Arbre BIC optimal et taux d'erreur

Gilbert Ritschard
Dept Econométrie, Université de Genève

Atelier DKQ, EGC, janvier 2005

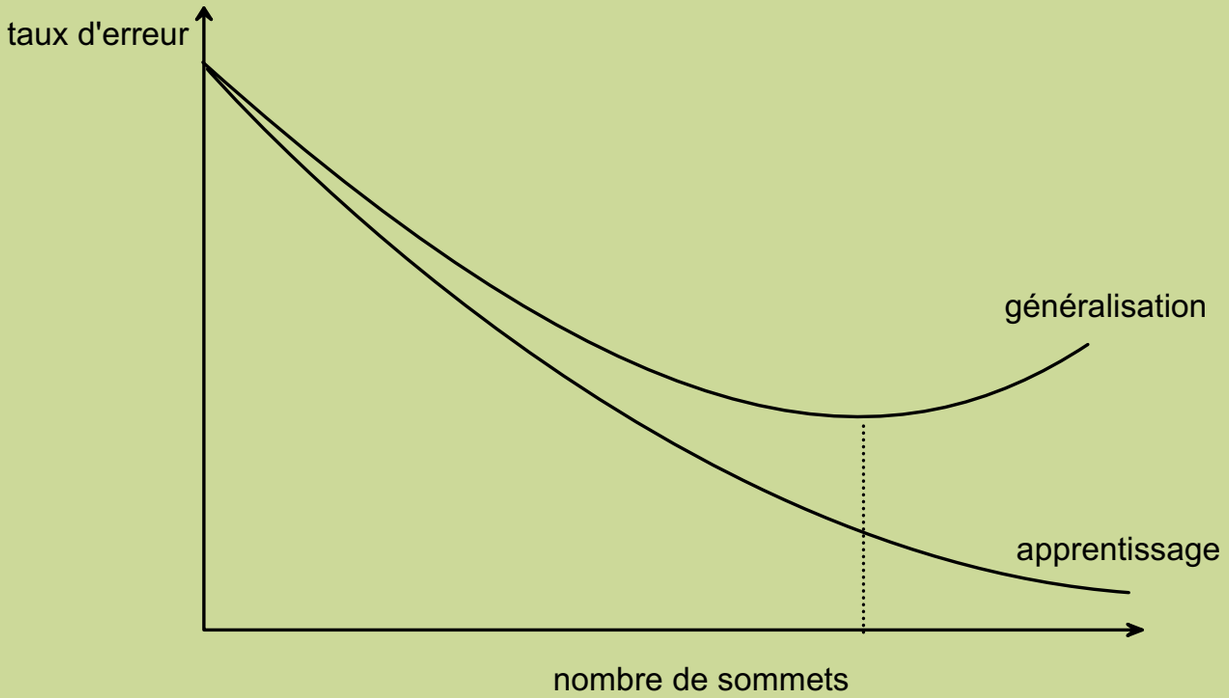
Plan

1 Introduction
2 Le critère BIC pour arbre d'induction
3 Example de relation entre BIC et erreur
4 Protocole de vérification de la conjecture
5 Conclusion

http://mephisto.unige.ch

1 Introduction

Qualité d'un arbre induit en général évaluée par
taux d'erreur en généralisation



Conjecture

Peut-on déterminer a priori (sans échantillon test)
quel arbre induit minimisera (en moyenne) le taux d'erreur en généralisation ?

Conjecture : Oui, l'arbre qui minimise le critère BIC (Ritschard and Zighed, 2004) minimise en moyenne le taux d'erreur en généralisation.

Objectifs

- Déviance et critère BIC (rappel)
- Remarques sur le calcul de la déviance
- Illustration de la conjecture sur l'exemple du Titanic
- Protocole d'une étude empirique pour démontrer la conjecture

2 Le critère BIC pour arbre d'induction

BIC = déviance pénalisée pour la complexité (nbre de paramètres)

$$\text{BIC}(m) \quad = \quad D(m) + p \ln n$$

p nbre de paramètres (complexité)

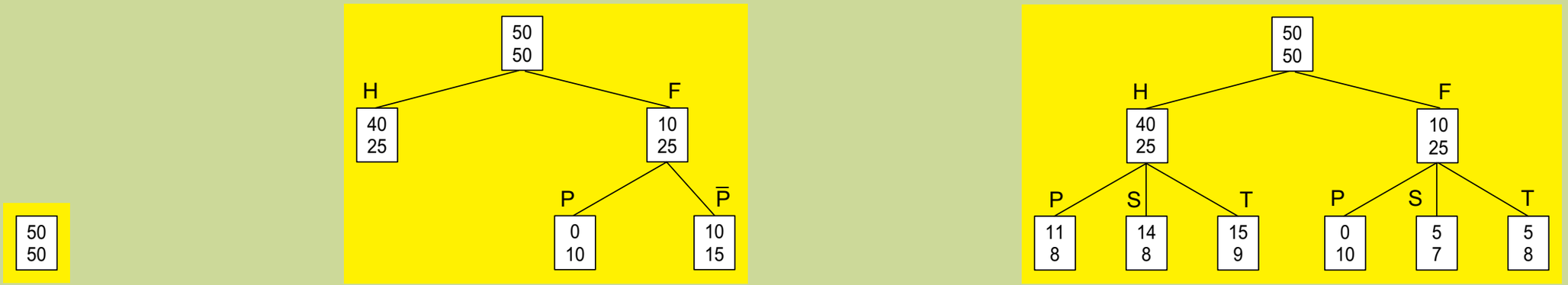
n nombre de données (apprentissage)

2.1 Déviance

2.2 Remarques sur calcul de la déviance

2.3 Remarques sur nombre de paramètres

2.1 Déviance



50		
50		

$\leftrightarrow$

40	0	10
25	10	15

$\leftrightarrow$

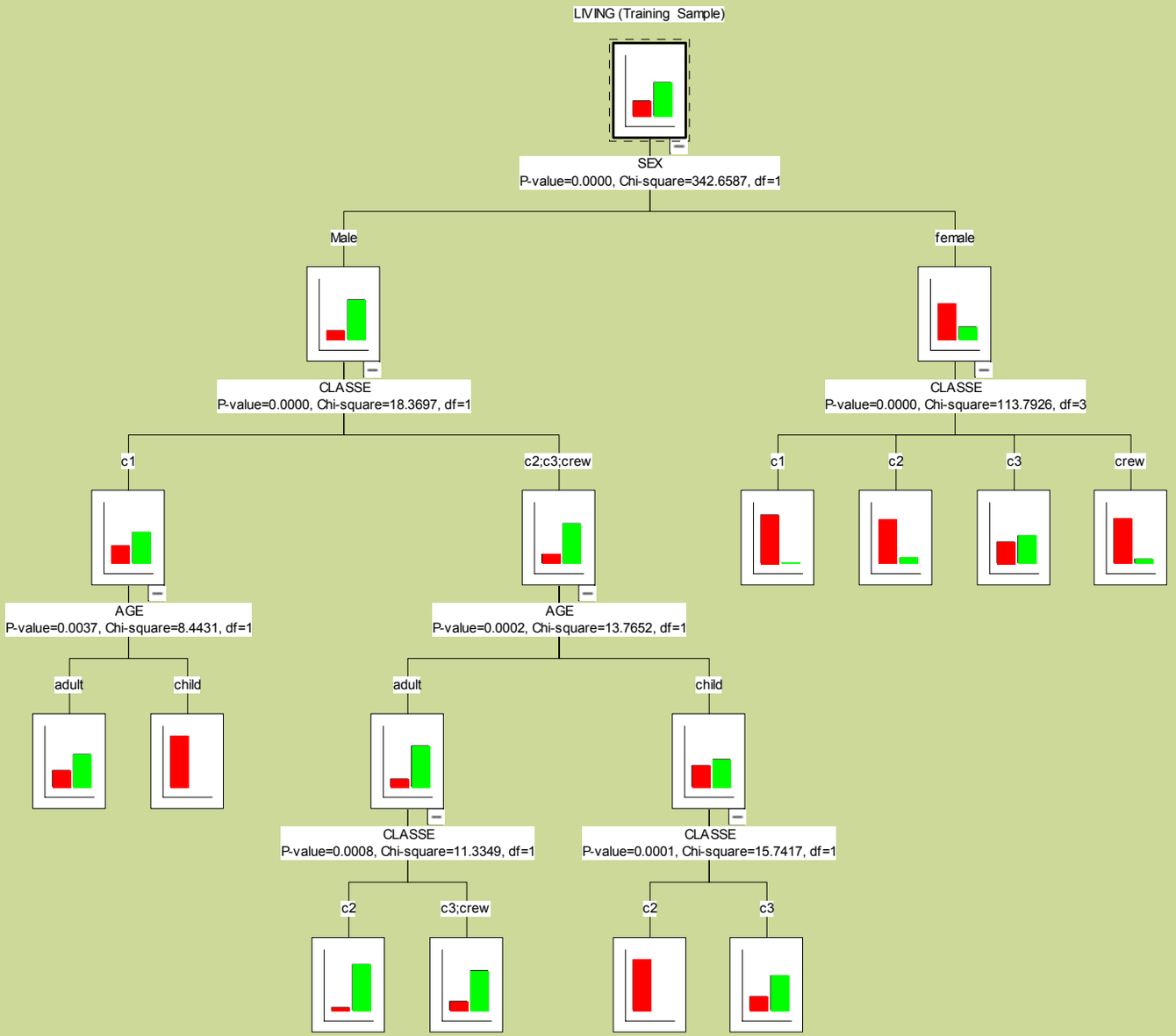
11	14	15	0	5	5
8	8	9	10	7	8

$D(m_0|m)$ $D(m)$

indépendance — modèle ajusté — tableau cible

nœud inital — graphe induit — arbre maximal

$D(m_0)$



Arbre induit, échantillon d'apprentissage, $n = 1659$

Titanic : Table cible et effectifs prédits

	table cible (n_{ij})		table prédite ($\hat{n}_{ij}$)		
	yes	no	yes	no	total
MAc1	45	88	45	88	133
MAc2	10	114	10	114	124
MAc3	59	289	67.6175	280.3825	348
MAc4	132	503	123.3825	511.6175	635
MCc1	4	0	4	0	4
MCc2	8	0	8	0	8
MCc3	10	23	10	23	33
FAc1	112	4	112.0342	3.9658	116
FAc2	66	11	67.375	9.625	77
FAc3	62	71	59.5	73.5	133
FAc4	15	2	15	2	17
FCc1	1	0	0.9658	0.0342	1
FCc2	11	0	9.625	1.375	11
FCc3	6	13	8.5	10.5	19
total	541	1118	541	1118	1659

Titanic : Table croisant la variable à prédire avec les feuilles

	table T_m		
	yes	no	total
MAc1	45	88	133
MAc2	10	114	124
MAc3,c4	191	792	983
MCc1	4	0	4
MCc2	8	0	8
MCc3	10	23	33
FA,Cc1	113	4	117
FA,Cc2	77	11	88
FA,Cc3	68	84	152
FAc4	15	2	17
total	541	1118	1659

2.2 Remarques sur calcul de la déviance

$T = (n_{ij})$ tableau $\ell \times c$ cible :

ℓ lignes = catégories de la variable à prédire

c colonnes = profils différents en termes des prédicteurs

$\hat{T} = (\hat{n}_{ij})$ tableau $\ell \times c$ prédit par l'arbre

Total de chaque colonne réparti selon distribution de la feuille contenant le profil correspondant.

$$D(m) = -2 \sum_{i=1}^{\ell} \sum_{j=1}^{c} n_{ij} \ln \left(\frac{\hat{n}_{ij}}{n_{ij}} \right)$$

Difficulté : construction des tableaux T et $\hat{T}$ car c peut être très grand

Principe de construction de la table cible

Table prédite $\hat{T}$

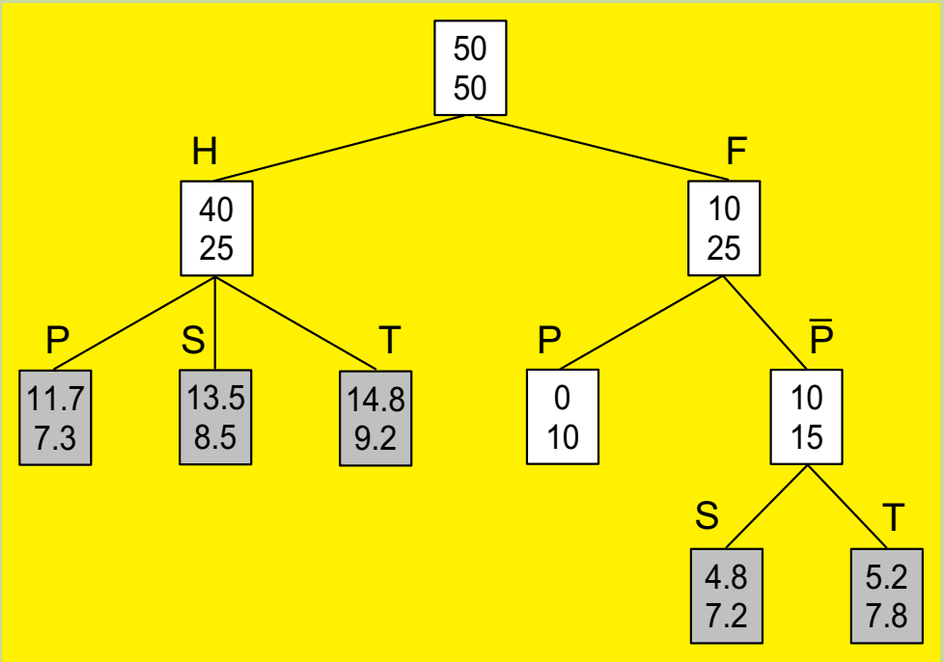
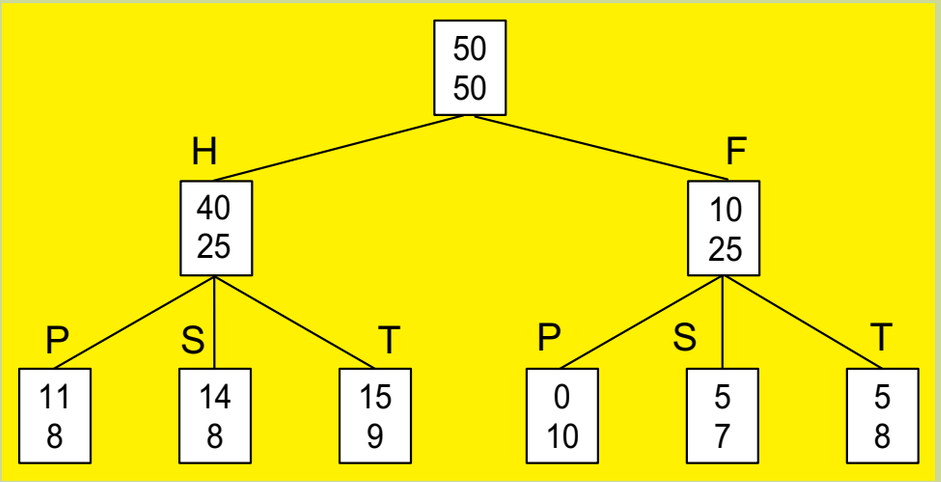


Table cible T



Déviance et rapport de vraisemblance

$D(m_0|m)$ = statistique du khi-2 du rapport de vraisemblance
pour test indépendance sur tableau associé à l'arbre induit.

$D(m_0)$ = statistique du khi-2 du rapport de vraisemblance
pour test indépendance sur tableau cible.

Ces deux valeurs s'obtiennent avec les logiciels statistiques (SPSS, SAS, ...)

On obtient la déviance partielle par différence

$$D(m) = D(m_0) - D(m_0|m)$$

2.3 Remarques sur nombre de paramètres

$$\text{BIC}(m) = D(m) + p \ln n$$

La pénalisation dépend du nombre p de paramètres.

Pour les arbres, $p = (\ell - 1)q + c$ (voir Ritschard and Zighed, 2004)

Cependant, le BIC est défini à une constante additive près

- on peut « oublier » le c
- si p augmente de 1, d les degrés de liberté de $D(m)$ diminue de 1
  $\Rightarrow$ définition équivalente

$$\text{BIC}(m) = D(m) - d \ln n$$

d différence entre les d.l. de $D(m_0)$ et $D(m_0|m)$,
les deux khi-2 du rapport de vraisemblance pour test d'indépendance.

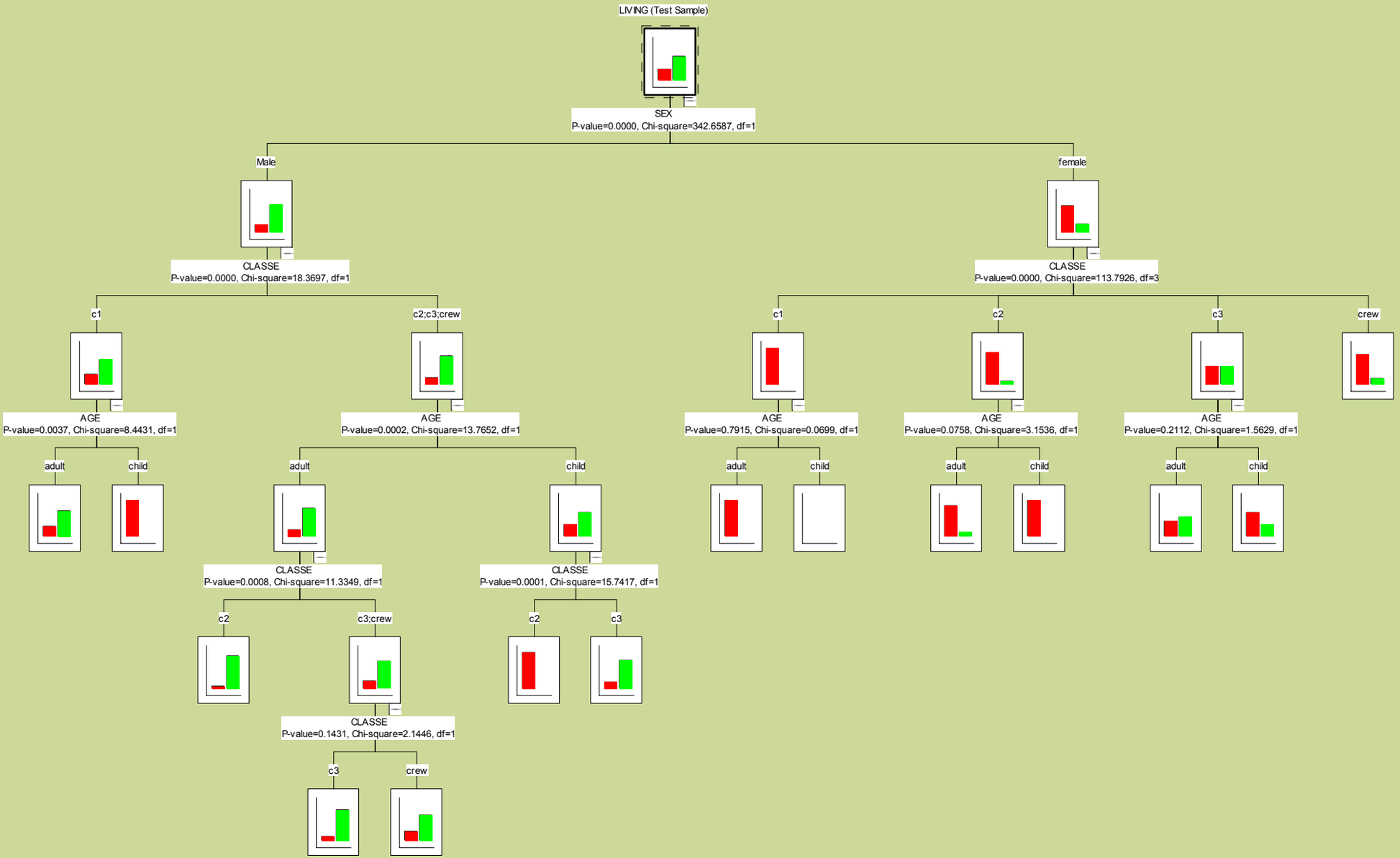
3 Example de relation entre BIC et erreur

2201 données du Titanic partitionnées en
- 1659 données d'apprentissage (∼25%)
- 542 données tests (∼75%)

Génération de 12 arbres

par élagages successifs à partir de l'arbre saturé

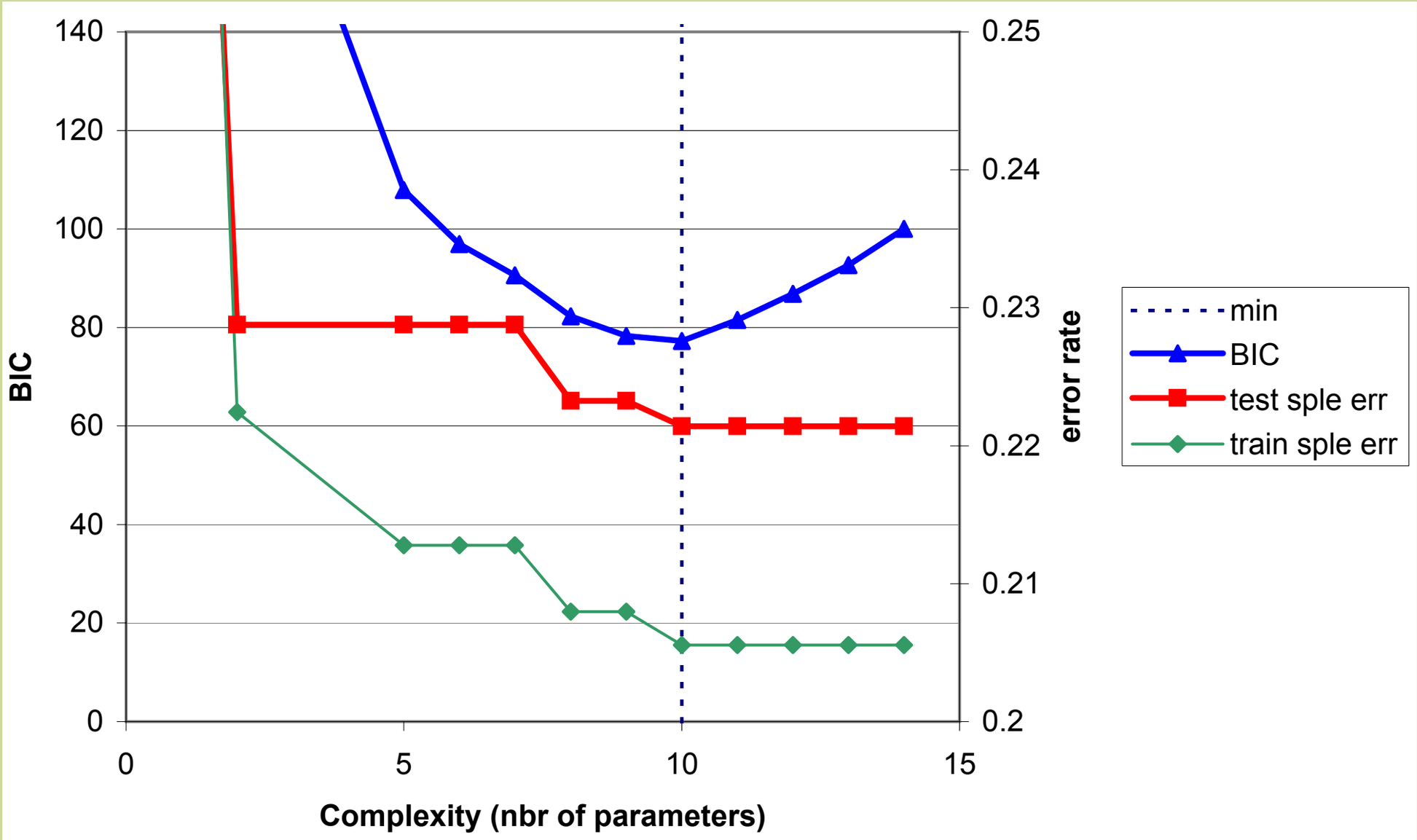
Calcul pour chacun des 12 arbres
- BIC sur données d'apprentissage
- taux d'erreur sur données d'apprentissage
- taux d'erreur sur données tests



Arbre saturé, échantillon test, $n = 542$

Qualité des modèles successifs

regroup.	model	p	D(m)	d	BIC	taux d'erreur test	taux d'erreur apprent.
	saturated	14	0	0	100	0.221	0.206
A,C | F,c1	m1	13	0.07	1	92.66	0.221	0.206
A,C | F,c3	m2	12	1.63	2	86.81	0.221	0.206
c3,c4 | M,A	m3	11	3.78	3	81.54	0.221	0.206
A,C | F,c2	m4	10	6.93	4	77.28	0.221	0.206
A,C | M,c1	m5	9	15.37	5	78.30	0.223	0.208
c2,c3c4 | M,A	m6	8	26.71	6	82.23	0.223	0.208
c2,c3 | M,C	m7	7	42.45	7	90.55	0.229	0.213
A,C | M,c2c3c4	m8	6	56.22	8	96.90	0.229	0.213
c1,c2c3c4 | M	m9	5	74.59	9	107.86	0.229	0.213
c1,c2,c3,c4 | F	m10	2	188.38	12	199.41	0.229	0.222
tout	indep	1	531.04	13	534.66	0.314	0.326



BIC sur échantillon d'apprentissage et taux d'erreur

4 Protocole de vérification de la conjecture

Plusieurs structures (cas de figure)

Pour chaque structure, générer :

- 1 échantillon d'apprentissage (taille selon structure postulée)
- 100 échantillons tests

Détermination d'une séries d'arbres (10-20) de complexité variable sur l'échantillon apprentissage

Calcul, pour chaque arbre, du taux moyen d'erreur sur les 100 échantillons tests

Comparer l'évolution du BIC et du taux d'erreur en généralisation en fonction de la complexité (paramètres)

Principale difficulté : trouver un logiciel générant les arbres

- dans lequel on peut implanter le calcul du BIC
- et qui peut être appelé itérativement par la procédure de test

5 Conclusion

Critère BIC pour arbre

- introduit pour déterminer l'arbre le plus adéquat du point de vue descriptif
- notre conjecture : pertinent également dans l'optique de la classification
  L'arbre avec le plus petit BIC devrait générer la plus petite erreur en généralisation
- La vérification reste à faire

Principale difficulté : trouver un logiciel générant les arbres

- dans lequel on peut implanter le calcul du BIC
- et qui peut être appelé itérativement par la procédure de test

Références

Ritschard, G. and D. A. Zighed (2004). Qualité d'ajustement d'arbres d'induction. *Revue des nouvelles technologies de l'information E-1*, 45–67.