Examen Master SICOM, Cours Identification
Luc Pronzato, Décembre 2004 — Correction

1a) L'estimateur des moindres carrés est donné par $\hat{\theta}^k = (\mathbf{X}_k^\top \mathbf{X}_k)^{-1}\mathbf{X}_k^\top \mathbf{y}_1^k$.

Sa moyenne vaut $\mathtt{E}\{\hat{\theta}^k\} = (\mathbf{X}_k^\top \mathbf{X}_k)^{-1}\mathbf{X}_k^\top \mathtt{E}\{\mathbf{y}_1^k\} = (\mathbf{X}_k^\top \mathbf{X}_k)^{-1}\mathbf{X}_k^\top \mathbf{X}_k\bar{\theta} = \bar{\theta}$.

Sa covariance vaut $\mathbf{C}_k = \mathtt{E}\{(\hat{\theta}^k - \bar{\theta})(\hat{\theta}^k - \bar{\theta})^\top\} = (\mathbf{X}_k^\top \mathbf{X}_k)^{-1}\mathbf{X}_k^\top \mathtt{E}\{(\mathbf{y}_1^k - \mathbf{X}_k\bar{\theta})(\mathbf{y}_1^k - \mathbf{X}_k\bar{\theta})^\top\}\mathbf{X}_k(\mathbf{X}_k^\top \mathbf{X}_k)^{-1} = (\mathbf{X}_k^\top \mathbf{X}_k)^{-1}\mathbf{X}_k^\top \sigma^2 \mathbf{I}_k \mathbf{X}_k (\mathbf{X}_k^\top \mathbf{X}_k)^{-1}$ avec $\mathbf{I}_k$ la matrice identité de dimension k, d'où $\mathbf{C}_k = \sigma^2(\mathbf{X}_k^\top \mathbf{X}_k)^{-1}$.

1b)

$$\begin{aligned}
\hat{\theta}^{k+1} &= \mathbf{M}_{k+1}^{-1}\mathbf{X}_{k+1}^\top \mathbf{y}_1^{k+1} \\
&= \mathbf{M}_{k+1}^{-1}(\mathbf{X}_k^\top \mathbf{y}_1^k + y_{k+1}\mathbf{r}_{k+1})\,.
\end{aligned}$$

$\mathbf{M}_{k+1} = \mathbf{M}_k + \mathbf{r}_{k+1}\mathbf{r}_{k+1}^\top$, d'où

$$\begin{aligned}
\hat{\theta}^{k+1} &= \mathbf{M}_{k+1}^{-1}(\mathbf{M}_k\hat{\theta}^k + y_{k+1}\mathbf{r}_{k+1}) \\
&= \mathbf{M}_{k+1}^{-1}[(\mathbf{M}_{k+1} - \mathbf{r}_{k+1}\mathbf{r}_{k+1}^\top)\hat{\theta}^k + y_{k+1}\mathbf{r}_{k+1}] \\
&= \hat{\theta}^k + \mathbf{M}_{k+1}^{-1}\mathbf{r}_{k+1}\underbrace{(y_{k+1} - \mathbf{r}_{k+1}^\top\hat{\theta}^k)}_{e_{k+1}(\hat{\theta}^k)}
\end{aligned}$$

1c)

$$\mathbf{M}_{k+1}^{-1} = \mathbf{M}_k^{-1} - \frac{\mathbf{M}_k^{-1}\mathbf{r}_{k+1}\mathbf{r}_{k+1}^\top\mathbf{M}_k^{-1}}{1 + \mathbf{r}_{k+1}^\top\mathbf{M}_k^{-1}\mathbf{r}_{k+1}}$$

d'où

$$\mathbf{M}_{k+1}^{-1}\mathbf{r}_{k+1} = \frac{\mathbf{M}_k^{-1}\mathbf{r}_{k+1}}{1 + \mathbf{r}_{k+1}^\top\mathbf{M}_k^{-1}\mathbf{r}_{k+1}}$$

et

$$\hat{\theta}^{k+1} = \hat{\theta}^k + \frac{\mathbf{M}_k^{-1}\mathbf{r}_{k+1}}{1 + \mathbf{r}_{k+1}^\top\mathbf{M}_k^{-1}\mathbf{r}_{k+1}} e_{k+1}(\hat{\theta}^k)\,.$$

Comme $\mathbf{M}_k^{-1} = \mathbf{C}_k/\sigma^2$ pour tout k, on obtient

$$\begin{aligned}
\mathbf{C}_{k+1} &= \mathbf{C}_k - \frac{\mathbf{C}_k\mathbf{r}_{k+1}\mathbf{r}_{k+1}^\top\mathbf{C}_k}{\sigma^2 + \mathbf{r}_{k+1}^\top\mathbf{C}_k\mathbf{r}_{k+1}} \\
\hat{\theta}^{k+1} &= \hat{\theta}^k + \frac{\mathbf{C}_k\mathbf{r}_{k+1}}{\sigma^2 + \mathbf{r}_{k+1}\mathbf{C}_k\mathbf{r}_{k+1}} e_{k+1}(\hat{\theta}^k)\,.
\end{aligned}$$

2a) L'estimateur obtenu par LMS est plus facile à mettre en œuvre que celui des moindres carrés car il ne demande pas la manipulation de matrices.

On a $\mathtt{E}\{\tilde{\theta}^{k+1}\} = \mathtt{E}\{\tilde{\theta}^k\} + \lambda_{k+1}\mathbf{r}_{k+1}\mathbf{r}_{k+1}^\top(\bar{\theta} - \mathtt{E}\{\tilde{\theta}^k\}) = \bar{\theta}$ (puisque l'on suppose que $\mathtt{E}\{\tilde{\theta}^k\} = \bar{\theta}$).

De même,

$$\begin{aligned}
\mathbf{V}_{k+1} &= \mathrm{Var}\{(\mathbf{I}_p - \lambda_{k+1}\mathbf{r}_{k+1}\mathbf{r}_{k+1}^\top)\tilde{\theta}^k + \lambda_{k+1}\mathbf{r}_{k+1}\mathbf{r}_{k+1}^\top\bar{\theta} + \lambda_{k+1}\mathbf{r}_{k+1}\varepsilon_{k+1}\} \\
&= (\mathbf{I}_p - \lambda_{k+1}\mathbf{r}_{k+1}\mathbf{r}_{k+1}^\top)\mathbf{C}_k(\mathbf{I}_p - \lambda_{k+1}\mathbf{r}_{k+1}\mathbf{r}_{k+1}^\top) + \lambda_{k+1}^2\mathbf{r}_{k+1}\mathbf{r}_{k+1}^\top\sigma^2 \\
&= \mathbf{C}_k - \lambda_{k+1}\mathbf{r}_{k+1}\mathbf{r}_{k+1}^\top\mathbf{C}_k - \lambda_{k+1}\mathbf{C}_k\mathbf{r}_{k+1}\mathbf{r}_{k+1}^\top + \lambda_{k+1}^2\mathbf{r}_{k+1}\mathbf{r}_{k+1}^\top(\mathbf{r}_{k+1}^\top\mathbf{C}_k\mathbf{r}_{k+1} + \sigma^2)\,.
\end{aligned}$$

2b) Pour $\|\mathbf{r}_{k+1}\| = 1$ on obtient $\mathrm{trace}(\mathbf{V}_{k+1}) = \mathrm{trace}(\mathbf{C}_k) - 2\lambda_{k+1}\mathbf{r}_{k+1}^\top\mathbf{C}_k\mathbf{r}_{k+1} + \lambda_{k+1}^2(\mathbf{r}_{k+1}^\top\mathbf{C}_k\mathbf{r}_{k+1} + \sigma^2)$, qui est minimum pour

$$\lambda_{k+1} = \lambda_{k+1}^* = \frac{\mathbf{r}_{k+1}^\top\mathbf{C}_k\mathbf{r}_{k+1}}{\mathbf{r}_{k+1}^\top\mathbf{C}_k\mathbf{r}_{k+1} + \sigma^2} = \frac{1}{1 + \rho_{k+1}}\,.$$

Pour $\lambda_{k+1} = \lambda^*_{k+1}$, on obtient

$$\text{trace}(\mathbf{V}_{k+1}) = \text{trace}(\mathbf{C}_k) - \frac{(\mathbf{r}_{k+1}^\top \mathbf{C}_k \mathbf{r}_{k+1})^2}{\mathbf{r}_{k+1}^\top \mathbf{C}_k \mathbf{r}_{k+1} + \sigma^2} \,.$$

D'un autre coté, pour l'estimateur des moindres carrés on a

$$\text{trace}(\mathbf{C}_{k+1}) = \text{trace}(\mathbf{C}_k) - \frac{\mathbf{r}_{k+1}^\top \mathbf{C}_k^2 \mathbf{r}_{k+1}}{\mathbf{r}_{k+1}^\top \mathbf{C}_k \mathbf{r}_{k+1} + \sigma^2} \,.$$

L'inégalité de Cauchy-Schwarz $(\mathbf{u}^\top \mathbf{v})^2 \leq \|\mathbf{u}\|^2 \|\mathbf{v}\|^2$ donne pour $\mathbf{u} = \mathbf{r}_{k+1}$ et $\mathbf{v} = \mathbf{C}_k \mathbf{r}_{k+1}$

$$(\mathbf{r}_{k+1}^\top \mathbf{C}_k \mathbf{r}_{k+1})^2 \leq (\mathbf{r}_{k+1}^\top \mathbf{r}_{k+1})(\mathbf{r}_{k+1}^\top \mathbf{C}_k^2 \mathbf{r}_{k+1}) = \mathbf{r}_{k+1}^\top \mathbf{C}_k^2 \mathbf{r}_{k+1}$$

et donc trace$(\mathbf{C}_{k+1}) \leq$ trace$(\mathbf{V}_{k+1})$, l'estimateur des moindres carrés est donc plus performant que celui obtenu par LMS.

3a) Minimiser $\det(\mathbf{C}_{k+1})$ revient à maximiser $\det(\mathbf{M}_{k+1}) = \det(\mathbf{M}_k + \mathbf{r}_{k+1}\mathbf{r}_{k+1}^\top) = \det(\mathbf{M}_k)(1 + \mathbf{r}_{k+1}^\top \mathbf{M}_k^{-1} \mathbf{r}_{k+1})$. Il s'agit donc de choisir $\mathbf{r}_{k+1} = \mathbf{r}^*_{k+1}$ qui maximise $\mathbf{r}_{k+1}^\top \mathbf{M}_k^{-1} \mathbf{r}_{k+1} = \mathbf{r}_{k+1}^\top \mathbf{C}_k \mathbf{r}_{k+1}/\sigma^2$. Par conséquent $\mathbf{r}^*_{k+1} = \pm r \mathbf{u}_k$ avec $\mathbf{u}_k$ vecteur propre de $\mathbf{C}_k$ associé à sa valeur propre maximale.

3b) $\mathtt{E}\{\hat{y}^k(x)|x\} = \mathbf{r}^\top(x)\mathtt{E}\{\hat{\theta}^k\} = \mathbf{r}^\top(x)\bar{\theta}$, $\text{var}\{\hat{y}^k(x)|x\} = \mathbf{r}^\top(x)\text{Var}(\hat{\theta}^k)\mathbf{r}(x) = \mathbf{r}^\top(x)\mathbf{C}_k\mathbf{r}(x)$.

Le choix $\mathbf{r}^*_{k+1}$ correspond à observer là où la variance de la prédiction est maximale et donc où la prédiction est la plus incertaine.

3c) $\mathbf{C}_k$ est une matrice de covariance, et donc définie non-négative : $\mathbf{u}^\top \mathbf{C}_k \mathbf{u} \geq 0$ pour tout $\mathbf{u}$. En prenant $\mathbf{u} = (1\ 0)^\top$ puis $\mathbf{u} = (0\ 1)^\top$ on trouve que les termes diagonaux sont forcément positifs ou nuls. Aussi, $\det(\mathbf{C}_k) \geq 0$ et donc $|b| \leq ac$.

$$d(x,\mathbf{X}_k) = \text{var}\{\hat{y}^k(x)|x\} = \mathbf{r}^\top(x)\mathbf{C}_k\mathbf{r}(x) = (1\ x)\begin{pmatrix} c^2 & b \\ b & a^2 \end{pmatrix}\begin{pmatrix} 1 \\ x \end{pmatrix} = c^2 + 2bx + a^2x^2 \,,$$

qui forme une parabole convexe (tournée vers le haut). Le maximum pour $x \in [-1,1]$ est donc atteint en $x^* = \pm 1$ et $d(x^*,\mathbf{X}_k) = c^2 + a^2 + 2|b|$.

Puisque $\mathbf{C}_k = \sigma^2 \mathbf{M}_k^{-1}$ avec

$$\mathbf{M}_k = \sum_{i=1}^k \mathbf{r}(x_i)\mathbf{r}^\top(x_i) = \begin{pmatrix} k & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix} \,,$$

on a

$$c^2 = \sigma^2 \, \frac{\mu_2}{k\mu_2 - \mu_1^2} \,,\ a^2 = \sigma^2 \, \frac{k}{k\mu_2 - \mu_1^2} \ \text{ et } b = -\sigma^2 \, \frac{\mu_1}{k\mu_2 - \mu_1^2} \,.$$

En reportant dans $d(x^*,\mathbf{X}_k)$ on obtient

$$d(x^*,\mathbf{X}_k) = \sigma^2 \, \frac{\mu_2 + k + 2\mu_1}{k\mu_2 - \mu_1^2}$$

qui est minimum pour $\mu_1 = 0$ et $\mu_2 = k$ (notons que $-k \leq \mu_1 \leq k$ et $0 \leq \mu_2 \leq k$). Lorsque k est pair, ceci est réalisable en plaçant $k/2$ points x_i en +1 et $k/2$ en -1.

3d)

$$EQM\{\hat{y}^k(x)|x\} = \mathtt{E}\{(\hat{y}^k(x) - \mathbf{r}^\top(x)\bar{\theta})^2|x\} = \text{var}\{\hat{y}^k(x)|x\}$$

puisque $\mathtt{E}[\hat{y}^k(x)|x] = \mathbf{r}^\top(x)\bar{\theta}$. Par conséquent,

$$\Phi(\mathbf{X}_k) = \int EQM\{\hat{y}^k(x)|x\}\varphi(x)\,dx = \int \text{var}\{\hat{y}^k(x)|x\}\varphi(x)\,dx$$

ou encore
$$\Phi(\mathbf{X}_k) = \int \mathbf{r}^\top(x)\mathbf{C}_k\mathbf{r}(x)\varphi(x)\,dx = \text{trace}[\mathbf{C}_k\mathbf{V}_x]\,.$$

Lorsque φ est de moyenne nulle et de variance s^2,
$$\mathbf{V}_x = \begin{pmatrix} 1 & 0 \\ 0 & s^2 \end{pmatrix}$$
et
$$\Phi(\mathbf{X}_k) = \text{trace}[\mathbf{C}_k\mathbf{V}_x] = c^2 + s^2a^2 = \sigma^2\,\frac{\mu_2 + ks^2}{k\mu_2 - \mu_1^2}\,.$$
De nouveau $\Phi(\mathbf{X}_k)$ est minimum pour $\mu_1 = 0$ et $\mu_2 = k$, réalisable de la même manière qu'en c) lorsque k est pair.

Lorsque φ est de moyenne 1 et de variance s^2,
$$\mathbf{V}_x = \begin{pmatrix} 1 & 1 \\ 1 & 1+s^2 \end{pmatrix}$$
et
$$\Phi(\mathbf{X}_k) = \text{trace}[\mathbf{C}_k\mathbf{V}_x] = c^2 + 2b + (1+s^2)a^2 = \sigma^2\,\frac{\mu_2 - 2\mu_1 + k(1+s^2)}{k\mu_2 - \mu_1^2}\,.$$
On obtient donc
$$k\Phi(\mathbf{X}_k) = \sigma^2\,\frac{\mu_2/k - 2\mu_1/k + (1+s^2)}{\mu_2/k - (\mu_1/k)^2}$$
avec $\mu_2/k = 1$ et $\mu_1/k \to 1 - 2\alpha$ quand $k \to \infty$. Par conséquent,
$$\Psi(\alpha) = \lim_{k\to\infty} k\Phi(\mathbf{X}_k) = \sigma^2\,\frac{s^2 + 4\alpha}{4\alpha(1-\alpha)}\,,$$
qui tend vers $+\infty$ quand α tend vers 0^+ ou 1^-. La dérivée $\Psi'(\alpha)$ de $\Psi(\alpha)$ vaut
$$\Psi'(\alpha) = \sigma^2\,\frac{4\alpha^2 + 2s^2\alpha - s^2}{4\alpha^2(1-\alpha)^2}$$
qui a une seule racine positive $\alpha^*(s) = (s/4)(\sqrt{s^2+4} - s)$, α^* correspond donc à un minimum de Ψ.

Quand $s \to 0$, $\alpha^*(s) \to 0$; φ tend alors à être concentrée en $x = 1$ et placer toutes les observations en $x_i = 1$ ($\alpha = 0$) est optimum pour la précision de la prédiction en $x = 1$.

Quand $s \to \infty$, $\alpha^*(s) \to 1/2$: il faut placer la moitié des observations en +1 et l'autre moitié en -1.