Least Squares Methods

• Overdetermined linear equations

- $\mathbf{y} = \mathbf{A}\mathbf{x}$ where $\mathbf{A} \in \mathbf{R}^{m \times n}$ and $m > n$
- More equations than unknowns
- Cannot solve for **x** in most cases.

• Least squares solution of overdetermined linear equations

- Residual or error is $\mathbf{r} = \mathbf{A}\mathbf{x} - \mathbf{y}$.
- Find $\mathbf{x}_{ls} = \arg\min_{\mathbf{x} \in \mathbf{R}^n} \|\mathbf{r}\|^2$.
- $\mathbf{A}\mathbf{x}_{ls} \in R(\mathbf{A})$ and $\mathbf{A}\mathbf{x}_{ls}$ is closest to **y**.
- $\mathbf{A}\mathbf{x}_{ls}$ is projection of **y** on $R(\mathbf{A})$.
- Assume **A** is full rank. From

$$\frac{d\|\mathbf{r}\|^2}{d\mathbf{x}} = \frac{d}{d\mathbf{x}}\left(\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{y}^T\mathbf{A}\mathbf{x} + \mathbf{y}^T\mathbf{y}\right) = 2\mathbf{x}^T\mathbf{A}^T\mathbf{A} - 2\mathbf{y}^T\mathbf{A} = 0,$$

$$\mathbf{x}_{ls} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{y} = \mathbf{A}^+\mathbf{y}$$

- The optimal residual is $\mathbf{r}^* = \mathbf{A}\mathbf{x}_{ls} - \mathbf{y} = \left(\mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T - \mathbf{I}\right)\mathbf{y}$.
- Pseudo-inverse $\mathbf{A}^+$ is a left inverse of **A**; $\mathbf{A}^+\mathbf{A} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{A} = \mathbf{I}$
- Projection of **y** on $R(\mathbf{A})$ is $\mathbf{A}\mathbf{x}_{ls} = \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{y} = \mathbf{P}_\mathbf{A}\mathbf{y}$ where projection matrix is

  $\mathbf{P}_\mathbf{A} = \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T$.
- Orthogonality principle: $\mathbf{r} \perp R(\mathbf{A})$, i.e., $\forall \mathbf{A}\mathbf{z} \in R(\mathbf{A})$,

$$\langle \mathbf{r}, \mathbf{A}\mathbf{z} \rangle = \mathbf{y}^T\left(\mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T - \mathbf{I}\right)^T\mathbf{A}\mathbf{z} = 0$$

• Least squares estimation

- Model: $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v}$

  (1) **x** is what we want to estimate.

  (2) **y** is sensor measurements.

  (3) **v** is unknown noise or measurement error

- Least squares estimate, $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{y}\|^2 = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{y}$

  (1) If $\mathbf{v} = \mathbf{0}$, $\hat{\mathbf{x}} = \mathbf{x}$. $\Rightarrow$ unbiased
  (2) Linear estimator, i.e., $\hat{\mathbf{x}} = \mathbf{By}$ for some $\mathbf{B}$
  (3) $\hat{\mathbf{x}}$ is the *best linear unbiased estimator (BLUE)*.

• Least squares data fitting

- Preparation

  (1) Functions $f_1, \cdots, f_n : S \to \mathbf{R}$, $S \subseteq \mathbf{R}^n$: basis functions or regressors

  (2) $\{(\mathbf{s}_i, g_i)\}_{i=1}^{m}$ with $\mathbf{s}_i \in S$ and $m > n$: data or measurements

- Problem: find a linear combination of functions that fits data, i.e.,

$$\sum_{j=1}^{n} x_j f_j(\mathbf{s}_i) \approx g_i \quad \text{for } i = 1, 2, \ldots, m$$

- Least squares solution: $\mathbf{x}^* = \arg\min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{g}\|^2 = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{g}$

  (1) $\mathbf{x} = [x_1 \cdots x_n]^T$

  (2) $\mathbf{g} = [g_1 \cdots g_m]^T$

  (3) $\mathbf{A} = \left[a_{ij}\right] = \left[f_j(\mathbf{s}_i)\right] \in \mathbf{R}^{m \times n}$

- Fitting function: $f_{ls}(\mathbf{s}) = x_1 f_1(\mathbf{s}) + \cdots + x_n f_n(\mathbf{s})$
- Choosing the value of n: plot $\|\mathbf{r}\| = \|\mathbf{Ax} - \mathbf{g}\|$ as a function of n
- Least squares polynomial fitting for $S \subseteq \mathbf{R}$

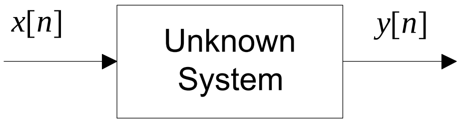
  (1) $f_j(s) = s^{j-1}$

  (2) $a_{ij} = s_i^{j-1}$ (Vandermonde matrix)

- Applications
  (1) Interpolation, extrapolation, smoothing of data
  (2) Simple and approximate model of data

• Least squares system identification

- Problem: find a model for unknown system from input-output data



- Assume a model: for example, MA(n)

$$y[n]=\sum_{i=0}^{n} w_i x[n-i]$$

- Collect input-output data with $N > (n+1)$

(1) $$\mathbf{Aw}=\begin{bmatrix} x[n] & x[n-1] & \cdots & x[0] \\ x[n+1] & x[n] & \cdots & x[1] \\ \vdots & \vdots & \ddots & \vdots \\ x[n+N-1] & x[n+N-2] & \cdots & x[N-1] \end{bmatrix}\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}$$

(2) $$\mathbf{y}=\begin{bmatrix} y[n] \\ y[n+1] \\ \vdots \\ y[n+N-1] \end{bmatrix}$$

- Least squares solution: $\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\mathbf{Aw}-\mathbf{y}\|^2 = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{y}$

- Choosing the order of the model (n)
  - (1) Larger n
    - (a) Small error for a particular set of data (smaller bias)
    - (b) Try to fit even noise (overfit or overmodeling)
    - (c) Poor generalization and poor predictive ability (larger variance)
  - (2) Cross-validation: use different set of data and monitor the change of error on this data set

• Multi-objective least squares

- Two (competing) objectives

(1) $J_1 = \|\mathbf{Ax}-\mathbf{y}\|^2$

(2) $J_2 = \|\mathbf{Bx}-\mathbf{z}\|^2$

- Weighted sum objective

(1) $J = J_1 + \mu J_2 = \|\mathbf{Ax}-\mathbf{y}\|^2 + \mu\|\mathbf{Bx}-\mathbf{z}\|^2$ with $\mu \ge 0$

(2) $\|\mathbf{Ax}-\mathbf{y}\|^2+\mu\|\mathbf{Bx}-\mathbf{z}\|^2=\left\|\begin{bmatrix}\mathbf{A}\\ \sqrt{\mu}\mathbf{B}\end{bmatrix}\mathbf{x}-\begin{bmatrix}\mathbf{y}\\ \sqrt{\mu}\mathbf{z}\end{bmatrix}\right\|^2=\|\tilde{\mathbf{A}}\mathbf{x}-\tilde{\mathbf{y}}\|^2$

- Least squares solution: $\hat{\mathbf{x}}=\left(\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\right)^{-1}\tilde{\mathbf{A}}^T\tilde{\mathbf{y}}=\left(\mathbf{A}^T\mathbf{A}+\mu\mathbf{B}^T\mathbf{B}\right)^{-1}\left(\mathbf{A}^T\mathbf{y}+\mu\mathbf{B}^T\mathbf{z}\right)$

• Regularized least squares

- With $\mathbf{B}=\mathbf{I}$ and $\mathbf{z}=\mathbf{0}$, $J=J_1+\mu J_2=\|\mathbf{Ax}-\mathbf{y}\|^2+\mu\|\mathbf{x}\|^2$

(1) There is a penalty for large $\mathbf{x}$.
(2) It stabilizes the algorithm.
(3) It works for any $\mathbf{A}$.
(4) Tikhonov regularization

- Least squares solution: $\hat{\mathbf{x}}=\left(\mathbf{A}^T\mathbf{A}+\mu\mathbf{I}\right)^{-1}\mathbf{A}^T\mathbf{y}$

(1) Useful when $\mathbf{A}^T\mathbf{A}$ is ill-conditioned.
(2) Useful when we know $\mathbf{x}$ is small.
(3) Useful when model is accurate only for small $\mathbf{x}$.

• Underdetermined linear equations

- $\mathbf{y}=\mathbf{Ax}$ where $\mathbf{A}\in\mathbf{R}^{m\times n}$ with $m<n$
- More unknowns than equations
- $\mathbf{x}$ is under specified and there are many solutions.
- Assuming rank($\mathbf{A}$) = n, for each $\mathbf{y}\in\mathbf{R}^m$

(1) Set of solutions is $\{\mathbf{x}:\mathbf{Ax}=\mathbf{y}\}=\{\mathbf{x}+\mathbf{z}:\mathbf{Ax}=\mathbf{y},\mathbf{z}\in N(\mathbf{A})\}$

(2) Solution has $\dim N(\mathbf{A})=n-m$ degree of freedom
(3) What is the best solution?

• Minimum norm solution

- $\mathbf{y}=\mathbf{Ax}$ where $\mathbf{A}\in\mathbf{R}^{m\times n}$ with $m<n$ and rank($\mathbf{A}$) = n

- Minimum norm solution: $\tilde{\mathbf{x}}=\mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}\mathbf{y}$

- For any $\mathbf{x}$ such that $\mathbf{Ax}=\mathbf{y}$,

(1) $\|\mathbf{x}-\tilde{\mathbf{x}},\tilde{\mathbf{x}}\|=(\mathbf{x}-\tilde{\mathbf{x}})^T\tilde{\mathbf{x}}=0$, i.e., $(\mathbf{x}-\tilde{\mathbf{x}})\perp\tilde{\mathbf{x}}$

(2) $\|\mathbf{x}\|^2 \geq \|\tilde{\mathbf{x}}\|^2$

- Orthogonality condition: $\tilde{\mathbf{x}} \perp N(\mathbf{A})$

- $\tilde{\mathbf{x}}$ is projection of $\mathbf{0}$ on the solution set $\{\mathbf{x} : \mathbf{Ax} = \mathbf{y}\} = \{\mathbf{x} + \mathbf{z} : \mathbf{Ax} = \mathbf{y}, \mathbf{z} \in N(\mathbf{A})\}$

- $\mathbf{A}^+ = \mathbf{A}^T\left(\mathbf{AA}^T\right)^{-1}$ is pseudo-inverse of $\mathbf{A}$

- Derivation

(1) Formulation: $\min_{\mathbf{x}} \mathbf{x}^T\mathbf{x}$ subject to $\mathbf{Ax} = \mathbf{y}$

(2) Lagrange multiplier, $L(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{x}^T\mathbf{x} + \boldsymbol{\lambda}^T\left(\mathbf{Ax} - \mathbf{y}\right)$

(3) Optimality conditions

(a) $\dfrac{\partial L}{\partial \mathbf{x}} = 2\mathbf{x}^T + \boldsymbol{\lambda}^T\mathbf{A} = \mathbf{0} \;\Rightarrow\; \mathbf{x} = -\dfrac{\mathbf{A}^T\boldsymbol{\lambda}}{2}$

(b) $\dfrac{\partial L}{\partial \boldsymbol{\lambda}} = \left(\mathbf{Ax} - \mathbf{y}\right)^T = \mathbf{0} \;\Rightarrow\; \boldsymbol{\lambda} = -2\left(\mathbf{AA}^T\right)^{-1}\mathbf{y}$

(4) $\tilde{\mathbf{x}} = \mathbf{A}^T\left(\mathbf{AA}^T\right)^{-1}\mathbf{y}$

- Comparing regularized least squares solution,

$$\left(\mathbf{A}^T\mathbf{A} + \mu\mathbf{I}\right)^{-1}\mathbf{A}^T \xrightarrow[\mu \to 0]{} \mathbf{A}^T\left(\mathbf{AA}^T\right)^{-1}$$

Least Squares Estimation

- **Linear model:** $\mathbf{y}=\mathbf{x}+\mathbf{n}$ and $\mathbf{x}=\mathbf{H}\boldsymbol{\theta}=\left[\mathbf{h}_1,\mathbf{h}_2,\cdots,\mathbf{h}_p\right]\boldsymbol{\theta}=\sum_{i=1}^{p}\theta_i\mathbf{h}_i$ or $x_i=\tilde{\mathbf{h}}_i^T\boldsymbol{\theta}$

- $\mathbf{y}=\left[y_1y_2\cdots y_N\right]^T\in\mathbf{R}^N$: measurements, known

- $\mathbf{x}=\left[x_1x_2\cdots x_N\right]^T\in\mathbf{R}^N$: model output, signal component, unknown

- $\mathbf{H}=\begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1p} \\ h_{21} & h_{22} & \cdots & h_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N1} & h_{N2} & \cdots & h_{Np}\end{bmatrix}\in\mathbf{R}^{N\times p}$: model structure, assumed to be known

- $\boldsymbol{\theta}=\left[\theta_1\theta_2\cdots\theta_p\right]^T\in\mathbf{R}^p$: model parameter, unknown

- $\mathbf{n}=\left[n_1n_2\cdots n_N\right]^T\in\mathbf{R}^N$: measurement error, noise component, unknown or known statistics

- **Least squares solution**

- $\mathbf{y}=\mathbf{H}\boldsymbol{\theta}+\mathbf{n}$ and $e^2=\mathbf{n}^T\mathbf{n}=\left(\mathbf{y}-\mathbf{H}\boldsymbol{\theta}\right)^T\left(\mathbf{y}-\mathbf{H}\boldsymbol{\theta}\right)=tr\left[\left(\mathbf{y}-\mathbf{H}\boldsymbol{\theta}\right)\left(\mathbf{y}-\mathbf{H}\boldsymbol{\theta}\right)^T\right]$

- $\frac{\partial}{\partial\boldsymbol{\theta}}e^2=2\mathbf{H}^T\left(\mathbf{y}-\mathbf{H}\boldsymbol{\theta}\right)$ and $\frac{\partial^2}{\partial\boldsymbol{\theta}^2}e^2=2\mathbf{H}^T\mathbf{H}\geq 0 \;\Rightarrow\; \hat{\boldsymbol{\theta}}=\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y}=\mathbf{G}_{\mathbf{H}}^{-1}\mathbf{H}^T\mathbf{y}$

- Normal equation: $\mathbf{H}^T\mathbf{H}\hat{\boldsymbol{\theta}}=\mathbf{H}^T\mathbf{y}$ or $\mathbf{G}_{\mathbf{H}}\hat{\boldsymbol{\theta}}=\mathbf{H}^T\mathbf{y}$

- Gram matrix or Grammian, $\mathbf{G}_{\mathbf{H}}=\mathbf{H}^T\mathbf{H}$ is nonsingular iff $\{\mathbf{h}_i\}_{i=1}^p$ are linearly independent.

- Let $\langle\mathbf{H}\rangle=\text{span}\{\mathbf{h}_i\}_{i=1}^p$ and assume $\langle\mathbf{H}\rangle\oplus\langle\mathbf{A}\rangle=\mathbf{R}^N$ and $\langle\mathbf{H}\rangle\perp\langle\mathbf{A}\rangle$

- Projections

  (a) $\hat{\mathbf{x}}=\mathbf{H}\hat{\boldsymbol{\theta}}=\mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y}=\mathbf{P}_{\mathbf{H}}\mathbf{y}$, $\mathbf{P}_{\mathbf{H}}=\mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T$: projection onto $\langle\mathbf{H}\rangle$

(b) $\hat{\mathbf{n}} = \mathbf{y} - \hat{\mathbf{x}} = (\mathbf{I}_N - \mathbf{P_H})\mathbf{y} = \mathbf{P_A y}$, $\mathbf{P_A} = \mathbf{I}_N - \mathbf{P_H}$: orthogonal onto $\langle \mathbf{A} \rangle$

(c) $\mathbf{P_H}^T = \mathbf{P_H}$ and $\mathbf{P_A}^T = \mathbf{P_A}$: symmetric

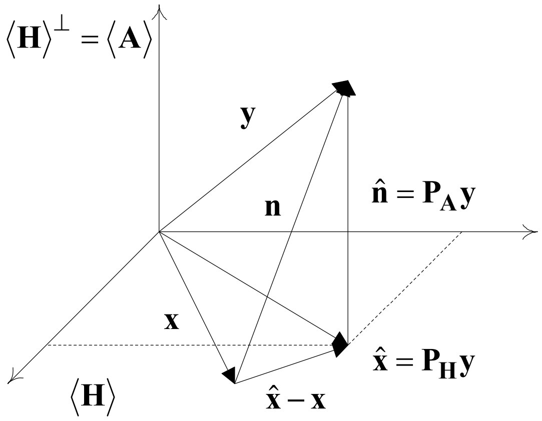
(d) $\mathbf{P_H}^2 = \mathbf{P_H P_H} = \mathbf{P_H}$ and $\mathbf{P_A}^2 = \mathbf{P_A P_A} = \mathbf{P_A}$: idempotent

(e) $\mathbf{P_H P_A} = \mathbf{P_A P_H} = \mathbf{0}$: orthogonal

(f) $\mathbf{P_H} + \mathbf{P_A} = \mathbf{I}_N$: decomposition of identity

(g) Also note that $\mathbf{P_H H} = \mathbf{H}$, $\mathbf{P_A H} = \mathbf{0}$, $\mathbf{P_H x} = \mathbf{x}, \forall \mathbf{x} \in \langle \mathbf{H} \rangle$, $\mathbf{P_A x} = \mathbf{0}, \forall \mathbf{x} \in \langle \mathbf{H} \rangle$

• **Geometry**



- $\langle \mathbf{H} \rangle = \text{span}\{\mathbf{h}_i\}_{i=1}^{p}$: signal subspace with $\mathbf{P_H} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$

- $\langle \mathbf{A} \rangle = \text{span}\{\mathbf{a}_i\}_{i=1}^{N-p}$: orthogonal subspace with $\mathbf{P_A} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$

  (a) Construct $\mathbf{A} = [\mathbf{a}_1 \mathbf{a}_2 \cdots \mathbf{a}_{N-p}] \in \mathbf{R}^{N\times(N-p)}$ so that

  (b) $\mathbf{a}_i^T \mathbf{h}_j = 0, \forall i = 1, \cdots, (N-p), \forall j = 1, \cdots, p \Leftrightarrow \mathbf{A}^T\mathbf{H} = \mathbf{0}$

- $\mathbf{I}_N = \mathbf{P_H} + \mathbf{P_A} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T + \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$

- $\forall \mathbf{y} \in \mathbf{R}^N, \mathbf{y} = \mathbf{Iy} = \mathbf{P_H y} + \mathbf{P_A y} = \hat{\mathbf{x}} + \hat{\mathbf{n}}$

  (a) $\hat{\mathbf{x}} \in \langle \mathbf{H} \rangle, \hat{\mathbf{x}} = \mathbf{P_H y} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y} = \mathbf{H}\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}$

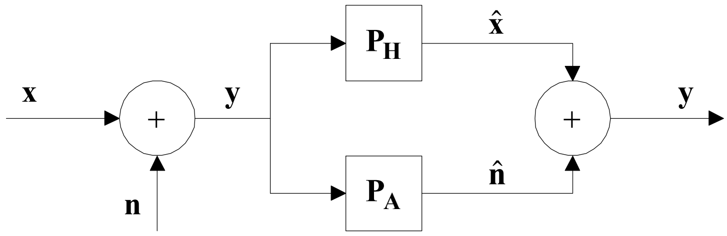
  (b) $\hat{\mathbf{n}} \in \langle \mathbf{A} \rangle, \hat{\mathbf{n}} = \mathbf{P_A y} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} = \mathbf{A}\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\phi}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$

• Orthogonality

- $\forall \mathbf{y} \in \mathbf{R}^N, \mathbf{y} = \mathbf{I}\mathbf{y} = \mathbf{P_H}\mathbf{y} + \mathbf{P_A}\mathbf{y} = \hat{\mathbf{x}} + \hat{\mathbf{n}} \;\Rightarrow\; \hat{\mathbf{n}}^T\hat{\mathbf{x}} = \mathbf{y}^T\mathbf{P_A}\mathbf{P_H}\mathbf{y} = 0$
- $\hat{\mathbf{n}}^T\hat{\mathbf{n}} = \mathbf{y}^T\mathbf{P_A}\mathbf{y} = \mathbf{y}^T\left(\mathbf{I}_N - \mathbf{P_H}\right)\mathbf{y} = \mathbf{y}^T\mathbf{y} - \hat{\mathbf{x}}^T\hat{\mathbf{x}} \;\Rightarrow\; \mathbf{y}^T\mathbf{y} = \hat{\mathbf{x}}^T\hat{\mathbf{x}} + \hat{\mathbf{n}}^T\hat{\mathbf{n}}$
- $\mathbf{n} = \hat{\mathbf{n}} + \left(\hat{\mathbf{x}} - \mathbf{x}\right)$: orthogonal decomposition of $\mathbf{n}$

(a)
$$\begin{aligned}\mathbf{n}^T\mathbf{n} &= \left(\mathbf{y} - \mathbf{x} + \hat{\mathbf{x}} - \hat{\mathbf{x}}\right)^T\left(\mathbf{y} - \mathbf{x} + \hat{\mathbf{x}} - \hat{\mathbf{x}}\right) = \left[\hat{\mathbf{n}} - \left(\mathbf{x} - \hat{\mathbf{x}}\right)\right]^T\left[\hat{\mathbf{n}} - \left(\mathbf{x} - \hat{\mathbf{x}}\right)\right] \\ &= \hat{\mathbf{n}}^T\hat{\mathbf{n}} + \left(\mathbf{x} - \hat{\mathbf{x}}\right)^T\left(\mathbf{x} - \hat{\mathbf{x}}\right) \ge \hat{\mathbf{n}}^T\hat{\mathbf{n}}\end{aligned}$$

(b) $\hat{\mathbf{n}} = \mathbf{P_A}\mathbf{y} = \mathbf{y} - \hat{\mathbf{x}}$: minimum norm, i.e., least squares



• Example: complex exponential modes analysis

- Model: $x(t) = \sum_{i=1}^{p} \theta_i e^{j\omega_i t} = \theta_1 e^{j\omega_1 t} + \theta_2 e^{j\omega_2 t} + \cdots \theta_p e^{j\omega_p t}$, i.e., sum of p complex exponentials
- From $y(t) = x(t) + n(t)$, take N measurements at $t = kT$, k = 0, 1, 2, …, (N – 1). WLOG, assume T = 1.
- $$\mathbf{y} = \begin{bmatrix} y(0) \\ y(1) \\ y(2) \\ \vdots \\ y(N-1) \end{bmatrix} = \begin{bmatrix} \theta_1 + \theta_2 + \cdots \theta_p \\ \theta_1 e^{j\omega_1} + \theta_2 e^{j\omega_2} + \cdots \theta_p e^{j\omega_p} \\ \theta_1 e^{j\omega_1 2} + \theta_2 e^{j2\omega_2} + \cdots \theta_p e^{j2\omega_p} \\ \vdots \\ \theta_1 e^{j(N-1)\omega_1} + \theta_2 e^{j(N-1)\omega_2} + \cdots \theta_p e^{j(N-1)\omega_p} \end{bmatrix} + \begin{bmatrix} n(0) \\ n(1) \\ n(2) \\ \vdots \\ n(N-1) \end{bmatrix}, \text{ or}$$

$$\mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ e^{j\omega_1} & e^{j\omega_2} & e^{j\omega_3} & \cdots & e^{j\omega_p} \\ e^{j2\omega_1} & e^{j2\omega_2} & e^{j2\omega_3} & \cdots & e^{j2\omega_p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e^{j(N-1)\omega_1} & e^{j(N-1)\omega_2} & e^{j(N-1)\omega_3} & \cdots & e^{j(N-1)\omega_p} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \vdots \\ \theta_p \end{bmatrix} + \begin{bmatrix} n(0) \\ n(1) \\ n(2) \\ \vdots \\ n(N-1) \end{bmatrix}, \text{ or}$$

$$\mathbf{y} = \mathbf{x} + \mathbf{n} = \mathbf{H}\boldsymbol{\theta} + \mathbf{n}$$

- Least squares solution: $\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y}$ and $\hat{\mathbf{x}} = \mathbf{H}\hat{\boldsymbol{\theta}} = \mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y} = \mathbf{P}_\mathbf{H}\mathbf{y}$

• Example: polynomial curve fitting

- Model: $x(t) = \sum_{i=1}^{p} \theta_i t^{i-1} = \theta_1 + \theta_2 t + \theta_3 t^2 + \cdots \theta_p t^{p-1}$, i.e., polynomial
- From $y(t) = x(t) + n(t)$, take N measurements at $t = kT$, $k = 1, 2, \ldots, N$. WLOG, assume $T = 1$.
- $$\mathbf{y} = \begin{bmatrix} y(1) \\ y(2) \\ y(3) \\ \vdots \\ y(N) \end{bmatrix} = \begin{bmatrix} \theta_1 + \theta_2 + \cdots \theta_p \\ \theta_1 + \theta_2 2 + \cdots \theta_p 2^{p-1} \\ \theta_1 + \theta_2 3 + \cdots \theta_p 3^{p-1} \\ \vdots \\ \theta_1 + \theta_2 2 + \cdots \theta_p N^{p-1} \end{bmatrix} + \begin{bmatrix} n(1) \\ n(2) \\ n(3) \\ \vdots \\ n(N) \end{bmatrix}, \text{ or}$$

$$\mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 2^2 & \cdots & 2^{p-1} \\ 1 & 3 & 3^2 & \cdots & 3^{p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & N & N^2 & \cdots & N^{p-1} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \vdots \\ \theta_p \end{bmatrix} + \begin{bmatrix} n(1) \\ n(2) \\ n(3) \\ \vdots \\ n(N) \end{bmatrix}, \text{ or } \mathbf{y} = \mathbf{x} + \mathbf{n} = \mathbf{H}\boldsymbol{\theta} + \mathbf{n}$$

- Least squares solution: $\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y}$ and $\hat{\mathbf{x}} = \mathbf{H}\hat{\boldsymbol{\theta}} = \mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y} = \mathbf{P}_\mathbf{H}\mathbf{y}$

• Recursive least squares (RLS)

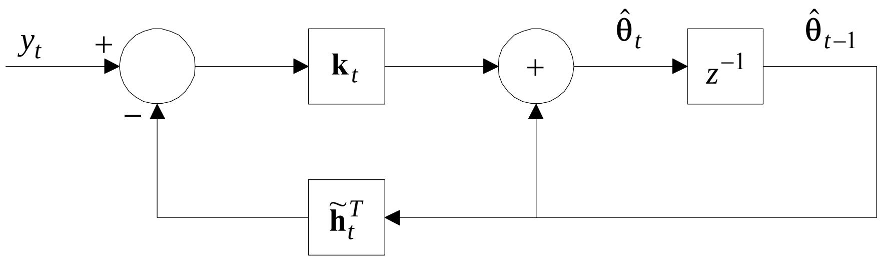
- $$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{n} \Leftrightarrow \begin{bmatrix} \mathbf{y}_{t-1} \\ \hline y_t \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{t-1} \\ \hline \tilde{\mathbf{h}}_t^T \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} + \begin{bmatrix} \mathbf{n}_{t-1} \\ \hline n_t \end{bmatrix} \Leftrightarrow \begin{cases} \mathbf{y}_{t-1} = \mathbf{H}_{t-1}\boldsymbol{\theta} + \mathbf{n}_{t-1} \\ y_t = \tilde{\mathbf{h}}_t^T\boldsymbol{\theta} + n_t \end{cases}$$
- $\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{n} \Leftrightarrow y_t = \tilde{\mathbf{h}}_t^T\boldsymbol{\theta} + n_t$ for $t = 1, 2, 3, \ldots$
- Define $\mathbf{P}_t^{-1} = \mathbf{G}_{\mathbf{H}_t} = \mathbf{H}_t^T\mathbf{H}_t = \mathbf{H}_{t-1}^T\mathbf{H}_{t-1} + \tilde{\mathbf{h}}_t\tilde{\mathbf{h}}_t^T = \mathbf{P}_{t-1}^{-1} + \tilde{\mathbf{h}}_t\tilde{\mathbf{h}}_t^T$.
- $$\mathbf{H}_t^T\mathbf{y}_t = \begin{bmatrix} \mathbf{H}_{t-1} \\ \hline \tilde{\mathbf{h}}_t^T \end{bmatrix}^T \begin{bmatrix} \mathbf{y}_{t-1} \\ \hline y_t \end{bmatrix} = \mathbf{H}_{t-1}^T\mathbf{y}_{t-1} + \tilde{\mathbf{h}}_t y_t$$

- LS solution is $\hat{\theta}_t = \mathbf{P}_t \mathbf{H}_t^T \mathbf{y}_t = \left(\mathbf{P}_{t-1}^{-1} + \tilde{\mathbf{h}}_t \tilde{\mathbf{h}}_t^T\right)^{-1} \left(\mathbf{H}_{t-1}^T \mathbf{y}_{t-1} + \tilde{\mathbf{h}}_t y_t\right)$.

- From matrix inversion lemma, $\left(\mathbf{P}_{t-1}^{-1} + \tilde{\mathbf{h}}_t \tilde{\mathbf{h}}_t^T\right)^{-1} = \mathbf{P}_{t-1} - \gamma_t \mathbf{P}_{t-1} \tilde{\mathbf{h}}_t \tilde{\mathbf{h}}_t^T \mathbf{P}_{t-1}$ and $\gamma_t^{-1} = 1 + \tilde{\mathbf{h}}_t^T \mathbf{P}_{t-1} \tilde{\mathbf{h}}_t$ or $\gamma_t \tilde{\mathbf{h}}_t^T \mathbf{P}_{t-1} \tilde{\mathbf{h}}_t = 1 - \gamma_t$

- $\hat{\theta}_t = \left(\mathbf{P}_{t-1} - \gamma_t \mathbf{P}_{t-1} \tilde{\mathbf{h}}_t \tilde{\mathbf{h}}_t^T \mathbf{P}_{t-1}\right)\left(\mathbf{H}_{t-1}^T \mathbf{y}_{t-1} + \tilde{\mathbf{h}}_t y_t\right) = \hat{\theta}_{t-1} + \gamma_t \mathbf{P}_{t-1} \tilde{\mathbf{h}}_t \left(y_t - \tilde{\mathbf{h}}_t^T \hat{\theta}_{t-1}\right)$

- Define $\mathbf{k}_t = \gamma_t \mathbf{P}_{t-1} \mathbf{h}_t$



- Summary of RLS

  (1) Initialization: $\mathbf{P}_0 = \mathbf{I}_p$ and $\hat{\theta}_0 = \mathbf{0}_p$

  (2) $\gamma_t^{-1} = 1 + \tilde{\mathbf{h}}_t^T \mathbf{P}_{t-1} \tilde{\mathbf{h}}_t$

  (3) $\mathbf{k}_t = \gamma_t \mathbf{P}_{t-1} \tilde{\mathbf{h}}_t$

  (4) $\hat{\theta}_t = \hat{\theta}_{t-1} + \mathbf{k}_t \left(y_t - \tilde{\mathbf{h}}_t^T \hat{\theta}_{t-1}\right)$

  (5) $\mathbf{P}_t = \mathbf{P}_{t-1} - \gamma_t \mathbf{P}_{t-1} \tilde{\mathbf{h}}_t \tilde{\mathbf{h}}_t^T \mathbf{P}_{t-1}$

• Weighted least squares

- Problem: $\min_{\theta} (\mathbf{y} - \mathbf{H}\theta)^T \mathbf{W} (\mathbf{y} - \mathbf{H}\theta)$ with nonsingular symmetric $\mathbf{W} \in \mathbf{R}^{N\times N}$

- $\dfrac{\partial}{\partial \theta} = 0 \;\Rightarrow\; \mathbf{H}^T \mathbf{W}(\mathbf{y} - \mathbf{H}\theta) = \mathbf{0}$'

- Solution: $\hat{\theta} = \left(\mathbf{H}^T \mathbf{W} \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{W} \mathbf{y}$

- Choice of **W**

(1) If $\mathbf{n} : \mathrm{N}\left[\mathbf{0}, \mathbf{R}\right]$, $\mathbf{W} = \mathbf{R}^{-1}$.

(2) $\mathbf{W} = \mathrm{diag}\left(w_1, \cdots, w_N\right)$ where $w_i = \dfrac{1}{SNR_i}$.