# State Estimation of Linear and Nonlinear Dynamic Systems

## Part I: Linear Systems with Gaussian Noise

James B. Rawlings and Fernando V. Lima

Department of Chemical and Biological Engineering
University of Wisconsin–Madison

AICES Regional School
RWTH Aachen
March 10, 2008

## Outline

1. Introduction to State Estimation
2. Probability Concepts
   - Normal Distributions
3. State Estimation of Linear Systems
   - Kalman Filter (KF)
   - Least Squares (LS) Estimation
   - Connections between KF and LS
   - Limitations of These Approaches
4. Conclusions
5. Additional Reading

## Introduction to State Estimation

- Definition: estimate the system states ($x$) from measurements ($y$)
- Why estimate $x$?
  - $x$ is required to model the system
  - $y$ is a set of economically or conveniently measurable variables
    - Usually a subset of $x$
  - $x$ is corrupted with process noise ($w$) and $y$ with sensor noise ($v$)
- Challenge of State Estimation
  - Determine a good state estimate in the face of noisy and incomplete output measurements
- Probability Theory
  - Necessary to develop an optimal state estimator
  - Used to model fluctuations in the data

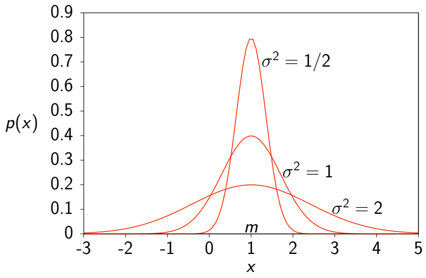
## Probability Concepts: Normal Distribution

### Definition (Normal Distribution)

The Normal or Gaussian distribution of a random variable $x$ is characterized by its mean $m$ and variance $\sigma^2$ and is given by

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

- Shorthand notation $x \sim N(m, \sigma^2)$
- Example of a normal distribution with
  - $m = 1$
  - $\sigma^2 = 1/2$, 1 and 2



## Multivariate Normal Distribution

- Let $x \sim N(m, P)$:

$$p(x) = \frac{1}{(2\pi)^{n/2} \left|P\right|^{1/2}} \exp\left[-\frac{1}{2}(x-m)^T P^{-1}(x-m)\right]$$

in which

- $x \in \mathbb{R}^n$ is a vector of random variables
- $p(x)$ is the probability density function
- $m \in \mathbb{R}^n$ is the mean
- $P \in \mathbb{R}^{n\times n}$ is the covariance matrix

## Multivariate Normal Example

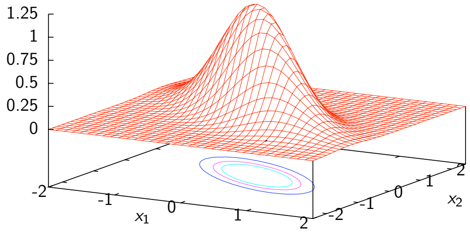


Figure displays multivariate normal for:

$$P^{-1} = \begin{bmatrix} 3.5 & 2.5 \\ 2.5 & 4.0 \end{bmatrix}, \qquad m = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

# Special Features of Normal Distributions (Fact 1)

**Definition (Statistical Independence)**

Two random variables $x$ and $y$ are statistically independent if

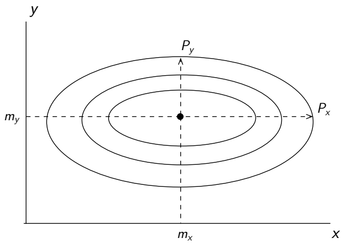
$$p(x, y) = p(x)p(y)$$

where $p(x, y)$ is the joint distribution of $x$ and $y$

- Joint densities of independent normals (Fact 1)
  - The joint density of two independent normals $x \sim N(m_x, P_x)$ and $y \sim N(m_y, P_y)$ is

$$p(x, y) = N(m_x, P_x)N(m_y, P_y)$$

$$p\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = N\left(\begin{bmatrix} m_x \\ m_y \end{bmatrix}, \begin{bmatrix} P_x & 0 \\ 0 & P_y \end{bmatrix}\right)$$



# Special Features of Normal Distributions (Fact 2)

- Linear transformation of a normal (Fact 2)

$$\text{For } x \sim N(m, P) \text{ and } y = Ax + b \Rightarrow y \sim N(\underbrace{Am + b}_{m_y}, \underbrace{APA^T}_{P_y})$$

- Example: For the distribution

$$P = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix} \qquad m = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

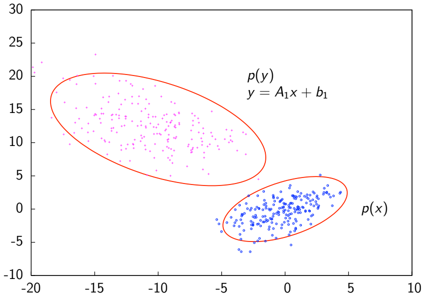
  - Consider two cases of linear transformations
    1. $$A_1 = \begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix} \qquad b_1 = \begin{bmatrix} -10 \\ 12 \end{bmatrix}$$
    2. $$A_2 = \begin{bmatrix} 1 & 1 \end{bmatrix} \qquad b_2 = \begin{bmatrix} -10 \end{bmatrix}$$

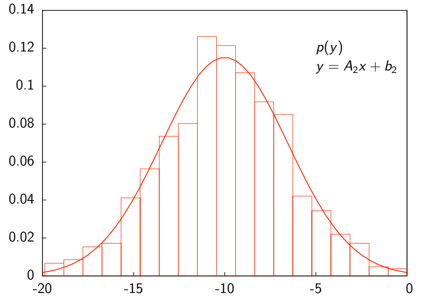
# Fact 2: Example Solution

- The transformed covariance matrix and mean for each case are

1.
$$P_y = A_1 P A_1^T = \begin{bmatrix} 12 & -6 \\ -6 & 12 \end{bmatrix}$$
$$m_y = A_1 m + b_1 = \begin{bmatrix} -10 \\ 12 \end{bmatrix}$$

2.
$$P_y = A_2 P A_2^T = \begin{bmatrix} 12 \end{bmatrix}$$
$$m_y = A_2 m + b_2 = \begin{bmatrix} -10 \end{bmatrix}$$





# Conditional Density Function

## Definition (Conditional Density)

Let $x$ and $y$ be jointly distributed with density $p(x, y)$. Assume that a specific realization of $y$ has been observed. The conditional density function is defined as

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

- Example: Rolling of a single die
  - Consider the following possible outcomes for $x$ and $y$

    $x = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$ and $y = \{\text{even, odd}\}$
- Calculate the probability $p(1|\text{odd})$ of having 1 as an outcome given that we know that the outcome is odd

## Solution (Use the definition of conditional density)

$$p(1, odd) = 1/6 \ and \ p(odd) = 3/6 \Rightarrow p(1|odd) = \frac{1/6}{3/6} = 1/3$$

# Special Features of Normal Distributions (Fact 3)

- Conditional densities of normal joint densities are also normal (Fact 3)

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} m_x \\ m_y \end{bmatrix}, \begin{bmatrix} P_x & P_{xy} \\ P_{yx} & P_y \end{bmatrix}\right) \Rightarrow p(x|y) \sim N(m, P)$$

$$\begin{aligned} m &= m_x + P_{xy}P_y^{-1}(y - m_y) \\ P &= P_x - P_{xy}P_y^{-1}P_{yx} \end{aligned}$$

# Why Normal Distributions?

**Theorem (Central Limit Theorem (CLT))**

*Consider a sequence $x_1$, $x_2$, ..., $x_n$ of independent and identically distributed random variables with respective densities $p_i(x) (i = 1, \dots, n)$. As n increases, the distribution of $\overline{x} = (x_1 + x_2 + \cdots x_n)/n$ tends to a normal curve, regardless of the shape of the original densities $p_i(x)$.*

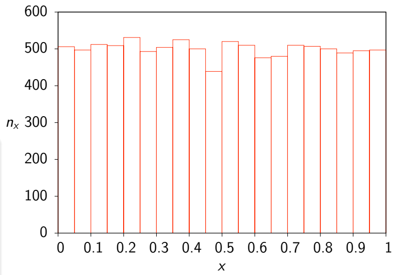
## Example: Adding Ten Uniformly Distributed Random Variables

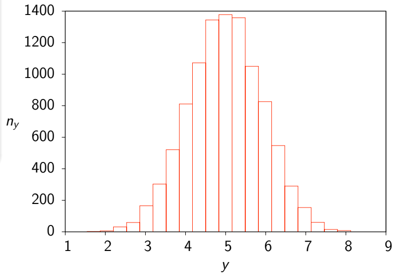
- Consider ten independent random variables and their sum

$$
\begin{aligned}
x &= [x_1, x_2, \ldots x_{10}]^T; \quad x \sim U(0,1) \\
y &= x_1 + x_2 + \cdots x_{10}
\end{aligned}
$$

- Using Fact 2 above and the CLT

$$y \text{ approximates } N(5, 5/6)$$





## State Estimation of Linear Systems: Kalman Filter (KF)

- Consider the linear, time invariant model with Gaussian noise

$$
\begin{aligned}
x(k+1) &= Ax(k) + w(k) \\
y(k) &= Cx(k) + v(k)
\end{aligned}
$$

$$w \sim N(0, Q) \qquad v \sim N(0, R) \qquad x(0) \sim N(\overline{x}(0), Q(0))$$

- The parameters of the initial state distribution, $\overline{x}(0)$ and $Q(0)$, are usually not known and often assumed

- Starting with the upcoming measurement $y(0)$, we want to determine the state estimate $\widehat{x}(0)$

## Step 1: Combining the Measurement $y(0)$

- The measurement $y(0)$ satisfies

$$\begin{bmatrix} x(0) \\ y(0) \end{bmatrix} = \begin{bmatrix} I & 0 \\ C & I \end{bmatrix} \begin{bmatrix} x(0) \\ v(0) \end{bmatrix}$$

- Assuming $v(0)$ is independent of $x(0)$, then from Fact 1

$$p\left(\begin{bmatrix} x(0) \\ v(0) \end{bmatrix}\right) = N\left(\begin{bmatrix} \overline{x}(0) \\ 0 \end{bmatrix}, \begin{bmatrix} Q(0) & 0 \\ 0 & R \end{bmatrix}\right)$$

- Since the pair $(x(0), y(0))$ is a linear transformation of $(x(0), v(0))$, then from Fact 2

$$p\left(\begin{bmatrix} x(0) \\ y(0) \end{bmatrix}\right) = N\left(\begin{bmatrix} \overline{x}(0) \\ C\overline{x}(0) \end{bmatrix}, \begin{bmatrix} Q(0) & Q(0)C^T \\ CQ(0) & CQ(0)C^T + R \end{bmatrix}\right)$$

## Step 1: Combining the Measurement $y(0)$ (cont'd)

- Using the conditional density result (Fact 3)

$$p\left(x(0)|y(0)\right) = N\left(m, P\right)$$

in which

$$\begin{aligned} m &= \overline{x}(0) + L(0)\left(y(0) - C\overline{x}(0)\right) \\ L(0) &= Q(0)C^T(CQ(0)C^T + R)^{-1} \\ P &= Q(0) - Q(0)C^T(CQ(0)C^T + R)^{-1}CQ(0) \end{aligned}$$

- The *optimal* state estimate is the value of $x(0)$ that maximizes $p(x(0)|y(0))$
  - For a normal, that is the mean and we choose $\widehat{x}(0) = m$
  - The change in variance after measurement ($Q(0)$ to $P(0) = P$) quantifies the information increase by obtaining $y(0)$

## Step 2: Forecasting the State Evolution

- We forecast from 0 to 1 using the model: $x(1) = Ax(0) + w(0)$

$$x(1) = \begin{bmatrix} A & I \end{bmatrix} \begin{bmatrix} x(0) \\ w(0) \end{bmatrix}$$

- Assuming $w(0)$ is independent of $x(0)$ and $y(0)$, then from Fact 1

$$p\left( \begin{bmatrix} x(0) \\ w(0) \end{bmatrix} \middle| y(0) \right) = N\left( \begin{bmatrix} \widehat{x}(0) \\ 0 \end{bmatrix}, \begin{bmatrix} P(0) & 0 \\ 0 & Q \end{bmatrix} \right)$$

- Using again the linear transformation result (Fact 2)

$$p(x(1)|y(0)) = N(\widehat{x}^-(1), P^-(1))$$

in which

$$\widehat{x}^-(1) = A\widehat{x}(0)$$
$$P^-(1) = AP(0)A^T + Q$$

## Recursion yields the KF update equations...

- Now we are ready to add measurement $y(1)$
  - calculate $p(x(1)|y(1))$ and forecast forward one more step
- We proceed adding measurements followed by forecasting
  - until we calculate the entire state distribution
    - this recursion yields the KF update equations

## Kalman Filter — Summary

- For a measurement trajectory $Y(k) = \{y(0), y(1), \dots y(k)\}$
  - we can compute the conditional density function exactly

$$\begin{aligned} p(x(k)|Y(k-1)) &= N(\widehat{x}^-(k), P^-(k)) && \text{(before } y(k)\text{)} \\ p(x(k)|Y(k)) &= N(\widehat{x}(k), P(k)) && \text{(after } y(k)\text{)} \end{aligned}$$

in which

$$\begin{aligned} \widehat{x}(k) &= \widehat{x}^-(k) + L(k)\left(y(k) - C\widehat{x}^-(k)\right) \\ L(k) &= P^-(k)C^T(CP^-(k)C^T + R)^{-1} \\ P(k) &= P^-(k) - P^-(k)C^T(CP^-(k)C^T + R)^{-1}CP^-(k) \end{aligned}$$

- Then we forecast the state evolution

$$\begin{aligned} \widehat{x}^-(k+1) &= A\widehat{x}(k) \\ P^-(k+1) &= AP(k)A^T + Q \end{aligned}$$

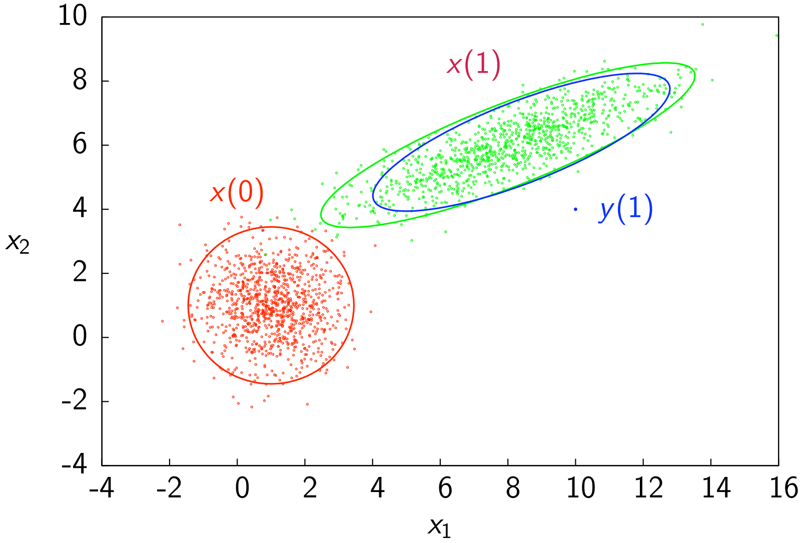
## Knowledge of Uncertainty

$P(k)$ provides a measure of the goodness of the estimate. It can be argued that knowledge of $P(k)$ is just as important as knowing the estimate $\widehat{x}(k)$ itself.

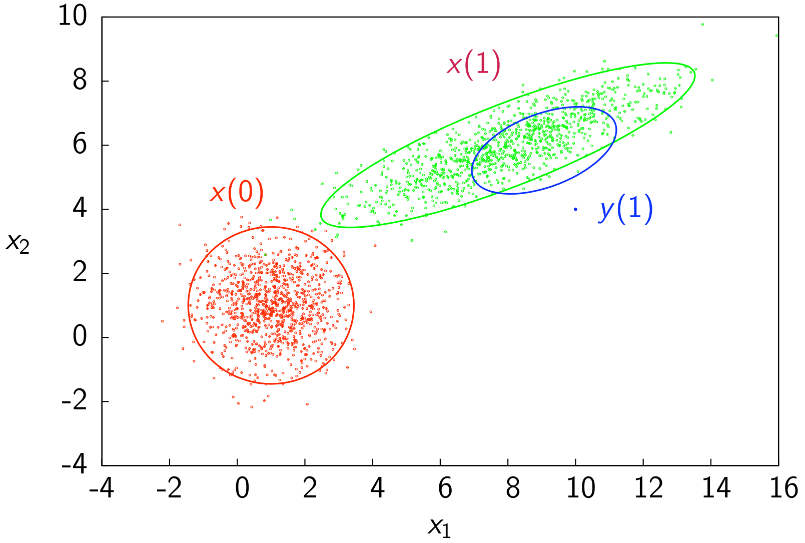
> An estimate is meaningless unless one knows how good it is.

—A.H. Jazwinski
Stochastic Processes and Filtering Theory (1970)

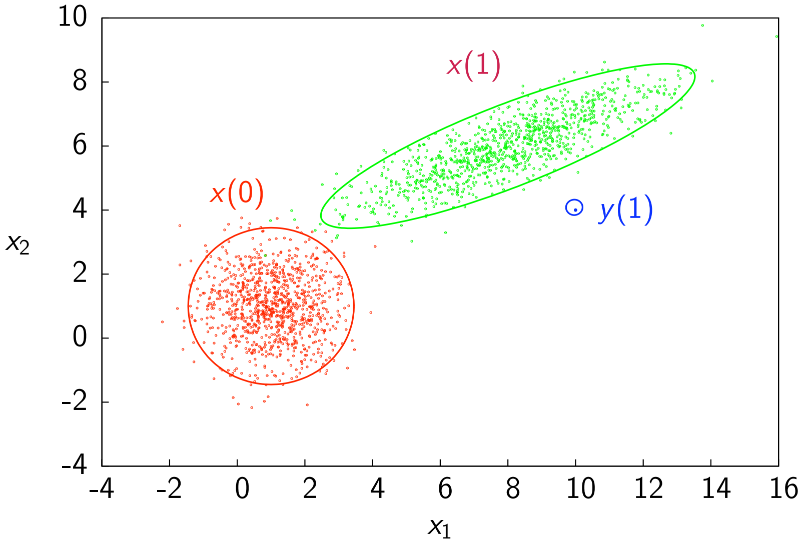
## Large $R$, ignore the measurement, trust the model forecast



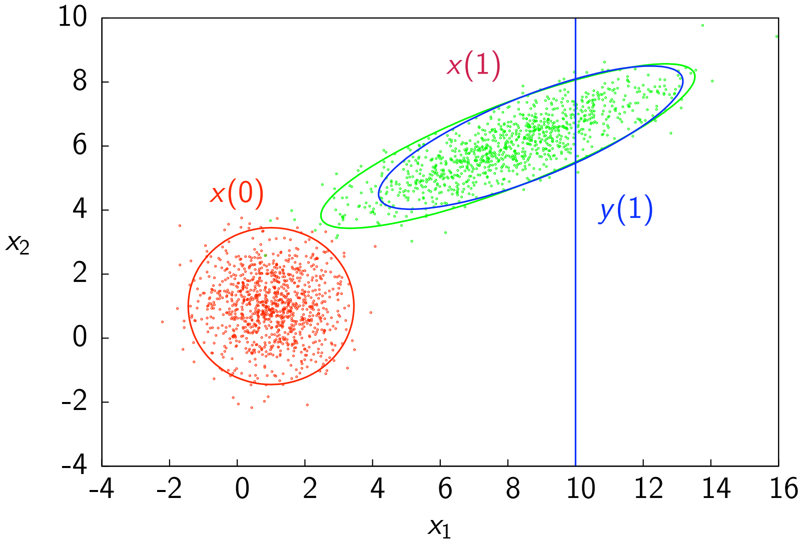
## Medium $R$, blend the measurement and the forecast



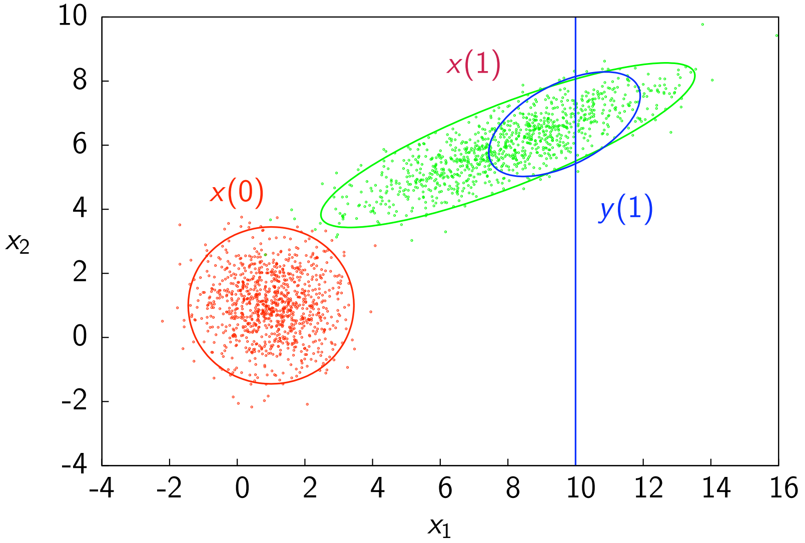
# Small $R$, trust the measurement, override the forecast



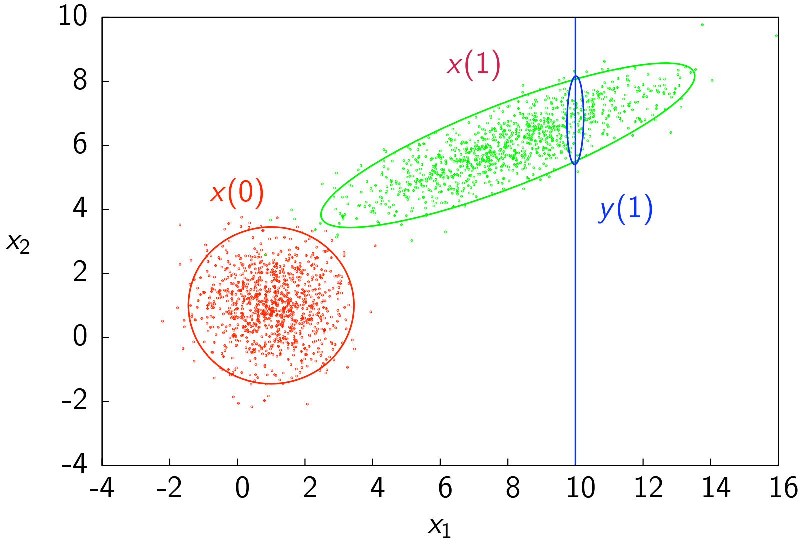
# Large $R$, $y$ measures $x_1$ only



# Medium $R$, $y$ measures $x_1$ only



# Small $R$, $y$ measures $x_1$ only



## Least Squares (LS) Estimation

> *One of the most important problems in the application of mathematics to the natural sciences is to choose the best of these many combinations, i.e., the combination that yields values of the unknowns that are least subject to errors.*

Theory of the Combination of Observations Least Subject to Errors.
C.F. Gauss, 1821.
G.W. Stewart Translation, 1995, p. 31.

## LS Formulation for Unconstrained Linear Systems

- Recall the unconstrained linear state space model

$$\begin{aligned} x(k+1) &= Ax(k) + w(k) \\ y(k) &= Cx(k) + v(k) \end{aligned}$$

- The state estimation problem is formulated as a deterministic LS optimization problem

$$\min_{x(0),\ldots,x(T)} \Phi(X(T))$$

# LS Formulation: Objective Function

- A reasonably flexible choice for objective function is

$$\Phi(X(T)) = \|x(0) - \overline{x}(0)\|^2_{(Q(0))^{-1}} + \sum_{k=0}^{T-1} + \|x(k+1) - Ax(k)\|^2_{Q^{-1}} + \sum_{k=0}^{T} \|y(k) - Cx(k)\|^2_{R^{-1}}$$

- Heuristic selection of Q and R
  - $Q \gg R$: trust the measurement, override the model forecast
  - $R \gg Q$: ignore the measurement, trust the model forecast

# Solution of LS Problem by Forward Dynamic Programming

**Step 1:** Adding the measurement at time $k$

$$\begin{aligned} P(k) &= P^-(k) - P^-(k)C^T(CP^-(k)C^T + R)^{-1}CP^-(k) && \text{(covariance)} \\ L(k) &= P^-(k)C^T(CP^-(k)C^T + R)^{-1} && \text{(gain)} \\ \widehat{x}(k) &= \widehat{x}^-(k) + L(k)(y(k) - C\widehat{x}^-(k)) && \text{(estimate)} \end{aligned}$$

**Step 2:** Propagating the model to time $k+1$

$$\begin{aligned} \widehat{x}^-(k+1) &= A\widehat{x}(k) && \text{(estimate)} \\ P^-(k+1) &= Q + AP(k)A^T && \text{(covariance)} \\ (\widehat{x}^-(0), P^-(0)) &= (\overline{x}(0), Q(0)) && \text{(initial condition)} \end{aligned}$$

**Same result as KF!**

## Probabilistic Estimation versus Least Squares

> *The recursive least squares approach was actually inspired by probabilistic results that automatically produce an equation of evolution for the estimate (the conditional mean). In fact, much of the recent least squares work did nothing more than rederive the probabilistic results (perhaps in an attempt to understand them). As a result, much of the least squares work contributes very little to estimation theory.*
> *—A.H. Jazwinski*
> *Stochastic Processes and Filtering Theory (1970)*

## Probability and Estimation

Probabilistic justification of least squares estimation.

- Either an arbitrary assumption on errors: Normal distribution. Gauss.
- Or an asymptotic argument: central limit theorem. Justifies least squares for infinitely large data sets regardless of error distribution. Laplace.

> *The connection of probability theory with a special problem in the combination of observations was made by Laplace in 1774 . . . Laplace's work described above was the beginning of a game of intellectual leapfrog between Gauss and Laplace that spanned several decades, and it is not easy to untangle their relative contributions. The problem is complicated by the fact that the two men are at extremes stylistically. Laplace is slapdash and lacks rigor, even by the standards of the time, while Gauss is reserved, often to the point of obscurity. Neither is easy to read.*
> *– G.W. Stewart, 1995, p. 214*

# Kalman Filter and Least Squares: Comparison

- Kalman Filter (Probabilistic)
  - Offers more insights on the comparison of different state estimators
    - based on the variance of their estimate error
  - Choice of $Q$ and $R$ is part of the model
  - Superior for unconstrained linear systems
- Least Squares
  - Objective function, although reasonable, is ad hoc
  - Choice of $Q$ and $R$ is arbitrary
  - Advantageous for significantly more complex models

# Limitations of These Approaches

- What about constraints?
  - Concentrations, particle size distributions, pressures, temperatures are positive.
  - Using this extra information provides more accurate estimates.
  - Projecting the unconstrained KF estimates to the feasible region is an ad hoc solution that does not satisfy the model.
- What about nonlinear models?
  - Almost all physical models in chemical and biological applications are nonlinear differential equations or nonlinear Markov processes.
  - Linearizing the nonlinear model and using the standard update formulas (extended Kalman filter) is the standard industrial approach.

# Conclusions

Here we have learned...

- Concepts of Probability Theory
  - Normal distributions and linear transformations
  - Joint and conditional densities
  - Independence and Central Limit Theorem
- Two state estimation techniques for linear systems
  - Kalman Filter and Least Squares
    - ★ Both provide the same result for unconstrained systems
    - ★ Do not apply to constrained or nonlinear systems

# Additional Reading

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, New York, third edition, 2003.

C. F. Gauss. *Theory of the Combination of Observations Least Subject to Errors: Part One, Part Two, Supplement*. SIAM, Philadelphia, 1995. ISBN 0-89871-347-1. Translated by G. W. Stewart.

A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970.

A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, Inc., second edition, 1984.