# Évolution de la stabilité de la sélection de variables en fonction de la taille d'échantillon et de la dimension

David Dernoncourt, Blaise Hanczar, Jean-Daniel Zucker

# Évolution de la stabilité de la sélection de variables en fonction de la taille d'échantillon et de la dimension.

David Dernoncourt[1,2,3], Blaise Hanczar[4], Jean-Daniel Zucker[1,2,3,5]

[1] Université Pierre et Marie Curie - Paris 6, Centre de Recherche des Cordeliers, UMR S 872, Paris, F-75006 France
david.dernoncourt-@-yahoo.com et jdzucker-@-gmail.com

[2] Université Paris Descartes, UMR S 872, Paris, F-75006 France

[3] INSERM, U872, Nutriomique, Equipe 7, Paris, F-75006 France

[4] LIPADE, Université Paris Descartes, 45 rue des Saint-Pères, 75006 Paris, France
hanczar_blaise-@-yahoo.fr

[5] IRD/UPMC UMI 209 UMMISCO, Centre IRD de l'Île de France, Bondy, F-93143 France

**Résumé** : La sélection de variables est une étape importante lors de la construction d'un classificateur sur des données de grande dimension. Lorsque le nombre d'observations est faible, cette sélection a tendance à être instable, au point qu'il est courant d'observer que sur deux jeux de données différents mais traitant d'un problème similaire, les variables sélectionnées ne se recoupent presque pas. Pourtant, ce problème de la stabilité semble encore peu étudié. Dans cet article, nous présentons des méthodes de quantification de la stabilité, puis nous en étudions les variations en fonction de divers paramètres (dimensionalité, nombre d'observations, distribution des variables, seuil de sélection) sur des données artificielles, avant de réaliser ces mesures sur des données réelles d'expression génique (données puces).

**Mots-clés** : Sélection de variables, stabilité, petits échantillons.

## 1. Introduction

Les tâches de classification dans lesquelles le nombre $D$ de variables est largement supérieur à la taille $N$ de l'échantillon sont un problème de plus en plus fréquent et sont devenues un champ de recherche à part entière (Hastie *et al.*, 2009). Par exemple, en biologie, des données "puces" contiennent

l'expression simultanée de dizaines de milliers de gènes, et des données métagénomiques l'expression de quelques millions de gènes... généralement mesurée sur (au plus) quelques centaines de patients. Dimensionalité élevée et petite taille de l'échantillon représentent un défi pour les techniques de classification, car tous deux augmentent le risque d'overfitting et diminuent la précision des classificateurs (Jain & Chandrasekaran, 1982). En outre, la dimensionalité élevée peut augmenter le temps de calcul de façon excessive, car les classificateurs ne s'adaptent généralement pas très bien à un très grand nombre de variables. Pour faire face à ces problèmes, la sélection de variables est utilisée pour réduire la dimensionalité des données.

La sélection de variables consiste à retirer les variables non pertinentes ou redondantes de l'ensemble de variables initial $\mathcal{F} = \{f_1, f_2, ..., f_{|\mathcal{F}|=D}\}$, de façon à conserver un sous-ensemble $S \subset \mathcal{F}$ ne contenant que des variables informatives et utiles pour la classification. Les méthodes de sélection peuvent être divisées en trois catégories : les filtres, les wrappers et les méthodes intégrées (Saeys *et al.*, 2007). Il est généralement admis que les wrappers ou les méthodes intégrées devraient être préférées si cela est techniquement possible (Pudil & Somol, 2008), cependant, sur des données de très haute dimensionnalité, les filtres restent la méthode de choix pour des raisons de faisabilité, c'est pourquoi nous nous focaliserons sur eux

L'objectif principal de la sélection de variables est de produire une signature robuste et précise de la classe à prédire. Une bonne signature ne doit pas être surajustée aux données disponibles et doit être exportable aux autres jeux de données traitant d'un même problème de classification. Ces conditions ne sont pas respectées si le sous-ensemble de variables sélectionnées est très inconstant. De nombreux exemples dans la littérature montrent que dans des conditions "petit échantillon - haute dimension", la sélection de variables est instable. Par exemple, dans (Miecznikowski *et al.*, 2010), cinq tâches de classification portant sur un problème semblable (pronostic de cancer du sein à partir de données d'expression géniques) sont réalisées sur cinq jeux de données différents, résultant en des résultats très variables au niveau de l'analyse individuelle des gènes. Plusieurs autres études, telles que (Ein-Dor *et al.*, 2006) and (Haury *et al.*, 2011), ont souligné la difficulté d'obtenir une signature reproductible sur des données puces avec de petits échantillons. Cette difficulté de trouver un sous-ensemble commun de prédicteurs entre des jeux de données différents mais semblables, ou même entre des sous-ensembles différents d'un même jeu de données, soulève le problème de la stabilité de sélection de variables.

Relativement peu d'études ont traité de ce problème, et la plupart ont mis l'accent sur la comparaison des stabilités de différentes méthodes de sélection sans explorer comment différentes variations dans les données peuvent affecter cette stabilité. En outre, celles-ci ont souvent utilisé des mesures de stabilité pouvant être biaisées par la proportion de variables sélectionnées (la plupart des mesures de stabilité sont artificiellement augmentées quand la proportion de variables sélectionnées augmente) ou par la quantité de variables non sélectionnées (certaines mesures prennent en compte la stabilité des variables non sélectionnées, et peuvent donc être excessivement élevées sur des jeux de données contenant une large proportion de variables non pertinentes et faciles à éliminer). Dans ce papier, nous présentons deux mesures de stabilité complémentaires et non biaisées, puis nous effectuons une analyse empirique de la stabilité de la sélection (et de la précision de la classification) sur des données artificielles et réelles (puces à ADN). Cela nous a permis de clarifier l'influence des diverses caractéristiques du jeu de données (taille d'échantillon, seuil de sélection, nombre de variables, distribution des variables) sur la stabilité de la sélection. Nos simulations soulignent le manque de stabilité sur les données hautes dimensions avec petit échantillon, et ses conséquences sur les performances de classification.

## 2. Mesures de stabilité

La stabilité d'une méthode de sélection est définie dans (Kalousis *et al.*, 2007) comme *la robustesse du choix de variables qu'elle produit face à des jeux d'apprentissage différents générés à partir d'une même distribution*. Pour évaluer cette robustesse, différentes mesures de stabilité ont été décrites. Nous suivrons la taxonomie présentée dans (Somol & Novovičová, 2010), qui distingue :

- des mesures *feature-focused* versus *subset-focused* : les premières évaluent la fréquence de sélection des variables parmi l'ensemble des sous-ensembles sélectionnés, considéré comme un tout, les secondes évaluent les similarités dans chaque paire de sous-ensembles. Ces deux types fournissant des informations complémentaires, nous utiliserons une mesure de chaque.
- des mesures *selection-registering* versus *selection-exclusion-registering* : les premières ne considèrent que la stabilité des variables sélectionnées, les secondes tiennent aussi compte de la stabilité des variables exclues. Sur des jeux de données où un grand nombre de variables sont non-

pertinentes et facile à exclure, les mesures *selection-exclusion-registering* seront fortement biaisé vers le haut. Nous nous intéresserons donc aux mesures *selection-registering*.

– des mesures *subset-size-biased* versus *subset-size-unbiased* : les premières produisent des valeurs délimitées plus étroitement que $[0; 1]$, avec en particulier une borne basse qui augmente fortement avec la proportion de variables sélectionnées. Les secondes sont ajustées pour être délimitées par $[0; 1]$. Pour une meilleure généralisation, nous utiliserons des mesures *subset-size-unbiased*.

### 2.1. La cohérence pondérée relative (*relative weighted consistency*)

Parmi les mesures classées dans la taxonomie de Somol and Novovičová, une seule est à la fois *selection-registering* et *subset-size-unbiased* : la *relative weighted consistency* $CW_{rel}$ (Somol & Novovičová, 2010). Elle est construite à partir d'une mesure *subset-size-biased*, la cohérence pondérée $CW$, corrigée pour être délimitée par $[0; 1]$ quelle que soit la proportion de variables sélectionnées. Une valeur de 0 indique l'instabilité la plus forte possible, une valeur de 1 la stabilité la plus forte possible (sous-ensembles tous identiques).

Soient $\mathcal{S} = \{S_1, S_2, ..., S_\omega\}$ un système de $\omega$ sous-ensembles obtainus par $\omega$ exécution de la sélection sur différents échantillons, $\Omega = \sum_{i=1}^{\omega} |S_i|$ la somme des cardinalités des sous-ensembles $S_i$ et $F_f$ le nombre d'occurrences de la variable $f \in \mathcal{F}$ dans $\mathcal{S}$. $CW$ est définie par :

$$CW(\mathcal{S}) = \sum_{f \in X} \frac{F_f}{\Omega} \cdot \frac{F_f - 1}{\omega - 1} \tag{1}$$

$CW_{rel}$ s'obtient en ajustant $CW$ par ses valeurs minimales et maximales :

$$CW_{rel}(\mathcal{S}, \mathcal{F}) = \frac{CW(\mathcal{S}) - CW_{min}(\Omega, \omega, \mathcal{F})}{CW_{max}(\Omega, \omega) - CW_{min}(\Omega, \omega, \mathcal{F})} \tag{2}$$

### 2.2. L'index de Tanimoto moyen partiellement ajusté

$CW_{rel}$ est *feature-focused*, nous avons donc cherché une mesure *subset-focused* pour la compléter. Les mesures de stabilité définies dans (Kuncheva, 2007) et (Krížek *et al.*, 2007) sont toutes deux *subset-focused*, mais ne peuvent s'utiliser que sur des sous-ensembles de même cardinalité. Nous avons retenu l'index de Tanimoto moyen $ATI$ (*Average Tanimoto Index*), introduit dans

(Somol & Novovičová, 2010). $ATI$ est une généralisation de la mesure de similarité de Kalousis $S_S$ entre deux ensembles $S_i$ et $S_j$ (Kalousis *et al.*, 2005) :

$$S_S(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \tag{3}$$

Cet index est calculé sur toutes les paires de sous-ensembles, puis moyenné :

$$ATI(\mathcal{S}) = \frac{2}{\omega(\omega-1)} \sum_{i=1}^{\omega-1} \sum_{j=i+1}^{\omega} S_S(S_i, S_j) \tag{4}$$

$ATI$ est *subset-focused* et *selection-registering*, mais aussi *subset-size-biased*. Nous proposons une correction de cet index, l'index de Tanimoto moyen partiellement ajusté, défini par :

$$ATI_{PA}(\mathcal{S}) = \frac{ATI(\mathcal{S}) - ATI_{exp}(\mathcal{S})}{ATI_{max}(\mathcal{S}) - ATI_{exp}(\mathcal{S})} \tag{5}$$

où $ATI_{max}$ est la valeur maximale possible d'$ATI$ et $ATI_{exp}$ est l'espérance d'$ATI$ quand les sous-ensembles sont définis aléatoirement. Comme nous utilisons une méthode de sélection qui sélectionne un nombre prédéfini de variables, $ATI_{max} = CW_{max} = 1$. Pour $ATI_{exp}$, nous avons utilisé une approximation déterminée expérimentalement, fonction de la proportion de variables sélectionnées. Ainsi la correction que nous réalisons dans $ATI_{PA}$ diffère de celle appliquée dans $CW_{rel}$ : $CW_{rel}$ est ajusté sur la valeur minimale possible, alors qu'$ATI_{PA}$ est ajusté sur la valeur attendue. Il est donc possible pour $ATI_{PA}$ de prendre des valeurs négatives si la sélection est pire que l'aléatoire. Toutes les valeurs négatives seront considérées comme égales à 0, ainsi l'index reste dans l'intervalle $[0; 1]$.

### 2.3. Mesures basées sur la corrélation

$ATI$ et $CW$ se focalisent sur la stabilité des variables sélectionnées. Bien que cet aspect soit important pour la découverte de connaissances, dans le but d'évaluer les méthodes de sélection la stabilité du score sur l'ensemble des variables est intéressante également. Les mesures *selection-exclusion-registering* sont trop biaisées quand la proportion de variables exclues est élevée. Par contre, les corrélations des scores et des rangs fournissent un aperçu plus équilibré. Nous avons donc utilisé la corrélation moyenne des scores $\overline{S_W}$ et la corrélation moyenne des rangs $\overline{S_R}$, telles que décrites dans (Kalousis *et al.*, 2005).

## 3. Design expérimental

Afin d'évaluer l'impact de la taille d'échantillon, du nombre de variable, du seuil de sélection et de la distribution des variables sur la stabilité de la sélection de variables, nous utilisons d'abord des données artificielles. Le modèle simple choisi nous permet de mesurer la stabilité de la sélection sous des variations contrôlées de ces paramètres. L'autre raison du choix de ces données simplistes est la volonté de se placer dans un cadre "idéal", dans le sens où on ne pourrait s'attendre qu'à une stabilité moindre sur des données réelles, plus complexes. Dans un second temps, nous étudions l'impact de la taille de l'échantillon sur la stabilité de la sélection sur des données réelles.

### 3.1. Données artificielles

Les distributions utilisées pour générer les jeux de données étaient constituées d'un nombre variable ($D \in [50; 10000]$) de variables aléatoires indépendantes. Les jeux d'apprentissage contenaient $N \in [25; 10000]$ exemples, couvrant des ratio $N/D$ de 0.01 à 10, ce qui couvre et excède le range de ce ratio sur nos données réelles. Chacune des deux classes suit une distribution normale définie respectivement par $\mathcal{N}(\mu, \sigma^2)$ et $\mathcal{N}(-\mu, \sigma^2)$, où $\mu$ est un vecteur de moyenne tel que $|\mu| = D$ et l'écart-type est $\sigma = 1$ pour toutes les variables. Les $\mu_i$ étaient tirés d'une distribution triangulaire (densité de probabilité : $f(x) = 2 - 2x$ for $x \in [0; 1]$). Afin d'obtenir des formes différentes de densité de probabilité, simulant des dispersions et pertinences variables des variables, nous avons ensuite élevé $\mu$ à une puissance $\gamma$ ($\mu_i = \mu_i^\gamma, \gamma \in [1; 10]$). Enfin, $\mu$ a été réduit afin que $\mathcal{F}$ corresponde à une erreur de Bayes ($\epsilon_{Bayes}$) spécifique ou que le $\mu_i$ ait une valeur spécifique $\mu_{imax}$. Dans nos expériences nous avons choisi $\epsilon_{Bayes} = 0.10$ ou $\mu_{imax} = 0.15$.

Le score utilisé pour classer les variables sur le jeu d'apprentissage était la valeur absolue du t-score. Puis les $d$ variables ayant le score le plus élevé étaient sélectionnées. Nous avons choisi le t-test parce qu'il devrait être optimal sur de telles variables, indépendantes et normales.

Pour diverses combinaisons des paramètres $N$, $D$, $d$ et $\gamma$, 100 jeux d'apprentissage étaient générés. Sur chacun, un classificateur par analyse discriminante linéaire (LDA) a été construit après sélection de variables. Chaque classificateur a ensuite été testé sur un jeu contenant 10000 exemples. En dehors des mesures de stabilité décrites en section 2., nous avons mesuré l'erreur de classification moyenne et la fréquence de sélection de chaque variable, et

calculé deux erreurs de Bayes : $\epsilon_{BayesOptimal}$ et $\epsilon_{BayesObs}$. $\epsilon_{BayesOptimal}$ est le taux d'erreur d'un classificateur de Bayes sur le meilleur sous-ensemble de $d$ variables. Il représente la meilleure classification possible, sélectionnant les $d$ meilleures variables et construisant un classificateur idéal. Cette valeur permet d'évaluer la difficulté du problème et sert de point de repère pour évaluer la sélection et la classification. $\epsilon_{BayesObs}$ est le taux d'erreur d'un classificateur de Bayes sur une sélection donnée. Cette valeur permet d'évaluer la qualité de la sélection de variables. Plus $\epsilon_{BayesObs}$ est proche de $\epsilon_{BayesOptimal}$, meilleure est la sélection.

### 3.2. Données réelles

Nous avons utilisé deux jeux de données puces publiques, relatifs aux cancers du poumon (Bhattacharjee *et al.*, 2001) ($D = 2000$, $N = 203$, $N/D \approx 0.10$) et du sein (van de Vijver *et al.*, 2002) ($D = 2000$, $N = 295$, $N/D \approx 0.15$). Pour chaque jeu, pour différentes valeurs de $N$, 200 jeux d'apprentissage ont été générés par tirage aléatoire sans remise. Sur chacun, les mêmes méthodes de sélection et classificateur que sur les données artificielles ont été appliquées. Les classificateurs ont ensuite été testés sur les échantillons non inclus dans le jeu d'apprentissage correspondant. Nous avons mesuré la stabilité de la sélection (deux mesures différentes : l'une sur toutes les jeux à la fois, l'autre moyenne des stabilités dans chaque paire jeu d'apprentissage - jeu test), et l'erreur de classification moyenne.

## 4. Résultats

### 4.1. Données artificielles

Dans cet ensemble de simulations, nous présentons la stabilité de la sélection et les performances de classification en fonction des paramètres.

La figure 1 montre un aperçu de la stabilité du score des variables dans deux conditions extrêmes : un très petit échantillon ($N = 50$, à gauche), et un grand échantillon ($N = 5000$, à droite). Sur le petit échantillon, les scores des variables (figure 1a)) varient peu avec $\mu_i$, et bien que les variables les plus informatives aient des scores un peu plus élevés que les moins informatives, leurs scores varient approximativement sur le même intervalle. Cela contraste avec le cas du grand échantillon, où les variables les plus pertinentes ont des scores sur [8 ;12], loin des moins pertinentes, qui restent dans l'intervalle [0 ;3] et sont donc faciles à distinguer.

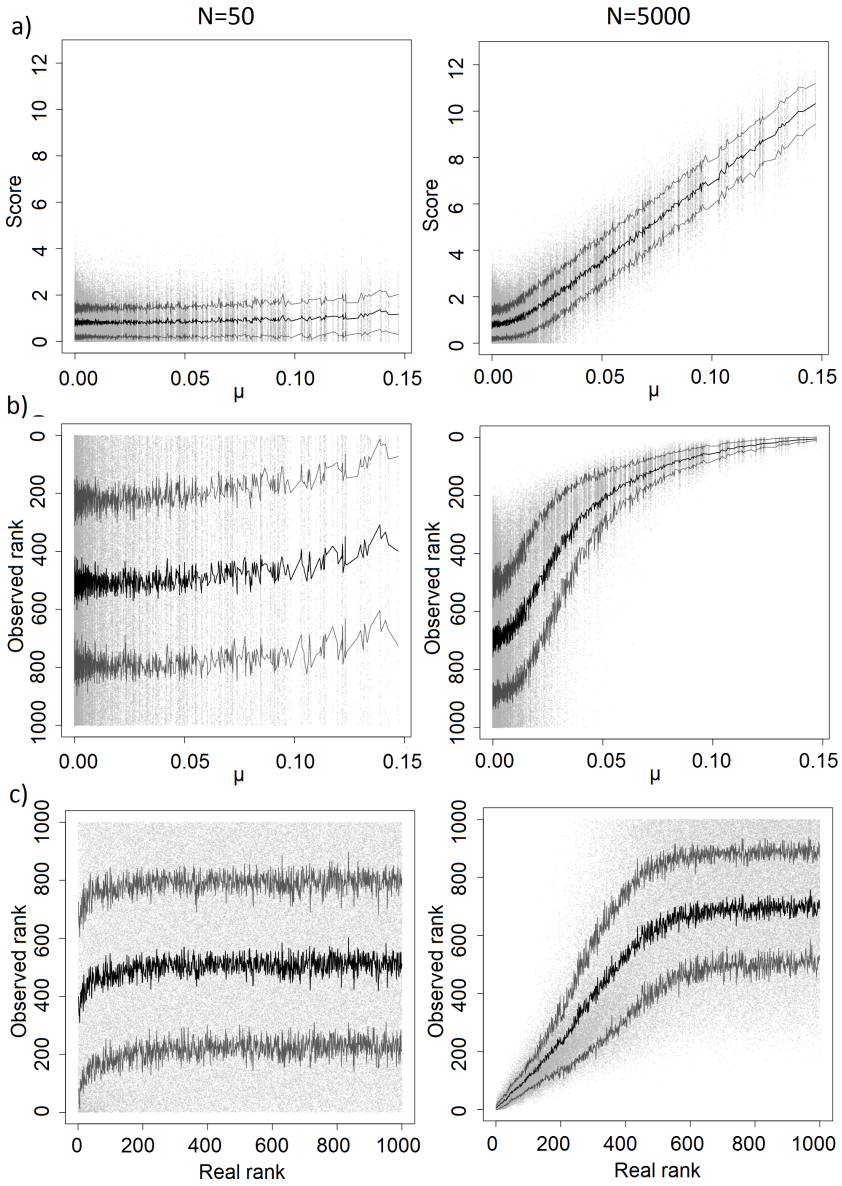


FIGURE 1: Données artificielles, avec $D = 1000$, $d = 100$, $\gamma = 2$ and $N = 50$ (gauche) or $N = 5000$ (droite). a) score observé en fonction de $\mu_i$, b) rang observé en fonction de $\mu_i$, c) rang observé en fonction du rang réel. Un point par variable et jeu d'apprentissage. Courbe noire : moyenne par variable. Courbe grise : moyenne par variable $\pm$ écart-type.

Les rangs en résultant reflètent l'instabilité des scores. La figure 1b) représente les rangs observés en fonction de $\mu_i$, la figure 1c) fournit une visualisation différente (rangs observés en fonction des rangs réels). En raison de la façon dont notre modèle a été conçu, les variables les moins pertinentes peuvent être considérées comme du bruit. Ainsi, un mauvais classement parmi les pires variables n'est pas un mauvais résultat. Par contre, dans le cas du petit échantillon, même les vraies meilleures variables ont un rang moyen à peine plus élevé que les autres variables. Pour plus de 90% des variables restantes, le rang attribué n'est que du bruit, comme en témoignent le nuage de point et les écarts-types (courbes grises) sur la figure 1. Dans le cas du grand échantillon, les vraies meilleures variables sont classées bien plus précisément, bien qu'il persiste un certain bruit, et seule la moitié basse des variables a un rang principalement aléatoire.

La figure 2 présente l'évolution de la probabilité de sélection des variables en fonction de leur vrai $\mu_i$. On constate que sur un petit échantillon (figure 2a)), la probabilité pour les variables les plus pertinentes d'être sélectionnées n'atteint pas 35%, et pas une seule variable parmi les moins pertinentes n'a une probabilité nulle d'être sélectionnée. Sur un grand échantillon (figure 2b)), la sélection est bien plus précise : à peu près toutes les variables avec $\mu_i > 0.10$ sont (presque) toujours sélectionnées, à peu près toutes les variables avec $\mu_i < 0.05$ sont (presque) toujours éliminées, et il persiste une zone de doute entre ces deux seuils. La figure 2c) montre l'évolution de ces courbes de probabilité quand $N$ varie de 25 à 10000 : à mesure que la taille d'échantillon augmente, la forme logistique ressort de plus en plus nettement, illustrant comme la sélection devient de plus en plus précise. Mais il faut tout de même atteindre environ 1000 observations pour que la méthode de sélection soit capable de sélectionner les variables les plus pertinentes avec une sensibilité parfaite.

La figure 3 présente l'évolution des mesures de stabilité en fonction de différents paramètres des données. La stabilité est considérablement influencée par la taille d'échantillon $N$, avec des valeurs proches de zéro pour 100 observations et une augmentation rapide lorsque plus d'observations sont ajoutées au jeux d'apprentissage, jusqu'à plus de 0.6 pour $AIT_{PA}$ pour $N = 10000$. Elle est aussi influencée par le nombre de variables $D$, avec des valeurs plutôt élevées (0.4 à 0.6) quand le jeu de données ne contient que 100 observations et 50 variables, mais se rapprochant rapidement de zéro dès 1000 variables.

Dans une moindre mesure, le seuil de sélection $d$ influence aussi la stabilité. Par exemple, $CW_{rel}$ est minimal lorsqu'on sélection très peu de va-

riables, puis atteint un maximum quand on sélectionne 150-180 variables, et enfin diminue progressivement quand on continue d'ajouter des variables. La forme de cette courbe illustre la difficulté à identifier de façon fiable même les variables les plus pertinentes : essayer de ne conserver que les 2 meilleures variables produit des résultats très instables, alors qu'essayer de conserver les 50 meilleures variables permettra sans doute d'inclure de façon fiable les 5-10 meilleures variables, entraînant une meilleure stabilité. À noter que dans cette configuration, $\overline{S_W}$ et $\overline{S_R}$ ne varient pas puisqu'ils ne tiennent pas compte du fait qu'une variable soit sélectionnée ou non.

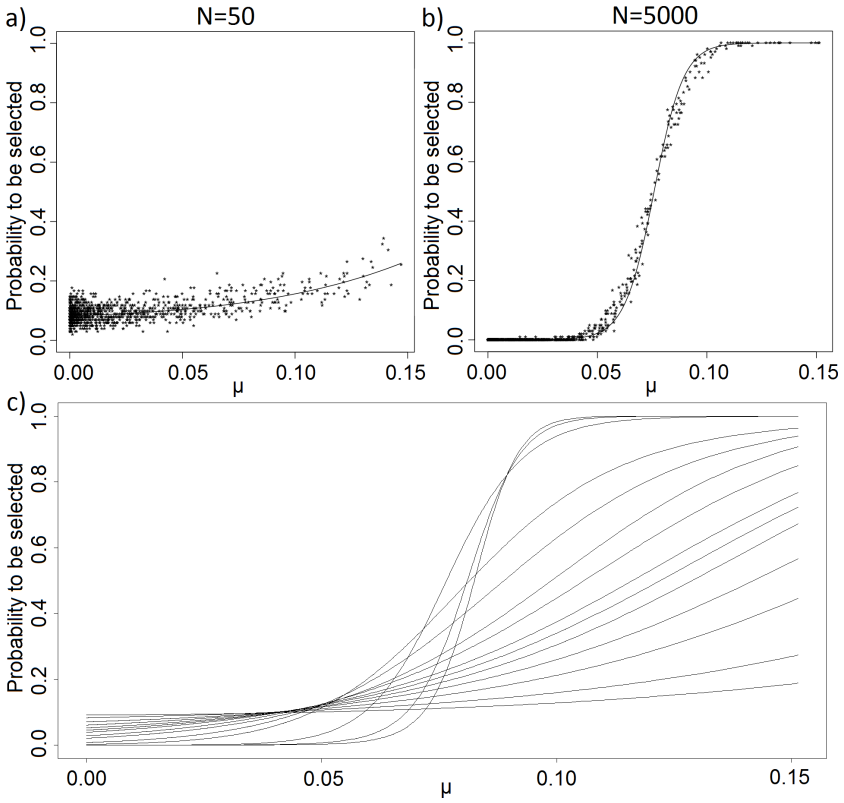


FIGURE 2: Probabilité pour une variable d'être sélectionnée en fonction de $\mu_i$. Données artificielles, $D = 1000$, $d = 100$ et $\gamma = 2$. a) $N = 50$ b) $N = 5000$, un point par variable, courbe obtenue par régression logistique. c) $N$ variant de 25 (courbe la plus basse à $\mu_i = 0.15$) à 10000 (courbe atteignant 1 le plus tôt). La forme logistique est de plus en plus marquée quand $N$ augmente.

La distribution des variables a également une influence sur la stabilité : la stabilité est minimale quand les $\mu_i$ on une distribution triangulaire, et augmente avec $\gamma$, mais suivant un schéma différent. $\overline{S_W}$ augmente toujours : cette mesure n'est pas pénalisée par les difficultés de classement ni par l'instabilité de la sélection finale, et bénéficie des valeurs extrêmes prises par les variables : les variables au $\mu_i$ initialement proche de 0 ne perdent pas vraiment en corrélation des scores lorsqu'elles sont tassées encore plus proches de 0, alors que les variables au $\mu_i$ plus élevé bénéficient d'un meilleur isolement, plus loin de 0. $\overline{S_R}$ augmente d'abord, puis diminue après avoir atteint un maximum autour de $\gamma = 5$ : cette mesure profite d'abord d'une dispersion accentuée des variables à $\mu_i$ élevé ou intermédiaire, mais à un certain point cet effet est neutralisé par une difficulté accrue à classer les variables à $\mu_i$ intermédiaire (comme nous gardons une erreur de Bayes constante, plus la distribution est étirée plus les variables moyennement pertinentes deviennent dures à identifier), qui finissent par trop se rapprocher de 0. $CW_{rel}$ et $ATI_{PA}$, qui sélectionnent les variables à partir d'une limite de rang, évoluent en conséquence de $\overline{S_R}$, avec une diminution retardée car elles ne sont affectées que par les top $d$ rangs. Il est probable qu'une méthode de sélection optimisant la taille du sous-ensemble aurait une stabilité plus influencée par la distribution des données, parce qu'elle éliminerait alors les variables devenant trop peu pertinentes tout en gardant les plus pertinentes, de moins en moins nombreuses mais de plus en plus faciles à identifier.

La figure 4 présente les taux d'erreur de classification obtenus à partir des sélections présentées en figure 3. La courbe pointillée indique $\epsilon_{BayesOptimal}$, le meilleur taux d'erreur possible sur le jeu de données avec $d$ variables (sélection des $d$ vraies meilleures variables et construction d'un classificateur idéal avec), la courbe grise $\epsilon_{BayesObs}$, le meilleur taux d'erreur possible en construisant un classificateur idéal sur les variables sélectionnées, la courbe noire le taux d'erreur observé. Le taux d'erreur et l'erreur de Bayes sur les variables sélectionnées augmentent quand la taille d'échantillon diminue. Pour les plus petits échantillons, cette augmentation est très marquée et on constate un large écart entre $\epsilon_{BayesOptimal}$ et $\epsilon_{BayesObs}$, indiquant une mauvaise sélection. Le taux d'erreur est aussi influencé par l'augmentation du nombre de variables (figure 4b)) : sur le petit échantillon testé ($N = 100$), avec seulement 50 variables l'erreur de classification est inférieure à 20% (pour une erreur de Bayes optimale supérieure à 16%), mais quand on atteint 2500 variables le taux d'erreur est supérieur à 40%, avec une erreur de Bayes sur la sélection de plus de 30%. C'est une conséquence de la dilution de l'information :

comme nous avons gardé une $\epsilon_{BayesOptimal}$ constante, quand la dimensionalité augmente, l'information est étalée sur des variables plus faibles, difficiles à sélectionner, donc $\epsilon_{BayesObs}$ augmente. Il est de surcroît difficile d'entraîner un classificateur sur ces variables plus faibles, donc le taux d'erreur de classification augmente.

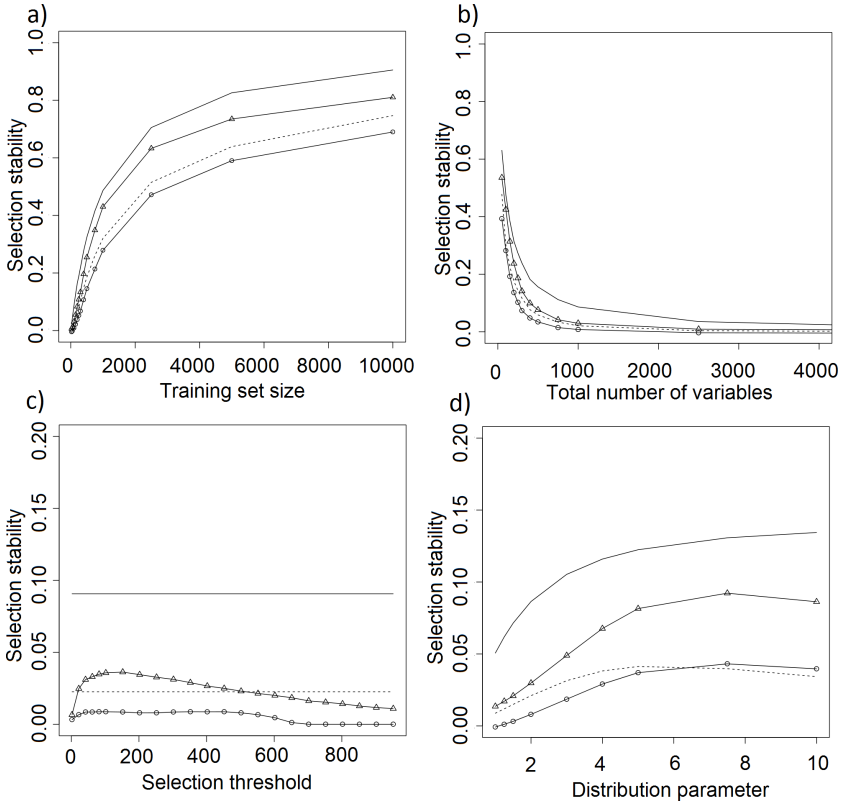


FIGURE 3: Evolution des mesures de stabilité $CW_{rel}$ (triangles), $ATI_{PA}$ (cercles), $\overline{S_W}$ (ligne continue) and $\overline{S_R}$ (pointilles) selon : a) $N \in [25; 10000]$ b) $D \in [50; 10000]$ c) $d \in [2; 1000]$ et d) $\gamma \in [1; 10]$. Quand ils ne sont pas le paramètre d'intérêt, les paramètres sont : $N = 100$, $D = 1000$, $d = D \cdot 10\%$, $\gamma = 2$.

L'évolution du taux d'erreur avec le seuil de sélection $d$ (figure 4c)) peut sembler plus surprenante. En particulier, la diminution régulière de l'erreur de Bayes sur la sélection peut donner l'impression que, quand on augmente le seuil, la sélection continue d'inclure les variables pertinentes dans l'ordre.

Bien entendu, ce n'est pas le cas : plus on élargit le seuil de sélection, plus on inclut des variables faiblement pertinentes, de façon de plus en plus aléatoire (comme on peut le déduire des mesures de stabilité vues précédemment). Même pour de faibles seuils, la sélection ne contient pas nécessairement les variables les plus pertinentes. Ceci est illustré par la croissance rapide de la distance entre $\epsilon_{BayesOptimal}$ et $\epsilon_{BayesObs}$. Quand la courbe de $\epsilon_{BayesObs}$ atteint enfin celle de $\epsilon_{BayesOptimal}$, ce n'est pas parce que la sélection est bonne mais parce que toutes les variables sont sélectionnées. C'est pourquoi l'erreur de classification diminue peu quand $d > 200$ (le temps de calcul, par contre, augmente substantiellement) : l'information contenue dans les variables les plus pertinentes est noyée par les variables peu pertinentes mais sélectionnées. Les résultats sont similaires lorsqu'on augmente la taille d'échantillon à $N = 1000$, mais avec des taux d'erreur moindre et des stabilités augmentées.

Le paramètre $\gamma$ semble avoir plus d'influence sur le taux d'erreur que sur la stabilité. Plus $\gamma$ est élevé, plus le taux d'erreur est bas (figure 4d)) : quand $\gamma$ augmente, le nombre de variables très pertinentes diminue mais leur pouvoir de discrimination augmente, permettant une meilleure sélection (cet effet est dilué par l'augmentation de l'instabilité des autres variables lorsqu'on observe les mesures de stabilité, mais ressort quand on observe comme $\epsilon_{BayesObs}$ se rapproche de $\epsilon_{BayesOptimal}$) et une meilleure précision de classification. Cependant, bien que le taux d'erreur de classification s'améliore, il s'améliore moins vite que $\epsilon_{BayesObs}$. Une explication à cette différence est que, malgré l'amélioration sur les variables les plus pertinentes, le classificateur est toujours pénalisé par les variables peu pertinentes restantes.

### 4.2. Données réelles

Le taux d'erreur diminue exponentiellement quand la taille d'échantillon augmente sur les données cancer du poumon (13.5% pour $N = 20$, 8% pour $N = 50$, 5% pour $N = 150$), plus linéairement sur les données cancer du sein (38.2% pour $N = 20$, 37.4% pour $N = 100$, 36.3% pour $N = 200$)

Le tableau 1 présente les mesures de stabilité et taux d'erreur obtenus sur les données réelles avec 50 et 100 sujets dans le jeu d'apprentissage. Les mesures de stabilité présentées ont été calculées sur des jeux à l'intersection nulle (des paires de jeux d'apprentissage et de test), et ne sont donc pas biaisées par des observations communes entre les jeux. Les données cancer du poumon semblent plus faciles que les données cancer du sein, comme en témoigne son plus faible taux d'erreur et sa meilleure stabilité. Les mesures de

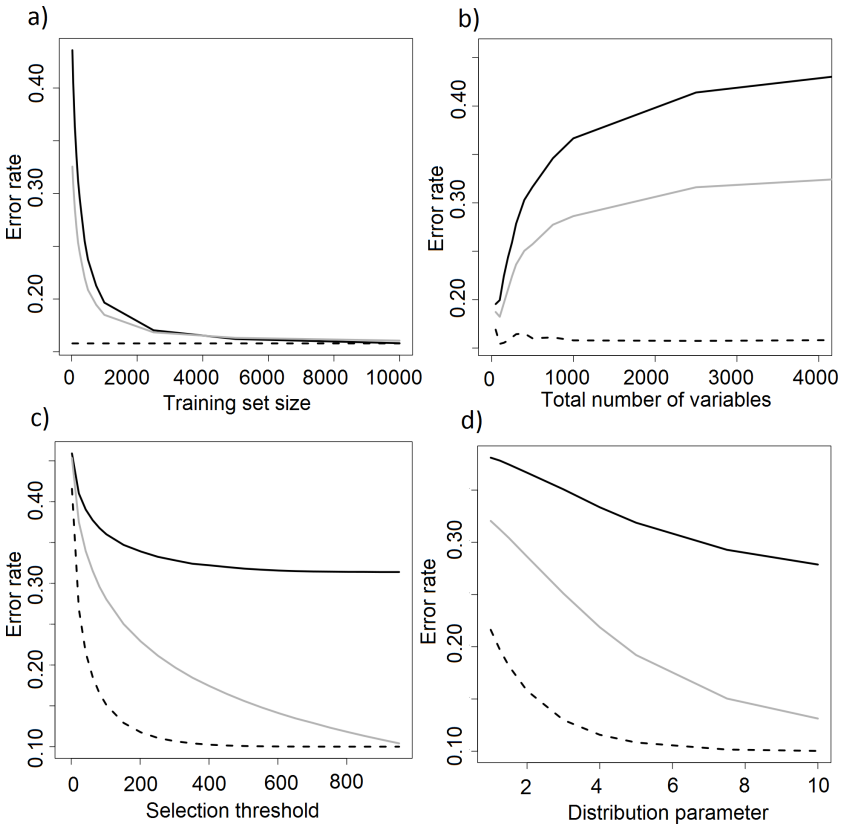


FIGURE 4: Evolution du taux d'erreur (noir), $\epsilon_{BayesObs}$ (gris) et $\epsilon_{BayesOptimal}$ (pointillés) selon : a) $N \in [25; 10000]$ b) $D \in [50; 10000]$ c) $d \in [2; 1000]$ et d) $\gamma \in [1; 10]$. Quand ils ne sont pas le paramètre d'intérêt, les paramètres sont : $N = 100$, $D = 1000$, $d = D \cdot 10\%$, $\gamma = 2$.

TABLE 1: Erreurs de classification et stabilité de sélection sur les données cancers du sein (van de Vijver) et du poumon (Bhattacharjee)

| **Mesure** | **Cancer du sein** | | **Cancer du poumon** | |
|---|---|---|---|---|
| | **N=50 N/D=0.025** | **N=100 N/D=0.05** | **N=50 N/D=0.025** | **N=100 N/D=0.05** |
| Error rate | 38.2% | 37.4% | 8.0% | 6.1% |
| $CW_{rel}$ | 0.20 | 0.26 | 0.46 | 0.51 |
| $ATI_{PA}$ | 0.06 | 0.10 | 0.26 | 0.30 |
| $\overline{S_R}$ | 0.09 | 0.14 | 0.51 | 0.58 |
| $\overline{S_W}$ | 0.33 | 0.41 | 0.81 | 0.85 |

stabilités, bien que différentes en valeur, suivent une même tendance, opposée au taux d'erreur. La stabilité augmente avec la taille d'échantillon, mais reste globalement basse, bien que plus élevée que sur nos données artificielles de dimension similaire. C'est particulièrement étonnant dans le cas des données cancer du sein, qui ont une stabilité plus élevée mais un taux d'erreur semblable comparé à nos données artificielles. Ces résultats confirment l'impact de la taille d'échantillon sur la stabilité de la sélection.

## 5. Conclusion

Dans ce papier, nous avons étudié la performance de la sélection de variables et en particulier sa stabilité dans un contexte haute dimension - petit échantillon. Nous avons utilisé des mesures de stabilité existantes et introduit $ATI_{PA}$, une modification de la mesure $ATI$ ajustée pour éviter un biais sur le nombre de variables sélectionnées. Nous avons ensuite étudié les relations entre la stabilité de la sélection et diverses caractéristiques des données. Bien que la stabilité soit influencée par la distribution des variables et le seuil de sélection, la taille d'échantillon et la dimension sont les paramètres les plus importants. La sélection est hautement instable lorsqu'on augmente le nombre de variables et réduit la taille d'échantillon à des valeurs comparables à ce qui est communément observé dans les problèmes $N << D$. Dans nos expériences, un ratio $N/D = 1$ permettait d'obtenir un $CW_{rel}$ autour de 0.5. Bien qu'il soit possible d'obtenir une meilleure stabilité en utilisant une erreur de Bayes plus basse, ces résultats vont dans le même sens que (Ioannidis, 2005) et (Ein-Dor *et al.*, 2006), qui suggèrent que des milliers d'observations sont nécessaires pour obtenir une sélection stable sur des données puces. Il semble également plus aisé d'obtenir de bonnes performances de classification qu'une bonne stabilité. Pour aller plus loin, il serait intéressant de réaliser de tests similaires sur d'autres méthodes de sélection, en particulier sur celles qui optimisent la cardinalité de la sélection et celles conçues pour améliorer la stabilité, telles que la *Complementary Pairs Stability Selection*.

## Références

BHATTACHARJEE A., RICHARDS W. G. & ET AL. (2001). Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(24), 13790–13795.

EIN-DOR L., ZUK O. & DOMANY E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences*, **103**(15), 5923–5928.

HASTIE T., TIBSHIRANI R. & FRIEDMAN J. (2009). *The Elements of Statistical Learning, 2nd ed.* Springer Series in Statistics. New York, NY, USA : Springer New York Inc.

HAURY A.-C., GESTRAUD P. & VERT J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE*, **6**(12), e28210.

IOANNIDIS J. P. (2005). Microarrays and molecular research : noise discovery ? *Lancet*, **365**, 454–455.

JAIN A. K. & CHANDRASEKARAN B. (1982). 39 dimensionality and sample size considerations in pattern recognition practice. *Handbook of Statistics*, **2**, 835–855.

KALOUSIS A., PRADOS J. & HILARIO M. (2005). Stability of feature selection algorithms. In *ICDM*, p. 218–225 : IEEE Computer Society.

KALOUSIS A., PRADOS J. & HILARIO M. (2007). Stability of feature selection algorithms : a study on high-dimensional spaces. *Knowl. Inf. Syst.*, **12**, 95–116.

KRÍŽEK P., KITTLER J. & HLAVÁC V. (2007). Improving stability of feature selection methods. In W. KROPATSCH, M. KAMPEL & A. HANBURY, Eds., *Computer Analysis of Images and Patterns*, volume 4673 of *Lecture Notes in Computer Science*, p. 929–936. Springer Berlin / Heidelberg.

KUNCHEVA L. I. (2007). A stability index for feature selection. In V. DEVEDZIC, Ed., *Artificial Intelligence and Applications*, p. 421–427 : IASTED/ACTA Press.

MIECZNIKOWSKI J. C., WANG D., LIU S., SUCHESTON L. & GOLD D. (2010). Comparative survival analysis of breast cancer microarray studies identifies important prognostic genetic pathways. *BMC Cancer*, **10**, 573.

PUDIL P. & SOMOL P. (2008). Identifying the most informative variables for decision-making problems - a survey of recent approaches and accompanying problems. *Acta Oeconomica Pragensia*, **2008**(4), 37–55.

SAEYS Y., INZA I. & LARRAÑAGA P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**(19), 2507–2517.

SOMOL P. & NOVOVIČOVÁ J. (2010). Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. *IEEE Trans. Pattern Anal. Mach. Intell.*, **32**(11), 1921–1939.

VAN DE VIJVER M. J., HE Y. D., VAN 'T VEER L. J. & ET AL. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, **347**(25), 1999–2009.