Application du shrinkage aux modèles de Cox

Présentation des enjeux et de la théorie

Corentin Choisy (Inserm UMR 1246 SPHERE)

April 26, 2023

SPHERE methodS in Patient-centered outcomes & HEalth ResEarch

Nantes Université

Table des matières

1 Le problème

▶ Le problème

▶ Shrinkage: principe
Shrinkage global
Shrinkage par paramètre
Shrinkage conjoint

▶ Plan de validation

Contexte

1 Le problème

- Deux étapes majeures dans la construction de notre modèle prédictif:
  - Sélection des variables $X_1, ..., X_K$ à inclure dans le modèle final
  - Estimation des paramètres correspondants $\beta_1, ..., \beta_K$
- Modèle final estimé sur les données ayant servi à la sélection des variables
  $\Rightarrow$ X_j a + de chances d'être retenu si β_j est surestimé sur les données que sous-estimé

Biais

1 Le problème

- Deux biais introduits:
  - **Biais de compétition des variables:** $\beta_1, ..., \beta_K$ tendent à être surestimés
  - **Biais de règle d'arrêt:** Des variables sont exclues à tort car leur coefficient est sous-estimé sur les données

Des solutions ?

- Corriger les valeurs de $\beta_1, ..., \beta_K \Rightarrow$ shrinkage
- Estimer le modèle sur d'autres données (par exemple échantillon de validation)
  - **Problème:** Si l'échantillon sur lequel on estime le modèle est un échantillon représentatif de la même population source que l'échantillon ayant servi à la sélection, les coefficients peuvent toujours être surestimés

    $\Rightarrow$ Les coefficients doivent être corrigés $\Rightarrow$ shrinkage

Table des matières

2 Shrinkage: principe

▶ Le problème

▶ Shrinkage: principe
Shrinkage global
Shrinkage par paramètre
Shrinkage conjoint

▶ Plan de validation

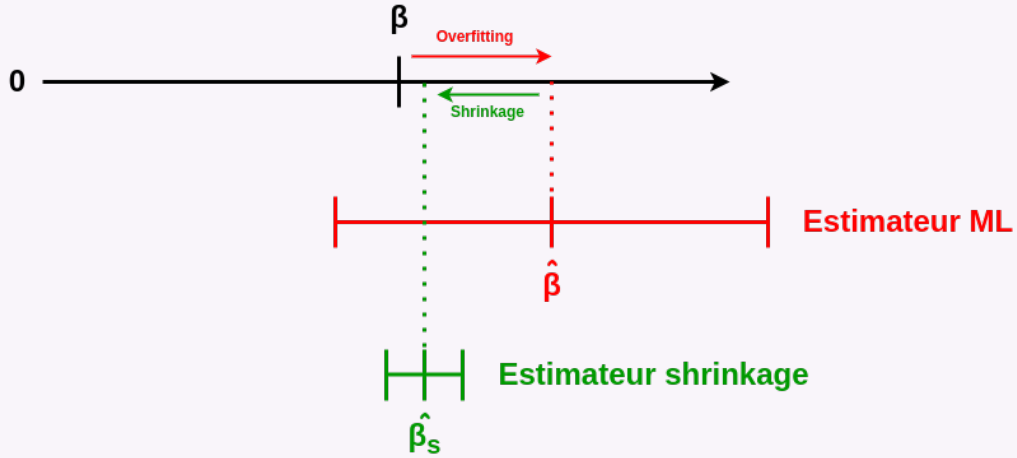
Principe général

2 Shrinkage: principe

- L'estimation des paramètres d'un modèle par maximum de vraisemblance se fait toujours en acceptant un compromis biais-variance.
  - Généralement, on choisit l'estimateur **sans biais à variance minimale**
- Le shrinkage consiste en un choix de compromis différent acceptant un **biais vers 0** de l'estimateur contre une **réduction de la variance**.
- Ici, ce bais est avantageux car nos estimateurs sont enflés du fait du biais de compétition des variables.

Illustration

2 Shrinkage: principe



Shrinkage global

2 Shrinkage: principe

- Le shrinkage global calcule un facteur de réduction des coefficients commun à tous les estimateurs du modèle.
- **Inconvénient:** Ce facteur ”moyen” peut sur-corriger certains coefficients et en sous-corriger d’autres.
- Calculé par cross-validation (voir diapo suivante)

Shrinkage global

2 Shrinkage: principe

Notons $X = (X_1, ..., X_K)$ le vecteur des K variables inclues dans le modèle. On peut écrire le modèle de Cox ainsi, pour chaque individu i:

$$h_i(t) = \lambda_0(t) w_i = \lambda_0(t) e^{X_i \beta}$$

où β est obtenu en maximisant la vraisemblance partielle:

$$\mathcal{L}(\beta) = \prod_{i=1}^{N} \left(\frac{w_i}{\sum_{R_i} w_j} \right)^{d_i}$$

Avec d_i l'indicatrice d'évènement et R_i l'ensemble des individus à risque au temps t_i. Notons $\ell_i(\beta)$ la contribution individuelle de l'individu i à la log-vraisemblance correspondante.

Shrinkage global

2 Shrinkage: principe

On a, avec $\ell_{(-i)}(\beta)$ la log-vraisemblance, maximisée par $\hat{\beta}_{(-i)}$, du modèle avec l'individu i supprimé:

$$\ell_i(\beta) = \ell(\beta) - \ell_{(-i)}(\beta)$$

Avec $\sum_{i=1}^{N} \ell_i(\beta) = \ell(\beta)$, il vient que:

$$cvl = \sum_{i=1}^{N} \ell_i(\hat{\beta}_{(-i)})$$

est une mesure de la capacité prédictive du modèle.

Shrinkage global

2 Shrinkage: principe

Supposons maintenant que l'on calcule un modèle de Cox sur les mêmes données avec pour seule variable dépendante l'indice pronostique $X_i\hat{\beta}_{(-i)}$ pour chaque individu i. En notant $\ell^*(c)$ la log-vraisemblance de ce modèle avec c le coefficient de sa seule variable et $\ell(c)$ la log-vraisemblance du modèle avec pour seule covariable $X_i\hat{\beta}$, on remarque:

- $\ell^*(c) < \ell(c), \forall c$, indiquant que le fit à de nouvelles données est toujours moins bon que celui aux données utilisées pour calculer le modèle
- Par définition de $\hat{\beta}$, $\ell(c)$ est maximisée par $c = 1$ et $\ell(1)$ est une mesure du fit du modèle aux données, tandis que $\ell^*(1)$ mesure la capacité prédictive du modèle.
- La maximisation de $\ell^*(c)$ par $\hat{c}$ permet donc d'obtenir un coefficient c généralement inférieur à 1 permettant de ramener les β à des valeurs reflétant la capacité prédictive du modèle.

Shrinkage global - Résumé

1. Calcul du modèle N fois en retirant un sujet à chaque fois (jackknife), afin de calculer les $\hat{\beta}_{(-i)}$
2. Calcul du modèle de Cox avec pour variable $X_i\hat{\beta}_{(-i)}$. Le coefficient c de ce modèle donne le facteur de shrinkage.
3. On retient comme coefficients les $c\beta_j; j = 1, ..., K$

Shrinkage par paramètre

De façon à éviter la sur-correction de certains paramètres, on propose de calculer un facteur de shrinkage par variable en modifiant les étapes 2 et 3.

2. Calcul du modèle de Cox avec K variables $X_{ij}\hat{\beta}_j^{(-i)}; j = 1, ..., K$. Les coefficients $c_1, ..., c_K$ de ce modèle donnent les facteurs de shrinkage.
3. On retient comme coefficients les $\gamma_j = c_j\beta_j; j = 1, ..., K$

Problème: Si des variables très corrélées (y compris des termes d'interaction) sont présentes dans le modèle, les coefficients corrigés pour ces variables ne sont pas interprétables.

Shrinkage conjoint

2 Shrinkage: principe

Le shrinkage conjoint est un compromis entre le shrinkage global et par paramètres. Il propose de donner un facteur de shrinkage commun pour chaque sous-groupe de variables corrélées entre elles $J_1, ..., J_h$, mutuellement exclusifs et collectivement exhaustifs. On modifie la procédure ainsi:

2. Calcul du modèle de Cox à h covariables définies par $\eta_{ig} = \sum_{j \in J_g} x_{ij} \hat{\beta}_j^{(-i)}$ permettant d'obtenir les facteurs de shrinkage $c_1, ..., c_h$
3. Pour chaque $J_g; g = 1, ..., h$, le coefficient retenu pour les variables appartenant au groupe g sont les $\gamma_j = c_g \beta_j$

Table des matières

3 Plan de validation

▶ Le problème

▶ Shrinkage: principe
Shrinkage global
Shrinkage par paramètre
Shrinkage conjoint

▶ Plan de validation

Appliquer le shrinkage ?

3 Plan de validation

- Dans le cadre de notre projet, il semble donc intéressant d'appliquer un shrinkage conjoint qui permettra une évaluation du modèle plus réaliste dans la validation interne si un problème de calibration est détecté.
- Cette démarche peut sembler dispensable si nous pouvons bénéficier d'une validation externe, bien qu'elle permette également, en théorie, une réduction de la variance des coefficients.

Courbe de calibration

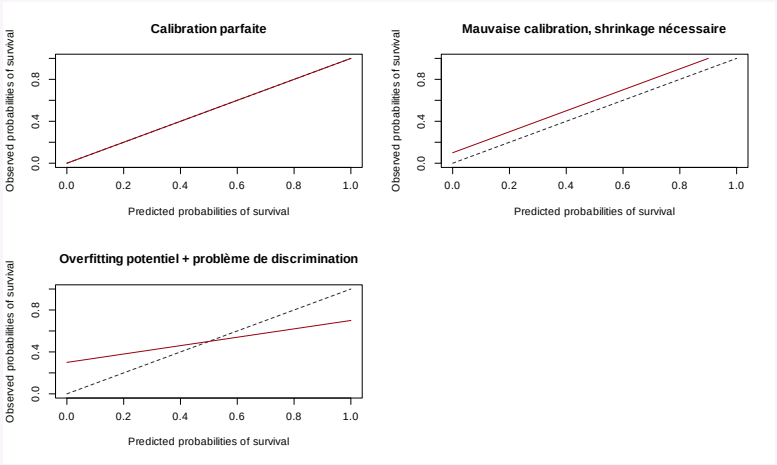
3 Plan de validation

- Représente la courbe lissée (par loess ou interpolation) des probabilités de survie (ou mortalité) à un horizon temporel t donné, en fonction des probabilités prédites. Une calibration parfaite correspond donc à la première bissectrice.
- Caractérisée par 2 paramètres:
  - **Intercept:** Pour une pente de 1, un intercept strictement négatif (pour la survie) indique une tendance à sous-estimer ou surestimer les probabilités de survie, ce qui est symptomatique de l'overfitting (coefficients β surestimés) et encourage l'application du shrinkage.
  - **Pente:** Une pente de 1 indique que la calibration des prédictions est la même pour tous les individus. Une pente différente de 1 indique donc que le modèle a tendance à sous-estimer ou sur-estimer les probabilités de survie pour certaines strates de risque (à ce regard, c'est une mesure de discrimination).

  Une pente de 1 associée à un intercept de 0 sont donc les indicateurs d'un modèle parfaitement calibré.

Typologie des graphes de calibration

3 Plan de validation



Plan de validation proposé

3 Plan de validation

1. Calibration
   - 1.1 Calibration plot
   - 1.2 Calibration intercept & slope $\Rightarrow$ shrinkage si besoin
2. Discrimination
   - 2.1 Courbe ROC temps dépendante (discrimination de l'outcome binaire)
   - 2.2 Calibration slope (discrimination sur la probabilité de survie)
   - 2.3 Discrimination slope (boxplot des probas de survie prédites à t pour Evt=0 et Evt=1)
   - 2.4 Courbe ROC selon les strates de receveur: AUC espérée minimale chez les receveurs à faible risque et maximale chez ceux à fort risque.
3. Performance globale
   - 3.1 Brier scrore

Application du shrinkage aux modèles de Cox