Introduction au Machine Learning

C2 - Méthodes linéaires pour la régression

Lucas Gnecco Heredia

LAMSADE - Université Paris-Dauphine
lucas.gnecco-heredia@dauphine.psl.eu - C602

* *Remerciement spécial à Florian Yger*

Plan

1. Rappel sur le modèle linéaire et les moindres carrés (ESL Ch 3, Sec 3.1 - 3.2)
   - Propriétés de l'estimateur de moindres carrés
2. Sélection de variables (ESL Sec 3.3)
3. Régularisation (ESL Sec 3.4) (IML Ch 6)
   - Ridge
   - Lasso

1. Rappel sur le modèle linéaire et les moindres carrés (ESL Ch 3, Sec 3.1 - 3.2)
   - Propriétés de l'estimateur de moindres carrés
2. Sélection de variables (ESL Sec 3.3)
3. Régularisation (ESL Sec 3.4) (IML Ch 6)

Modèle linéaire

On parle de modèle linéaire quand on modélise la relation entre $\mathcal{X}$ et $\mathcal{Y}$ avec une fonction linéaire en $\mathcal{X}$:

$$y \approx \beta_0 + \sum_{k=1}^{p} x^k \beta^k = \beta_0 + x^T \beta$$

Modèle linéaire

On parle de modèle linéaire quand on modélise la relation entre $\mathcal{X}$ et $\mathcal{Y}$ avec une fonction linéaire en $\mathcal{X}$:

$$y \approx \beta_0 + \sum_{k=1}^{p} x^k \beta^k = \beta_0 + x^T \beta$$

Souvent on va augmenter x avec une composante constante de 1 et inclure aussi l'intercepte β_0 dans le vecteur β. Ceci donne une expression plus compacte pour le même modèle

$$y \approx x^T \beta$$

Modèle linéaire

On parle de modèle linéaire quand on modélise la relation entre $\mathcal{X}$ et $\mathcal{Y}$ avec une fonction linéaire en $\mathcal{X}$:

$$y \approx \beta_0 + \sum_{k=1}^{p} x^k \beta^k = \beta_0 + x^T \beta$$

Souvent on va augmenter x avec une composante constante de 1 et inclure aussi l'intercepte β_0 dans le vecteur β. Ceci donne une expression plus compacte pour le même modèle

$$y \approx x^T \beta$$

Apprendre un modèle linéaire revient à trouver le β.

Pourquoi un modèle linéaire ?

- Simple à apprendre et à comprendre, très étudié.

Pourquoi un modèle linéaire ?

- Simple à apprendre et à comprendre, très étudié.
- Interprétable : l'effet de chaque variable sur l'output et facile à interpréter et à comparer

Pourquoi un modèle linéaire ?

- Simple à apprendre et à comprendre, très étudié.
- Interprétable : l'effet de chaque variable sur l'output et facile à interpréter et à comparer
- Peut être complexifié en créant une expansion des variables tout en gardant la simplicité du modèle linéaire (exemple : régression polynomiale)

Les variables à utiliser

Dans notre problème d'apprentissage, nos variables $\mathcal{X}$ sont déjà définies, mais on peut toujours enrichir notre ensemble de variables avec, par exemple :

Les variables à utiliser

Dans notre problème d'apprentissage, nos variables $\mathcal{X}$ sont déjà définies, mais on peut toujours enrichir notre ensemble de variables avec, par exemple :

- Transformations des variables originales (log, racine carré, carré)

Les variables à utiliser

Dans notre problème d'apprentissage, nos variables $\mathcal{X}$ sont déjà définies, mais on peut toujours enrichir notre ensemble de variables avec, par exemple :

- Transformations des variables originales (log, racine carré, carré)
- Pour les variables qualitatives, un encodage peut être proposé. Le choix d'encodage dépend du type de variable.

Les variables à utiliser

Dans notre problème d'apprentissage, nos variables $\mathcal{X}$ sont déjà définies, mais on peut toujours enrichir notre ensemble de variables avec, par exemple :

- Transformations des variables originales (log, racine carré, carré)
- Pour les variables qualitatives, un encodage peut être proposé. Le choix d'encodage dépend du type de variable.
- Interactions entre variables ($x_i \cdot x_j$)

Apprendre un modèle linéaire

Étant donné un jeu de données $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \subseteq \mathbb{R}^p \times \mathbb{R}$, ils existent plusieurs méthodes pour apprendre un modèle linéaire, dont le plus populaire est les moindres carrés

1. De l'anglais Residual Sum of Squares

Apprendre un modèle linéaire

Étant donné un jeu de données $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \subseteq \mathbb{R}^p \times \mathbb{R}$, ils existent plusieurs méthodes pour apprendre un modèle linéaire, dont le plus populaire est les moindres carrés

On cherche $\hat{\beta}$ qui minimise la fonction de coût quadratique $RSS(\beta)$ [1] sur l'ensemble de données

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} RSS(\beta) = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - x_i^T \beta \right)^2$$

1. De l'anglais Residual Sum of Squares

Moindres carrés en forme matricielle

On construit la matrice X, de taille $N \times (p+1)$, en mettant comme ligne i le vecteur x_i^T.
Y, de taille $N \times 1$, contient les y_i.

Moindres carrés en forme matricielle

On construit la matrice X, de taille $N \times (p+1)$, en mettant comme ligne i le vecteur x_i^T.
Y, de taille $N \times 1$, contient les y_i.
Avec cette notation, on peut réécrire les expressions importantes :

- La relation entre les données x_i, β et y_i, avant décrite pour chaque point comme $x_i^T \beta = y_i$, devient une seule expression pour tous les points : $Y = X\beta$

Moindres carrés en forme matricielle

On construit la matrice X, de taille $N \times (p+1)$, en mettant comme ligne i le vecteur x_i^T.
Y, de taille $N \times 1$, contient les y_i.
Avec cette notation, on peut réécrire les expressions importantes :

- La relation entre les données x_i, β et y_i, avant décrite pour chaque point comme $x_i^T \beta = y_i$, devient une seule expression pour tous les points : $Y = X\beta$
- La somme d'erreurs au carré, *c.-à-d.* $\sum_{i=1}^{N} \left(y_i - x_i^T \beta\right)^2$ devient $(Y - X\beta)^T (Y - X\beta)$, ou aussi $\|(Y - X\beta)\|_2^2$

Solution des moindres carrés

Le problème de moindres carrés, avec la notation matricielle, admet la solution suivante

$$\hat{\beta} = \left(X^T X\right)^+ X^T Y$$

A^+ est la pseudo-inverse de Moore-Penrose de A. Si A est inversible, A^+ est l'inverse de A.

1. Rappel sur le modèle linéaire et les moindres carrés (ESL Ch 3, Sec 3.1 - 3.2)
   - Propriétés de l'estimateur de moindres carrés
2. Sélection de variables (ESL Sec 3.3)
3. Régularisation (ESL Sec 3.4) (IML Ch 6)

Hypothèses

On fera l'hypothèse que le modèle linéaire est le modèle correct pour l'espérance de Y, et que les déviations observées suivent une distribution normale centrée avec variance σ^2

Hypothèses

On fera l'hypothèse que le modèle linéaire est le modèle correct pour l'espérance de Y, et que les déviations observées suivent une distribution normale centrée avec variance σ^2

$$Y = X^T \beta + \epsilon, \ \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Hypothèses

On fera l'hypothèse que le modèle linéaire est le modèle correct pour l'espérance de Y, et que les déviations observées suivent une distribution normale centrée avec variance σ^2

$$Y = X^T\beta + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Ces hypothèses nous donneront des résultats sur notre estimateur (moindres carrés) $\hat{\beta}$

Propriétés de $\hat{\beta}$

- $\hat{\beta} \sim \mathcal{N}(\beta, (X^T X)^{-1} \sigma^2)$
  - Chaque $\hat{\beta}_j$ est un estimateur non biaisé de β_j
  - Chaque $\hat{\beta}_j$ suit une loi normale !
- On doit estimer σ^2 aussi. On a l'estimateur non biaisé suivant qui suit une loi χ^2 :

$$\hat{\sigma}^2 = \frac{1}{N - p - 1} \sum_{i}^{N} (y_i - \hat{y}_i)^2$$

- $\hat{\beta}$ and $\hat{\sigma}$ sont indépendants

Propriétés de $\hat{\beta}$ (cont)

Nous pouvons faire des tests d'hypothèse sur la nullité de $\hat{\beta}_j$

- $\hat{\beta}_j \sim \mathcal{N}(\beta_j, v_i\sigma^2)$ où v_i est le i^{eme} élément de la diagonale de $(X^TX)^{-1}$
- On standardise avec l'estimateur de σ^2, *i.e* $z_j = \frac{\hat{\beta}_j - 0}{\sqrt{v_i}\hat{\sigma}}$ et alors z_j suit une loi t_{N-p-1}
- On fait un test d'hypothèse traditionnel !

Propriétés de $\hat{\beta}$ (cont)

D'autres tests nous permettent de tester l'impact d'un groupe de variables. Pour ceci, on utilise le statistique F (lié à la distribution de Fisher–Snedecor)

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)}$$

où RSS_i est la somme d'erreurs carrées après avoir appris le modèle avec $p_i + 1$ paramètres, $p_1 > p_0$, avec $p_1 - p_0$ paramètres mis à zéro.

Conclusion (pour l'instant)

Avec un modèle linéaire, la méthode de moindres carrés et certaines hypothèses sur la relation entre $\mathcal{X}$ et $\mathcal{Y}$ nous pouvons déjà :

Conclusion (pour l'instant)

Avec un modèle linéaire, la méthode de moindres carrés et certaines hypothèses sur la relation entre $\mathcal{X}$ et $\mathcal{Y}$ nous pouvons déjà :

1. Apprendre un modèle simple qui capture une relation linéaire (ou non) entre $\mathcal{X}$ et $\mathcal{Y}$

Conclusion (pour l'instant)

Avec un modèle linéaire, la méthode de moindres carrés et certaines hypothèses sur la relation entre $\mathcal{X}$ et $\mathcal{Y}$ nous pouvons déjà :

1. Apprendre un modèle simple qui capture une relation linéaire (ou non) entre $\mathcal{X}$ et $\mathcal{Y}$
2. Avoir des hypothèses sur l'estimateur résultant qui nous permettent d'interpréter son impact sur le résultat et comparer les variables initiales.

Conclusion (pour l'instant)

Avec un modèle linéaire, la méthode de moindres carrés et certaines hypothèses sur la relation entre $\mathcal{X}$ et $\mathcal{Y}$ nous pouvons déjà :

1. Apprendre un modèle simple qui capture une relation linéaire (ou non) entre $\mathcal{X}$ et $\mathcal{Y}$
2. Avoir des hypothèses sur l'estimateur résultant qui nous permettent d'interpréter son impact sur le résultat et comparer les variables initiales.

Peut-on faire mieux ? Avons-nous une méthode parfaite pour faire la régression ?

Limites de la méthode de moindres carrés

Deux raisons pour ne pas rester satisfaits avec le simple moindres carrés :

1. Les moindres carrés donnent un estimateur non biaisé, mais avec une grande variance ! Ceci peut endommager la performance du modèle.
2. On veut plus d'interprétabilité ! Et plus de sélection de variables !

1 Rappel sur le modèle linéaire et les moindres carrés (ESL Ch 3, Sec 3.1 - 3.2)

2 Sélection de variables (ESL Sec 3.3)

3 Régularisation (ESL Sec 3.4) (IML Ch 6)

Best subset, le cas idéal

Tester tous les sous ensembles possibles de k variables et prendre celui qui minimise l'erreur. Tester après pour différentes valeurs de k

Best subset, le cas idéal

Tester tous les sous ensembles possibles de k variables et prendre celui qui minimise l'erreur. Tester après pour différentes valeurs de k

Très coûteux !

Pour un k fixé, on a $\mathcal{O}(p^k)$ sous ensembles de taille k. Pour chaque sous ensemble, il faut faire toute la procédure !

Best subset, le cas idéal

Tester tous les sous ensembles possibles de k variables et prendre celui qui minimise l'erreur. Tester après pour différentes valeurs de k

Très coûteux !

Pour un k fixé, on a $\mathcal{O}(p^k)$ sous ensembles de taille k. Pour chaque sous ensemble, il faut faire toute la procédure !
Si on teste pour tout sous ensemble, c'est 2^p régressions !

Best subset, le cas idéal

Tester tous les sous ensembles possibles de k variables et prendre celui qui minimise l'erreur. Tester après pour différentes valeurs de k

Très coûteux !

Pour un k fixé, on a $\mathcal{O}(p^k)$ sous ensembles de taille k. Pour chaque sous ensemble, il faut faire toute la procédure !
Si on teste pour tout sous ensemble, c'est 2^p régressions !

Il faut trouver une méthode qui marche pour p grand...

Best subset, le cas idéal

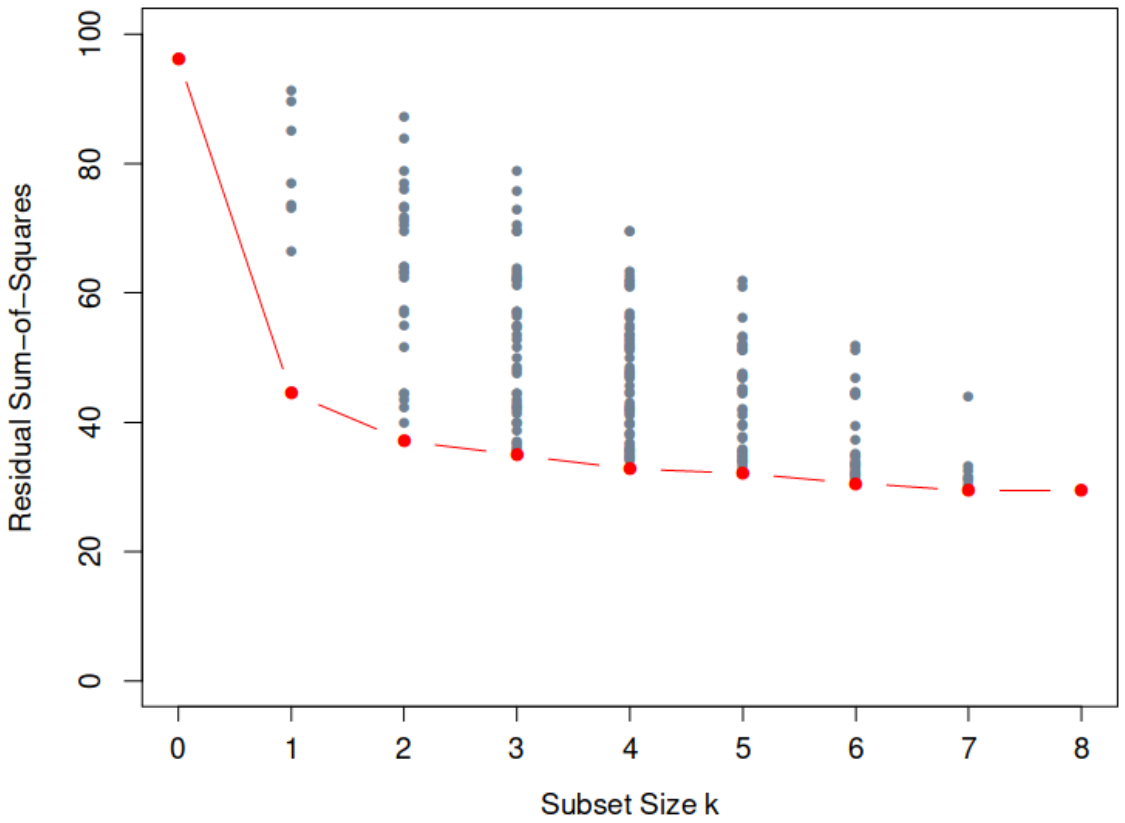

Figure – [Has+09]

Forward Stepwise selection

Commencer avec une fonction constante et ajouter des variables, une à la fois, pour augmenter la performance.

Forward Stepwise selection

Commencer avec une fonction constante et ajouter des variables, une à la fois, pour augmenter la performance.

- Marche pour p grand grâce aux algorithmes qui permettent de passer d'une étape à la prochaine de manière efficace
- Comme on ne teste pas toutes les options, forcement on aura une solution sous-optimale, mais avec moins de variance. (+ bias, - variance)
- Fonctionne même si $p > N$

Backward Stepwise selection

Commencer avec le modèle complet, rejeter à chaque fois la variable avec le moindre impact (en termes de Z-scores que l'on a vus).

Backward Stepwise selection

Commencer avec le modèle complet, rejeter à chaque fois la variable avec le moindre impact (en termes de Z-scores que l'on a vus).

- Fonctionne pour $N > p$

Backward Stepwise selection

Commencer avec le modèle complet, rejeter à chaque fois la variable avec le moindre impact (en termes de Z-scores que l'on a vus).

- Fonctionne pour $N > p$
- A nouveau + biais, - variance

Bilan

On a vu très rapidement deux méthodes pour apprendre un modèle linéaire en faisant de la sélection de variables pour avoir plus de biais, mais moins de variance, et plus d'interprétabilité.

On a vu qu'une petite partie

Ils existent plusieurs autres ! La littérature en statistique est dense et souvent pas très étudié dans ces jours.

1. Rappel sur le modèle linéaire et les moindres carrés (ESL Ch 3, Sec 3.1 - 3.2)

2. Sélection de variables (ESL Sec 3.3)

3. Régularisation (ESL Sec 3.4) (IML Ch 6)
   - Ridge
   - Lasso

Quoi de neuf ?

La sélection de modèle vue avant et très discrète, et parfois très coûteuse.

Quoi de neuf ?

La sélection de modèle vue avant et très discrète, et parfois très coûteuse.
La régularisation (ou méthodes de rétraction ou rétrécissement) offre une alternative plus continue, souvent moins coûteuse.

1. Rappel sur le modèle linéaire et les moindres carrés (ESL Ch 3, Sec 3.1 - 3.2)

2. Sélection de variables (ESL Sec 3.3)

3. Régularisation (ESL Sec 3.4) (IML Ch 6)
   - Ridge
   - Lasso

Régression Ridge

Ridge

Pénalité à la taille des coefficients de $\hat{\beta}$ en norme 2

$$\hat{\beta}^{ridge} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{N} \left(y_i - x_i^T \beta \right) + \lambda \|\beta\|_2^2$$

Régression Ridge

Ridge

Pénalité à la taille des coefficients de $\hat{\beta}$ en norme 2

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - x_i^T \beta \right) + \lambda \|\beta\|_2^2$$

Dans la pratique

Régression Ridge

Ridge

Pénalité à la taille des coefficients de $\hat{\beta}$ en norme 2

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - x_i^T \beta \right) + \lambda \|\beta\|_2^2$$

Dans la pratique

1. Il faut standardiser et centrer les données

Régression Ridge

Ridge

Pénalité à la taille des coefficients de $\hat{\beta}$ en norme 2

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - x_i^T \beta \right) + \lambda \|\beta\|_2^2$$

Dans la pratique

1. Il faut standardiser et centrer les données
2. On estime le biais β_0 séparément avec $\hat{\beta}_0^{ridge} = \frac{1}{N} \sum_1^N y_i$

Régression Ridge

Ridge

Pénalité à la taille des coefficients de $\hat{\beta}$ en norme 2

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - x_i^T \beta \right) + \lambda \|\beta\|_2^2$$

Dans la pratique

1. Il faut standardiser et centrer les données
2. On estime le biais β_0 séparément avec $\hat{\beta}_0^{ridge} = \frac{1}{N} \sum_1^N y_i$
3. Avec les données centrées, on résout le problème linéaire sans biais (p paramètres au lieu de $p+1$) pour les autres coefficients

Régression Ridge (solution)

Somme d'erreurs carrées

$$RSS(\beta, \lambda) = (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$$

Très similaire au cas classique

Régression Ridge (solution)

Somme d'erreurs carrées

$$RSS(\beta, \lambda) = (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta$$

Très similaire au cas classique

Solution

$$\hat{\beta}^{ridge} = \left(X^T X + \lambda I\right)^{-1} X^T Y$$

Régression Ridge (solution)

Somme d'erreurs carrées

$$RSS(\beta, \lambda) = (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta$$

Très similaire au cas classique

Solution

$$\hat{\beta}^{ridge} = \left(X^T X + \lambda I\right)^{-1} X^T Y$$

Très similaire au cas classique, mais maintenant la matrice à inverser $X^T X + \lambda I$ est "meilleure"

Régression Ridge (solution)

Somme d'erreurs carrées

$$RSS(\beta, \lambda) = (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta$$

Très similaire au cas classique

Solution

$$\hat{\beta}^{ridge} = \left(X^T X + \lambda I\right)^{-1} X^T Y$$

Très similaire au cas classique, mais maintenant la matrice à inverser $X^T X + \lambda I$ est "meilleure"
Lien avec *PCA*, plus d'importance aux composantes principales...

1. Rappel sur le modèle linéaire et les moindres carrés (ESL Ch 3, Sec 3.1 - 3.2)

2. Sélection de variables (ESL Sec 3.3)

3. Régularisation (ESL Sec 3.4) (IML Ch 6)
   - Ridge
   - Lasso

Régression Lasso

Lasso

Pénalité à la taille des coefficients de $\hat{\beta}$ en norme 1

$$\hat{\beta}^{ridge} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{N} \left(y_i - x_i^T \beta \right) + \lambda \sum_{j=1}^{p} |\beta_j|$$

Régression Lasso

Lasso

Pénalité à la taille des coefficients de $\hat{\beta}$ en norme 1

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - x_i^T \beta \right) + \lambda \sum_{j=1}^{p} |\beta_j|$$

Dans la pratique

Mêmes considérations pratiques que pour Ridge (centrer, standardiser, traiter β_0 séparément)

Régression Lasso (solution)

Cette fois, pas de solution en forme close !

Régression Lasso (solution)

Cette fois, pas de solution en forme close !
Ils existent algorithmes pour calculer la solution de manière efficace, *i.e.* ISTA.

Intuitions

Lasso fait une sélection de variables en promouvant la parcimonie de la solution.

Intuitions sur Ridge et Lasso

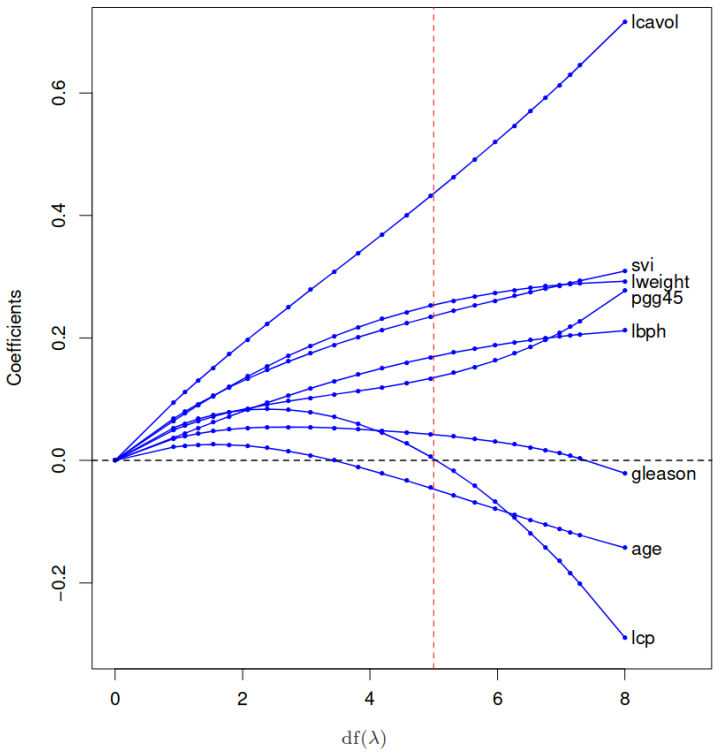

Figure – Ridge [Has+09]

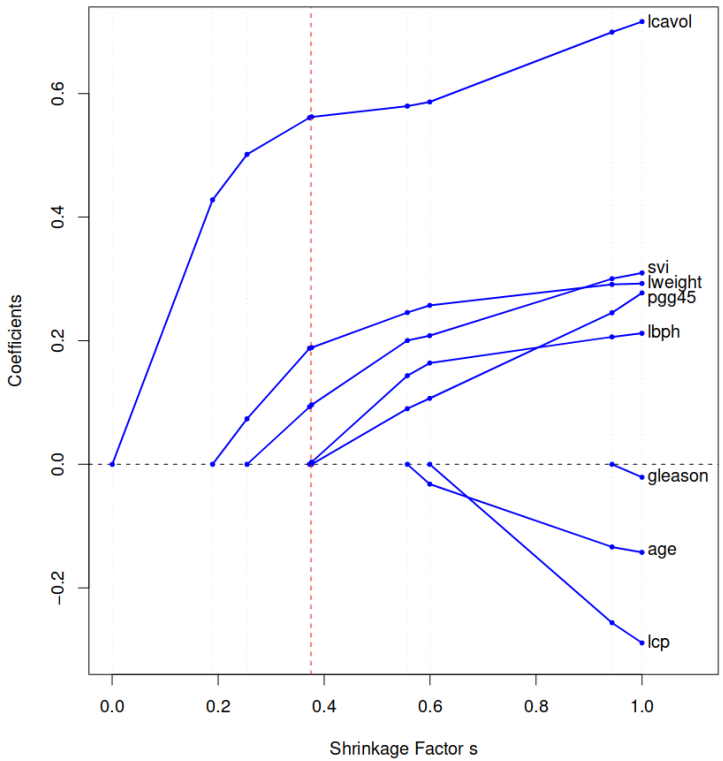

Figure – Lasso [Has+09]

Intuitions sur Ridge et Lasso (cont)

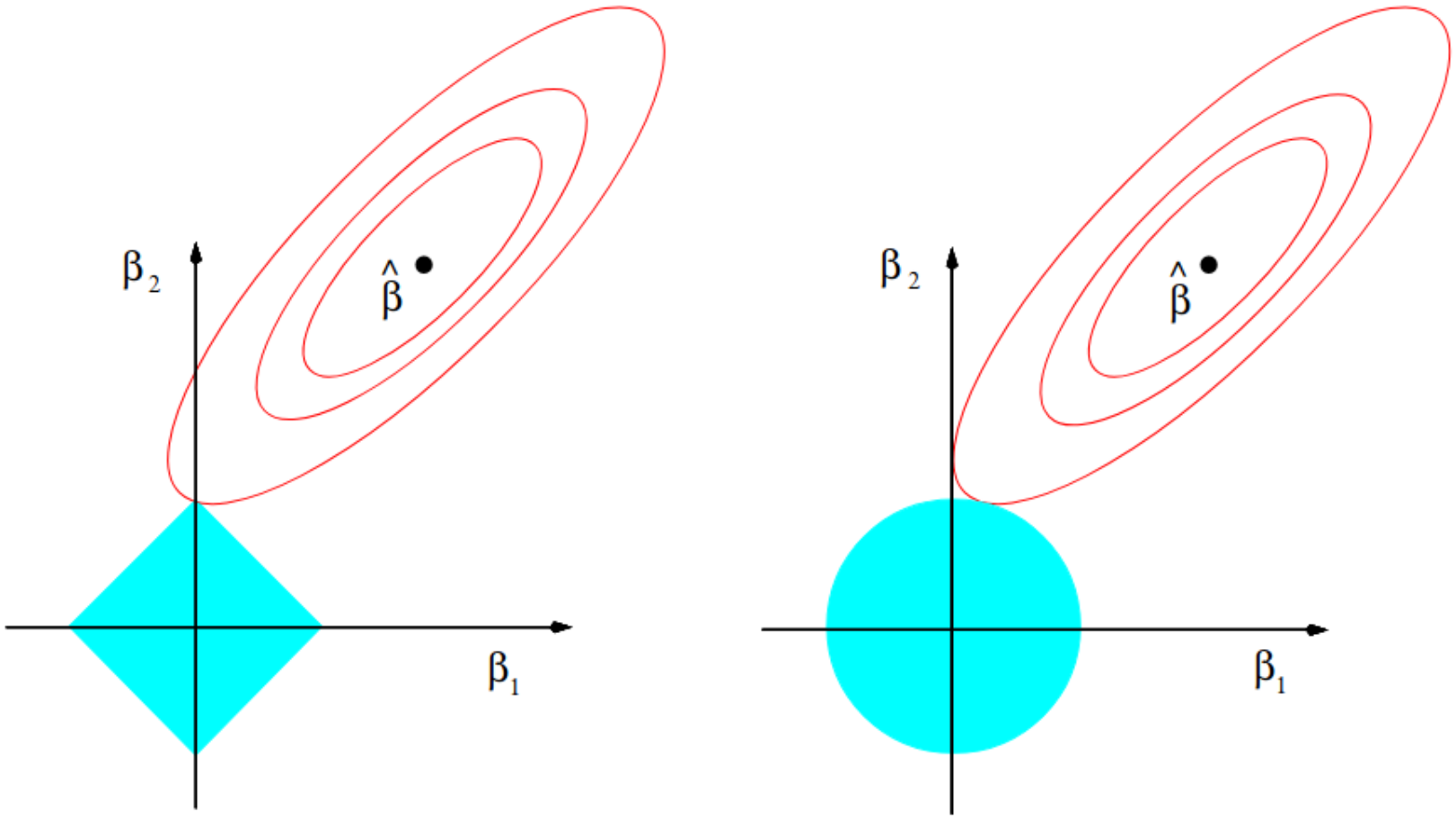

Figure – [Has+09]

Plus sur les normes q

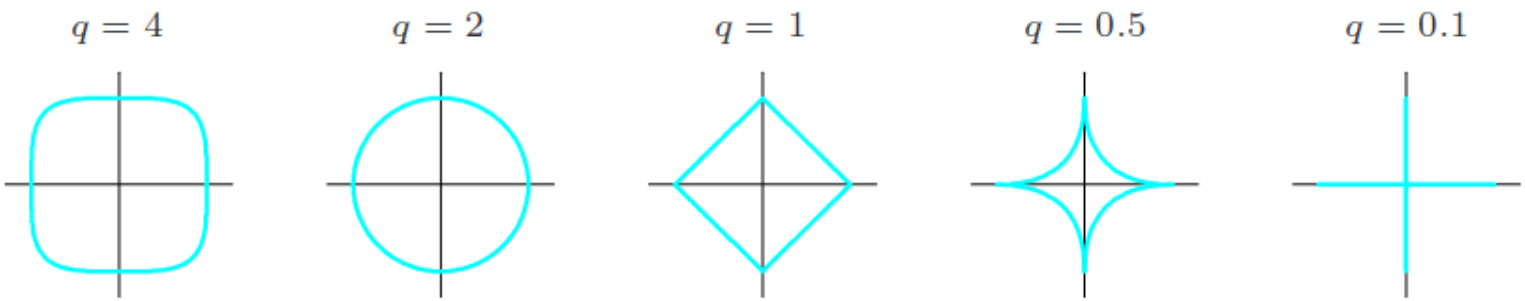

FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

Figure – [Has+09]

Elastic net (intuition)

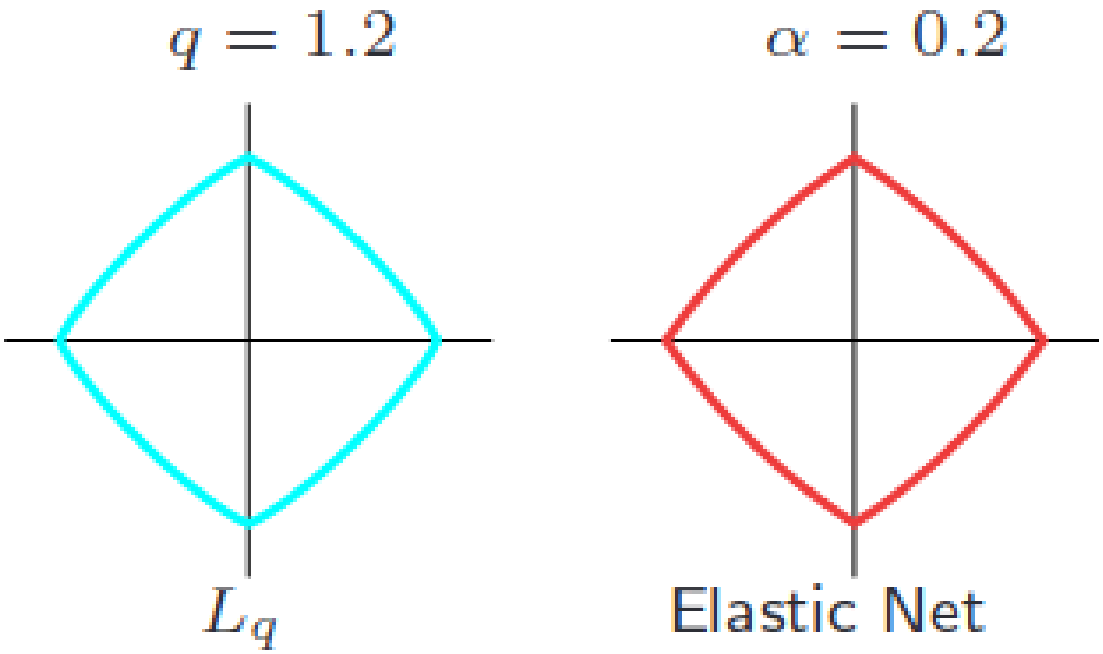

Figure – [Has+09]