2.5 Linear MMSE Estimation

- Not all estimators are linear.

  Not all mmse estimators are linear.

  Requiring the estimator to be linear may cost accuracy.

  But... combining linear estimation with mmse criterion has interesting properties:
  - the linear coefficients and resulting variances depend only on second order statistics
  - for Gaussian statistics, linear is optimum
    - mmse
    - MAP

- For Gaussian, we already have the estimator

$$\underline{\mu} = BD^{-1}\underline{x}$$

  What's left to discover?

  We'll derive linear mmse estimators for any statistics, Gaussian or not. More in Section 3 of the course, where we go through Haykin Chap 5

2.5.1 Start with scalars

$$\hat{\Theta} = h^* x \quad , \quad e = \Theta - \hat{\Theta} \qquad \text{assume zero means}$$

minimize $\sigma_e^2 = E\left[(\Theta - h^* x)(\Theta^* - h x^*)\right]$

differentiate wrt h

$$\frac{d\sigma_e^2}{dh}$$

not analytic ??

- Generalized derivative wrt complex variable

$$\frac{d}{dz} = \frac{1}{2}\left(\frac{\partial}{\partial u} - j\frac{\partial}{\partial v}\right) \qquad z = u + jv$$

Try some obvious tests

$$\frac{dz}{dz} = \frac{1}{2}\left(\frac{\partial z}{\partial u} - j\frac{\partial z}{\partial v}\right) = \frac{1}{2}(1 - j\cdot j) = 1$$

$$\frac{dz^2}{dz} = \frac{1}{2}\left(\frac{\partial z^2}{\partial u} - j\frac{\partial z^2}{\partial v}\right) = \frac{1}{2}\left(2z\frac{\partial z}{\partial u} - j2z\frac{\partial z}{\partial v}\right)$$

$$= 2z$$

$$\frac{dz^*}{dz} = \frac{1}{2}\left(\frac{\partial z^*}{\partial u} - j\frac{\partial z^*}{\partial v}\right) = \frac{1}{2}(1 - j(-j)) = 0$$

Consistent with conventional derivative for any analytic function (C-R conditions $\frac{\partial f_r}{\partial u} = \frac{\partial f_i}{\partial v}$, $\frac{\partial f_r}{\partial v} = -\frac{\partial f_i}{\partial u}$)

- Back to estimation:

$$\frac{d\sigma_e^2}{dh} = -E\left[(\theta - h^* x)x^*\right] = -\sigma_{\theta x}^2 + h^*\sigma_x^2$$

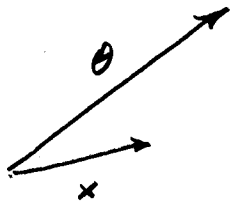
$$= -E[ex^*] = -\sigma_{ex}^2$$

This is the derivative, or gradient. Set it to zero.

$$h^* = \frac{\sigma_{\theta x}^2}{\sigma_x^2} \qquad \sigma_{ex}^2 = 0 \quad \text{or the orthogonality condition}$$

$$\hat{\theta} = \frac{\sigma_{\theta x}^2}{\sigma_x^2} x$$

- Geometric illustration



2.5.2 Multidimensional $\underline{x}$

$\hat{\theta} = \underline{h}^+ \underline{x}$ $\quad e = \theta - \underline{h}^+ \underline{x}$

$\sigma_e^2 = E\left[(\theta - \underline{h}^+ \underline{x})(\theta^* - \underline{x}^+ \underline{h})\right] = \sigma_\theta^2 - \underline{h}^+ \underline{p} - \underline{p}^+ \underline{h} + \underline{h}^+ R_x \underline{h}$

where $\underline{p} = E[\theta^* \underline{x}]$, $R_x = E[\underline{x}\,\underline{x}^+]$

Several components, so take derivative wrt $\underline{h}$

$$\frac{d\sigma_e^2}{d\underline{h}} = \begin{bmatrix} \frac{\partial \sigma_e^2}{\partial h_1} \\ \vdots \\ \frac{\partial \sigma_e^2}{\partial h_M} \end{bmatrix}$$

Read conventions in Haykin Appendix B carefully

$= - E\left[(\theta - \underline{h}^+ \underline{x})\underline{x}^*\right] = - \underline{p}^* + R_x^* \underline{h}^*$

$= - E[e \underline{x}^*] = - \sigma_{e\underline{x}}^2$

conjugating, setting to zero

$\left(\frac{d\sigma_e^2}{d\underline{h}}\right)^* = \underline{0} \Rightarrow R_x \underline{h} = \underline{p}$ $\quad \sigma_{\underline{x}e}^2 = \underline{0}$

normal equations $\quad$ orthogonality conditions

Equivalent conditions for optimum:

- The normal equations with covariance matrix of basis of estimate $\hat{\theta}$, and covariance vector of basis of estimate and quantity being estimated
- The estimation error is uncorrelated with (orthogonal to) basis of estimate

Minimized error:

$$\sigma_e^2 = \sigma_\theta^2 - \underline{h}^+\underline{p} - \underline{p}^+\underline{h} + \underline{h}^+ R_x \underline{h} \quad \longleftarrow \quad \underline{h} = R_x^{-1}\underline{p}$$

$$= \sigma_\theta^2 - \underline{p}^+ R_x^{-1}\underline{p} - \underline{p}^+ R_x^{-1}\underline{p} + \underline{p}^+ R_x^{-1} R_x R_x^{-1}\underline{p}$$

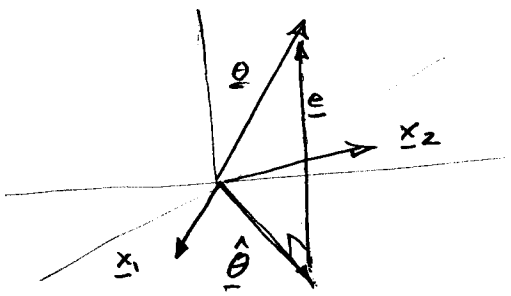
$$= \sigma_\theta^2 - \underline{p}^+ R_x^{-1}\underline{p}$$

$$= \sigma_\theta^2 - \sigma_{\hat\theta}^2 \quad \text{since } \sigma_{\hat\theta}^2 = E[|\hat\Theta|^2] = E[\underline{h}^+ \underline{x}\underline{x}^+ \underline{h}] = \underline{p}^+ R_x^{-1}\underline{p}$$

Relate to earlier $\underline{\mu} = BD^{-1}\underline{x}$ for Gaussian:

$$\begin{bmatrix}\theta \\ \underline{x}\end{bmatrix} \quad R = \left[\begin{array}{c|c} \sigma_\theta^2 & \underline{p}^+ \\ \hline \underline{p} & R_x \end{array}\right] \qquad \hat\Theta = \underline{h}^+\underline{x} = \underline{p}^+ R_x^{-1}\underline{x} = BD^{-1}\underline{x}$$

- Geometric view of multidimensional



→ $\underline{x}_1, \underline{x}_2$ shown in plane

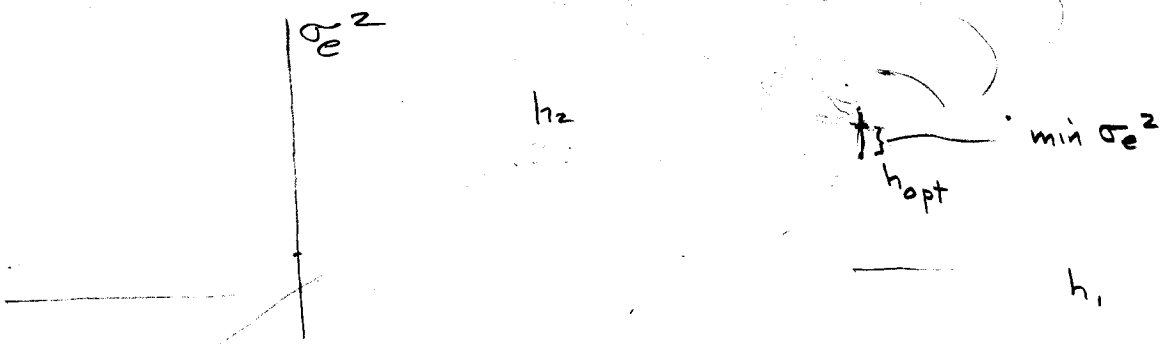
→ and $\hat{\underline{\theta}}$ lies in subspace spanned by them – a projection of $\hat\Theta$ onto subspace

→ the error $\underline{e}$ is orthogonal to the basis of estimate $\{\underline{x}_1, \underline{x}_2\}$, to $\hat{\underline{\theta}}$ and to the entire subspace

→ by Pythagoras $\sigma_e^2 = \sigma_\theta^2 - \sigma_{\hat\theta}^2$

The mse is a quadratic function of $\underline{h}$ (easier to sketch if real)

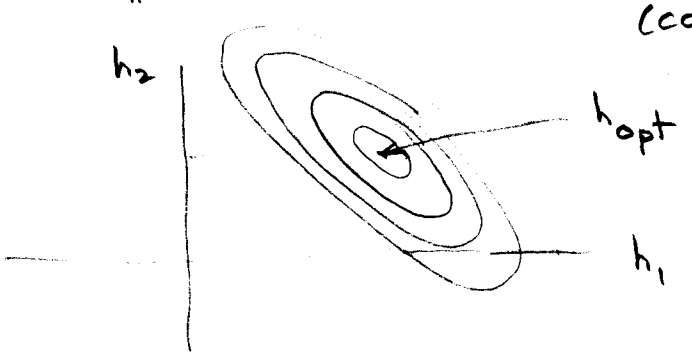
$$\sigma_e^2 = \sigma_\theta^2 - 2\underline{p}^T\underline{h} + \underline{h}^T R_x \underline{h}$$



We have the gradient of the error surface already

$$\nabla_h \sigma_e^2 = -\underline{p} + R_x \underline{h}$$

so can do steepest-descent (conjugate gradient) iteration.



top view, contour lines. eigenvectors, eigenvalues?

- A convenient representation:

$$\underline{h} = \underline{h}_{opt} + \underline{\epsilon} = R_x^{-1}\underline{p} + \underline{\epsilon}$$

Then

$$\sigma_e^2 = \sigma_\theta^2 - (\underline{p}^+ R_x^{-1} + \underline{\epsilon}^+)\underline{p} - \underline{p}^+(R_x^{-1}\underline{p} + \underline{\epsilon}) + (\underline{p}^+ R_x^{-1} + \underline{\epsilon}^+) R_x (R_x^{-1}\underline{p} + \underline{\epsilon})$$

$$= \sigma_\theta^2 - \underline{p}^+ R_x^{-1}\underline{p} + \underline{\epsilon}^+ R_x \underline{\epsilon}$$

$$= \sigma_{e\,min}^2 + \underline{\epsilon}^+ R_x \underline{\epsilon}$$ very simple!

2.5.3 Multidimensional $\underline{\theta}$ and $\underline{x}$

$\hat{\underline{\theta}} = H^+ \underline{x}$ so $H = [\underline{h}_1 | \underline{h}_2 | \cdots | \underline{h}_K]$ $\underline{h}_i$ is column of coeffs to est θ_i

$$\underline{e} = \underline{\theta} - H^+ \underline{x}$$

minimize $J = E[\underline{e}^+ \underline{e}] = \sum_{i=1}^{K} E[|e_i|^2] = \sum_{i=1}^{K} \sigma_{e_i}^2$

The $\sigma_{e_i}^2$ are all non-negative, no opportunity for trade off among them, so decomposes to K separate optimizations

$$\min_{\underline{h}_i} \sigma_{e_i}^2 \Rightarrow \underline{h}_i = R_x^{-1}\underline{p}_i \quad \text{where } \underline{p}_i = E[\theta_i^* \underline{x}]$$

$$P = [\underline{p}_1 | \cdots | \underline{p}_K] = E[\underline{x}\,\underline{\theta}^+]$$

Therefore $H = [\underline{h}_1 | \cdots | \underline{h}_K] = R_x^{-1}[\underline{p}_1 | \underline{p}_2 | \cdots | \underline{p}_K]$

$= R_x^{-1} P$

Relate to earlier $\underline{\mu} = BD^{-1}\underline{x}$ for Gaussian

$$\begin{bmatrix} \underline{\theta} \\ \underline{x} \end{bmatrix} \quad R = \left[\begin{array}{c|c} R_\theta & P^+ \\ \hline P & R_x \end{array}\right] \qquad \underline{\hat{\theta}} = H^+\underline{x} = P^+ R_x^{-1}\underline{x} = BD^{-1}\underline{x}$$

projection of entire subspace spanned by $\{\theta_i\}$ onto subspace spanned by $\{x_k\}$

- Note mse can be written

$$J = E[\underline{e}^+\underline{e}] = E[(\underline{\theta}^+ - \underline{x}^+H)(\underline{\theta} - H^+\underline{x})] = \sum_{i=1}^{K} \sigma_{e_i}^2$$

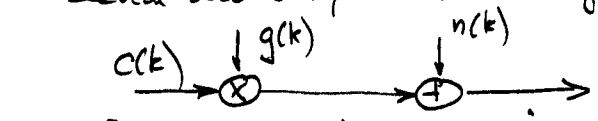
$$= E[\underline{\theta}^+\underline{\theta}] - E[\underline{x}^+H\underline{\theta}] - E[\underline{\theta}^+H^+\underline{x}] + E[\underline{x}^+HH^+\underline{x}]$$

$$= tr[R_\theta] - tr[P^+H] - tr[H^+P] + tr[H^+R_x H]$$

$$= tr[R_\theta - P^+H - H^+P + H^+R_x H]$$

2.5.4 Example: Pilot Symbol Assisted Modulation

Send data symbols through a fading channel (e.g., mobile)



$r(k) = g(k)c(k) + n(k)$

$SNR = \sigma_g^2/\sigma_n^2$

$\{+1,-1\}$ for simplicity — Gaussian process — Gaussian process

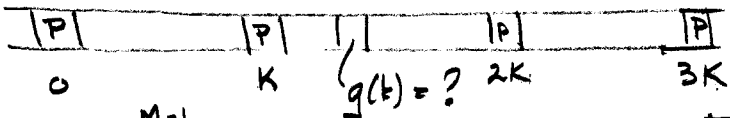
Problem: $g(k)$ is complex Gaussian, so $c(k)$ could emerge with any phase, including reversal.

Solution: form an estimate $\hat{g}(k)$ of $g(k)$. Then do this:

$$r(k)\hat{g}^*(k) = \underbrace{g(k)\hat{g}^*(k)}_{\text{almost positive real}} c(k) + n(k)\hat{g}^*(k) \qquad \text{reversible}$$

Here's how to get estimate $\hat{g}(k)$:

Make every K^{th} symbol a +1, $c(iK) = 1$ (pilot symbols)



$$\hat{g}(k) = \sum_{i=0}^{M-1} h_k^*(i)\, r(iK) \quad \text{or} \quad \hat{\Theta}(k) = \underline{h}_k^{\dagger}\, \underline{x}, \qquad \Theta = g(k)$$

a linear interpolator or linear estimate

where $\underline{x} = \begin{bmatrix} r(0) \\ r(K) \\ \vdots \\ r(K(M-1)) \end{bmatrix}$

the mmse estimate of $g(k)$ uses

$$\underline{h}_k = R_x^{-1} \underline{p}_k \qquad R_x = \overline{\underline{x}\,\underline{x}^{\dagger}} \qquad R_{x_{ij}} = R_g((i-j)K) + R_n((i-j)K)$$

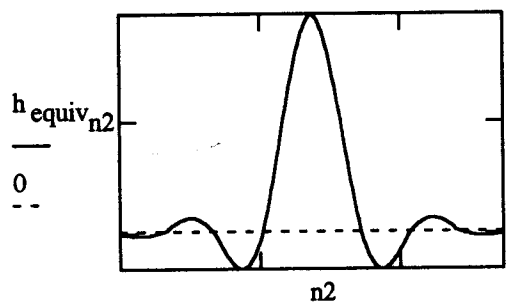
autocorr. fcns, not matrices

$$\underline{p}_k = E\left[g(k)^* \underline{x}\right] \qquad \left[\underline{p}_k\right]_i = R_g(k - iK)$$

and the mmse estimate of all the $g(k)$:

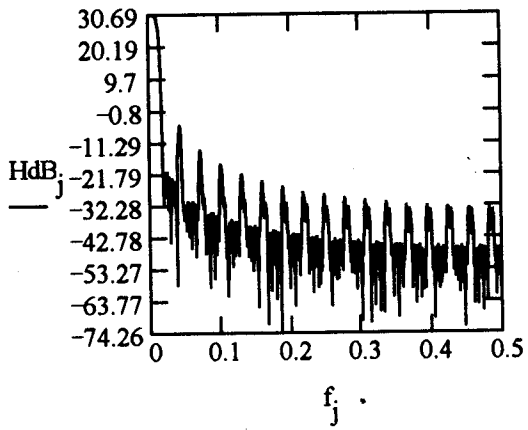
$$H = R_x^{-1} P \qquad \underline{\hat{\Theta}} = \begin{bmatrix} \hat{g}(K+1) \\ \vdots \\ \hat{g}(2K-1) \end{bmatrix} = P^{\dagger} R_x^{-1} \underline{x}$$

Interesting fact — the coefficients can be considered as a polyphase filter; deinterleaved, it looks like this



For $R_g(i) = \sigma_g^2 J_0(2\pi i f_D T)$

$R_n(i) = \sigma_n^2 \delta(i)$



Frequency response shows periodic nulls characteristic of interpolation filters

2.5.5 Partial Correlation

- Last topic on linear mmse estimation for a while — promise!
- Frequently, two or more quantities to be estimated are themselves correlated; e.g. $\begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix}$ has R_θ

  If they are both being estimated by $\underline{x}$, an interesting question is how much of their correlation was due to an $\underline{x}$ component — i.e., how correlated are they after removal of their projections on X subspace?
- Use tuples $\underline{\theta}_1$ has K_1 components, $\underline{\theta}_2$ has K_2 components, and $\underline{x}$ has M components.

$$\underline{e}_1 = \underline{\theta}_1 - \underline{\hat{\theta}}_1 = \underline{\theta}_1 - P^+ R_x^{-1}\underline{x} = \underline{\theta}_1 - E[\underline{\theta}_1 \underline{x}^+] R_x^{-1}\underline{x}$$

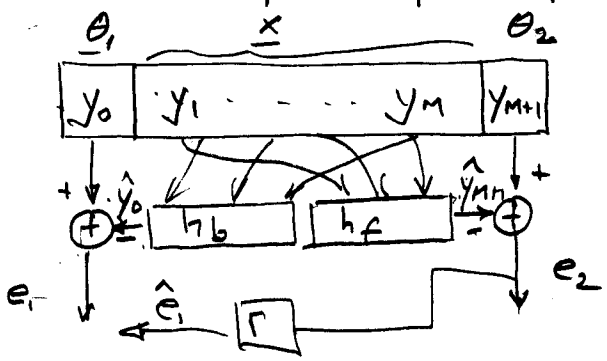
$$\underline{e}_2 = \underline{\theta}_2 - E[\underline{\theta}_2 \underline{x}^+] R_x^{-1}\underline{x}$$

Look at covariance matrix $E[\underline{e}_1 \underline{e}_2^+]$ $\quad K_1 \times K_2$

Perhaps even project $\underline{e}_1$ onto $\underline{e}_2$

$$\underline{\hat{e}}_1 = \Gamma \underline{e}_2 \qquad \Gamma = \underbrace{E[\underline{e}_1 \underline{e}_2^+]\, E[\underline{e}_2 \underline{e}_2^+]^{-1}}_{\text{PARCOR coefficients}}$$

- The PARCOR coefficients appear in forward and backward linear prediction (especially in speech work)



helps in forming a recursive estimate

- If vectors are Gaussian, then e_1, e_2 related to conditional pdfs of y_0, y_{M+1}:

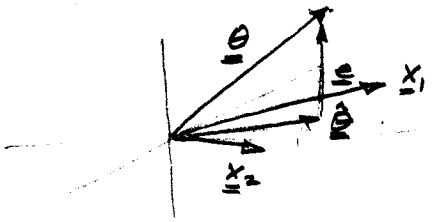
$$p_{e_1}(e_1) = p_{\theta_1|\underline{x}}(e_1 + \hat{y}_0 | \underline{x}) \qquad p_{e_1 e_2}(e_1 e_2) = p_{\theta_1,\theta_2|\underline{x}}(e_1 + \hat{y}_0, e_2 + \hat{y}_{M+1} | \underline{x})$$

may be conditionally independent.

2.5.6 General View of Linear Estimation

- Really the last topic on linear estimation, quadratic criterion (I was just kidding last time)
- In an inner product space, we want to approximate $\underline{\theta}$ as a linear combination of basis vectors $\underline{x}_1, \ldots, \underline{x}_M$ — that is, $\hat{\underline{\theta}}$ lies in subspace spanned by $\{\underline{x}_i\}$.

$$\hat{\underline{\theta}} = \sum_{i=1}^{M} h_i^* \underline{x}_i \qquad \underline{e} = \underline{\theta} - \hat{\underline{\theta}}$$



Criterion — minimize length (norm) of $\underline{e}$
equivalently, squared norm.
Essential features: - error depends linearly on h coefficients
- criterion is quadratic function of error

- For any $\underline{\theta}$ in the subspace, we get coeffs of $\underline{\theta}$ by inner prod with successive $\underline{x}_k$:

$$(\underline{x}_k, \underline{\theta}) = \sum_{i=1}^{M} h_i (\underline{x}_k, \underline{x}_i) \qquad k = 1 \ldots M$$

Note: h_i lost its conjugate — see distributive rules for inner prod

$$G\,\underline{h} = \underline{p} \qquad \text{elementary linear algebra}$$

$[G]_{k,i} = (\underline{x}_k, \underline{x}_i)$ Gram matrix of inner prods of basis functions with each other

$[\underline{p}]_k = (\underline{x}_k, \underline{\theta})$ tuple of inner prods of basis functions with vector $\underline{\theta}$

- But $\underline{\theta}$ may or may not be in the subspace, so we minimize

$$\|\underline{e}\|^2 = (\underline{e}, \underline{e}) = \underbrace{(\underline{\theta} - \hat{\underline{\theta}}, \underline{e})}_{①} = \underbrace{(\underline{\theta} - \hat{\underline{\theta}}, \underline{\theta} - \hat{\underline{\theta}})}_{②}$$

Consider ① $\|\underline{e}\|^2 = (\underline{\theta}, \underline{e}) - \sum_{i=1}^{M} h_i (\underline{x}_i, \underline{e})$

$$\frac{d}{d\underline{h}} \|\underline{e}\|^2 = -\begin{bmatrix} (\underline{x}_1, \underline{e}) \\ \vdots \\ (\underline{x}_M, \underline{e}) \end{bmatrix} = \underline{0}$$

Condition for min norm: error is orthog to all basis vectors, hence to all members of the subspace.

Consider (2) $= (\underline{e} - \sum_i h_i^* \underline{x}_i, \underline{e} - \sum_i h_i^* \underline{x}_i)$

$$\|\underline{e}\|^2 = (\underline{e},\underline{e}) - \sum_i h_i^* (\underline{x}_i, \underline{e}) - \sum_i h_i (\underline{e}, \underline{x}_i) + \sum_i \sum_k h_i^* h_k (\underline{x}_i, \underline{x}_k)$$

Just for a change, find opt by

$$\frac{d}{d\underline{h}^*} = -\begin{bmatrix} (\underline{x}_1, \underline{e}) \\ \vdots \\ (\underline{x}_M, \underline{e}) \end{bmatrix} + \begin{bmatrix} \sum_k h_k (\underline{x}_1, \underline{x}_k) \\ \vdots \\ \sum_k h_k (\underline{x}_M, \underline{x}_k) \end{bmatrix} = \underline{0}$$

or $\quad -\underline{p} + G\underline{h} = 0 \qquad G\underline{h} = \underline{p}$

This has same form as equations for $\underline{h}$ when $\underline{e}$ is in the space spanned by $\{\underline{x}_i\}$

$$\underline{p} = \begin{bmatrix} (\underline{x}_1, \underline{e}) \\ \vdots \\ (\underline{x}_M, \underline{e}) \end{bmatrix}$$

in the space (just a special case)

$$\underline{p} = \begin{bmatrix} (\underline{x}_1, \underline{e}) \\ \vdots \\ (\underline{x}_M, \underline{e}) \end{bmatrix}$$

may not be in the space.

Finally the minimized $\|\underline{e}\|$. For any $\underline{h}$,

$$\|\underline{e}\|^2 = (\underline{e},\underline{e}) - \underline{h}^+ \underline{p} - \underline{p}^+ \underline{h} + \underline{h}^+ G \underline{h}$$

Subst $\underline{h} = G^{-1}\underline{p}$, and

$$\|\underline{e}\|^2 = \|\underline{e}\|^2 - \underline{p}^+ G^{-1} \underline{p}$$

but $\|\hat{\underline{e}}\|^2 = (\sum_i h_i \underline{x}_i, \sum_i h_i \underline{x}_i) = \underline{h}^+ G \underline{h} \longrightarrow \underline{p}^+ G^{-1} \underline{p}$ at optimum

so $\min \|\underline{e}\|^2 = \|\underline{e}\|^2 - \|\hat{\underline{e}}\|^2$

Pythagoras