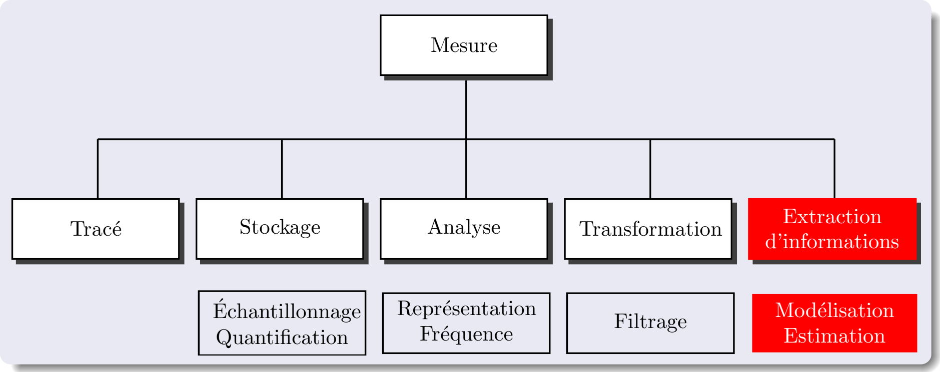
Estimation

Yves Goussard

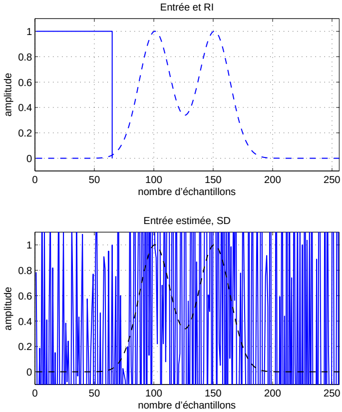
GBM6103A

24 septembre 2014

Traitement de signaux biomédicaux



Exemple : effet des incertitudes



Plan

Position du problème

Mise en situation

Exemple : estimation d'une quantité scalaire x

- Estimation de x par une moyenne : hypothèses sous-jacentes sur
  - la structure du système
  - le comportement statistiques de certaines grandeurs
- Éléments importants
  - mise en équation du problème
  - utilisation d'un *contraste* entre bruit et quantité à estimer
  - possibilité d'introduire des informations sur x sous forme probabiliste

Position du problème

Résumé

Démarche

- Mise en équation du problème
  - partie déterministe
  - partie probabiliste
    - incertitudes (bruit)
    - informations *a priori sur* x
- Estimation : utilisation du contraste entre x et les incertitudes
  - taille
  - comportement

Comment ?

Méthodologie de l'estimation

Démarche générale

- Mise en équation du problème
  - partie déterministe : structure du modèle
  - partie probabiliste
    - incertitudes (bruit) : $f_B(\boldsymbol{b})$
    - informations *a priori sur* $\boldsymbol{x}$: $f_{\boldsymbol{X}}(\boldsymbol{x})$
  - on peut calculer $f_{\boldsymbol{Z}|\boldsymbol{X}=x}(\boldsymbol{z})$
- Estimation
  - on en déduit $f_{\boldsymbol{X}|\boldsymbol{Z}=z}(\boldsymbol{x}) = \dfrac{f_{\boldsymbol{Z}|\boldsymbol{X}=x}(\boldsymbol{z})f_{\boldsymbol{X}}(\boldsymbol{x})}{f_{\boldsymbol{Z}}(\boldsymbol{z})}$ qui représente la totalité de l'information disponible sur $\boldsymbol{X}$
  - détermination d'*une* estimée (estimateur ponctuel) : maximum de $f_{\boldsymbol{X}|\boldsymbol{Z}=z}(\boldsymbol{x})$, moyenne *a posteriori*, ...

Exemples

- Estimation d'une quantité scalaire très bruitée : voir `demo_caraest.m`
- Estimation d'un retard (échographie) : voir `demo_corr.m`

Estimateurs classiques (1)

Maximum *a posteriori* (MAP)

- $\hat{x} = \arg\max_{x} \left\{ f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z}}(x) \right\} = \arg\max_{x} \left\{ f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}}(z)\ f_{\boldsymbol{X}}(x) \right\}$
- $\hat{x}$: valeur de $\boldsymbol{x}$ « la plus probable » compte tenu de la valeur des mesures
- Estimation : problème d'optimisation

Maximum de vraisemblance (MV)

- MAP avec $f_{\boldsymbol{X}}(x)$ constante $\Longrightarrow \hat{x} = \arg\max_{x} \left\{ f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}}(z) \right\}$

Moyenne *a posteriori*

- $\hat{x} = \mathrm{E}\left[\boldsymbol{X}|\boldsymbol{Z} = z\right] = \displaystyle\int x\ f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z}}(x)\ dx$
- Robuste
- Interprétation : erreur quadratique moyenne minimale
- Mise en œuvre difficile

Estimateurs classiques (2)

Estimateur linéaire d'erreur quadratique moyenne minimale

- $\hat{x} = w * g \quad ; \quad \mathrm{E}\left[(\hat{x} - x)^2\right]$ minimale
- Hypothèse gaussienne implicite
- Mise en œuvre par filtre de Kalman ou de Wiener

Estimateurs empiriques

- $\hat{x} = \Psi(z)$
- Choix de Ψ ?
- Qualité de l'estimateur

Comment choisir un estimateur ?

Caractéristiques des estimateurs (1)

Position du problème

- Tous les estimateurs ponctuels sont de la forme $\hat{x} = \Psi(z)$
- $\hat{\boldsymbol{X}}$ aléatoire car $\boldsymbol{Z}$ aléatoire
- Comportement statistique de $\hat{\boldsymbol{X}}$?

Principales caractéristiques

- Comportement statistique de $\hat{\boldsymbol{X}}$ résumé par sa moyenne et sa matrice de covariance
- Moyenne : précision
- Matrice de covariance : fidélité

Caractéristiques des estimateurs (2)

Biais et variance

- Biais : $\boldsymbol{B} = \mathrm{E}\left[\hat{\boldsymbol{X}} - \boldsymbol{x}_v\right]$; $\boldsymbol{x}_v$: vraie valeur de $\boldsymbol{X}$
- Erreur quadratique moyenne : $\|\boldsymbol{B}\|^2 + \sum \mathrm{diagCov}\left[\hat{\boldsymbol{X}}\right]$

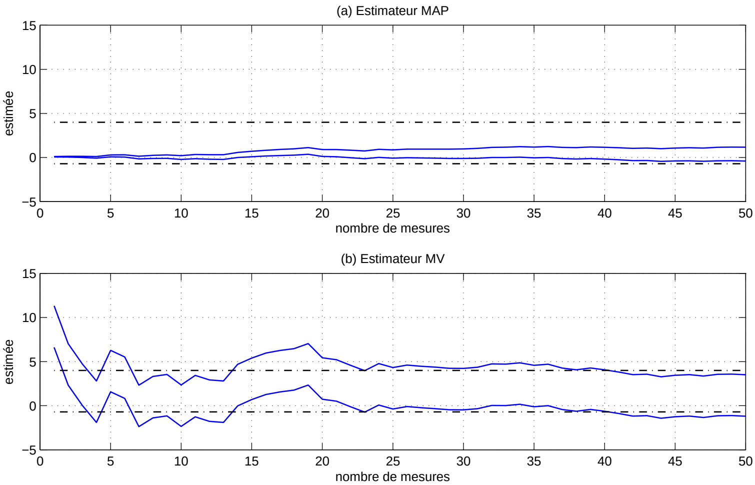
En général, nécessité d'un compromis entre biais et covariance

Comportement asymptotique

- L'erreur quadratique moyenne doit tendre vers 0 lorsque la taille de $\boldsymbol{Z}$ augmente à taille de $\boldsymbol{X}$ constante
- Pertinence à évaluer au cas par cas

Compromis biais-variance

Exemple : estimation d'une quantité scalaire dans un bruit fort



Estimation MAP dans le cas linéaire et gaussien (1)

Hypothèses

- $\boldsymbol{X}$ et $\boldsymbol{Z}$ de dimension finie
- Structure linéaire et bruit additif : $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{b}$
- Bruit gaussien ; $f_{\boldsymbol{B}} \ : \ \mathcal{N}(0, \sigma_{\boldsymbol{B}}^2 \bar{R}_{\boldsymbol{B}})$
- $\boldsymbol{X}$ gaussien ; $f_{\boldsymbol{X}} \ : \ \mathcal{N}(0, \sigma_{\boldsymbol{X}}^2 \bar{R}_{\boldsymbol{X}})$

Approche MAP

- $\{\boldsymbol{Z}|\boldsymbol{X} = \boldsymbol{x}\}$ gaussien ; $f_{\boldsymbol{Z}|\boldsymbol{X}=f} \ : \ \mathcal{N}(\boldsymbol{H}\boldsymbol{x}, \sigma_{\boldsymbol{B}}^2 \bar{R}_{\boldsymbol{B}})$
- $$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \left\{ \exp\left(-\frac{(\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x})^{\mathrm{t}} \bar{R}_{\boldsymbol{B}}^{-1} (\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x})}{2\sigma_{\boldsymbol{B}}^2} - \frac{\boldsymbol{x}^{\mathrm{t}} \bar{R}_{\boldsymbol{X}}^{-1} \boldsymbol{x}}{2\sigma_{\boldsymbol{X}}^2} \right) \right\}$$
  $$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \left\{ \frac{(\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x})^{\mathrm{t}} \bar{R}_{\boldsymbol{B}}^{-1} (\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x})}{2\sigma_{\boldsymbol{B}}^2} + \frac{\boldsymbol{x}^{\mathrm{t}} \bar{R}_{\boldsymbol{X}}^{-1} \boldsymbol{x}}{2\sigma_{\boldsymbol{X}}^2} \right\}$$

Estimation MAP dans le cas linéaire et gaussien (2)

Forme de l'estimateur

- Estimateur général : $\hat{x} = \left(\boldsymbol{H}^{\mathrm{t}} \bar{R}_{\boldsymbol{B}}^{-1} \boldsymbol{H} + \frac{\sigma_{\boldsymbol{B}}^2}{\sigma_{\boldsymbol{X}}^2} \bar{R}_{\boldsymbol{X}}^{-1}\right)^{-1} \boldsymbol{H}^{\mathrm{t}} \bar{R}_{\boldsymbol{B}}^{-1} \boldsymbol{z}$
- Si le bruit est blanc : $\hat{x} = \left(\boldsymbol{H}^{\mathrm{t}} \boldsymbol{H} + \frac{\sigma_{\boldsymbol{B}}^2}{\sigma_{\boldsymbol{X}}^2} \bar{R}_{\boldsymbol{X}}^{-1}\right)^{-1} \boldsymbol{H}^{\mathrm{t}} \boldsymbol{z}$
- Si $\boldsymbol{X}$ est déterministe : $\hat{x} = \left(\boldsymbol{H}^{\mathrm{t}} \bar{R}_{\boldsymbol{B}}^{-1} \boldsymbol{H}\right)^{-1} \boldsymbol{H}^{\mathrm{t}} \bar{R}_{\boldsymbol{B}}^{-1} \boldsymbol{z}$
- Deux cas précédents : $\hat{x} = \left(\boldsymbol{H}^{\mathrm{t}} \boldsymbol{H}\right)^{-1} \boldsymbol{H}^{\mathrm{t}} \boldsymbol{z}$ Moindres carrés

Choix d'un estimateur

- Dépend du contraste entre $\boldsymbol{x}$ et $\boldsymbol{b}$ et du « contenu informationnel » de $\boldsymbol{z}$
- Conditions favorables : moindres carrés OK
- Autrement : MAP
- Illustration : `demo_estRI.m`

Calcul pratique de la solution

Position du problème

Cas linéaire et gaussien

- $\hat{x}$ solution de $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{b}$
- Taille de $\boldsymbol{Q}$?
- Inversion de $\boldsymbol{Q}$?

Cas général

- $\hat{x}$ défini par : $\hat{x} = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$; $\hat{x}$ solution de $\nabla f(\boldsymbol{x}) = 0$
- Taille des grandeurs permettant le calcul de $f(\boldsymbol{x})$?
- Pas de forme explicite de $\hat{x}$

Approche itérative

- Suite de $\boldsymbol{x}_k$ *convergeant* vers $\hat{x}$
- Calcul *simple* de $\boldsymbol{x}_k$ à partir de $\boldsymbol{x}_{k-1}$ et $f(\boldsymbol{x})$

Optimisation sans contrainte (1)

Définition

- $\hat{x} = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$
- f deux fois continuement différentiable

Condition nécessaire d'optimalité

Si $\hat{x}$ est un minimum *local* de f, alors

- $\nabla f(\hat{x}) = 0$
- $\nabla^2 f(\hat{x})$ est semi-définie positive

Condition suffisante d'optimalité

Si

- $\nabla f(\hat{x}) = 0$
- $\nabla^2 f(\hat{x})$ est définie positive

alors $\hat{x}$ est un minimum *local* de f

Optimisation sans contrainte (2)

Interprétation

- Conditions d'optimalité relatives aux minima *locaux*
- Minimum local $=$ minimum global si $f(\boldsymbol{x})$ est *convexe*
- Cas général
  - $\nabla f(\boldsymbol{x})$ non linéaire par rapport à $\boldsymbol{x}$
  - $\nabla^2 f(\boldsymbol{x})$ varie en fonction de $\boldsymbol{x}$
  - $\nabla^2 f(\boldsymbol{x})$ semi-définie positive pour tout $\boldsymbol{x}$ si $f(\boldsymbol{x})$ est convexe
- Cas linéaire et gaussien
  - $f(\boldsymbol{x})$ est une forme quadratique semi-définie positive
  - $f(\boldsymbol{x})$ est convexe
  - $\nabla f(\boldsymbol{x})$ est linéaire par rapport à $\boldsymbol{x}$
  - $\nabla^2 f(\boldsymbol{x})$ est une matrice constante (indépendante de $\boldsymbol{x}$)

Méthodes itératives à directions de descente

Principe

En $\boldsymbol{x}_k$

- Choisir une direction de déplacement $\boldsymbol{d}_k$
- Choisir $\alpha_k > 0$ tel que $f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) < f(\boldsymbol{x}_k)$
- $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$

Intérêt

- Structure simple
- Choix de $\boldsymbol{d}_k$ et α_k généralement lié à $f(\boldsymbol{x}_k)$ et $\nabla f(\boldsymbol{x}_k)$

Difficultés

- Convergence vers $\hat{\boldsymbol{x}}$?
- Vitesse de convergence ?
- Règle d'arrêt pratique ?

Convergence globale, conditions de descente suffisante

Principe

- Choisir $\boldsymbol{d}_k$ suffisamment loin de l'orthogonalité à la plus grande pente
- Choisir α_k pour se rapprocher suffisamment du minimum de f dans la direction $\boldsymbol{d}_k$

Formalisation

Soit $f(\boldsymbol{x})$ deux fois continuement différentiable et vérifiant des conditions de régularité supplémentaires. Si $\exists\ (c_1, c_2)\ ;\ 0 < c_1 < c_2 < 1$ tels que

- $\forall k\ ;\ f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) - f(\boldsymbol{x}_k) \leq c_1 \alpha_k \boldsymbol{d}_k^{\mathrm{t}} \nabla f(\boldsymbol{x}_k)$
- $\forall k\ ;\ |\boldsymbol{d}_k^{\mathrm{t}} \nabla f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k)| \leq c_2 |\boldsymbol{d}_k^{\mathrm{t}} \nabla f(\boldsymbol{x}_k)|$

alors la suite $\boldsymbol{x}_k$ converge vers un minimum local de f (convergence globale)

Choix pratique de $\boldsymbol{d}_k$ et α_k ?

Méthodes pratiques

Directions de descente $\boldsymbol{d}_k$

- Quasi-Newton : $\boldsymbol{d}_k = -\boldsymbol{H}_k \nabla f(\boldsymbol{x}_k)$; $\boldsymbol{H}_k \approx (\nabla^2 f(\boldsymbol{x}_k))^{-1}$
  En général, $\boldsymbol{H}_k$ Calculé à partir de $\nabla f(\boldsymbol{x}_k)$, $\nabla f(\boldsymbol{x}_{k-1})$, ...
- Gradient conjugué : quasi-Newton simplifié
- Gradient : $\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)$

Pas α_k

Méthodes de *backtracking* assurant que les conditions de descente suffisante sont vérifiées

En pratique...

- Algorithmes tout faits
- Écriture d'une fonction retournant $f(\boldsymbol{x})$ et $\nabla f(\boldsymbol{x})$ pour tout $\boldsymbol{x}$
- Critère d'arrêt : seuil sur la norme de $\nabla f(\boldsymbol{x}_k)$

Cas linéaire et gaussien

Caractéristiques de $f(x)$

- $f(\boldsymbol{x})$ forme quadratique semi-définie positive, donc convexe :
$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{t}}\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^{\mathrm{t}}\boldsymbol{x} + c$$
- $\nabla f(\boldsymbol{x})$ linéaire par rapport à $\boldsymbol{x}$, $\nabla^2 f(\boldsymbol{x})$ constant
- $f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k)$ polynôme du 2^e degré en α_k.

Calcul explicite du α_k qui minimise $f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k)$ (pas optimal)

Conséquences pratiques

- Pas optimal et directions de descente du gradient ou du gradient conjugué : convergence globale
- Simplification importante (calcul direct de α_k)
- Algorithmes nécessitant seulement une fonction retournant $\boldsymbol{Q}\boldsymbol{x}$ et $\boldsymbol{b}$ pour tout $\boldsymbol{x}$
- Critère d'arrêt : souplesse accrue