Lecture 2

- Linear regression
- The least squares method
- Properties of the (deterministic) least squares method
- BLUE
- Computational aspects

Linear Regression

SI procedure: Collect data, *choose a model class*, *find the best model in the model class*, validation.

- Linear regression models. Models that are linearly parametrized.
  - Computationally simple.
  - Simple to implement.
  - Low memory consumption.
  - Common in signal processing. Ex. Echo cancellation.
- Original work by Gauss 1809.
- Starting point of system identification.

Linear Regression Cont'd

Model structure ($\mathcal{M}$):

$$y_m(t) = \boldsymbol{\varphi}^T(t)\boldsymbol{\theta}, \quad t = 1, \ldots, N \tag{1}$$

where $y_m(t)$ is the model output, $\boldsymbol{\varphi}(t) \in \mathbb{R}^{n\times 1}$ is a vector of known quantities and $\boldsymbol{\theta} \in \mathbb{R}^{n\times 1}$ is a vector of unknown quantities.

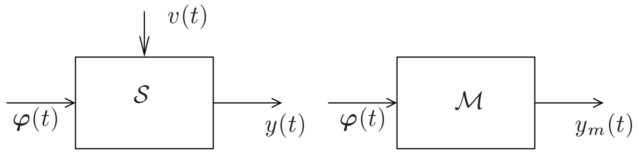
The model (1) can be compactly written as

$$\boldsymbol{Y}_m = \boldsymbol{\Phi}\boldsymbol{\theta}, \quad \boldsymbol{Y}_m = \begin{bmatrix} y_m(1) \\ \vdots \\ y_m(N) \end{bmatrix}, \boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\varphi}^T(1) \\ \vdots \\ \boldsymbol{\varphi}^T(N) \end{bmatrix} \tag{2}$$

- Linear regression can be used also for certain non-linear models.

Linear Regression Cont'd

Problem: Find an estimate of $\boldsymbol{\theta}$ given measurement $y(1), \boldsymbol{\varphi}(1), \ldots, y(N), \boldsymbol{\varphi}(N)$.



- Noiseless case ($v = 0$, $\mathcal{M} = \mathcal{S}$). Exact solution exists.
- What to do when noise $v(t)$ is present and $\mathcal{M} \neq \mathcal{S}$?

Least Squares (Optimization)

Introduce the equation error

$$\varepsilon(t) = y(t) - y_m(t) = y(t) - \boldsymbol{\varphi}^T(t)\boldsymbol{\theta}, \quad t = 1, \ldots, N$$

or compactly

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{Y}_m = \boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{\theta}$$

Least squares method: Choose $\boldsymbol{\theta}$ such that $\varepsilon^2(t)$ is small for all t:

$$\hat{\boldsymbol{\theta}}_{LS} = \arg\min_{\boldsymbol{\theta}} V(\boldsymbol{\theta}), \quad V(\boldsymbol{\theta}) = \frac{1}{2}\sum_{t=1}^{N}\varepsilon^2(t) = \frac{1}{2}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$$

Results: Assume that $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$ is invertible, then

$$\hat{\boldsymbol{\theta}}_{LS} = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\boldsymbol{Y} = \left(\sum_{t=1}^{N}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^T(t)\right)^{-1}\sum_{t=1}^{N}\boldsymbol{\varphi}(t)y(t)$$

Weighted least squares estimate:

$$\hat{\boldsymbol{\theta}}_{WLS} = \arg\min_{\boldsymbol{\theta}} V(\boldsymbol{\theta}), \quad V(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\varepsilon}^T\boldsymbol{W}\boldsymbol{\varepsilon}$$

$$\Rightarrow \quad \hat{\boldsymbol{\theta}}_{WLS} = \left(\boldsymbol{\Phi}^T\boldsymbol{W}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\boldsymbol{W}\boldsymbol{Y}$$

where $\boldsymbol{W}$ is symmetric ($\boldsymbol{W}^T = \boldsymbol{W}$) and positive definite.

Rem: $\boldsymbol{W} = \boldsymbol{I} \Rightarrow \hat{\boldsymbol{\theta}}_{WLS} = \hat{\boldsymbol{\theta}}_{LS}$.

Least Squares (Geometric Approach)

Model: $\boldsymbol{Y}_m = \boldsymbol{\Phi}\boldsymbol{\theta} = \sum_{i=1}^{n}\boldsymbol{\Phi}_i\theta_i$, where $\boldsymbol{\Phi} = \begin{bmatrix}\boldsymbol{\Phi}_1 & \cdots & \boldsymbol{\Phi}_n\end{bmatrix}$.

Measurements: $\boldsymbol{Y}$.

$\boldsymbol{Y}$ and $\boldsymbol{\Phi}_i$ are vectors in the vector space $\mathbb{R}^{N\times 1}$.

Objective: Find a linear combination of the vectors $\boldsymbol{\Phi}_i$, $i = 1, \ldots, n$ ($\boldsymbol{Y}_m$), that approximates $\boldsymbol{Y}$ as well as possible.

Solution: $\{\Phi_i\}_{i=1}^n$ span an n-dimensional subspace D_n. The best approximation of $\boldsymbol{Y}$ in D_n is the orthogonal projection of $\boldsymbol{Y}$ on D_n.

- Define the inner product: $<\boldsymbol{x}, \boldsymbol{y}> = \boldsymbol{x}^T\boldsymbol{y}$.
- The approximation error $\boldsymbol{Y} - \boldsymbol{Y}_m$ is orthogonal to $\boldsymbol{\Phi}_i$, $i = 1, \ldots, n$

  $$<\boldsymbol{\Phi}_i, \boldsymbol{Y} - \boldsymbol{Y}_m> = \boldsymbol{\Phi}_i^T(\boldsymbol{Y} - \boldsymbol{Y}_m) = 0, \quad i = 1, \ldots, m$$

  Consequently,

  $$\boldsymbol{\Phi}^T(\boldsymbol{Y} - \boldsymbol{Y}_m) = \boldsymbol{0}$$

- Estimated model: $\hat{\boldsymbol{Y}}_m = \boldsymbol{\Phi}\hat{\boldsymbol{\theta}}$ implies that

  $$\boldsymbol{\Phi}^T(\boldsymbol{Y} - \boldsymbol{\Phi}\hat{\boldsymbol{\theta}}) = 0 \quad \Rightarrow \hat{\boldsymbol{\theta}} = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\boldsymbol{Y} = \hat{\boldsymbol{\theta}}_{LS}$$

Rem: Using the scalar product $<\boldsymbol{x}, \boldsymbol{y}> = \boldsymbol{x}^T\boldsymbol{W}\boldsymbol{y}$ yields the weighted least squares estimate.

Least Squares (Statistical Properties)

To explore the properties of the least squares estimate we need to specify the system, *i.e.*, we need to make some assumptions about the generating data.

Assumptions:

- $\boldsymbol{\varphi}(t)$ is deterministic and known. (Quite restrictive assumption!)
- System: $y(t) = \boldsymbol{\varphi}^T(t)\boldsymbol{\theta}_0 + e(t)$, where $e(t)$ is a sequence of random variables, $\mathrm{E}\, e(t) = 0$ and $\mathrm{E}\, e(t)e(s) = R_{ts}$. Compactly written as

$$\boldsymbol{Y} = \boldsymbol{\Phi}\boldsymbol{\theta}_0 + \boldsymbol{e}, \qquad \mathrm{E}\, \boldsymbol{e} = \boldsymbol{0}, \quad \mathrm{E}\, \boldsymbol{e}\boldsymbol{e}^T = \boldsymbol{R}$$

Rem: If $\boldsymbol{R} = \lambda^2\boldsymbol{I}$ then $e(t)$ is white noise with variance λ^2.

Least Squares (Statistical Properties) - Results

- The (weighted) least squares estimate is *unbiased*:
  - $\mathrm{E}\, \hat{\boldsymbol{\theta}}_{WLS} = \boldsymbol{\theta}_0$
- Covariance matrix, $\operatorname{cov} \hat{\boldsymbol{\theta}} = \mathrm{E}\, (\hat{\boldsymbol{\theta}} - \mathrm{E}\, \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \mathrm{E}\, \hat{\boldsymbol{\theta}})^T$:
  - $\operatorname{cov} \hat{\boldsymbol{\theta}}_{WLS} = [\boldsymbol{\Phi}^T\boldsymbol{W}\boldsymbol{\Phi}]^{-1}\boldsymbol{\Phi}^T\boldsymbol{W}\boldsymbol{R}\boldsymbol{W}\boldsymbol{\Phi}[\boldsymbol{\Phi}^T\boldsymbol{W}\boldsymbol{\Phi}]^{-1}$
  - $\operatorname{cov} \hat{\boldsymbol{\theta}}_{LS} = [\boldsymbol{\Phi}^T\boldsymbol{\Phi}]^{-1}\boldsymbol{\Phi}^T\boldsymbol{R}\boldsymbol{\Phi}[\boldsymbol{\Phi}^T\boldsymbol{\Phi}]^{-1}$
  - $\boldsymbol{R} = \lambda^2\boldsymbol{I} \Rightarrow$
    $\operatorname{cov} \hat{\boldsymbol{\theta}}_{LS} = \frac{\lambda^2}{N}[\frac{1}{N}\boldsymbol{\Phi}^T\boldsymbol{\Phi}]^{-1} = \frac{\lambda^2}{N}\left[\frac{1}{N}\sum_{t=1}^{N}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^T(t)\right]^{-1}$
- If $e(t)$ is Gaussian distributed $e(t) \sim N(0, \boldsymbol{R})$, $\boldsymbol{\Phi}$ deterministic, then $\hat{\boldsymbol{\theta}}_{WLS} \sim N(\boldsymbol{\theta}_0, \operatorname{cov} \hat{\boldsymbol{\theta}}_{WLS})$. (Holds for finite N)
- $\hat{\boldsymbol{\theta}}_{WLS}$ is very often consistent: $\hat{\boldsymbol{\theta}}_{WLS} \to \boldsymbol{\theta}_0$, $N \to \infty$.

Def: The estimate $\hat{\boldsymbol{\theta}}_1$ is statistically more efficient than $\hat{\boldsymbol{\theta}}_2$ if

$$\operatorname{cov} \hat{\boldsymbol{\theta}}_1 \leq \operatorname{cov} \hat{\boldsymbol{\theta}}_2$$

Question: Which choice of $\boldsymbol{W}$ will minimize $\operatorname{cov} \hat{\boldsymbol{\theta}}_{WLS}$?

Result: The choice $\boldsymbol{W} = \boldsymbol{R}^{-1}$ yields optimal accuracy:

- $\hat{\boldsymbol{\theta}}_{WLS} = \left(\boldsymbol{\Phi}^T\boldsymbol{R}^{-1}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\boldsymbol{R}^{-1}\boldsymbol{Y}$
- $\operatorname{cov} \hat{\boldsymbol{\theta}}_{WLS} = [\boldsymbol{\Phi}^T\boldsymbol{R}^{-1}\boldsymbol{\Phi}]^{-1}$

In this case $\hat{\boldsymbol{\theta}}_{WLS}$ is known as the BLUE (best linear unbiased estimator) or the Gauss-Markov estimate.

BLUE

- BLUE = Best Linear Unbiased Estimator.
- White noise, $\boldsymbol{R} = \lambda^2\boldsymbol{I}$. BLUE yields the same estimate as the unweighted least squares method.
- If $e(t)$ is Gaussian, then BLUE yields the best possible estimate! If $e(t)$ is non-Gaussian, then there might exist better non-linear estimates.
- BLUE can be derived also for singular $\boldsymbol{R}$.

Computational Aspects

The least squares solution ($\mathbf{\Phi} \in \mathbb{R}^{N \times n}$)

- $\hat{\boldsymbol{\theta}}_{LS} = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\boldsymbol{Y} = \left(\sum_{t=1}^{N}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^T(t)\right)^{-1}\sum_{t=1}^{N}\boldsymbol{\varphi}(t)y(t)$

is unsuitable for numerical implementation.

Alternatives: Avoid the inverse!

- The normal equations: $(\mathbf{\Phi}^T\mathbf{\Phi})\hat{\boldsymbol{\theta}}_{LS} = \mathbf{\Phi}^T\boldsymbol{Y}$.
- Solve an overdetermined linear system of equations: $\boldsymbol{Y} = \mathbf{\Phi}\hat{\boldsymbol{\theta}}_{LS}$. (Recall that $\boldsymbol{Y} - \boldsymbol{Y}_m = \boldsymbol{Y} - \mathbf{\Phi}\boldsymbol{\theta}$ should be small.)
  - QR factorizations
  - SVD factorizations

QR factorization: Let $\mathbf{\Phi} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q} \in \mathbb{R}^{N \times N}$ is orthogonal ($\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$) and $\boldsymbol{R} \in \mathbb{R}^{N \times n}$ is upper triangular. Then, instead of solving

$$\boldsymbol{Y} = \mathbf{\Phi}\boldsymbol{\theta}$$

we can equally well solve

$$\boldsymbol{Q}^T\boldsymbol{Y} = \boldsymbol{Q}^T\mathbf{\Phi}\boldsymbol{\theta} = \boldsymbol{R}\boldsymbol{\theta}$$

which is easy due to the structure of $\boldsymbol{R}$.

- Requires more computations than solving the normal equations.
- Less sensitive to rounding errors.

Rem: MATLAB: $\hat{\boldsymbol{\theta}} = \mathbf{\Phi}\backslash\boldsymbol{Y}$

Conclusions

- Regression models describes a large class of dynamic systems (linear w.r.t the parameters).
- The least squares method is fundamental in system identification, and can be derived from various starting points.
- We have assumed that $\mathbf{\Phi}$ is a known and deterministic matrix. Problems when this matrix is a function of $u(t)$ and $y(t)$ (ex. ARX-model).