duration $\tau$. $\mathbf{N}$ has a Poisson distribution

$$\Pr\left[\mathbf{N} = N|\theta\right] \;=\; (\theta\tau)^N \,\frac{e^{-\theta\tau}}{N!}\;, \quad N = 0, 1, 2, \ldots$$

  where the parameter $\theta$ represents the average number of messages passing per second by the node we are considering.

- This average frequency has itself an a priori distribution (over the different nodes in the network) which we take to be exponential with average value $m_\theta = 1/\lambda$:

$$f(\theta) \;=\; \begin{cases} \lambda\, e^{-\lambda\,\theta} & , \quad \theta > 0 \\ 0 & , \quad \theta \leq 0\;. \end{cases}$$

# Ex: Poisson Process (2)

- In order to find $\widehat{\theta}_{MMSE}$, we shall compute the a posteriori distribution of $\theta$ given the measurement $N$. Using Bayes' rule we get

$$f(\theta|N) = \frac{\Pr\left[\mathbf{N} = N|\theta\right]\ f(\theta)}{\Pr\left[\mathbf{N} = N\right]} = \frac{1}{\Pr\left[\mathbf{N} = N\right]}\lambda\frac{\tau^N}{N!}\theta^N e^{-\theta(\tau+\lambda)} = k(N)\theta^N e^{-\theta(\tau+\lambda)},\ \theta > 0$$

  where $k(N)$ depends on $N$ but not on $\theta$. Note that $f(\theta|N) = 0,\ \theta \leq 0$.

- In order to determine $k(N)$, we use the constraint

$$\int_0^\infty f(\theta|N)d\theta \ = \ 1 \ = \ k(N)\ \underbrace{\int_0^\infty \theta^N e^{-\theta(\tau+\lambda)}d\theta}_{g(N)}\ .$$

  Hence $k(N) = 1/g(N)$. We shall calculate $g(N)$ by partial integration:

$$g(N) \ = \ \left[\frac{\theta^N e^{-\theta(\tau+\lambda)}}{-(\tau+\lambda)}\right]_0^\infty + \frac{N}{\tau+\lambda}\int_0^\infty \theta^{N-1}e^{-\theta(\tau+\lambda)}d\theta = \frac{N}{\tau+\lambda}g(N-1) = \cdots = \frac{N!}{(\tau+\lambda)^N}g(0)$$

  where

$$g(0) \ = \ \int_0^\infty e^{-\theta(\tau+\lambda)}d\theta \ = \ \left[\frac{e^{-\theta(\tau+\lambda)}}{-(\tau+\lambda)}\right]_0^\infty \ = \ \frac{1}{\tau+\lambda}$$

# Ex: Poisson Process (3)

- which finally leads to

$$g(N) \;=\; \frac{N!}{(\tau+\lambda)^{N+1}}\,, k(N) \;=\; 1/g(N) \;=\; \frac{(\tau+\lambda)^{N+1}}{N!}\,.$$

- Now we can calculate

$$\widehat{\theta}_{MMSE} \;=\; E\left[\theta|N\right] \;=\; \int_0^\infty \theta\, f(\theta|N)\, d\theta$$

$$= \frac{1}{g(N)} \int_0^\infty \theta^{N+1} e^{-\theta(\tau+\lambda)} d\theta \;=\; \frac{g(N+1)}{g(N)} \;=\; \frac{N+1}{\tau+\lambda}\,.$$

# Ex: Poisson Process (4)

- To determine $\widehat{\theta}_{MAP}$, consider

$$\ln f(\theta|N) \;=\; -\ln g(N) + N \ln \theta - \theta(\tau + \lambda) \; .$$

Differentiation w.r.t. $\theta$ yields

$$\frac{\partial}{\partial \theta} \ln f(\theta|Y) \;=\; 0 \;=\; \frac{N}{\theta} - (\tau + \lambda) \; .$$

So we obtain

$$\widehat{\theta}_{MAP} \;=\; \frac{N}{\tau + \lambda} \;\neq\; \frac{N+1}{\tau + \lambda} \;=\; \widehat{\theta}_{MMSE} \; .$$

- Note however that if $\tau \gg \lambda$ (observation duration much longer than the a priori expected time between observations), then with high probability $N \gg 1$ and hence $\widehat{\theta}_{MAP} \approx \widehat{\theta}_{MMSE} \approx \frac{N}{\tau}$, which is simply the sample average of the number of messages per second.