Mathematics for Robotics (ROB-GY 6013 Section A)

- **Week 12:**
  - Minimum Variance Estimator
  - Peek at Kalman Filter

Estimation Problem

- You try to measure the length of a table that you *secretly* know to be exactly $1.0\ \mathrm{m}$.
- Measure twice.
- Given measurements $y_1 = 0.9\ \mathrm{m}$ and $y_2 = 1.1\ \mathrm{m}$, estimate the true length x.
- Using our notation:

$$y = Cx + \varepsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = Cx + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \qquad y = \begin{bmatrix} 0.9 \\ 1.1 \end{bmatrix}, C = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \varepsilon = \begin{bmatrix} -0.1 \\ 0.1 \end{bmatrix}$$

- Here C is just ones, because there is a direct one to one relationship between measured length and actual length (as opposed to thermometer level and temperature)

Estimation Problem

- Estimation is easy: Just average the two measurements. Works perfectly here!

$$\hat{x} = \frac{1}{2} y_1 + \frac{1}{2} y_2 = 1.0$$

- In our notation: $\hat{x} = Ky$ $\quad \hat{K} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix}$

- Now I ask, is this choice of $\hat{K}$ really the ***best*** estimator for all possible situations?
  - Can I do better?
  - What if the errors of my two measurements have different variances? (shakier hands or measured by different people)
  - Should I still just average the measurements?

Estimation Problem

- I pose this simple example to guide you from being lost in the matrix
- What we are looking for in BLUE and MVE is a thoughtful approach to estimation problems and how to incorporate information about the world into mathematical form

Best Linear Unbiased Estimator (BLUE)

- **Goal:** How to choose the **weight matrix** in an **overdetermined** problem
- **Model:** $y = Cx + \varepsilon,$ (C is linearly independent)
  - **Measurement** (model output) $y \in \mathbb{R}^m$
  - **State** (model input) $x \in \mathbb{R}^n$ *unknown, deterministic*
  - **Noise** (output) $\varepsilon \in \mathbb{R}^m$ *stochastic,* $E\{\varepsilon\} = 0, \operatorname{cov}\{\varepsilon, \varepsilon\} = E\{\varepsilon\varepsilon^T\} = Q > 0$

$$\hat{x} = Ky \qquad E\{\hat{x} - x\} = 0 \qquad \operatorname{Var}(\hat{x} - x) = E\{(\hat{x} - x)^T (\hat{x} - x)\}$$

holds for all $x \in \mathbb{R}^n$

Minimizes variance

- **Find:** $\hat{K}$

$$\boxed{\hat{K} = (C^T Q^{-1} C)^{-1} C^T Q^{-1} \qquad \operatorname{cov}(\hat{x} - x) = (C^T Q^{-1} C)^{-1}}$$

Best Linear Unbiased Estimator (BLUE)

- **BLUE** is weighted least squares, where the weight matrix $S = Q^{-1}$ and A is C

Best Linear Unbiased Estimator (BLUE)

- **BLUE** is weighted least squares, where the weight matrix $S = Q^{-1}$ and A is C
- When you solve a **deterministic** weighted least squares problem, you implicitly assume that the uncertainty in your measurements has zero mean and covariance matrix Q

Best Linear Unbiased Estimator (BLUE)

- **BLUE** is weighted least squares, where the weight matrix $S = Q^{-1}$ and A is C
- When you solve a **deterministic** weighted least squares problem, you implicitly assume that the uncertainty in your measurements has zero mean and covariance matrix Q
- **Final remark:**
  - Weighted least squares was derived with normal equations for ***over***determined systems of linear equations
  - **BLUE** was derived with normal equations for ***under***determined systems of linear equations

Minimizing Variance Step in BLUE derivation

- Recall, the variance of a vector is the sum of the variances of its components

$$\operatorname{var}(X) = \sum_{i=1}^{p} \operatorname{var}(X_i)$$

$$\begin{aligned}\mathcal{E}\{(\widehat{x} - x)^\top (\widehat{x} - x)\} &= \operatorname{tr} \mathcal{E}\{K\varepsilon\varepsilon^\top K^\top\} \\ &= \operatorname{tr}(KQK^\top).\end{aligned}$$

$$\widehat{K} = \underset{KC=I}{\arg\min}\, \mathcal{E}\{(\widehat{x} - x)^\top (\widehat{x} - x)\} \iff \widehat{K} = \underset{KC=I}{\arg\min}\, \operatorname{tr}(KQK^\top)$$

Minimizing Variance Step in BLUE derivation

- Recall, the variance of a vector is the sum of the variances of its components

$$\mathrm{var}(X) = \sum_{i=1}^{p} \mathrm{var}(X_i)$$

$$\begin{aligned}\mathcal{E}\{(\widehat{x} - x)^\top (\widehat{x} - x)\} &= \mathrm{tr}\,\mathcal{E}\{K\varepsilon\varepsilon^\top K^\top\} \\ &= \mathrm{tr}(KQK^\top).\end{aligned}$$

$$\widehat{K} = \arg\min_{KC=I} \boxed{\mathcal{E}\{(\widehat{x} - x)^\top (\widehat{x} - x)\}} \iff \widehat{K} = \arg\min_{KC=I} \mathrm{tr}(KQK^\top)$$

The Unexpected Inner Product

- How to design an inner product that is somehow connected to variance?

The Unexpected Inner Product

- How to design an inner product that is somehow connected to variance?

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \end{bmatrix} \in \mathbb{R}^m$$

The Unexpected Inner Product

- How to design an inner product that is somehow connected to variance?

$$x = \begin{bmatrix} \boxed{x_1} \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \boxed{\varepsilon_m} \end{bmatrix} \in \mathbb{R}^m$$

- Each is a random variable, and hence are ***functions***

The Unexpected Inner Product

- How to design an inner product that is somehow connected to variance?

$$x = \begin{bmatrix} \boxed{x_1} \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \boxed{\varepsilon_m} \end{bmatrix} \in \mathbb{R}^m$$

- Each is a random variable, and hence are ***functions***
- $(\mathcal{X}, \mathbb{R})$ is a **vector space**, where $\mathcal{X} = \operatorname{span}\{x_1, \ldots, x_n, \varepsilon_1, \ldots, \varepsilon_m\}$

The Unexpected Inner Product

- How to design an inner product that is somehow connected to variance?

$$x = \begin{bmatrix} \boxed{x_1} \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \boxed{\varepsilon_m} \end{bmatrix} \in \mathbb{R}^m$$

- Each is a random variable, and hence are ***functions***
- $(\mathcal{X}, \mathbb{R})$ is a **vector space**, where $\mathcal{X} = \text{span}\{x_1, \ldots, x_n, \varepsilon_1, \ldots, \varepsilon_m\}$
- $(\mathcal{X}, \mathbb{R}, < \cdot\,, \cdot >)$ is an **inner product space**, where $< z_1, z_2 > := E\{z_1 z_2\}$

Minimum Variance Estimator (MVE)

- **Model:** $y = Cx + \varepsilon,\;\; y \in \mathbb{R}^m,\;\; x \in \mathbb{R}^n$, and $\varepsilon \in \mathbb{R}^m$

Minimum Variance Estimator (MVE)

- **Model:** $y = Cx + \varepsilon,\ \ y \in \mathbb{R}^m,\ \ x \in \mathbb{R}^n$, and $\varepsilon \in \mathbb{R}^m$

C does not have to be linearly independent
but something else has to be…

Minimum Variance Estimator (MVE)

- **Model:** $y = Cx + \varepsilon, \; y \in \mathbb{R}^m, \; x \in \mathbb{R}^n$, and $\varepsilon \in \mathbb{R}^m$
  - Stochastic assumptions:
    - **Zero means:** $E\{x\} = 0, E\{\varepsilon\} = 0$
    - **Covariances:** $E\{\varepsilon\varepsilon^T\} = Q, E\{xx^T\} = P, E\{\varepsilon x^T\} = 0$

C does not have to be linearly independent
but something else has to be…

Minimum Variance Estimator (MVE)

- **Model:** $y = Cx + \varepsilon,\ \ y \in \mathbb{R}^m,\ \ x \in \mathbb{R}^n$, and $\varepsilon \in \mathbb{R}^m$

  C does not have to be linearly independent *but something else has to be…*

  - Stochastic assumptions:
    - **Zero means:** $E\{x\} = 0$, $E\{\varepsilon\} = 0$

      x can be shifted to have zero mean
    - **Covariances:** $E\{\varepsilon\varepsilon^T\} = Q$, $E\{xx^T\} = P$, $E\{\varepsilon x^T\} = 0$

      State and noise are uncorrelated

Minimum Variance Estimator (MVE)

- **Model:** $y = Cx + \varepsilon,\ \ y \in \mathbb{R}^m,\ \ x \in \mathbb{R}^n$, and $\varepsilon \in \mathbb{R}^m$
  - Stochastic assumptions:
    - **Zero means:** $E\{x\} = 0$, $E\{\varepsilon\} = 0$
    - **Covariances:** $E\{\varepsilon\varepsilon^T\} = Q$, $E\{xx^T\} = P$, $E\{\varepsilon x^T\} = 0$

C does not have to be linearly independent
but something else has to be…

x can be shifted to have zero mean

State and noise are uncorrelated

- Assumption: $Q \geq 0$, $P \geq 0$, and $CPC^T + Q > 0$

Minimum Variance Estimator (MVE)

- **Model:** $y = Cx + \varepsilon,\; y \in \mathbb{R}^m,\; x \in \mathbb{R}^n$, and $\varepsilon \in \mathbb{R}^m$

  C does not have to be linearly independent *but something else has to be…*

  - Stochastic assumptions:
    - **Zero means:** $E\{x\} = 0$, $E\{\varepsilon\} = 0$

      x can be shifted to have zero mean

    - **Covariances:** $E\{\varepsilon\varepsilon^T\} = Q$, $E\{xx^T\} = P$, $E\{\varepsilon x^T\} = 0$

      State and noise are uncorrelated

- Assumption: $Q \geq 0$, $P \geq 0$, and $CPC^T + Q > 0$

  This is related to that something else…

Minimum Variance Estimator (MVE)

- **Model:** $y = Cx + \varepsilon,\ \ y \in \mathbb{R}^m,\ \ x \in \mathbb{R}^n$, and $\varepsilon \in \mathbb{R}^m$

  C does not have to be linearly independent *but something else has to be…*

  - Stochastic assumptions:
    - **Zero means:** $E\{x\} = 0$, $E\{\varepsilon\} = 0$

      x can be shifted to have zero mean

    - **Covariances:** $E\{\varepsilon\varepsilon^T\} = Q$, $E\{xx^T\} = P$, $E\{\varepsilon x^T\} = 0$

      State and noise are uncorrelated

- Assumption: $Q \geq 0$, $P \geq 0$, and $CPC^T + Q > 0$

  This is related to that something else…

- **New Idea:** $(\mathcal{X}, \mathbb{R}, < \cdot\ , \cdot >)$ is an **inner product space**, where

$$\mathcal{X} = \operatorname{span}\{x_1, \ldots, x_n, \varepsilon_1, \ldots, \varepsilon_m\} \text{ and } < z_1, z_2 > := E\{z_1 z_2\}$$

Any linear estimate is unbiased

- For any $\hat{x} = Ky$, the zero mean assumption $E\{x\} = 0$ guarantees unbiasedness
- For **BLUE**, we needed to impose $KC = I$

$$E\{\widehat{x} - x\} = E\{Ky - x\} = E\{KCx + K\varepsilon - x\} = (KC - I)E\{x\} + KE\{\varepsilon\} = 0$$

More preparation

- Evaluating the inner product for MVE:

$$E\{z_1 z_2\} = \begin{cases} P_{ij} & z_1 = x_i, z_2 = x_j \\ Q_{ij} & z_1 = \varepsilon_i, z_2 = \varepsilon_j \\ 0 & z_1 = x_i, z_2 = \varepsilon_j \\ 0 & z_1 = \varepsilon_i, z_2 = x_j \end{cases}$$

- **Define:** $M = \operatorname{span}\{y_1, \ldots, y_m\} \subset \mathcal{X}$
    - Important that the y's are **linearly independent** (which they are if and only if $CPC^T + Q > 0$)

Proof

- Proof detailed in Section 5.3.2

Result: Minimum Variance Estimator

- **Goal:** How to choose the **weight matrix** in an **overdetermined** problem
- **Model:** $y = Cx + \varepsilon,$ (C does not have to be linearly independent)
  - **Measurement** (model output) $y \in \mathbb{R}^m$
  - **State** (model input) $x \in \mathbb{R}^n$ *stochastic,* $E\{x\} = 0,\ \mathrm{cov}\{x, x\} = E\{xx^T\} = P > 0$
  - **Noise** (output) $\varepsilon \in \mathbb{R}^m$ *stochastic,* $E\{\varepsilon\} = 0,\ \mathrm{cov}\{\varepsilon, \varepsilon\} = E\{\varepsilon\varepsilon^T\} = Q > 0$

    state and noise are uncorrelated $E\{xx^T\} = 0$

$$\hat{x} = Ky \qquad E\{\hat{x} - x\} = 0 \qquad \mathrm{Var}(\hat{x} - x) = E\{(\hat{x} - x)^T(\hat{x} - x)\}$$

holds for all $x \in \mathbb{R}^n$

Minimizes variance

- **Find:** $\hat{K}$

$$\boxed{\hat{K} = PC^T(CPC^T + Q)^{-1} \qquad \mathrm{cov}(\hat{x} - x) = P - PC^T(CPC^T + Q)^{-1}CP}$$

Observations

- $\operatorname{cov}(\begin{bmatrix} x \\ y \end{bmatrix}) = E\{\begin{bmatrix} x \\ Cx+\varepsilon \end{bmatrix}\begin{bmatrix} x^T & x^T C^T + \varepsilon^T \end{bmatrix}\} = \begin{bmatrix} P & PC^T \\ PC^T & CPC^T + Q \end{bmatrix}$

$$\boxed{\hat{K} = PC^T(CPC^T+Q)^{-1} \qquad \operatorname{cov}(\hat{x}-x) = P - PC^T(CPC^T+Q)^{-1}CP}$$

Observations

- $\mathrm{cov}(\begin{bmatrix} x \\ y \end{bmatrix}) = E\{\begin{bmatrix} x \\ Cx+\varepsilon \end{bmatrix}\begin{bmatrix} x^T & x^T C^T + \varepsilon^T \end{bmatrix}\} = \begin{bmatrix} P & PC^T \\ PC^T & CPC^T + Q \end{bmatrix}$
- Schur Complement of $\mathrm{cov}(\begin{bmatrix} x \\ y \end{bmatrix})$ is $\mathrm{cov}(\hat{x} - x) = P - PC^T (CPC^T + Q)^{-1} CP$

$$\boxed{\hat{K} = PC^T (CPC^T + Q)^{-1} \qquad \mathrm{cov}(\hat{x} - x) = P - PC^T (CPC^T + Q)^{-1} CP}$$

Observations

- $\mathrm{cov}(\begin{bmatrix} x \\ y \end{bmatrix}) = E\{\begin{bmatrix} x \\ Cx+\varepsilon \end{bmatrix}\begin{bmatrix} x^T & x^T C^T + \varepsilon^T \end{bmatrix}\} = \begin{bmatrix} P & PC^T \\ PC^T & CPC^T + Q \end{bmatrix}$

- Schur Complement of $\mathrm{cov}(\begin{bmatrix} x \\ y \end{bmatrix})$ is $\mathrm{cov}(\hat{x}-x) = P - \boxed{PC^T(CPC^T+Q)^{-1}CP}$

Reduction in P due to extra information

$$\boxed{\hat{K} = PC^T(CPC^T+Q)^{-1} \qquad \mathrm{cov}(\hat{x}-x) = P - PC^T(CPC^T+Q)^{-1}CP}$$

Observations

- $\mathrm{cov}(\begin{bmatrix} x \\ y \end{bmatrix}) = E\{\begin{bmatrix} x \\ Cx+\varepsilon \end{bmatrix}\begin{bmatrix} x^T & x^T C^T + \varepsilon^T \end{bmatrix}\} = \begin{bmatrix} P & PC^T \\ PC^T & CPC^T + Q \end{bmatrix}$

- Schur Complement of $\mathrm{cov}(\begin{bmatrix} x \\ y \end{bmatrix})$ is $\mathrm{cov}(\hat{x} - x) = P - \boxed{PC^T (CPC^T + Q)^{-1} CP}$

Reduction in P due to extra information

Apply Matrix Inversion Lemma $(A + BCD)^{-1} = A^{-1} - A^{-1}B\left(DA^{-1}B + C^{-1}\right)^{-1} DA^{-1}$

$$\boxed{\hat{K} = PC^T (CPC^T + Q)^{-1} \qquad \mathrm{cov}(\hat{x} - x) = P - PC^T (CPC^T + Q)^{-1} CP}$$

Observations

- $\mathrm{cov}(\begin{bmatrix} x \\ y \end{bmatrix}) = E\{\begin{bmatrix} x \\ Cx+\varepsilon \end{bmatrix}\begin{bmatrix} x^T & x^T C^T + \varepsilon^T \end{bmatrix}\} = \begin{bmatrix} P & PC^T \\ PC^T & CPC^T + Q \end{bmatrix}$

- Schur Complement of $\mathrm{cov}(\begin{bmatrix} x \\ y \end{bmatrix})$ is $\mathrm{cov}(\hat{x} - x) = P - \boxed{PC^T (CPC^T + Q)^{-1} CP}$

Reduction in P due to extra information

Apply Matrix Inversion Lemma $(A + BCD)^{-1} = A^{-1} - A^{-1}B\left(DA^{-1}B + C^{-1}\right)^{-1} DA^{-1}$

$$\boxed{\hat{K} = (C^T Q^{-1} C + P^{-1})^{-1} C^T Q^{-1} \qquad \mathrm{cov}(\hat{x} - x) = P - PC^T (CPC^T + Q)^{-1} CP}$$

Observations

- $\text{cov}(\begin{bmatrix} x \\ y \end{bmatrix}) = E\{\begin{bmatrix} x \\ Cx+\varepsilon \end{bmatrix}\begin{bmatrix} x^T & x^T C^T + \varepsilon^T \end{bmatrix}\} = \begin{bmatrix} P & PC^T \\ PC^T & CPC^T + Q \end{bmatrix}$

- Schur Complement of $\text{cov}(\begin{bmatrix} x \\ y \end{bmatrix})$ is $\text{cov}(\hat{x} - x) = P - \boxed{PC^T (CPC^T + Q)^{-1} CP}$

Reduction in P due to extra information

$$\boxed{\hat{K} = (C^T Q^{-1} C + P^{-1})^{-1} C^T Q^{-1} \qquad \text{cov}(\hat{x} - x) = P - PC^T (CPC^T + Q)^{-1} CP}$$

Observations

- $\text{cov}(\begin{bmatrix} x \\ y \end{bmatrix}) = E\{\begin{bmatrix} x \\ Cx+\varepsilon \end{bmatrix}\begin{bmatrix} x^T & x^T C^T + \varepsilon^T \end{bmatrix}\} = \begin{bmatrix} P & PC^T \\ PC^T & CPC^T + Q \end{bmatrix}$

- Schur Complement of $\text{cov}(\begin{bmatrix} x \\ y \end{bmatrix})$ is $\text{cov}(\hat{x}-x) = P - \boxed{PC^T(CPC^T+Q)^{-1}CP}$

Reduction in P due to extra information

- Comparison with **BLUE** $\boxed{\hat{K} = (C^T Q^{-1} C)^{-1} C^T Q^{-1}}$

$$\boxed{\hat{K} = (C^T Q^{-1} C + P^{-1})^{-1} C^T Q^{-1} \qquad \text{cov}(\hat{x}-x) = P - PC^T(CPC^T+Q)^{-1}CP}$$

Observations

- $\mathrm{cov}(\begin{bmatrix} x \\ y \end{bmatrix}) = E\{\begin{bmatrix} x \\ Cx+\varepsilon \end{bmatrix}\begin{bmatrix} x^T & x^T C^T + \varepsilon^T \end{bmatrix}\} = \begin{bmatrix} P & PC^T \\ PC^T & CPC^T + Q \end{bmatrix}$
- Schur Complement of $\mathrm{cov}(\begin{bmatrix} x \\ y \end{bmatrix})$ is $\mathrm{cov}(\hat{x}-x) = P - \boxed{PC^T(CPC^T+Q)^{-1}CP}$

Reduction in P due to extra information

- Comparison with **BLUE** $\boxed{\hat{K} = (C^T Q^{-1} C)^{-1} C^T Q^{-1}}$

Same when $P^{-1} = 0$ (zero information about x!)

$$\boxed{\hat{K} = (C^T Q^{-1} C + P^{-1})^{-1} C^T Q^{-1} \qquad \mathrm{cov}(\hat{x}-x) = P - PC^T(CPC^T+Q)^{-1}CP}$$

Observation

- For **BLUE** to exist: $\dim(y) \geq \dim(x)$
- For **MVE** to exist: $CPC^T + Q > 0$.
  - we can have $\dim(y) < \dim(x)$

Introduction to more probability

- Recap: probability density functions for a random vector

Definition 5.16 $X : \Omega \to \mathbb{R}^p$ *is a* ***continuous random vector*** *if there exists a* ***density*** $f_X : \mathbb{R}^p \to [0, \infty)$ *such that,*

$$\forall\, x \in \mathbb{R}^P,\ \ P(\{X \leq x\}) = \int_{-\infty}^{x_p} \dots \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f_X(\bar{x}_1, \bar{x}_2 \dots \bar{x}_p) d\bar{x}_1 d\bar{x}_2 \dots d\bar{x}_p.$$

More generally, for all $A \subset \mathbb{R}^p$ *such that the indicator function* I_A *has bounded variation,*

$$P(\{X \in A\}) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I_A(\bar{x}_1, \bar{x}_2 \dots \bar{x}_p) f_X(\bar{x}_1, \bar{x}_2 \dots \bar{x}_p) d\bar{x}_1 d\bar{x}_2 \dots d\bar{x}_p.$$

Notation 5.17 *The notation* $X \sim f$ *is read as* X *is distributed with density* f *or that* X *is a random vector with density* f*.*

Definition 5.18 *(Moments) Suppose* $g : \mathbb{R}^p \to \mathbb{R}^k$

$$\mathcal{E}\{g(X)\} := \int_{\mathbb{R}^p} g(x) f_X(x) dx := \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} g(x_1, \dots, x_p) f_X(x_1, \dots, x_p) dx_1 \dots dx_p$$

Random Vector X composed of: Two Random Variables X_1, X_2

- Random vector: $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$
- Density: $f_X(x_1, x_2)$
- Mean: $\mu = E\{X\} = \begin{bmatrix} E\{X_1\} \\ E\{X_2\} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$
- Covariance: $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = E\left\{ \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix}^T \right\}$

Random Vector X composed of: Two Random ***VECTORS*** X_1, X_2

- Random vector: $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$
- ***Joint Density:*** $f_X(x) = f_{\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}}(x_1, x_2) = f_{X_1 X_2}(x_1, x_2)$
- Mean: $\mu = E\{X\} = \begin{bmatrix} E\{X_1\} \\ E\{X_2\} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$
- Covariance: $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = E\left\{ \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix}^T \right\}$

Random Vector X composed of: Two Random ***<u>VECTORS</u>*** X_1, X_2 (e.g., state, sensor value)

- Random vector: $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$
- ***Joint Density:*** $f_X(x) = f_{\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}}(x_1, x_2) = f_{X_1 X_2}(x_1, x_2)$
- Mean: $\mu = E\{X\} = \begin{bmatrix} E\{X_1\} \\ E\{X_2\} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$
- Covariance: $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = E\left\{ \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix}^T \right\}$

Random Vector X composed of: Two Random ***VECTORS*** X_1, X_2 (e.g., state, sensor value)

- Random vector: $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$
- ***Joint Density:*** $f_X(x) = f_{\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}}(x_1, x_2) = f_{X_1 X_2}(x_1, x_2)$
- Mean: $\mu = E\{X\} = \begin{bmatrix} E\{X_1\} \\ E\{X_2\} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$
- Covariance: $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = E\left\{ \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix}^T \right\}$

$$= E\left\{ \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix} \begin{bmatrix} (X_1 - \mu_1)^T & (X_2 - \mu_2)^T \end{bmatrix} \right\}$$

$$= E\left\{ \begin{bmatrix} (X_1 - \mu_1)(X_1 - \mu_1)^T & (X_1 - \mu_1)(X_2 - \mu_2)^T \\ (X_2 - \mu_2)(X_1 - \mu_1)^T & (X_2 - \mu_2)(X_2 - \mu_2)^T \end{bmatrix} \right\}$$

Random Vector X composed of: Two Random ***VECTORS*** X_1, X_2 (e.g., state, sensor value)

- Random vector: $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$
- ***Joint Density:*** $f_X(x) = f_{\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}}(x_1, x_2) = f_{X_1 X_2}(x_1, x_2)$
- Mean: $\mu = E\{X\} = \begin{bmatrix} E\{X_1\} \\ E\{X_2\} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$
- Covariance: $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = E\left\{ \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix}^T \right\}$

$$= E\left\{ \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix} \begin{bmatrix} (X_1 - \mu_1)^T & (X_2 - \mu_2)^T \end{bmatrix} \right\}$$

$$= E\left\{ \begin{bmatrix} (X_1 - \mu_1)(X_1 - \mu_1)^T & (X_1 - \mu_1)(X_2 - \mu_2)^T \\ (X_2 - \mu_2)(X_1 - \mu_1)^T & (X_2 - \mu_2)(X_2 - \mu_2)^T \end{bmatrix} \right\}$$

$\Sigma_{12} = \Sigma_{21}^\top = cov(X_1, X_2) = \mathcal{E}\{(X_1 - \mu_1)(X_2 - \mu_2)^\top\}$ is also called the **correlation** of X_1 and X_2

Random Vector X composed of: Two Random ***VECTORS*** X_1, X_2 (e.g., state, sensor value)

- Random vector: $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$
- ***Joint Density:*** $f_X(x) = f_{\left[\begin{array}{c} X_1 \\ X_2 \end{array}\right]}(x_1, x_2) = f_{X_1 X_2}(x_1, x_2)$
- X_1, X_2 do not have to be the same size: $X_1 : \Omega \to R^n$ and $X_2 : \Omega \to R^m$

Random Vector X composed of: Two Random ***VECTORS*** X_1, X_2 (e.g., state, sensor value)

- Random vector: $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$
- ***Joint* Density:** $f_X(x) = f_{\left[\begin{smallmatrix} X_1 \\ X_2 \end{smallmatrix}\right]}(x_1, x_2) = f_{X_1 X_2}(x_1, x_2)$
- X_1, X_2 do not have to be the same size: $X_1 : \Omega \to R^n$ and $X_2 : \Omega \to R^m$
- ***Marginal* Density:**

$$f_{X_1}(x_1) := \int_{-\infty}^{\infty} f_{X_1 X_2}(x_1, x_2) dx_2$$

$$:= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1 X_2}(\underbrace{\bar{x}_1, \ldots, \bar{x}_n}_{x_1}, \underbrace{\bar{x}_{n+1}, \cdots, \bar{x}_{n+m}}_{x_2}) \underbrace{d\bar{x}_{n+1} \cdots d\bar{x}_{n+m}}_{dx_2}$$

$$f_{X_2}(x_2) := \int_{-\infty}^{\infty} f_{X_1 X_2}(x_1, x_2) dx_1$$

$$:= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1 X_2}(\underbrace{\bar{x}_1, \ldots, \bar{x}_n}_{x_1}, \underbrace{\bar{x}_{n+1}, \cdots, \bar{x}_{n+m}}_{x_2}) \underbrace{d\bar{x}_1 \cdots d\bar{x}_n}_{dx_1}$$

Independence

- Random vectors X_1 and X_2 are **independent** if and only if their joint density factors

$$f_X(x) = f_{X_1 X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$$

- Consider flipping coins

Correlation

- Random vectors X_1 and X_2 are **uncorrelated** if and only if their “cross covariance” or “correlation” is zero

$$cov(X_1, X_2) := \mathcal{E}\{(X_1 - \mu_1)(X_2 - \mu_2)^\top\} = 0_{n\times m}$$

- Independence is stronger than zero correlation
  - Independence → zero correlation
  - Converse not true in general!

Definition: Conditional Probability

- Consider two events $A, B \in \mathcal{F}$, with $P(B) > 0$. The conditional probability of A given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Definition: Conditional Probability

- Consider two events $A, B \in \mathcal{F}$, with $P(B) > 0$. The conditional probability of A given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- The conditional probability of event A given that event B occurred is how we “fuse” the two pieces of information

Definition: Conditional Probability

- Consider two events $A, B \in \mathcal{F}$, with $P(B) > 0$. The conditional probability of A given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- The conditional probability of event A given that event B occurred is how we “fuse” the two pieces of information

- X_1 given $X_2 = x_2$ is (still) a random vector.
  - Density function: $f_{X_1|X_2}(x_1 \mid x_2)$

Definition: Conditional Probability

- Consider two events $A, B \in \mathcal{F}$, with $P(B) > 0$. The conditional probability of A given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \begin{array}{l} \textbf{“Joint”} \\ \textbf{“Marginal”} \end{array}$$

- The conditional probability of event A given that event B occurred is how we “fuse” the two pieces of information

- X_1 given $X_2 = x_2$ is (still) a random vector.
  - Density function: $f_{X_1 \mid X_2}(x_1 \mid x_2)$

Definition: Conditional Probability

- Consider two events $A, B \in \mathcal{F}$, with $P(B) > 0$. The conditional probability of A given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

"Joint"

"Marginal"

- The conditional probability of event A given that event B occurred is how we "fuse" the two pieces of information

- X_1 given $X_2 = x_2$ is (still) a random vector.
  - Density function:

$$f_{X_1|X_2}(x_1 \mid x_2) := \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}$$

Conditional Random Vectors

- **Conditional density function:** $$f_{X_1|X_2}(x_1 \mid x_2) := \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}$$

- **Conditional mean** (as function of x_2): $$\mu_{X_1|X_2=x_2} := \mathcal{E}\{X_1 \mid X_2 = x_2\}$$ $$:= \int_{-\infty}^{\infty} x_1 f_{X_1|X_2}(x_1 \mid x_2) dx_1$$

- **Conditional covariance** (as function of x_2):

$$\Sigma_{X_1|X_2=x_2} := \mathcal{E}\{(X_1 - \mu_{X_1|X_2=x_2})(X_1 - \mu_{X_1|X_2=x_2})^\top \mid X_2 = x_2\}$$

$$:= \int_{-\infty}^{\infty} (X_1 - \mu_{X_1|X_2=x_2})(X_1 - \mu_{X_1|X_2=x_2})^\top f_{X_1|X_2}(x_1 \mid x_2) dx_1$$

Definition: Gaussian Random Variable

- A random variable X is normally distributed with mean μ and variance $\sigma^2 > 0$ if it has density

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

- The standard deviation is $\sigma > 0$

- Check that μ is the mean: $$\mu := \mathcal{E}\{X\} := \int_{\mathbb{R}} x f_X(x) dx := \int_{-\infty}^{\infty} x f_X(x) dx$$

- Check that σ^2 is the variance $$\sigma^2 := \mathcal{E}\{(X-\mu)^2\} := \int_{\mathbb{R}} (x-\mu)^2 f_X(x) dx := \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x) dx$$

Definition: Gaussian Random Vector

- A finite collection of random variables $X_1, X_2, \ldots, X_p$, or equivalently, the random vector

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$$

has a (non-degenerate) **multivariate normal distribution** with mean μ and covariance $\Sigma > 0$ if the joint density is given by

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

Definition: Gaussian Random Vector

- A finite collection of random variables $X_1, X_2, \ldots, X_p$, or equivalently, the random vector

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$$

Non-zero determinant

has a (non-degenerate) **multivariate normal distribution** with mean μ and covariance $\Sigma > 0$ if the joint density is given by

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^p \, |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right)$$

determinant

Definition: Gaussian Random Vector

- A finite collection of random variables $X_1, X_2, \ldots, X_p$, or equivalently, the random vector

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$$

Non-zero determinant

has a (non-degenerate) **multivariate normal distribution** with mean μ and covariance $\Sigma > 0$ if the joint density is given by

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

determinant

Compare with single Gaussian random variable $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$

Definition: Gaussian Marginal Densities/Distributions

- Each random variable X_i has a **univariate normal distribution** with mean μ_i and variance Σ_{ii},

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi\Sigma_{ii}}} e^{-\frac{(x_i - \mu_i)^2}{2\Sigma_{ii}}}$$

Variance of X_i

Variance of X_i

- No integrals needed!

Independence

- Gaussian random variables are very special in that they are independent if, and only if, they are uncorrelated.

  X_i and X_j are independent if, and only if, $\Sigma_{ij} = \Sigma_{ji} = 0$

- No need to check if joint density factors into marginal densities

Linear Combination

- Define a new random vector by $Y = AX + b$, with the rows of A linearly independent. Then Y is a Gaussian (normal) random vector with

$$\mathcal{E}\{Y\} = A\mu + b =: \mu_Y$$

$$\operatorname{cov}(Y, Y) = \mathcal{E}\{(Y - \mu_Y)(Y - \mu_Y)^\top\} = A\Sigma A^\top =: \Sigma_{YY}$$

Linear Combination

- Define a new random vector by $Y = AX + b$, with the rows of A linearly independent. Then Y is a Gaussian (normal) random vector with

$$\mathcal{E}\{Y\} = A\mu + b =: \mu_Y$$

$$\operatorname{cov}(Y, Y) = \mathcal{E}\{(Y - \mu_Y)(Y - \mu_Y)^\top\} = A\Sigma A^\top =: \Sigma_{YY}$$

- Note that $Y - \mu_Y = A(X - \mu)$:

$$\operatorname{cov}(Y, Y) = \mathcal{E}\{[A(X - \mu)][A(X - \mu)]^\top\} = A\mathcal{E}\{(X - \mu)(X - \mu)^\top\}A^\top = A\Sigma A^\top$$

Conditioning on Gaussian Random Vectors

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \begin{matrix} \in \mathbb{R}^n \\ \in \mathbb{R}^m \end{matrix}$$

$$\mu =: \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$\Sigma =: \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

$$\begin{aligned}
\mu_1 &= \mathcal{E}\{X_1\} \in \mathbb{R}^n \\
\mu_2 &= \mathcal{E}\{X_2\} \in \mathbb{R}^m \\
\Sigma_{11} &= \mathrm{cov}(X_1, X_1) \in \mathbb{R}^{n \times n} \\
\Sigma_{22} &= \mathrm{cov}(X_2, X_2) \in \mathbb{R}^{m \times m} \\
\Sigma_{12} &= \mathrm{cov}(X_1, X_2) \in \mathbb{R}^{n \times m} \\
\Sigma_{21} &= \mathrm{cov}(X_2, X_1) \in \mathbb{R}^{m \times n}
\end{aligned}$$

Conditioning on Gaussian Random Vectors

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \begin{matrix} \in \mathbb{R}^n \\ \in \mathbb{R}^m \end{matrix}$$

$$\mu_1 = \mathcal{E}\{X_1\} \in \mathbb{R}^n$$

$$\mu_2 = \mathcal{E}\{X_2\} \in \mathbb{R}^m$$

$$\Sigma_{11} = \mathrm{cov}(X_1, X_1) \in \mathbb{R}^{n\times n}$$

$$\mu =: \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$\Sigma_{22} = \mathrm{cov}(X_2, X_2) \in \mathbb{R}^{m\times m}$$

$$\Sigma =: \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

$$\Sigma_{12} = \mathrm{cov}(X_1, X_2) \in \mathbb{R}^{n\times m}$$

$$\Sigma_{21} = \mathrm{cov}(X_2, X_1) \in \mathbb{R}^{m\times n}$$

- Schur Complement: $\Sigma > 0$ if, and only if, $\Sigma_{22} > 0$ and $\Sigma_{11} - \Sigma_{12}(\Sigma_{22})^{-1}\Sigma_{21} > 0$.

Key Fact 1: Conditional Distributions of Gaussian Random Vectors

- Mean $\mu_{1|2} := \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$
- Covariance $\Sigma_{1|2} := \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$
- Proof: http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html

Key Fact 2: Conditional Independence

- Suppose we have 3 vectors X_1, X_2 and X_3 that are **jointly normally distributed** and X_1 and X_3 are each **independent** of X_2.

$$\Sigma = \begin{bmatrix} \Sigma_{11} & 0 & \Sigma_{13} \\ 0 & \Sigma_{22} & 0 \\ \Sigma_{13}^{\top} & 0 & \Sigma_{33} \end{bmatrix}$$

X_1 and X_2 are **conditionally independent** given X_3

Key Fact 2: Conditional Independence

- Suppose we have 3 vectors X_1, X_2 and X_3 that are **jointly normally distributed** and X_1 and X_3 are each **independent** of X_2.

$$\Sigma = \left[\begin{array}{cc|c} \Sigma_{11} & 0 & \Sigma_{13} \\ 0 & \Sigma_{22} & 0 \\ \hline \Sigma_{13}^{\top} & 0 & \Sigma_{33} \end{array}\right]$$

X_1 and X_2 are **conditionally independent** given X_3

Key Fact 2: Conditional Independence

- Suppose we have 3 vectors X_1, X_2 and X_3 that are **jointly normally distributed** and X_1 and X_3 are each **independent** of X_2.

$$\Sigma = \left[\begin{array}{cc|c} \Sigma_{11} & 0 & \Sigma_{13} \\ 0 & \Sigma_{22} & 0 \\ \hline \Sigma_{13}^\top & 0 & \Sigma_{33} \end{array}\right]$$

X_1 and X_2 are **conditionally independent** given X_3

- **Using Key Fact 1 for covariance:**

$$\left.\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right| X_3$$

$$\operatorname{cov}\left(\begin{bmatrix} X_{1|X_3} \\ X_{2|X_3} \end{bmatrix}, \begin{bmatrix} X_{1|X_3} \\ X_{2|X_3} \end{bmatrix}\right) = \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix} - \begin{bmatrix} \Sigma_{13} \\ 0 \end{bmatrix} \Sigma_{33}^{-1} \begin{bmatrix} \Sigma_{13}^\top & 0 \end{bmatrix}$$

$$= \begin{bmatrix} \Sigma_{11} - \Sigma_{13}\Sigma_{33}^{-1}\Sigma_{13}^\top & 0 \\ 0 & \Sigma_{22} \end{bmatrix}$$

Key Fact 3: Covariance of a Sum of Independent Normal Random Variables

- **Linear Combination:** $Y = AX_1 + BX_2$
- **Mean:** $\mu_Y = A\mu_1 + B\mu_2$
- **Covariance:** $\mathrm{cov}(Y, Y) = A\Sigma_{11}A^T + B\Sigma_{22}B^T.$

Key Fact 4

- Suppose that X, Y, and Z are jointly distributed random vectors with density f_{XYZ}.

$$(X|Z)\big|(Y|Z) \sim \frac{f_{(X|Z)(Y|Z)}}{f_{(Y|Z)}} = \frac{f_{XYZ}}{f_{YZ}} \sim X\Big|\begin{bmatrix} Y \\ Z \end{bmatrix}$$

Key Fact 4

- Suppose that X, Y, and Z are jointly distributed random vectors with density f_{XYZ}.

$$(X|Z)\big|(Y|Z) \sim \frac{f_{(X|Z)(Y|Z)}}{f_{(Y|Z)}} = \frac{f_{XYZ}}{f_{YZ}} \sim X\Big|\begin{bmatrix} Y \\ Z \end{bmatrix}$$

- **Proof:**

$$(X|Z)\big|(Y|Z) \sim \frac{f_{(X|Z)(Y|Z)}}{f_{(Y|Z)}} = \frac{f_{\left[\begin{array}{c} X \\ Y \end{array}\right]|Z}}{f_{Y|Z}} = \frac{\frac{f_{XYZ}}{f_Z}}{\frac{f_{YZ}}{f_Z}} = \frac{f_{XYZ}}{f_{YZ}} \sim X\Big|\begin{bmatrix} Y \\ Z \end{bmatrix}$$

Sneak Peek at Kalman Filter

Discrete-Time Kalman Filter

Model: Linear time-varying discrete-time system with "white[7]" Gaussian noise

$$
\begin{aligned}
x_{k+1} &= A_k x_k + G_k w_k, \quad x_0 \text{ initial condition} \\
y_k &= C_k x_k + v_k
\end{aligned}
$$

$x \in \mathbb{R}^n$, $w \in \mathbb{R}^p$, $y \in \mathbb{R}^m$, $v \in \mathbb{R}^m$. Moreover, the random vectors x_0, and, for $k \geq 0$, w_k, v_k are all independent[8] Gaussian (normal) random vectors.

Precise assumptions on the random vectors We'll denote $\delta_{kl} = 1 \iff k = l$ and $\delta_{kl} = 0, \quad k \neq l$.

- For all $k \geq 0,\ l \geq 0$, x_0, w_k, v_l are jointly Gaussian.
- w_k is a 0-mean white noise process: $\mathcal{E}\{w_k\} = 0$, and $\operatorname{cov}(w_k, w_l) = R_k \delta_{kl}$
- v_k is a 0-mean white noise process: $\mathcal{E}\{v_k\} = 0$, and $\operatorname{cov}(v_k, v_l) = Q_k \delta_{kl}$
- Uncorrelated noise processes: $\operatorname{cov}(w_k, v_l) = 0$
- The initial condition x_0 is uncorrelated with all other noise sequences.
- We denote the mean and covariance of x_0 by

$$
\bar{x}_0 = \mathcal{E}\{x_0\} \text{ and } P_0 = \operatorname{cov}(x_0) = \operatorname{cov}(x_0, x_0) = \mathcal{E}\{(x_0 - \bar{x}_0)(x_0 - \bar{x}_0)^\top\}
$$

Discrete-Time Kalman Filter

$$\operatorname{cov}\left(\begin{bmatrix} w_k \\ v_k \\ x_0 \end{bmatrix}, \begin{bmatrix} w_l \\ v_l \\ x_0 \end{bmatrix} \right) = \begin{bmatrix} R_k \delta_{kl} & 0 & 0 \\ 0 & Q_k \delta_{kl} & 0 \\ 0 & 0 & P_0 \end{bmatrix}, \quad \delta_{kl} = \begin{cases} 1 & k = l \\ 0 & k \neq l \end{cases}$$

Discrete-Time Kalman Filter

Definition of Terms:

$$\widehat{x}_{k|k} := \mathcal{E}\{x_k | y_0, \cdots, y_k\}$$

$$P_{k|k} := \mathcal{E}\{(x_k - \widehat{x}_{k|k})(x_k - \widehat{x}_{k|k})^\top | y_0, \cdots, y_k\}$$

$$\widehat{x}_{k+1|k} := \mathcal{E}\{x_{k+1} | y_0, \cdots, y_k\}$$

$$P_{k+1|k} := \mathcal{E}\{(x_{k+1} - \widehat{x}_{k+1|k})(x_{k+1} - \widehat{x}_{k+1|k})^\top | y_0, \cdots, y_k\}$$

Initial Conditions:

$$\widehat{x}_{0|-1} := \bar{x}_0 = \mathcal{E}\{x_0\}, \quad \text{and} \quad P_{0|-1} := P_0 = \operatorname{cov}(x_0)$$

Discrete-Time Kalman Filter

For $k \geq 0$

Measurement Update Step:

$$
\begin{aligned}
K_k &= P_{k|k-1} C_k^\top \left(C_k P_{k|k-1} C_k^\top + Q_k \right)^{-1} \quad \text{(Kalman Gain)} \\
\widehat{x}_{k|k} &= \widehat{x}_{k|k-1} + K_k \left(y_k - C_k \widehat{x}_{k|k-1} \right) \\
P_{k|k} &= P_{k|k-1} - K_k C_k P_{k|k-1}
\end{aligned}
$$

Time Update or Prediction Step:

$$
\begin{aligned}
\widehat{x}_{k+1|k} &= A_k \widehat{x}_{k|k} \\
P_{k+1|k} &= A_k P_{k|k} A_k^\top + G_k R_k G_k^\top
\end{aligned}
$$

End of For Loop (Just stated this way to emphasize the recursive nature of the filter)