

Présentation de la méthode de la vraisemblance empirique

J.Worms

Plan de l'exposé

- 1 Rappels de statistique paramétrique
 - Modèles paramétriques et vraisemblance
 - Rapport des maxima de vraisemblance
- 2 Vraisemblance empirique (cadre iid) : deux approches
 - Approche NPMLE (max de vraisemblance non-paramétrique)
 - Approche Minimum de Contraste
- 3 Cadres d'application, avantages, exemples

- 1 Rappels de statistique paramétrique
 - Modèles paramétriques et vraisemblance
 - Rapport des maxima de vraisemblance
- 2 Vraisemblance empirique (cadre iid) : deux approches
 - Approche NPMLE (max de vraisemblance non-paramétrique)
 - Approche Minimum de Contraste
- 3 Cadres d'application, avantages, exemples

Modèle paramétrique

X_1, \dots, X_n i.i.d. $\hookrightarrow \mathbb{R}^p$ définies sur (Ω, \mathcal{A}) , de loi μ

Modèle paramétrique :

$$\mu \in \mathcal{P} := \{ \mu_\theta : \theta \in \Theta \}$$

Objectif : estimer θ

\longrightarrow région de confiance ou test concernant θ ou $g(\theta)$

Moyen fréquent : exploiter normalité asymp. de l'estimateur $\hat{\theta}$

Question usuelle : qualité du procédé en non-asymptotique

Modèle paramétrique

X_1, \dots, X_n i.i.d. $\hookrightarrow \mathbb{R}^p$ définies sur (Ω, \mathcal{A}) , de loi μ

Modèle paramétrique :

$$\mu \in \mathcal{P} := \{ \mu_\theta : \theta \in \Theta \}$$

Objectif : estimer θ

\longrightarrow région de confiance ou test concernant θ ou $g(\theta)$

Moyen fréquent : exploiter normalité asymp. de l'estimateur $\hat{\theta}$

Question usuelle : qualité du procédé en non-asymptotique

Modèle paramétrique

X_1, \dots, X_n i.i.d. $\hookrightarrow \mathbb{R}^p$ définies sur (Ω, \mathcal{A}) , de loi μ

Modèle paramétrique :

$$\mu \in \mathcal{P} := \{ \mu_\theta : \theta \in \Theta \}$$

Objectif : estimer θ

→ région de confiance ou test concernant θ ou $g(\theta)$

Moyen fréquent : exploiter normalité asymp. de l'estimateur $\hat{\theta}$

Question usuelle : qualité du procédé en non-asymptotique

Modèle paramétrique

X_1, \dots, X_n i.i.d. $\hookrightarrow \mathbb{R}^p$ définies sur (Ω, \mathcal{A}) , de loi μ

Modèle paramétrique :

$$\mu \in \mathcal{P} := \{ \mu_\theta : \theta \in \Theta \}$$

Objectif : estimer θ

→ région de confiance ou test concernant θ ou $g(\theta)$

Moyen fréquent : exploiter normalité asymp. de l'estimateur $\hat{\theta}$

Question usuelle : qualité du procédé en non-asymptotique

Modèle paramétrique

X_1, \dots, X_n i.i.d. $\hookrightarrow \mathbb{R}^p$ définies sur (Ω, \mathcal{A}) , de loi μ

Modèle paramétrique :

$$\mu \in \mathcal{P} := \{ \mu_\theta : \theta \in \Theta \}$$

Objectif : estimer θ

→ région de confiance ou test concernant θ ou $g(\theta)$

Moyen fréquent : exploiter normalité asymp. de l'estimateur $\hat{\theta}$

Question usuelle : qualité du procédé en non-asymptotique

Maximum de vraisemblance

Paramètre fini-dimensionnel : $\theta \in \Theta$ avec $\dim \Theta = k < \infty$

Modèle dominé et vraisemblance

$$d\mu_\theta = f_\theta dm (\forall \theta)$$
$$\mathcal{L}(\theta) := \prod_{i=1}^n f_\theta(X_i)$$

Sous de "bonnes" conditions, $\mathcal{L}(\theta)$ est maximisée en un unique $\hat{\theta}$ appelé estimateur du maximum de vraisemblance (EMV) de θ .

Maximum de vraisemblance

Paramètre fini-dimensionnel : $\theta \in \Theta$ avec $\dim \Theta = k < \infty$

Modèle dominé et vraisemblance

$$d\mu_\theta = f_\theta dm (\forall \theta)$$
$$\mathcal{L}(\theta) := \prod_{i=1}^n f_\theta(X_i)$$

Sous de "bonnes" conditions, $\mathcal{L}(\theta)$ est maximisée en un unique $\hat{\theta}$ appelé estimateur du maximum de vraisemblance (EMV) de θ .

Maximum de vraisemblance

Paramètre fini-dimensionnel : $\theta \in \Theta$ avec $\dim \Theta = k < \infty$

Modèle dominé et vraisemblance

$$d\mu_\theta = f_\theta dm (\forall \theta)$$
$$\mathcal{L}(\theta) := \prod_{i=1}^n f_\theta(X_i)$$

Sous de “bonnes” conditions, $\mathcal{L}(\theta)$ est maximisée en un unique $\hat{\theta}$ appelé estimateur du maximum de vraisemblance (EMV) de θ .

Sous certaines conditions de régularité sur le modèle dominé, l'EMV est unique et asymptotiquement normal sous \mathbb{P}_θ ($\forall \theta$)

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_k(0, I_\theta^{-1})$$

où I_θ est la matrice d'information de Fisher.

➤ Région de confiance (ellipsoïdale)

$$\{ \theta \in \Theta / n I_\theta (\hat{\theta} - \theta) (\hat{\theta} - \theta) \leq r \}$$

où r est tq $\mathbb{P}[\chi^2(k) \geq r] = \alpha$.

➤ Test de $H_0 : \theta = \theta_0$

région critique de niv. asymp. $\alpha = \{ n I_{\theta_0} (\hat{\theta} - \theta_0) (\hat{\theta} - \theta_0) \geq r \}$

Sous certaines conditions de régularité sur le modèle dominé, l'EMV est unique et asymptotiquement normal sous \mathbb{P}_θ ($\forall \theta$)

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_k(0, I_\theta^{-1})$$

où I_θ est la matrice d'information de Fisher.

➔ Région de confiance (ellipsoïdale)

$$\{ \theta \in \Theta / n^t(\hat{\theta} - \theta)I_{\hat{\theta}}(\hat{\theta} - \theta) \leq r \}$$

où r est tq $\mathbb{P}[\chi^2(k) \geq r] = \alpha$.

➔ Test de $H_0 : \theta = \theta_0$

région critique de niv. asymp. $\alpha = \{ n^t(\hat{\theta} - \theta_0)I_{\theta_0}(\hat{\theta} - \theta_0) \geq r \}$

Sous certaines conditions de régularité sur le modèle dominé, l'EMV est unique et asymptotiquement normal sous \mathbb{P}_θ ($\forall \theta$)

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_k(0, I_\theta^{-1})$$

où I_θ est la matrice d'information de Fisher.

➔ Région de confiance (ellipsoïdale)

$$\{ \theta \in \Theta / n^t(\hat{\theta} - \theta)I_{\hat{\theta}}(\hat{\theta} - \theta) \leq r \}$$

où r est tq $\mathbb{P}[\chi^2(k) \geq r] = \alpha$.

➔ Test de $H_0 : \theta = \theta_0$

région critique de niv. asymp. $\alpha = \{ n^t(\hat{\theta} - \theta_0)I_{\theta_0}(\hat{\theta} - \theta_0) \geq r \}$

- 1 Rappels de statistique paramétrique
 - Modèles paramétriques et vraisemblance
 - Rapport des maxima de vraisemblance
- 2 Vraisemblance empirique (cadre iid) : deux approches
 - Approche NPMLE (max de vraisemblance non-paramétrique)
 - Approche Minimum de Contraste
- 3 Cadres d'application, avantages, exemples

Test du rapport des maxima de vraisemblance

Théorème (Wilks, 1938)

Si $\Theta_0 \subset \Theta$, $\dim \Theta = k$, $\dim \Theta_0 = l$, et

$$LR_0 := \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\theta)}$$

alors, sous $H_0 : \theta \in \Theta_0$, quand $n \rightarrow \infty$

$$-2 \log LR_0 \xrightarrow{\mathcal{L}} \chi^2(k - l)$$

d'où la région critique de niveau asymp. α

$$\{ -2 \log LR_0 \geq r \} \quad \text{où} \quad \mathbb{P}[\chi^2(k - l) \geq r] = \alpha$$

Région de confiance par inversion du LRT

Corollaire

Une région de confiance pour θ (de niv. asymp. $1 - \alpha$) est

$$\{ \theta \in \Theta / -2 \log LR(\theta) \leq r \} \text{ où } LR(\theta) := \frac{\mathcal{L}(\theta)}{\mathcal{L}(\hat{\theta})}$$

et $\mathbb{P}[\chi^2(k) \geq r] = \alpha$.

Région de confiance par inversion du LRT

Corollaire

Une région de confiance pour θ (de niv. asymp. $1 - \alpha$) est

$$\{ \theta \in \Theta / -2 \log LR(\theta) \leq r \} \text{ où } LR(\theta) := \frac{\mathcal{L}(\theta)}{\mathcal{L}(\hat{\theta})}$$

et $\mathbb{P}[\chi^2(k) \geq r] = \alpha$.

- ↪ vs région de Wald, région plus difficile à déterminer en pratique...
- ↪ mais souvent meilleure pour n faible (proba de couverture plus proche du niveau annoncé)
- ↪ et ne nécessitant aucune estimation de variance asymptotique

Région de confiance par inversion du LRT

Corollaire

Une région de confiance pour θ (de niv. asymp. $1 - \alpha$) est

$$\{ \theta \in \Theta / -2 \log LR(\theta) \leq r \} \text{ où } LR(\theta) := \frac{\mathcal{L}(\theta)}{\mathcal{L}(\hat{\theta})}$$

et $\mathbb{P}[\chi^2(k) \geq r] = \alpha$.

- ↪ vs région de Wald, région plus difficile à déterminer en pratique...
- ↪ mais souvent meilleure pour n faible (proba de couverture plus proche du niveau annoncé)
- ↪ et ne nécessitant aucune estimation de variance asymptotique

Région de confiance par inversion du LRT

Corollaire

Une région de confiance pour θ (de niv. asymp. $1 - \alpha$) est

$$\{ \theta \in \Theta / -2 \log LR(\theta) \leq r \} \text{ où } LR(\theta) := \frac{\mathcal{L}(\theta)}{\mathcal{L}(\hat{\theta})}$$

et $\mathbb{P}[\chi^2(k) \geq r] = \alpha$.

- ↪ vs région de Wald, région plus difficile à déterminer en pratique...
- ↪ mais souvent meilleure pour n faible (proba de couverture plus proche du niveau annoncé)
- ↪ et ne nécessitant aucune estimation de variance asymptotique

Sortir de la modélisation paramétrique ?

Maximum de Vraisemblance =

- méthodologie très répandue, très étudiée
- propriétés plaisantes (efficacité, invariance par transformation, ...)

mais sortir de la modélisation paramétrique est toujours désirable

→ des travaux existent sur la généralisation de la notion de vraisemblance en non-paramétrique (paramètre $\theta = \infty$ -dimensionnel), mais sont parsemés de nombreuses difficultés (fonction d'influence, détermination et estimation de la variance asymptotique,...)

Sortir de la modélisation paramétrique ?

Maximum de Vraisemblance =

- méthodologie très répandue, très étudiée
- propriétés plaisantes (efficacité, invariance par transformation, ...)

mais sortir de la modélisation paramétrique est toujours désirable

↪ des travaux existent sur la généralisation de la notion de vraisemblance en non-paramétrique (paramètre $\theta = \infty$ -dimensionnel), mais sont parsemés de nombreuses difficultés (fonction d'influence, détermination et estimation de la variance asymptotique,...)

- 1 Rappels de statistique paramétrique
 - Modèles paramétriques et vraisemblance
 - Rapport des maxima de vraisemblance
- 2 **Vraisemblance empirique (cadre iid) : deux approches**
 - **Approche NPMLE (max de vraisemblance non-paramétrique)**
 - Approche Minimum de Contraste
- 3 Cadres d'application, avantages, exemples

Vraisemblance non-paramétrique

On observe tjrs (X_1, \dots, X_n) iid $\hookrightarrow \mathbb{R}^p$ de loi μ

Modèle statistique : $\mu \in \mathcal{P} = \cup_{\theta \in \Theta} \mathcal{P}_\theta$ où

$\mathcal{P}_\theta = \{ \text{lois } \nu \text{ sur } \mathbb{R}^p \text{ vérifiant une condition relative à } \theta \}$

\Rightarrow 2 paramètres $\left\{ \begin{array}{l} \text{la loi } \mu, \text{ paramètre } \infty\text{-dimensionnel} \\ \theta = \theta(\mu), \text{ vu comme fonction de } \mu \end{array} \right.$

Définition

$$\begin{aligned} \text{NPL}(\nu) &:= \prod_{i=1}^n \nu(\{X_i\}) \\ &= \prod_{i=1}^n p_i \quad \text{si } \nu = \sum_{i=1}^n p_i \delta_{X_i} \end{aligned}$$

Vraisemblance non-paramétrique

On observe tjrs (X_1, \dots, X_n) iid $\hookrightarrow \mathbb{R}^p$ de loi μ

Modèle statistique : $\mu \in \mathcal{P} = \cup_{\theta \in \Theta} \mathcal{P}_\theta$ où

$\mathcal{P}_\theta = \{ \text{lois } \nu \text{ sur } \mathbb{R}^p \text{ vérifiant une condition relative à } \theta \}$

\Rightarrow 2 paramètres $\left\{ \begin{array}{l} \text{la loi } \mu, \text{ paramètre } \infty\text{-dimensionnel} \\ \theta = \theta(\mu), \text{ vu comme fonction de } \mu \end{array} \right.$

Définition

$$\begin{aligned} \text{NPL}(\nu) &:= \prod_{i=1}^n \nu(\{X_i\}) \\ &= \prod_{i=1}^n p_i \quad \text{si } \nu = \sum_{i=1}^n p_i \delta_{X_i} \end{aligned}$$

Vraisemblance non-paramétrique

On observe tjrs (X_1, \dots, X_n) iid $\hookrightarrow \mathbb{R}^p$ de loi μ

Modèle statistique : $\mu \in \mathcal{P} = \cup_{\theta \in \Theta} \mathcal{P}_\theta$ où

$\mathcal{P}_\theta = \{ \text{lois } \nu \text{ sur } \mathbb{R}^p \text{ vérifiant une condition relative à } \theta \}$

⇒ 2 paramètres $\left\{ \begin{array}{l} \text{la loi } \mu, \text{ paramètre } \infty\text{-dimensionnel} \\ \theta = \theta(\mu), \text{ vu comme fonction de } \mu \end{array} \right.$

Définition

$$\begin{aligned} \text{NPL}(\nu) &:= \prod_{i=1}^n \nu(\{X_i\}) \\ &= \prod_{i=1}^n p_i \quad \text{si } \nu = \sum_{i=1}^n p_i \delta_{X_i} \end{aligned}$$

Vraisemblance non-paramétrique

On observe tjrs (X_1, \dots, X_n) iid $\hookrightarrow \mathbb{R}^p$ de loi μ

Modèle statistique : $\mu \in \mathcal{P} = \cup_{\theta \in \Theta} \mathcal{P}_\theta$ où

$\mathcal{P}_\theta = \{ \text{lois } \nu \text{ sur } \mathbb{R}^p \text{ vérifiant une condition relative à } \theta \}$

\Rightarrow 2 paramètres $\left\{ \begin{array}{l} \text{la loi } \mu, \text{ paramètre } \infty\text{-dimensionnel} \\ \theta = \theta(\mu), \text{ vu comme fonction de } \mu \end{array} \right.$

Définition

$$\begin{aligned} \text{NP}\mathcal{L}(\nu) &:= \prod_{i=1}^n \nu(\{X_i\}) \\ &= \prod_{i=1}^n p_i \quad \text{si } \nu = \sum_{i=1}^n p_i \delta_{X_i} \end{aligned}$$

Vraisemblance non-paramétrique

La mesure empirique

$$\mu_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$$

maximise $NP\mathcal{L}(\cdot)$ sur l'espace des probas sur \mathbb{R}^p .

(Kiefer & Wolfowitz, 1956)

On veut maximiser $NP\mathcal{L}(\cdot)$ sur l'ensemble $\mathcal{P} = \cup_{\theta \in \Theta} \mathcal{P}_\theta$ où

$$\mathcal{P}_\theta := \{ \nu / \mathbb{E}_\nu(g(X, \theta)) = 0 \}$$

et $g : (\mathbb{R}^p, \Theta) \rightarrow \mathbb{R}^q$

- *moment condition model* : modèle **semiparamétrique**
- q est le nombre de contraintes imposées sur θ

Vraisemblance non-paramétrique

La mesure empirique

$$\mu_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$$

maximise $NP\mathcal{L}(\cdot)$ sur l'espace des probas sur \mathbb{R}^p .

(Kiefer & Wolfowitz, 1956)

On veut maximiser $NP\mathcal{L}(\cdot)$ sur l'ensemble $\mathcal{P} = \cup_{\theta \in \Theta} \mathcal{P}_\theta$ où

$$\mathcal{P}_\theta := \{ \nu / \mathbb{E}_\nu(g(X, \theta)) = 0 \}$$

et $g : (\mathbb{R}^p, \Theta) \rightarrow \mathbb{R}^q$

➤ *moment condition model* : modèle **semiparamétrique**

➤ q est le nombre de contraintes imposées sur θ

Vraisemblance non-paramétrique

La mesure empirique

$$\mu_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$$

maximise $NP\mathcal{L}(\cdot)$ sur l'espace des probas sur \mathbb{R}^p .

(Kiefer & Wolfowitz, 1956)

On veut maximiser $NP\mathcal{L}(\cdot)$ sur l'ensemble $\mathcal{P} = \cup_{\theta \in \Theta} \mathcal{P}_\theta$ où

$$\mathcal{P}_\theta := \{ \nu / \mathbb{E}_\nu(g(X, \theta)) = 0 \}$$

et $g : (\mathbb{R}^p, \Theta) \rightarrow \mathbb{R}^q$

- ➔ *moment condition model* : modèle **semiparamétrique**
- ➔ q est le nombre de contraintes imposées sur θ

Vraisemblance empirique

Définition

On définit la **vraisemblance empirique** de θ ayant observé (X_1, \dots, X_n) , par

$$\mathcal{EL}(\theta) := \sup_{\nu \in \mathcal{P}_\theta} N\mathcal{P}\mathcal{L}(\nu).$$

L'EMVE de $\theta = \theta(\mu)$ est défini comme

$$\hat{\theta}_{EL} \in \arg \max_{\theta \in \Theta} \mathcal{EL}(\theta)$$

Si S_n désigne le simplexe de \mathbb{R}^n , on a donc

$$\hat{\theta}_{EL} \in \arg \max_{\theta} \sup_{(p_i) \in S_n} \{ \prod_{i=1}^n p_i / \sum_{i=1}^n p_i g(X_i, \theta) = 0 \}$$

Remarque : $\theta \mapsto \mathcal{EL}(\theta)$ est une *vraisemblance profilée*.

Vraisemblance empirique

Définition

On définit la **vraisemblance empirique** de θ ayant observé (X_1, \dots, X_n) , par

$$\mathcal{EL}(\theta) := \sup_{\nu \in \mathcal{P}_\theta} NP\mathcal{L}(\nu).$$

L'EMVE de $\theta = \theta(\mu)$ est défini comme

$$\hat{\theta}_{EL} \in \arg \max_{\theta \in \Theta} \mathcal{EL}(\theta)$$

Si S_n désigne le simplexe de \mathbb{R}^n , on a donc

$$\hat{\theta}_{EL} \in \arg \max_{\theta} \sup_{(p_i) \in S_n} \{ \prod_{i=1}^n p_i / \sum_{i=1}^n p_i g(X_i, \theta) = 0 \}$$

Remarque : $\theta \mapsto \mathcal{EL}(\theta)$ est une *vraisemblance profilée*.

Vraisemblance empirique

Définition

On définit la **vraisemblance empirique** de θ ayant observé (X_1, \dots, X_n) , par

$$\mathcal{EL}(\theta) := \sup_{\nu \in \mathcal{P}_\theta} NP\mathcal{L}(\nu).$$

L'EMVE de $\theta = \theta(\mu)$ est défini comme

$$\hat{\theta}_{EL} \in \arg \max_{\theta \in \Theta} \mathcal{EL}(\theta)$$

Si S_n désigne le simplexe de \mathbb{R}^n , on a donc

$$\hat{\theta}_{EL} \in \arg \max_{\theta} \sup_{(p_i) \in S_n} \left\{ \prod_{i=1}^n p_i / \sum_{i=1}^n p_i g(X_i, \theta) = 0 \right\}$$

Remarque : $\theta \mapsto \mathcal{EL}(\theta)$ est une vraisemblance profilée.

Vraisemblance empirique

Définition

On définit la **vraisemblance empirique** de θ ayant observé (X_1, \dots, X_n) , par

$$\mathcal{EL}(\theta) := \sup_{\nu \in \mathcal{P}_\theta} NP\mathcal{L}(\nu).$$

L'EMVE de $\theta = \theta(\mu)$ est défini comme

$$\hat{\theta}_{EL} \in \arg \max_{\theta \in \Theta} \mathcal{EL}(\theta)$$

Si S_n désigne le simplexe de \mathbb{R}^n , on a donc

$$\hat{\theta}_{EL} \in \arg \max_{\theta} \sup_{(p_i) \in S_n} \left\{ \prod_{i=1}^n p_i / \sum_{i=1}^n p_i g(X_i, \theta) = 0 \right\}$$

Remarque : $\theta \mapsto \mathcal{EL}(\theta)$ est une *vraisemblance profilée*.

Calcul des estimateurs du MVE

Proposition

L'EMVE de $\theta = \theta(\mu)$ vérifie

$$\hat{\theta}_{EL} \in \arg \max_{\theta \in \Theta} \left\{ -\sum_{i=1}^n \log(1 + \langle \hat{\lambda}(\theta), g(X_i, \theta) \rangle) \right\}$$

où $\hat{\lambda}(\theta)$ désigne une solution dans \mathbb{R}^q de

$$\sum_{i=1}^n (1 + \langle \lambda, g(X_i, \theta) \rangle)^{-1} g(X_i, \theta) = 0$$

Si l'on pose $\hat{p}_i := \frac{1}{n} (1 + \langle \hat{\lambda}(\hat{\theta}_{EL}), g(X_i, \hat{\theta}_{EL}) \rangle)^{-1}$ alors

$$\hat{p}_{EL} := \sum_{i=1}^n \hat{p}_i \delta_{X_i}$$

est l'EMVE de la loi μ sous-jacente.

Calcul des estimateurs du MVE

Proposition

L'EMVE de $\theta = \theta(\mu)$ vérifie

$$\hat{\theta}_{EL} \in \arg \max_{\theta \in \Theta} \left\{ -\sum_{i=1}^n \log(1 + \langle \hat{\lambda}(\theta), g(X_i, \theta) \rangle) \right\}$$

où $\hat{\lambda}(\theta)$ désigne une solution dans \mathbb{R}^q de

$$\sum_{i=1}^n (1 + \langle \lambda, g(X_i, \theta) \rangle)^{-1} g(X_i, \theta) = 0$$

Si l'on pose $\hat{p}_i := \frac{1}{n} (1 + \langle \hat{\lambda}(\hat{\theta}_{EL}), g(X_i, \hat{\theta}_{EL}) \rangle)^{-1}$ alors

$$\hat{\mu}_{EL} := \sum_{i=1}^n \hat{p}_i \delta_{X_i}$$

est l'EMVE de la loi μ sous-jacente.

Expression de ces estimateurs ?

Dans le cas simple où $g(x, \theta) = x - \theta$ (i.e. $\theta = \mathbb{E}_\mu(X)$), on constate que

$$\hat{\theta}_{EL} = \bar{x} \quad \text{et} \quad \hat{\mu}_{EL} = \bar{x}$$

Dans le cas général, si

$$\sum_{i=1}^n g(X_i, \theta) = 0$$

admet une solution $\tilde{\theta}$ (qui est donc un *M-estimateur* de θ), alors on constate de la même façon que

$$\hat{\theta}_{EL} = \tilde{\theta} \quad \text{et} \quad \hat{\mu}_{EL} = \bar{x}$$

Quel intérêt alors cette "méthode" présente-t-elle ??

Expression de ces estimateurs ?

Dans le cas simple où $g(x, \theta) = x - \theta$ (i.e. $\theta = \mathbb{E}_\mu(X)$), on constate que

$$\hat{\theta}_{EL} = \bar{X}_n \quad \text{et} \quad \hat{\mu}_{EL} = \mu_n!$$

(avec $\hat{\lambda}(\bar{X}_n) = 0$ donc $\hat{p}_i = 1/n$).

Dans le cas général, si

$$\sum_{i=1}^n g(X_i, \theta) = 0$$

admet une solution $\tilde{\theta}$ (qui est donc un *M-estimateur* de θ), alors on constate de la même façon que

$$\hat{\theta}_{EL} = \tilde{\theta} \quad \text{et} \quad \hat{\mu}_{EL} = \mu_n$$

Quel intérêt alors cette "méthode" présente-t-elle ? ?

Expression de ces estimateurs ?

Dans le cas simple où $g(x, \theta) = x - \theta$ (i.e. $\theta = \mathbb{E}_\mu(X)$), on constate que

$$\hat{\theta}_{EL} = \bar{X}_n \quad \text{et} \quad \hat{\mu}_{EL} = \mu_n!$$

(avec $\hat{\lambda}(\bar{X}_n) = 0$ donc $\hat{p}_i = 1/n$).

Dans le cas général, si

$$\sum_{i=1}^n g(X_i, \theta) = 0$$

admet une solution $\tilde{\theta}$ (qui est donc un *M-estimateur* de θ), alors on constate de la même façon que

$$\hat{\theta}_{EL} = \tilde{\theta} \quad \text{et} \quad \hat{\mu}_{EL} = \mu_n$$

Quel intérêt alors cette "méthode" présente-t-elle ? ?

Expression de ces estimateurs ?

Dans le cas simple où $g(x, \theta) = x - \theta$ (i.e. $\theta = \mathbb{E}_\mu(X)$), on constate que

$$\hat{\theta}_{EL} = \bar{X}_n \quad \text{et} \quad \hat{\mu}_{EL} = \mu_n!$$

(avec $\hat{\lambda}(\bar{X}_n) = 0$ donc $\hat{p}_i = 1/n$).

Dans le cas général, si

$$\sum_{i=1}^n g(X_i, \theta) = 0$$

admet une solution $\tilde{\theta}$ (qui est donc un *M-estimateur* de θ), alors on constate de la même façon que

$$\hat{\theta}_{EL} = \tilde{\theta} \quad \text{et} \quad \hat{\mu}_{EL} = \mu_n$$

Quel intérêt alors cette “méthode” présente-t-elle ? ?

Rapport de vraisemblance empirique

On introduit le *rapport de vraisemblance empirique*

$$ELR(\theta) := \frac{\sup_{\nu \in \mathcal{P}_\theta} NP\mathcal{L}(\nu)}{\sup_{\nu} NP\mathcal{L}(\nu)} = \frac{\mathcal{E}\mathcal{L}(\theta)}{n^{-n}}$$

qui, comme on vient de le voir, coïncide souvent avec l'expression naturelle

$$ELR(\theta) = \frac{\mathcal{E}\mathcal{L}(\theta)}{\mathcal{E}\mathcal{L}(\hat{\theta}_{EL})}$$

Ceci donne

$$\begin{aligned} ELR(\theta) &= \sup_{(p_i)} \left\{ \prod_{i=1}^n (np_i) / (p_i) \in \mathcal{S}_n, \sum_{i=1}^n p_i g(X_i, \theta) = 0 \right\} \\ &= \prod_{i=1}^n (1 + \langle \hat{\lambda}(\theta), g(X_i, \theta) \rangle)^{-1} \end{aligned}$$

Rapport de vraisemblance empirique

On introduit le *rapport de vraisemblance empirique*

$$ELR(\theta) := \frac{\sup_{\nu \in \mathcal{P}_\theta} NP\mathcal{L}(\nu)}{\sup_{\nu} NP\mathcal{L}(\nu)} = \frac{\mathcal{E}\mathcal{L}(\theta)}{n^{-n}}$$

qui, comme on vient de le voir, coïncide souvent avec l'expression naturelle

$$ELR(\theta) = \frac{\mathcal{E}\mathcal{L}(\theta)}{\mathcal{E}\mathcal{L}(\hat{\theta}_{EL})}$$

Ceci donne

$$\begin{aligned} ELR(\theta) &= \sup_{(p_i)} \left\{ \prod_{i=1}^n (np_i) / (p_i) \in \mathcal{S}_n, \sum_{i=1}^n p_i g(X_i, \theta) = 0 \right\} \\ &= \prod_{i=1}^n (1 + \langle \hat{\lambda}(\theta), g(X_i, \theta) \rangle)^{-1} \end{aligned}$$

Rapport de vraisemblance empirique

On introduit le *rapport de vraisemblance empirique*

$$ELR(\theta) := \frac{\sup_{\nu \in \mathcal{P}_\theta} NP\mathcal{L}(\nu)}{\sup_{\nu} NP\mathcal{L}(\nu)} = \frac{\mathcal{E}\mathcal{L}(\theta)}{n^{-n}}$$

qui, comme on vient de le voir, coïncide souvent avec l'expression naturelle

$$ELR(\theta) = \frac{\mathcal{E}\mathcal{L}(\theta)}{\mathcal{E}\mathcal{L}(\hat{\theta}_{EL})}$$

Ceci donne

$$\begin{aligned} ELR(\theta) &= \sup_{(p_i)} \left\{ \prod_{i=1}^n (np_i) / (p_i) \in \mathcal{S}_n, \sum_{i=1}^n p_i g(X_i, \theta) = 0 \right\} \\ &= \prod_{i=1}^n (1 + \langle \hat{\lambda}(\theta), g(X_i, \theta) \rangle)^{-1} \end{aligned}$$

Analogie vraisemblance empirique du LRT

Dans le cas où $g(x, \theta) = x - \theta$ (i.e. $\theta = \mathbb{E}_\mu(X)$), on a

Théorème (Owen, 1991)

Si $\theta_0 = \mathbb{E}_\mu(X)$ et $\Sigma = \text{Cov}_\mu(X)$ est de rang $q \geq 1$, alors

$$-2 \log ELR(\theta_0) \xrightarrow{\mathcal{L}} \chi^2(q)$$

donc

$$C_{r,n} := \{ \theta \in \mathbb{R}^p / -2 \log ELR(\theta) \leq r \}$$

est un convexe de \mathbb{R}^p qui constitue une région de confiance de niveau asymptotique $1 - \alpha = \mathbb{P}[\chi^2(q) \leq r]$ pour θ .

De plus, si $\mathbb{E}(\|X\|^4) < \infty$, alors

$$|\mathbb{P}_\mu(\mu \in C_{r,n}) - \mathbb{P}(\chi^2(q) \leq r)| = o(n^{-1/2})$$

Rem : $-2 \log ELR(\theta_0) = 2 \sum_{i=1}^n \log(1 + \langle \hat{\lambda}(\theta_0), g(X_i, \theta_0) \rangle)$

Analogie vraisemblance empirique du LRT

Dans le cas où $g(x, \theta) = x - \theta$ (i.e. $\theta = \mathbb{E}_\mu(X)$), on a

Théorème (Owen, 1991)

Si $\theta_0 = \mathbb{E}_\mu(X)$ et $\Sigma = \text{Cov}_\mu(X)$ est de rang $q \geq 1$, alors

$$-2 \log ELR(\theta_0) \xrightarrow{\mathcal{L}} \chi^2(q)$$

donc

$$C_{r,n} := \{ \theta \in \mathbb{R}^p / -2 \log ELR(\theta) \leq r \}$$

est un convexe de \mathbb{R}^p qui constitue une région de confiance de niveau asymptotique $1 - \alpha = \mathbb{P}[\chi^2(q) \leq r]$ pour θ .

De plus, si $\mathbb{E}(\|X\|^4) < \infty$, alors

$$|\mathbb{P}_\mu(\mu \in C_{r,n}) - \mathbb{P}(\chi^2(q) \leq r)| = o(n^{-1/2})$$

Rem : $-2 \log ELR(\theta_0) = 2 \sum_{i=1}^n \log(1 + \langle \hat{\lambda}(\theta_0), g(X_i, \theta_0) \rangle)$

Analogie vraisemblance empirique du LRT

Dans le cas où $g(x, \theta) = x - \theta$ (i.e. $\theta = \mathbb{E}_\mu(X)$), on a

Théorème (Owen, 1991)

Si $\theta_0 = \mathbb{E}_\mu(X)$ et $\Sigma = \text{Cov}_\mu(X)$ est de rang $q \geq 1$, alors

$$-2 \log ELR(\theta_0) \xrightarrow{\mathcal{L}} \chi^2(q)$$

donc

$$C_{r,n} := \{ \theta \in \mathbb{R}^p / -2 \log ELR(\theta) \leq r \}$$

est un convexe de \mathbb{R}^p qui constitue une région de confiance de niveau asymptotique $1 - \alpha = \mathbb{P}[\chi^2(q) \leq r]$ pour θ .

De plus, si $\mathbb{E}(\|X\|^4) < \infty$, alors

$$\left| \mathbb{P}_\mu(\mu \in C_{r,n}) - \mathbb{P}(\chi^2(q) \leq r) \right| = o(n^{-1/2})$$

Rem : $-2 \log ELR(\theta_0) = 2 \sum_{i=1}^n \log(1 + \langle \hat{\lambda}(\theta_0), g(X_i, \theta_0) \rangle)$

Pour le modèle à condition de moments général, où $q \geq k$

$$\mu \in \mathcal{P} = \cup_{\theta} \mathcal{P}_{\theta} \quad \text{où} \quad \mathcal{P}_{\theta} := \{ \nu / \mathbb{E}_{\nu}(g(X, \theta)) = 0 \}$$

Qin & Lawless (1994) prouvent la normalité asymptotique

$$\sqrt{n}(\hat{\theta}_{EL} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (D'SD)^{-1}) \quad (\text{quand } \mu \in \mathcal{P}_{\theta})$$

où $\begin{cases} D = \mathbb{E}_{\mu}[\nabla_{\theta} g(X, \theta(\mu))] + \text{conditions de régularité sur } g \\ S = \mathbb{E}_{\mu}[g(X, \theta(\mu))g(X, \theta(\mu))^t] \text{ supposée définie positive.} \end{cases}$

Ils établissent également la convergence en loi

$$-2 \log ELR(\theta) \xrightarrow{\mathcal{L}} \chi^2(k)$$

quand $\mathbb{E}_{\mu}[g(X, \theta)] = 0$, et donnent d'autres résultats dans le cadre particulier où $q > k$ (risque de suridentification de θ).

Pour le modèle à condition de moments général, où $q \geq k$

$$\mu \in \mathcal{P} = \cup_{\theta} \mathcal{P}_{\theta} \quad \text{où} \quad \mathcal{P}_{\theta} := \{ \nu / \mathbb{E}_{\nu}(g(X, \theta)) = 0 \}$$

Qin & Lawless (1994) prouvent la normalité asymptotique

$$\sqrt{n}(\hat{\theta}_{EL} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (D'SD)^{-1}) \quad (\text{quand } \mu \in \mathcal{P}_{\theta})$$

où $\begin{cases} D = \mathbb{E}_{\mu}[\nabla_{\theta} g(X, \theta(\mu))] + \text{conditions de régularité sur } g \\ S = \mathbb{E}_{\mu}[g(X, \theta(\mu))g(X, \theta(\mu))^t] \text{ supposée définie positive.} \end{cases}$

Ils établissent également la convergence en loi

$$-2 \log ELR(\theta) \xrightarrow{\mathcal{L}} \chi^2(k)$$

quand $\mathbb{E}_{\mu}[g(X, \theta)] = 0$, et donnent d'autres résultats dans le cadre particulier où $q > k$ (risque de suridentification de θ).

- 1 Rappels de statistique paramétrique
 - Modèles paramétriques et vraisemblance
 - Rapport des maxima de vraisemblance
- 2 **Vraisemblance empirique (cadre iid) : deux approches**
 - Approche NPML (max de vraisemblance non-paramétrique)
 - **Approche Minimum de Contraste**
- 3 Cadres d'application, avantages, exemples

Rappels sur la divergence de Kullback-Leibler

Si μ et ν sont deux mesures de probabilité sur un même espace, on définit

$$K(\mu, \nu) = \begin{cases} \int \frac{d\mu}{d\nu} \log \frac{d\mu}{d\nu} d\nu = - \int \log \frac{d\nu}{d\mu} d\mu & \text{si } \mu \ll \nu \\ \infty & \text{sinon} \end{cases}$$

encore appelée entropie relative de μ par rapport à ν .

μ_n désignant la mesure empirique associée à $(X_i)_{i=1..n}$
on a donc

$$K(\mu_n, \nu) = \begin{cases} -\frac{1}{n} \sum_{i=1}^n \log(n p_i) & \text{si } \nu = \sum_{i=1}^n p_i \delta_{X_i} \text{ et } p_i > 0 (\forall i) \\ \infty & \text{sinon} \end{cases}$$

Rappels sur la divergence de Kullback-Leibler

Si μ et ν sont deux mesures de probabilité sur un même espace, on définit

$$K(\mu, \nu) = \begin{cases} \int \frac{d\mu}{d\nu} \log \frac{d\mu}{d\nu} d\nu = - \int \log \frac{d\nu}{d\mu} d\mu & \text{si } \mu \ll \nu \\ \infty & \text{sinon} \end{cases}$$

encore appelée entropie relative de μ par rapport à ν .

μ_n désignant la mesure empirique associée à $(X_i)_{i=1..n}$

on a donc

$$K(\mu_n, \nu) = \begin{cases} -\frac{1}{n} \sum_{i=1}^n \log(np_i) & \text{si } \nu = \sum_{i=1}^n p_i \delta_{X_i} \text{ et } p_i > 0 (\forall i) \\ \infty & \text{sinon} \end{cases}$$

L'EMVE vu comme un estimateur du minimum de contraste

On a

$$\begin{aligned}
 & -2 \log ELR(\theta) \\
 = & -2n \sup \left\{ \frac{1}{n} \sum_1^n \log(np_i) / (p_i) \in \mathcal{S}_n \text{ et } \sum_1^n p_i g(X_i, \theta) = 0 \right\} \\
 = & 2n \inf \left\{ K(\mu_n, \nu) / \mathbb{E}_\nu(g(X, \theta)) = 0 \right\} \\
 = & 2n \inf_{\nu \in \mathcal{P}_\theta} K(\mu_n, \nu)
 \end{aligned}$$

Ainsi

maximiser $\mathcal{EL}(\theta)$ = minimiser $K(\mu_n, \nu)$ pour $\nu \in \mathcal{P}$

↪ d'autres choix de divergence ?

avantage de celle de Kullback ?

L'EMVE vu comme un estimateur du minimum de contraste

On a

$$\begin{aligned} & -2 \log ELR(\theta) \\ = & -2n \sup \left\{ \frac{1}{n} \sum_1^n \log(np_i) / (p_i) \in \mathcal{S}_n \text{ et } \sum_1^n p_i g(X_i, \theta) = 0 \right\} \\ = & 2n \inf \left\{ K(\mu_n, \nu) / \mathbb{E}_\nu(g(X, \theta)) = 0 \right\} \\ = & 2n \inf_{\nu \in \mathcal{P}_\theta} K(\mu_n, \nu) \end{aligned}$$

Ainsi

maximiser $\mathcal{EL}(\theta)$ = minimiser $K(\mu_n, \nu)$ pour $\nu \in \mathcal{P}$

↪ d'autres choix de divergence ?
avantage de celle de Kullback ?

L'EMVE vu comme un estimateur du minimum de contraste

On a

$$\begin{aligned}
 & -2 \log ELR(\theta) \\
 = & -2n \sup \left\{ \frac{1}{n} \sum_1^n \log(np_i) / (p_i) \in \mathcal{S}_n \text{ et } \sum_1^n p_i g(X_i, \theta) = 0 \right\} \\
 = & 2n \inf \left\{ K(\mu_n, \nu) / \mathbb{E}_\nu(g(X, \theta)) = 0 \right\} \\
 = & 2n \inf_{\nu \in \mathcal{P}_\theta} K(\mu_n, \nu)
 \end{aligned}$$

Ainsi

maximiser $\mathcal{EL}(\theta)$ = minimiser $K(\mu_n, \nu)$ pour $\nu \in \mathcal{P}$

↪ d'autres choix de divergence ?
 avantage de celle de Kullback ?

- 1 Rappels de statistique paramétrique
 - Modèles paramétriques et vraisemblance
 - Rapport des maxima de vraisemblance
- 2 Vraisemblance empirique (cadre iid) : deux approches
 - Approche NPMLE (max de vraisemblance non-paramétrique)
 - Approche Minimum de Contraste
- 3 Cadres d'application, avantages, exemples

Estimation de paramètres fonctionnels

Pour l'estimation de paramètres fonctionnels généraux

$$\theta = T(\mu)$$

(i.e. pas seulement définis par $\mathbb{E}_\mu(g(X, \theta)) = 0$), il suffit de poser

$$\begin{aligned} \mathcal{EL}(\theta) &:= \sup_{\nu \ll \mu_n} \left\{ \prod_{i=1}^n (d\nu/d\mu_n)(X_i) / T(\nu) = \theta \right\} \\ -\frac{1}{n} \log ELR(\theta) &= \sup_{\nu \ll \mu_n} \left\{ K(\mu_n, \nu) / T(\nu) = \theta \right\} \end{aligned}$$

Dans ce cadre, des travaux existent sous l'hypothèse de la différentiabilité au sens de Fréchet ou de Hadamard, de la fonctionnelle T (voir Owen (1988), Bertail (2006) par exemple).

Estimation de paramètres fonctionnels

Pour l'estimation de paramètres fonctionnels généraux

$$\theta = T(\mu)$$

(i.e. pas seulement définis par $\mathbb{E}_\mu(g(X, \theta)) = 0$), il suffit de poser

$$\begin{aligned}\mathcal{EL}(\theta) &:= \sup_{\nu \ll \mu_n} \left\{ \prod_{i=1}^n (d\nu/d\mu_n)(X_i) / T(\nu) = \theta \right\} \\ -\frac{1}{n} \log ELR(\theta) &= \sup_{\nu \ll \mu_n} \left\{ K(\mu_n, \nu) / T(\nu) = \theta \right\}\end{aligned}$$

Dans ce cadre, des travaux existent sous l'hypothèse de la différentiabilité au sens de Fréchet ou de Hadamard, de la fonctionnelle T (voir Owen (1988), Bertail (2006) par exemple).

Quelques cadres d'application de la méthode EL

- ▶ modèles de mélange
 - ▶ estimation de quantiles
 - ▶ modèles linéaires, GLMs, partiellement linéaires
 - ▶ modèles de régression non-paramétrique
 - ▶ tests non-paramétrique d'adéquation, de symétrie, ...
 - ▶ statistique de données censurées
 - ▶ théorie des sondages
 - ▶ ...

Intérêt de la méthode

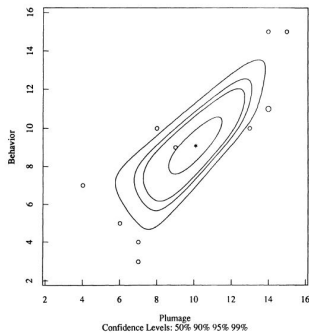
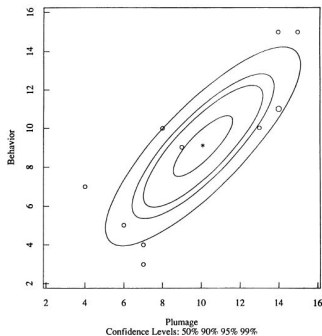
- propriétés du LRT mais en semi-paramétrique !
- les régions de confiance résultantes ont un "degré d'asymétrie" automatiquement adapté aux données

Intérêt de la méthode

- propriétés du LRT mais en semi-paramétrique !
- les régions de confiance résultantes ont un “degré d'asymétrie” automatiquement adapté aux données

Intérêt de la méthode

- propriétés du LRT mais en semi-paramétrique !
- les régions de confiance résultantes ont un “degré d'asymétrie” automatiquement adapté aux données



Intérêt de la méthode (suite)

→ les régions de confiance résultantes peuvent être “corrigées au sens de Bartlett”

→ efficacité asymptotique non paramétrique atteinte par l'EMVE et par des fonctionnelles de l'EMVE de μ

$$\xi = \mathbb{E}_\mu(\Psi(\theta(\mu), X)) \rightsquigarrow \hat{\xi} = \sum_{i=1}^n \hat{p}_i \Psi(\hat{\theta}_{EL}, X_i)$$

→ prise en compte facile de contraintes sur le paramètre

Intérêt de la méthode (suite)

- les régions de confiance résultantes peuvent être “corrigées au sens de Bartlett”
- efficacité asymptotique non paramétrique atteinte par l'EMVE et par des fonctionnelles de l'EMVE de μ

$$\xi = \mathbb{E}_\mu(\Psi(\theta(\mu), X)) \rightsquigarrow \hat{\xi} = \sum_{i=1}^n \hat{p}_i \Psi(\hat{\theta}_{EL}, X_i)$$

- prise en compte facile de contraintes sur le paramètre

Intérêt de la méthode (suite)

- les régions de confiance résultantes peuvent être “corrigées au sens de Bartlett”
- efficacité asymptotique non paramétrique atteinte par l'EMVE et par des fonctionnelles de l'EMVE de μ

$$\xi = \mathbb{E}_{\mu}(\Psi(\theta(\mu), X)) \rightsquigarrow \hat{\xi} = \sum_{i=1}^n \hat{p}_i \Psi(\hat{\theta}_{EL}, X_i)$$

- prise en compte facile de contraintes sur le paramètre

Intérêt de la méthode (suite)

- les régions de confiance résultantes peuvent être “corrigées au sens de Bartlett”
- efficacité asymptotique non paramétrique atteinte par l'EMVE et par des fonctionnelles de l'EMVE de μ

$$\xi = \mathbb{E}_{\mu}(\Psi(\theta(\mu), X)) \rightsquigarrow \hat{\xi} = \sum_{i=1}^n \hat{p}_i \Psi(\hat{\theta}_{EL}, X_i)$$

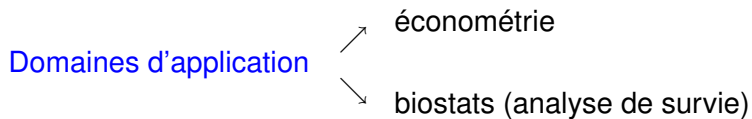
→ prise en compte facile de contraintes sur le paramètre

Intérêt de la méthode (suite)

- les régions de confiance résultantes peuvent être “corrigées au sens de Bartlett”
- efficacité asymptotique non paramétrique atteinte par l'EMVE et par des fonctionnelles de l'EMVE de μ

$$\xi = \mathbb{E}_{\mu}(\Psi(\theta(\mu), X)) \rightsquigarrow \hat{\xi} = \sum_{i=1}^n \hat{p}_i \Psi(\hat{\theta}_{EL}, X_i)$$

- prise en compte facile de contraintes sur le paramètre



Quelques références



Art Owen
Empirical Likelihood.
Chapman & Hall/CRC, 2001.



P. Bertail
Empirical likelihood in some semiparametric models.
Bernoulli, 12 (2) : 299–331, 2006.



J.H. Einmal & I.W. McKeage
Empirical likelihood based hypothesis testing.
Bernoulli, 9(2) : 267–290, 2003.



S. Chen & P. Hall
Smoothed empirical likelihood confidence intervals for quantiles.
Annals of Statistics, 21(3) : 1166–1181, 1993.



Y. Kitamura
Empirical likelihood methods in econometrics : theory and practice.
Cowles Foundation discussion paper, Working Paper n°1569, 2006.



Y. Kitamura
Empirical likelihood methods with weakly dependent processes.
Annals of Statistics, 25(5) : 2084–2102, 1997.



A. B. Owen
Empirical likelihood ratio confidence intervals for a single functional.
Biometrika, 75(2) : 237–249, 1988.



A. B. Owen
Empirical likelihood ratio confidence regions.
Annals of Statistics, 18(1) : 90–120, 1990.



A. B. Owen
Empirical likelihood for linear models.
Annals of Statistics, 19(4) : 1725–1747, 1991.



Y. S. Qin & J. Lawless
Empirical likelihood and general estimating equations.
Annals of Statistics, 22(1) : 300–325, 1994.



G. Qin & M. Tsao
Empirical likelihood based inference for the derivative of the nonparametric regression function
Bernoulli, 11(4) : 715–735, 2005.



S. S. Wilks
The large-sample distribution of the likelihood ratio for testing composite hypotheses.
Annals of Mathematical Statistics, 9 : 60–62, 1938.