Minimum Variance Estimator (MVE) & Modified Gram-Schmidt (MGS)

ROB 501
Nemcmiye Ozay
Presented by Andrew Wintenberg

Different Types of Estimators

Linear Model: $y = Cx + \varepsilon$

- y: measurements
- x: unknown quantity to estimate from measurements
- ε: error (usually due to measurement/sensor)

Estimator	Model of x	Model of ε
WLS	None	None
BLUE	None	Probabilistic
MVE	Probabilistic	Probabilistic

A probabilistic model (mean, covariance) of ε could come from testing sensors in a lab.
A probabilitic model for x could come from previous observations (e.g., to estimate the size of surrounding vehicles for autonomous-driving, we know the distribution of vehicle types).

Minimum Variance Estimator (MVE)

$$y = Cx + \varepsilon \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \end{bmatrix}$$

Assumptions

Uncorrelated

Mean $\quad \mathbb{E}(x) = 0 \qquad \mathbb{E}(\varepsilon) = 0$

Covariance $\quad \mathbb{E}(xx^T) = P \qquad \mathbb{E}(\varepsilon\varepsilon^T) = Q \qquad \overbrace{\mathbb{E}(x\varepsilon^T) = 0}^{\text{Uncorrelated}}$

$$CPC^T + Q > 0 \qquad (P \geq 0,\ Q \geq 0)$$

Objective: Find estimate $\hat{x}$ of x based on y $\quad (\hat{x} = g(y))$

minimize $\quad \mathbb{E}(\|\hat{x} - x\|_2^2)$

Assume $\mathbb{E}(\hat{x}-x)=0 \rightarrow \mathrm{Var}(\hat{x}-x) = \mathbb{E}(\|\hat{x}-x\|_2^2)$

$$\mathrm{Var}(z) = \mathbb{E}((z-\mathbb{E}(z))^2)$$

Linear Estimation $\hat{x} = Ky$ $\quad y = Cx + \varepsilon$

$$\mathbb{E}(\hat{x}-x) = \mathbb{E}(Ky - x) = \mathbb{E}(KCx + K\varepsilon - x)$$
$$= KC\,\mathbb{E}(x) + K\mathbb{E}(\varepsilon) - \mathbb{E}(x)$$
$$= 0$$

$$\mathbb{E}(\|\hat{x}-x\|_2^2) = \sum_{i=1}^{n} \mathbb{E}((\hat{x}_i - x_i)^2) \rightarrow$$ Like BLUE, separate into n optimizations.

x_i is a random variable

random variables are functions $\quad x_i : \underbrace{\Omega}_{\text{Probability Space}} \rightarrow \mathbb{R}$

Polynomials $\mathbb{P}(t)$ form an inner product space over $\mathbb{R}$

Basis $\{1, t, t^2, \dots\}$

$$\langle f, g \rangle = \int_{-1}^{1} f(t)g(t)\,dt$$

Define $\mathcal{X} = \text{span}(x_1, \dots, x_n, \varepsilon_1, \dots \varepsilon_m)$

Define for $z_1, z_2 \in \mathcal{X}$ $\langle z_1, z_2 \rangle = \mathbb{E}(z_1 \cdot z_2)$

<u>Observations</u>

i) $z_1 = \sum_{i=1}^{n} \alpha_i x_i + \sum_{j=1}^{m} \beta_j \varepsilon_j$

ii) $\mathbb{E}(z_1) = \sum_{i=1}^{n} \alpha_i \underbrace{\mathbb{E}(x_i)}_{0} + \sum_{j=1}^{m} \beta_j \underbrace{\mathbb{E}(\varepsilon_j)}_{0} = 0$

iii) $\text{Var}(z_1) = \mathbb{E}((z_1 - \underbrace{\mathbb{E}(z_1)}_{0})^2) = \mathbb{E}(z_1^2) = \langle z_1, z_1 \rangle$
$= \|z_1\|^2$

Remember $\hat{x} = Ky$

$y = Cx + \varepsilon$

$C = \begin{bmatrix} C_1 \\ \vdots \\ \vdots \\ C_m \end{bmatrix}$

$y_i = C_i \cdot x + \varepsilon_i$

$y_i \in \mathcal{X}$

Define $\mathcal{M} = \text{span}(y_1 \dots y_m)$

$\hat{x} = Ky \iff \hat{x}_i = K_i y \iff \hat{x}_i \in \mathcal{M}$

$K = \begin{bmatrix} K_1 \\ \vdots \\ K_n \end{bmatrix}$ $\left(\hat{x}_i = \sum_{j=1}^{m} K_{ij} y_j \right)$

Original

$$\min \mathbb{E}(\|\hat{x}-x\|_2^2)$$
$$\text{st} \quad \hat{x} = Ky$$
for some K

$$\longleftrightarrow$$

For each $i \in \{1 \dots n\}$

$$\min \|\hat{x}_i - x_i\|^2$$
$$\text{s.t.} \quad \hat{x}_i \in M$$

$\hat{x}_i$ is a linear function of $Y_1 \cdots Y_m$

$$\mathbb{E}((\hat{x}_i - x_i)\cdot(\hat{x}_i - x_i))$$

$$\hat{x}_i - x_i \perp M \implies \hat{x}_i = P_m(x_i)$$

Normal equations: $\hat{x}_i = \sum_{j=1}^{m} \alpha_j \cdot Y_j$ st. $G\alpha = \beta$

$$= \alpha^T Y \qquad \alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix}$$

$$G_{jk} = \langle Y_j, Y_k \rangle = \mathbb{E}(Y_j Y_k) \qquad Y_j = C_j \cdot x + \varepsilon_j$$

$$= \mathbb{E}((C_j x + \varepsilon_j)(C_k x + \varepsilon_k))$$

$a \in \mathbb{R}$ $a = a^T$

$$= \mathbb{E}((C_j x)(C_k x) + (C_j x)\varepsilon_k + \varepsilon_j(C_k x) + \varepsilon_k \varepsilon_j)$$

$(AB)^T = B^T A^T$

$$= \mathbb{E}((C_j x)\cdot(x^T C_k^T) + (C_j x)\varepsilon_k + \varepsilon_j(C_k x) + \varepsilon_j \varepsilon_k)$$

$$= C_j \mathbb{E}(x x^T) C_k^T + C_j \mathbb{E}(x \varepsilon_k) + C_k \mathbb{E}(x \varepsilon_j) + \mathbb{E}(\varepsilon_j \varepsilon_k)$$

$$= C_j \ P \ C_k^T + 0 + 0 + Q_{jk}$$

$$x\varepsilon_k = \begin{bmatrix} x_1\varepsilon_k \\ \vdots \\ x_n\varepsilon_k \end{bmatrix} \qquad Q = \mathbb{E}(\varepsilon\varepsilon^T)$$

$$G_{jk} = C_j P C_k^T + Q_{jk} \qquad G = \underline{CPC^T + Q}$$

$$C^T = \begin{bmatrix} c_1^T \cdots c_n^T \end{bmatrix}$$

From assumption - $G > 0$

$$G = \mathbb{E}(YY^T)$$

$$B_j = \langle y_j, x_i \rangle = \mathbb{E}(y_j x_i) \qquad (y_j = C_j x + \varepsilon_j)$$

$$= \mathbb{E}((C_j x + \varepsilon_j) x_i)$$

$$= \mathbb{E}(C_j x x_i + x_i \varepsilon_j) \qquad x x_i = \begin{bmatrix} x_1 x_i \\ \vdots \\ x_n x_i \end{bmatrix}$$

$$= C_j \mathbb{E}(x x_i) + \mathbb{E}(x_i \varepsilon_j)$$

$$= C_j \cdot P_i + 0$$

$$P = \mathbb{E}(xx^T)$$

$$B = C \cdot P_i$$

$$P = [P_1 \cdots P_n]$$

$$P_i = \mathbb{E}(x \cdot x_i)$$

To find $\hat{x}_i$

$$\hat{x}_i = \alpha^T y \quad \text{where } G\alpha = \beta \qquad G > 0$$

$$G = CPC^T + Q \qquad \beta = CP_i$$

$$\alpha = G^{-1} \cdot \beta \Rightarrow \hat{x}_i = \beta^T G^{-1^T} \cdot y$$

$P_i^T \to i^{th}$ row of P^T

$$\hat{x}_i = \underbrace{P_i^T C^T}_{\beta^T} \underbrace{(CPC^T + Q)^{-1}}_{G^{-1}} y$$

$$\hat{x} = \begin{bmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_n \end{bmatrix} = \underbrace{P C^T (CPC + Q)^{-1}}_{K} y$$

$$\hat{x} = Ky$$

MVE

MVE Remarks

1. Exercise: $\overbrace{\mathbb{E}\{(\hat{x}-x)(\hat{x}-x)^{\intercal}\}}^{\text{Posterior uncertainty in } x} = \overset{\text{Prior uncertainty in } x}{P} - \overbrace{PC^{\intercal}\left[CPC^{\intercal}+Q\right]^{-1}CP}^{\text{Reduction in uncertainty due to } y}$
2. The term $PC^{\intercal}\left[CPC^{\intercal}+Q\right]^{-1}CP$ represents the "value" of the measurements. It is the reduction in the variance of x given the measurements y.
3. If $Q>0$ and $P>0$, then from the Matrix Inversion Lemma

$$\boxed{\hat{x} = Ky = \left[C^{\intercal}Q^{-1}C + P^{-1}\right]^{-1} C^{\intercal}Q^{-1}y\,.}$$

This form is useful for comparing BLUE and MVE

BLUE vs. MVE

Assuming $Q > 0$ and $P > 0$,

- **BLUE**: $\hat{x} = \left[C^\intercal Q^{-1} C\right]^{-1} C^\intercal Q^{-1} \, y$
- **MVE**: $\hat{x} = \left[C^\intercal Q^{-1} C + P^{-1}\right]^{-1} C^\intercal Q^{-1} \, y$
- Hence, MVE $\to$ BLUE as $P^{-1} \to 0$.
- Roughly, $P^{-1} = 0$ occurs when $P = \infty I$, i.e., there is infinite covariance in x or we have no idea how x is distributed.
- For BLUE to exist, we need $\dim(y) \geq \dim(x)$
- For MVE to exist, we can have $\dim(y) < \dim(x)$ as long as

$$(CPC^\intercal + Q) > 0$$

Solution To Exercise

To find $\mathbb{E}\{(\hat{x}-x)(\hat{x}-x)^{\mathsf{T}}\}$, note that

$$\hat{x}-x = Ky - x = KCx + K\varepsilon - x = (KC-I)x + K\varepsilon\,,$$

and thus

$$(\hat{x}-x)(\hat{x}-x)^{\mathsf{T}} = (KC-I)xx^{\mathsf{T}}(KC-I)^{\mathsf{T}} + K\varepsilon\varepsilon^{\mathsf{T}}K^{\mathsf{T}} - 2(KC-I)x\varepsilon^{\mathsf{T}}K^{\mathsf{T}}\,.$$

Taking expectations and recalling that x and ε are uncorrelated

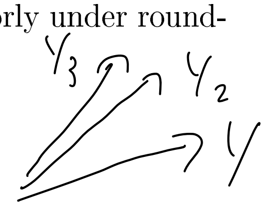
$$\begin{aligned}
\mathbb{E}\{(\hat{x}-x)(\hat{x}-x)^{\mathsf{T}}\} &= (KC-I)P(KC-I)^{\mathsf{T}} + KQK^{\mathsf{T}} \\
&= KCPC^{\mathsf{T}}K^{\mathsf{T}} + P - 2PC^{\mathsf{T}}K^{\mathsf{T}} + KQK^{\mathsf{T}} \\
&= P + K\left[CPC^{\mathsf{T}} + Q\right]K^{\mathsf{T}} - 2PC^{\mathsf{T}}K^{\mathsf{T}}\,.
\end{aligned}$$

Substituting $K = PC^{\mathsf{T}}\left[CPC^{\mathsf{T}} + Q\right]^{-1}$ yields

$$\mathbb{E}\{(\hat{x}-x)(\hat{x}-x)^{\mathsf{T}}\} = P - PC^{\mathsf{T}}\left[CPC^{\mathsf{T}} + Q\right]^{-1}CP$$

Gram Schmidt vs Modified Gram Schmidt

We have been using the classical Gram-Schmidt Algorithm. It behaves poorly under round-off error. Here is a standard example:



$$y^1 = \begin{bmatrix} 1 \\ \varepsilon \\ 0 \\ 0 \end{bmatrix}, y^2 = \begin{bmatrix} 1 \\ 0 \\ \varepsilon \\ 0 \end{bmatrix}, y^3 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \varepsilon \end{bmatrix}, \varepsilon > 0$$

Let $\{e^1, e^2, e^3, e^4\}$ be the standard basis vectors $\left(\text{Yes, } (e^i_j) = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} \right)$

We note that

$\langle y_1, y_2 \rangle = 1$

$$y^2 = y^1 + \varepsilon(e^3 - e^2)$$
$$y^3 = y^2 + \varepsilon(e^4 - e^3)$$

$\langle y_1, e^3 - e^2 \rangle = \begin{bmatrix} 1 & \varepsilon & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} = -\varepsilon$

and thus

$$\text{span}\{y^1, y^2\} = \text{span}\{y^1, (e^3 - e^2)\}$$
$$\text{span}\{y^1, y^2, y^3\} = \text{span}\{y^1, (e^3 - e^2), (e^4 - e^3)\}$$

Hence, GS applied to $\{y^1, y^2, y^3\}$ and $\{y^1, (e^3 - e^2), (e^4 - e^3)\}$ should produce the same orthonormal vectors. To check this, we go to MATLAB, and for $\varepsilon = 0.1$, we do indeed get the same results. You can verify this yourself. **However, with $\varepsilon = 10^{-8}$,**

$$\|Q_1 - Q_2\| = 0.5$$

where $Q_1 = [v^1, v^2, v^3]$ computed with Classical-GS for $\{y^1, y^2, y^3\}$ while $Q_2 = [v^1, v^2, v^3]$ computed with Classical-GS for $\{y^1, (e^3 - e^2), (e^4 - e^3)\}$. Hence we do NOT get the same result!

Classical Gram Schmidt Algorithm With Normalization: Initial data $\{y^1, \cdots, y^n\}$ linearly independent. Here, it is written slightly differently than in lecture:

$\langle y_1, y_1 \rangle = 1 + \varepsilon^2$

For $k = 1 : n$
    $v^k = y^k$
    For $j = 1 : k - 1$
        $v^k = v^k - \langle y^k, v^j \rangle v^j$
    End
    $v^k = \frac{v^k}{\|v^k\|}$
End

$v^k = P_{v^1 \cdots v^{k-1}}(y^k)$

$Q_1 = [v^1, v^2, v^3]$ computed with Classical-GS for $\{y^1, y^2, y^3\}$ while $Q_2 = [v^1, v^2, v^3]$ computed with Classical-GS for $\{y^1, (e^3 - e^2), (e^4 - e^3)\}$. R_1 shows that indeed, $\{y^1, y^2, y^3\}$ is 'nearly' linearly dependent while R_2 shows that $\{y^1, (e^3 - e^2), (e^4 - e^3)\}$ is 'quite' linearly independent.

```
>> DemoGramSchmidtProcess
Caluclations with Classical or Standard Gram Schmidt
Epsilon = 1e-08

Q1 =
```

q^2 q^3

$\langle q^2, q^3 \rangle = \frac{1}{2}$

```
    1.0000         0         0
    0.0000   -0.7071   -0.7071
         0    0.7071         0
         0         0    0.7071
```

$\frac{1}{\sqrt{2}}$

```
R1 =

    1.0000    1.0000    1.0000
         0    0.0000         0
         0         0    0.0000

Q2 =
```

q^2 q^3

$\langle q^2, q^3 \rangle = 0$

```
    1.0000    0.0000    0.0000
    0.0000   -0.7071   -0.4082
         0    0.7071   -0.4082
         0         0    0.8165

R2 =

    1.0000   -0.0000         0
         0    1.4142   -0.7071
         0         0    1.2247

norm(Q1-Q2)

ans =

    0.5176
```

There is a modification of the Gram Schmidt Algorithm that is much better for actual calculations. You do want to know about this! For your Final Exam, you **do not** have to know the Modified-GS Algorithm itself. **All you have to know for your Final Exam is that a Modified Gram Schmidt Algorithm exists and it provides better numerical results.**

Modified Gram Schmidt

For $k = 1 : n$
    $v^k = y^k$
End
For $i = 1 : n$
    $v^i = \frac{v^i}{\|v^i\|}$
    For $j = i + 1 : n$
        $v^j = v^j - \langle v^j, v^i \rangle v^i$
    End
End

$$\begin{bmatrix} y^1 & & \\ y^2 & y^2 - P_{y^1}(y^2) & \\ y^3 & y^3 - P_{y^1}(y^3) & \ddots \\ & & \\ y^n & y^n - P_{y^1}(y^2 & \end{bmatrix}$$

The demo code below in Canvas in the MATLAB folder

```
a=1e-8;
y1=[1 a 0 0]';
y2=[1 0 a 0]';
y3=[1 0 0 a]';

e1=[1 0 0 0]';
e2=[0 1 0 0]';
e3=[0 0 1 0]';
e4=[0 0 0 1]';

Y=[y1 y2 y3];

%Y=rand(4,4);

[Q1,R1]=GramSchmidtClassic(Y),  % Q1'*Q1-eye(3),

[Q2, R2] = GramSchmidtClassic([y1,-e2+e3,-e3+e4]),
```

```
disp('norm(Q1-Q2)')
norm(Q1-Q2)

pause


[Q3,R3]=GramSchmidtModified(Y),

[Q4,R4]=GramSchmidtModified([y1,-e2+e3,-e3+e4]),

disp('norm(Q3-Q4)')
norm(Q3-Q4)

pause


[Q5,R5]=GramSchmidtModified_MIT(Y),

[Q6,R6]=GramSchmidtModified_MIT([y1,-e2+e3,-e3+e4]),

disp('norm(Q5-Q6)')
norm(Q5-Q6)
```

$Q_3 = [v^1, v^2, v^3]$ computed with Modified-GS for $\{y^1, y^2, y^3\}$ while $Q_4 = [v^1, v^2, v^3]$ computed with Modified-GS for $\{y^1, (e^3 - e^2), (e^4 - e^3)\}$. R_3 shows that indeed, $\{y^1, y^2, y^3\}$ is 'nearly' linearly dependent while R_4 shows that $\{y^1, (e^3-e^2), (e^4-e^3)\}$ is 'quite' linearly independent.

```
Calculations with Modified Gram Schmidt
Epsilon = 1e-08

Q3 =

    1.0000         0         0
    0.0000   -0.7071   -0.4082
         0    0.7071   -0.4082
         0         0    0.8165


R3 =

    1.0000    1.0000    1.0000
         0    0.0000         0
         0         0    0.0000


Q4 =

    1.0000    0.0000    0.0000
    0.0000   -0.7071   -0.4082
         0    0.7071   -0.4082
         0         0    0.8165


R4 =

    1.0000   -0.0000         0
         0    1.4142   -0.7071
         0         0    1.2247

norm(Q3-Q4)

ans =

   8.1650e-09
```

Two GS Algorithms

Assume: $\{y^1, \ldots, y^n\}$ linearly independent

Classical Gram Schmidt

For $k = 1 : n$
    $v^k = y^k$
    For $j = 1 : k-1$
        $v^k = v^k - \langle y^k, v^j \rangle v^j$
    End
    $v^k = \frac{v^k}{\|v^k\|}$
End

Modified Gram Schmidt

For $k = 1 : n$
    $v^k = y^k$
End
For $i = 1 : n$
    $v^i = \frac{v^i}{\|v^i\|}$
    For $j = i+1 : n$
        $v^j = v^j - \langle v^j, v^i \rangle v^i$
    End
End

Comparison (not on any exam)

(a) Let $P_M(x)$ denote the orthogonal projection of x onto a subspace M.

(b) Classical GS: $v^1 = y^1$, and for $k \geq 2$, $v^k = y^k - P_M(y^k)$, where $M = \text{span}\{y^1, \cdots, y^{k-1}\} = \text{span}\{v^1, \cdots, v^{k-1}\}$ (optional: add in the normalization step)

(c) Modified GS:

- $v^1 = y^1$, and for $k \geq 2$, $\tilde{y}^k = y^k - P_M(y^k)$, where $M = \text{span}\{v^1\}$ (optional: add in the normalization step)
- $v^2 = \tilde{y}^2$, and for $k \geq 3$, $\tilde{y}^k = \tilde{y}^k - P_M(\tilde{y}^k)$, where $M = \text{span}\{v^2\}$ (optional: add in the normalization step)
- $v^3 = \tilde{y}^3$, and for $k \geq 4$, $\tilde{y}^k = \tilde{y}^k - P_M(\tilde{y}^k)$, where $M = \text{span}\{v^3\}$ (optional: add in the normalization step)
- etc.

You can learn more about this on the web.