Réseaux de neurones

Fabrice Rossi

http://apiacoa.org/contact.html

Université Paris-IX Dauphine

Plan du cours

1. Introduction aux réseaux de neurones
2. Rappels de probabilités et statistiques
3. Le modèle linéaire (et le perceptron simple)
4. Le modèle linéaire généralisé (et les réseaux RBF)
5. Les méthodes d'évaluation et de sélection de modèle
6. Les perceptrons multi-couches
7. Les k-moyennes
8. Les réseaux de Kohonen

Site du cours : `http://apiacoa.org/teaching/nn/`

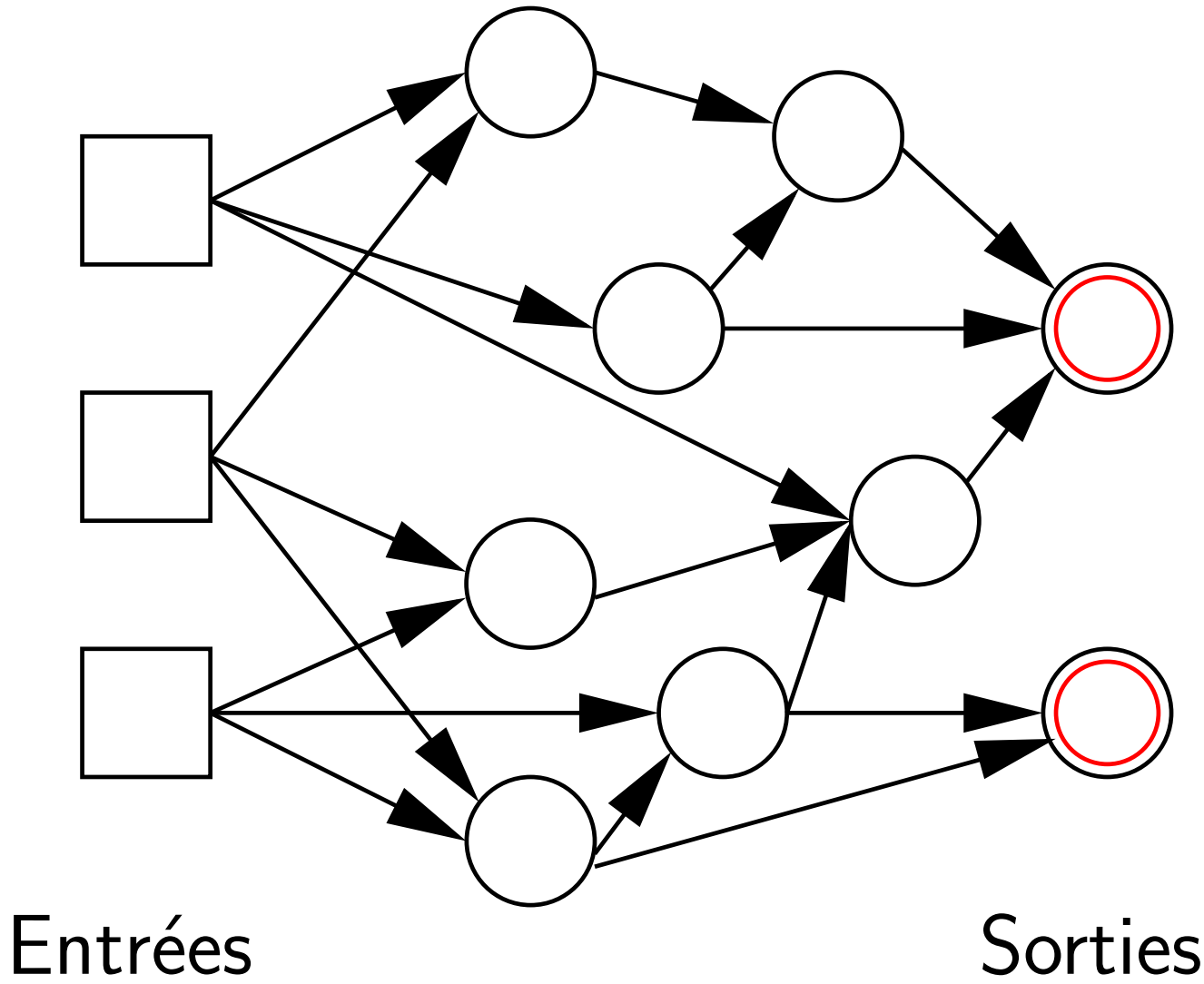
Plan de l'introduction

1. Qu'est qu'un réseau de neurones ?
2. Que peut-on faire avec un réseau de neurones ?
3. Pourquoi la modélisation des données est-elle un problème difficile ?
4. Modèle mathématique général

Réseaux de neurones

1. Ensemble de méthodes d'analyse et de traitement des données
2. Deux caractéristiques majeures :
   (a) combinaison (réseau) d'éléments simples (neurones)
   (b) non linéaire (par "opposition" aux méthodes classiques de l'analyse des données)
3. Vaguement issus de considérations biologiques
4. Essentiellement numériques, par opposition à l'IA symbolique (base de règles, raisonnement par cas, etc.)

Représentation graphique



Différents modèles

Les réseaux de neurones diffèrent selon :

- les neurones utilisés
- la structure du réseau
- le mode de calcul

Dans ce cours :

- les perceptrons multi-couches (Multi Layer Perceptron)
- les réseaux Radial Basis Function (RBF)
- les réseaux de Kohonen

Axes du cours :

- liens avec les méthodes classiques (Analyse des données)
- importance de la modélisation statistique

Analyse des données (*Data Mining*)

- classiquement : méthodes (en général linéaires) d'exploration des données (Analyse en Composantes Principales, régression linéaire, etc.)
- par extension : organisation et exploration des données :
  - **discrimination** : affectation de nouvelles données à des groupes connus (exemple : diagnostic médical)
  - "**régression**" : modélisation de relations fonctionnelles (exemple : niveau d'ozone demain en fonction du niveau d'aujourd'hui et de mesures météo comme la vitesse du vent, etc.)
  - **classification** : découverte de groupes dans des données (exemple : regroupement de profils de consommateurs)
- le but est soit d'apprendre par l'exemple (discrimination et régression), soit d'extraire de l'information (classification)

L'approche dite supervisée

On dispose d'un "professeur", c'est-à-dire un ensemble de données connues (les **exemples**) :

- en **discrimination**, des exemples de données provenant de chaque groupe (classe) :
  - des mesures biologiques (tension artérielle, numération sanguine, présence/absence de symptômes divers, etc.) pour des personnes malades et pour des personnes saines
  - des lettres et chiffres tracés par différentes personnes
- en **régression**, des exemples de la relation fonctionnelle :
  - le cours d'une action sur une période donnée
  - la consommation électrique d'une région sur une période donnée, associée à des mesures météorologiques sur la même période
- point commun : des observations et une cible (la classe, le cours de l'action dans le futur, etc.)

L'approche dite supervisée (2)

Le but principal est de **modéliser** la relation entre les observations et l'information cible, pour :

1. Estimer la valeur de la cible pour de nouvelles observations :
   - diagnostiquer un patient
   - reconnaître un code postal
   - prédire le cours d'une action
   - prédire la consommation électrique d'une région
2. Comprendre la relation observations/cibles :
   - déterminer les symptômes importants
   - comprendre les facteurs déterminants d'une forte consommation électrique

Les réseaux de neurones sont très efficaces pour le premier point, moins pour le second.

L'approche dite non supervisée

On ne dispose pas de "professeur", c'est-à-dire que les exemples ne sont pas déjà organisés :

- pas de classe, pas de cible
- but général : découvrir de la **régularité**, sous forme de **classes**, c'est-à-dire de groupes d'exemples qui se ressemblent
- but opérationnel :
  - groupes de consommateurs pour cibler une campagne marketing
  - groupes d'individus pour l'analyse sociologique, économique, etc.
- motivations :
  - comprendre les données en les simplifiant (analyse sur chaque groupe)
  - décrire des groupes

Nature des données

Dans la pratique, les données posent des problèmes :

- données fausses (erreur de mesure, tromperie, etc.)
- données incomplètes ou manquantes (absence de réponse, grosse erreur de mesure qui conduit à rejeter le résultat, etc.)
- données non exhaustives : presque toujours le cas (on se contente en général de sondages)

Les résultats sont donc “imprécis” :

- algorithmes exacts sur données réelles : le résultat est juste par rapport aux données, mais pas nécessairement par rapport à la réalité
- algorithmes heuristiques : on ne sait pas si l’algorithme donne le “bon” résultat, mais simplement qu’il est satisfaisant...

Modélisation statistique

Deux gros problèmes pratiques (fortement couplés) :

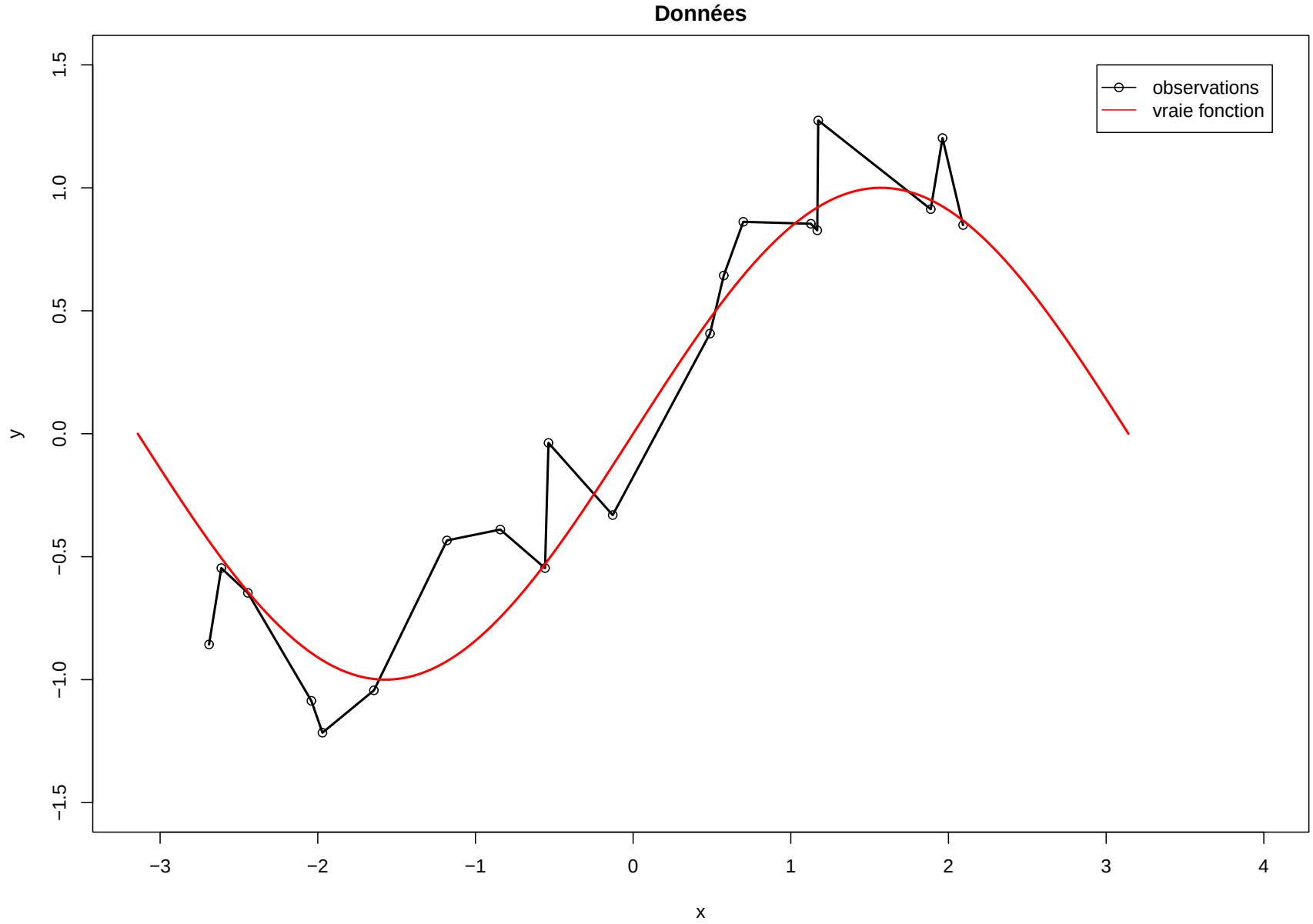
- mesure des performances :
  - $\Rightarrow$ bons résultats sur les données de départ, mais mauvais sur de nouvelles données
  - $\Rightarrow$ marges d'erreur, niveau de confiance
- choix du modèle :
  - $\Rightarrow$ dilemme biais/variance
  - $\Rightarrow$ biais : erreur systématique (par rapport à la réalité)
  - $\Rightarrow$ variance : sensibilité du modèle aux données
  - $\Rightarrow$ modèle puissant : faible biais, mais très sensible aux données
  - $\Rightarrow$ modèle faible : grand biais, mais moins sensible aux données

"Solution" : approche statistique en analyse des données.

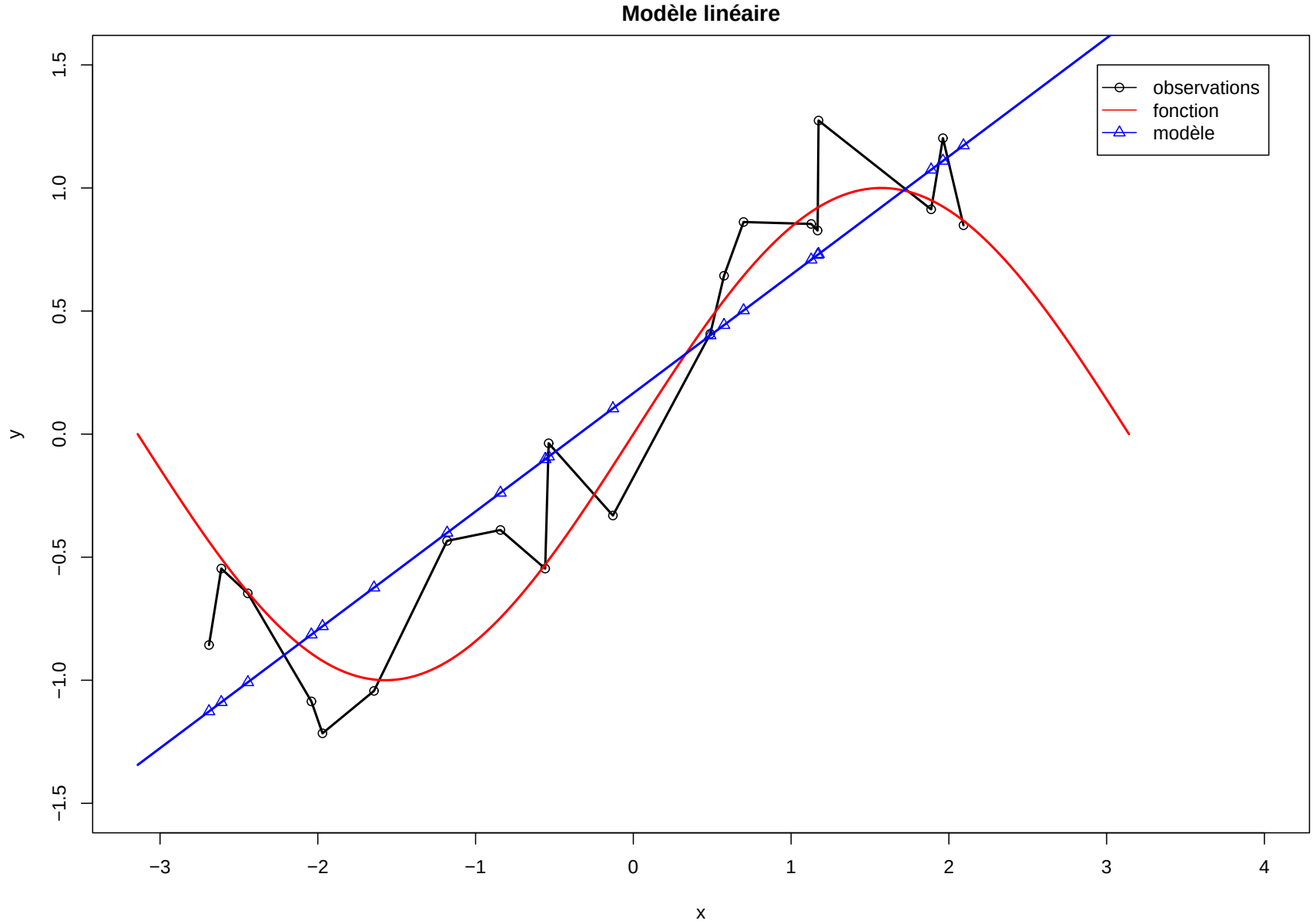
Un exemple de régression simple

- On possède une suite d'observations, les (x_i, y_i)
- On cherche à expliquer y_i grâce à x_i, c'est-à-dire trouver une fonction f telle que $y_i \simeq f(x_i)$ pour tout i
- Applications pratiques :
  - prédire demain en fonction d'aujourd'hui (météo, bourse, etc.)
  - évaluer le risque de défaillance d'un emprunteur en fonction de ses caractéristiques socioprofessionnelles
  - calculer le taux d'humidité du sol ou le niveau de maturité d'un champ de céréales en fonction d'une image radar de la zone concernée
  - etc.
- On peut trouver une approximation de f par des méthodes neuronales
- Difficulté principale : le bruit dans les observations

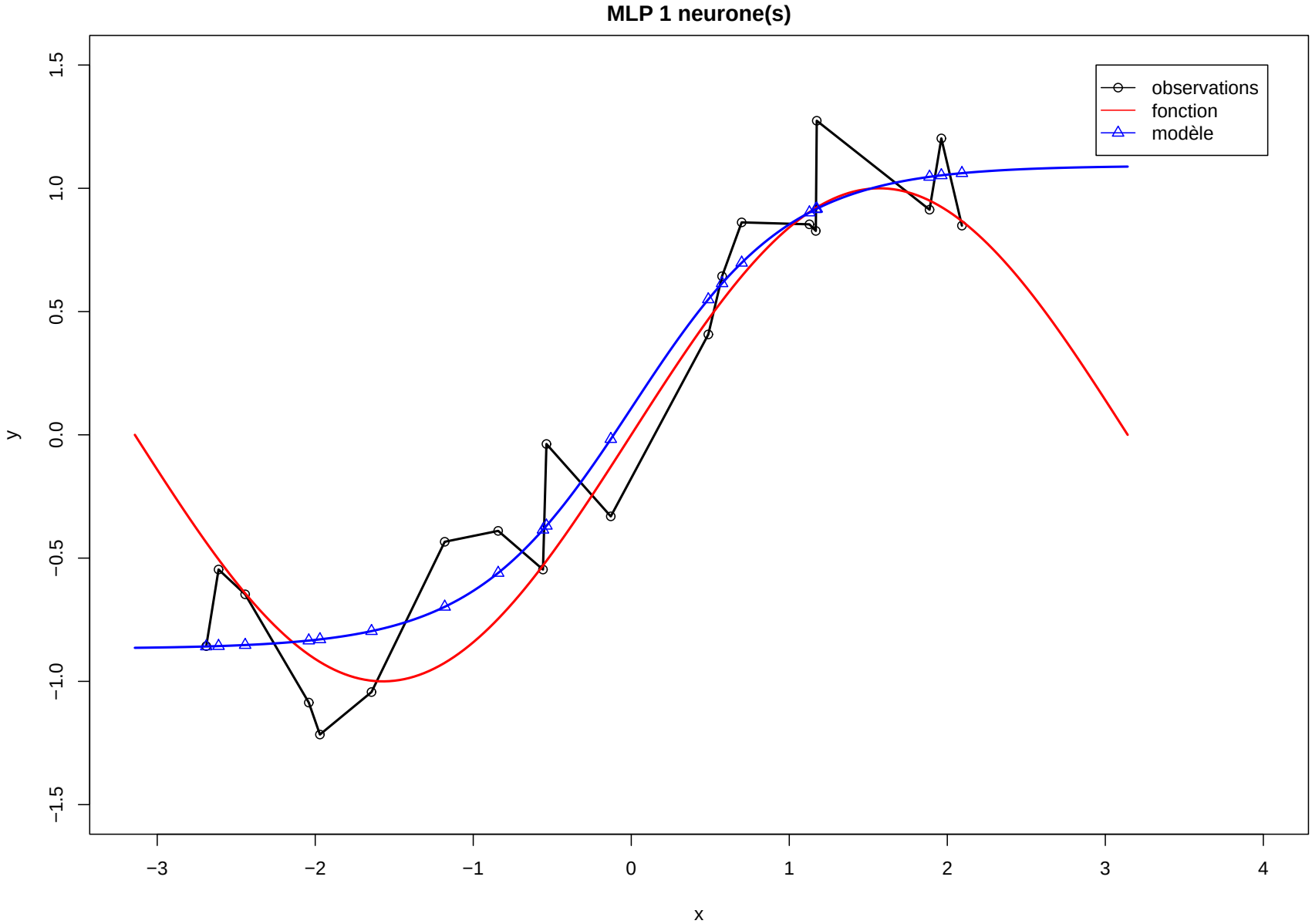
Les données


Données
observations
vraie fonction
y
x
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2
3
4

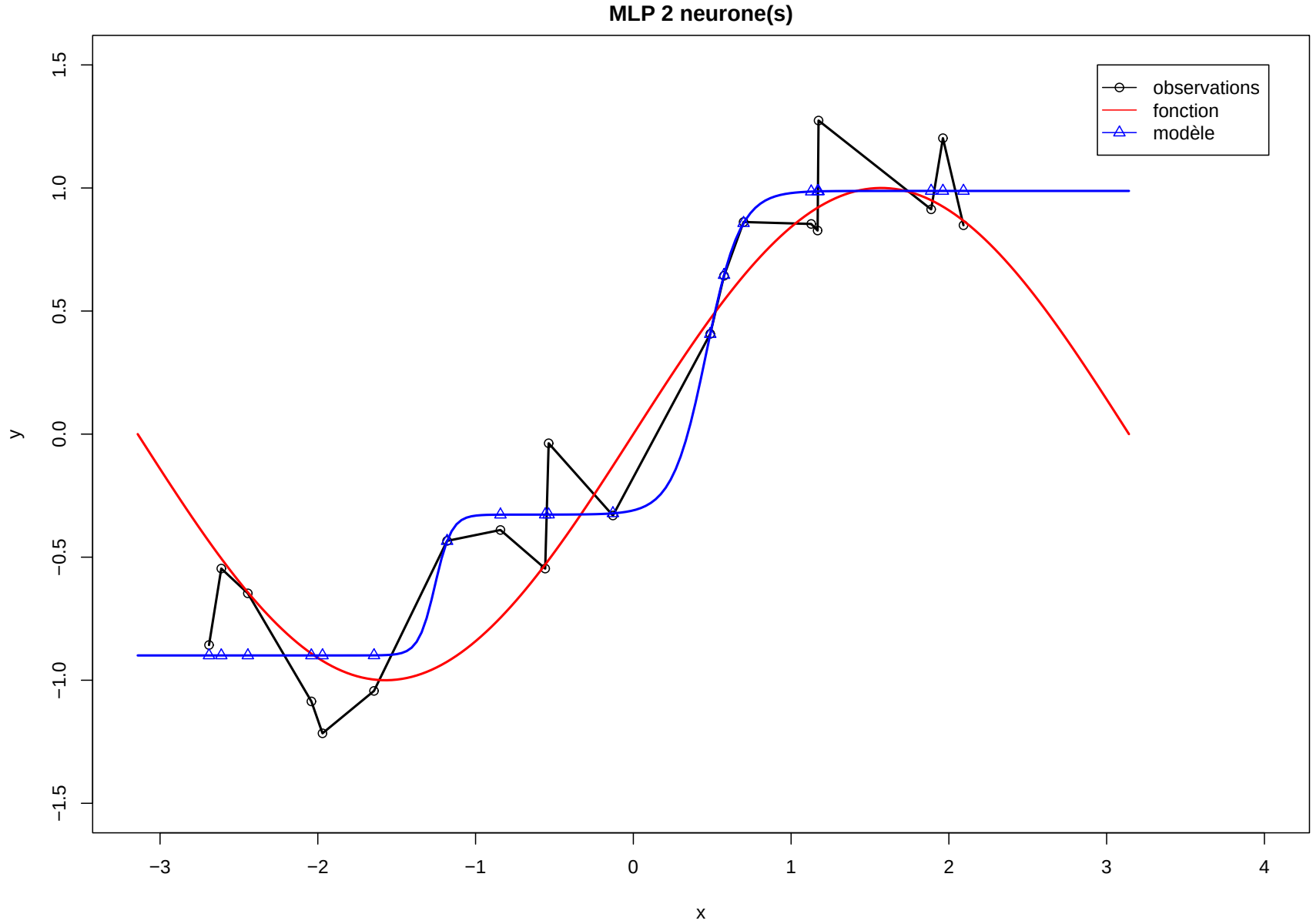
Comportement de divers modèles



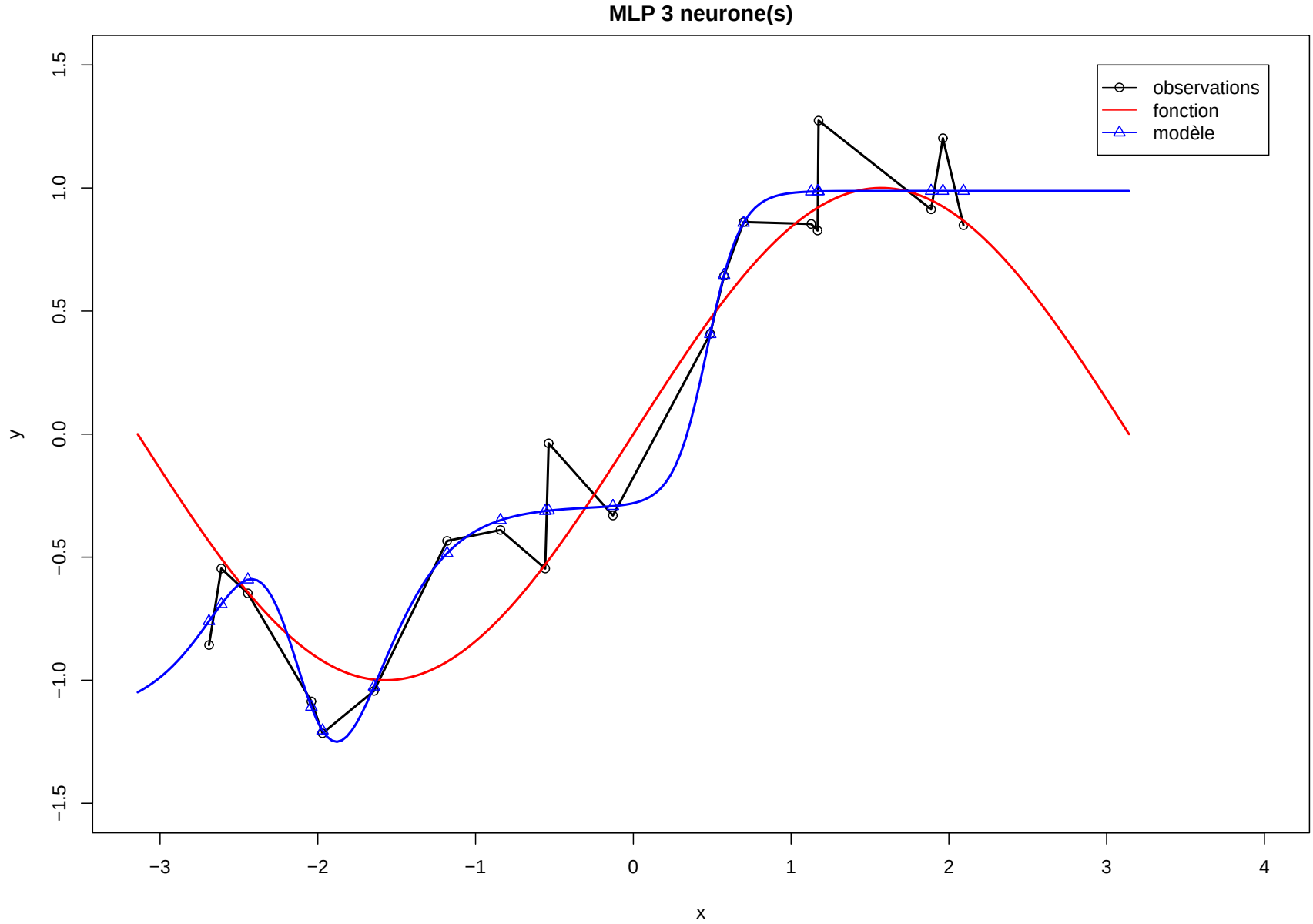
Comportement de divers modèles



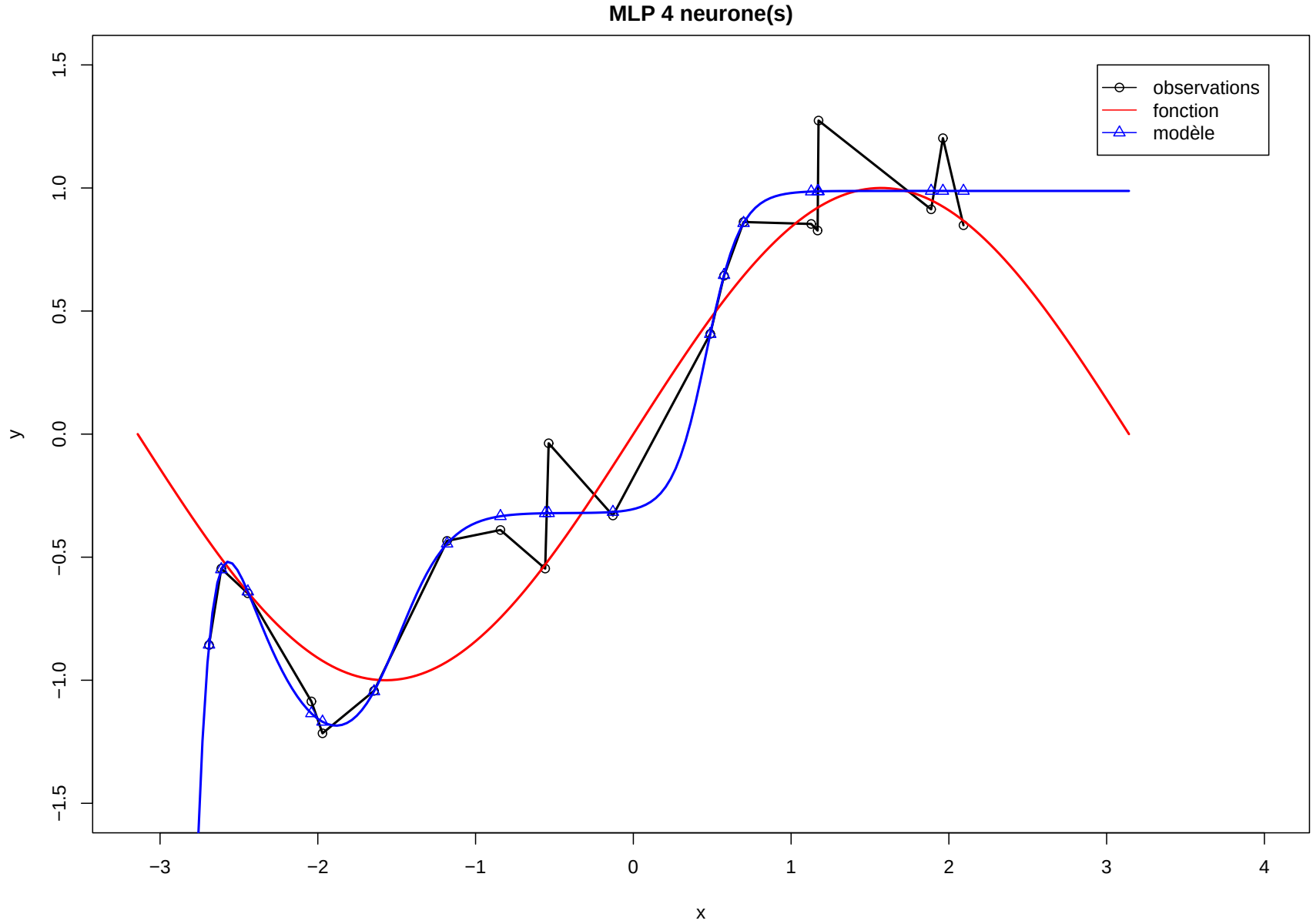
Comportement de divers modèles



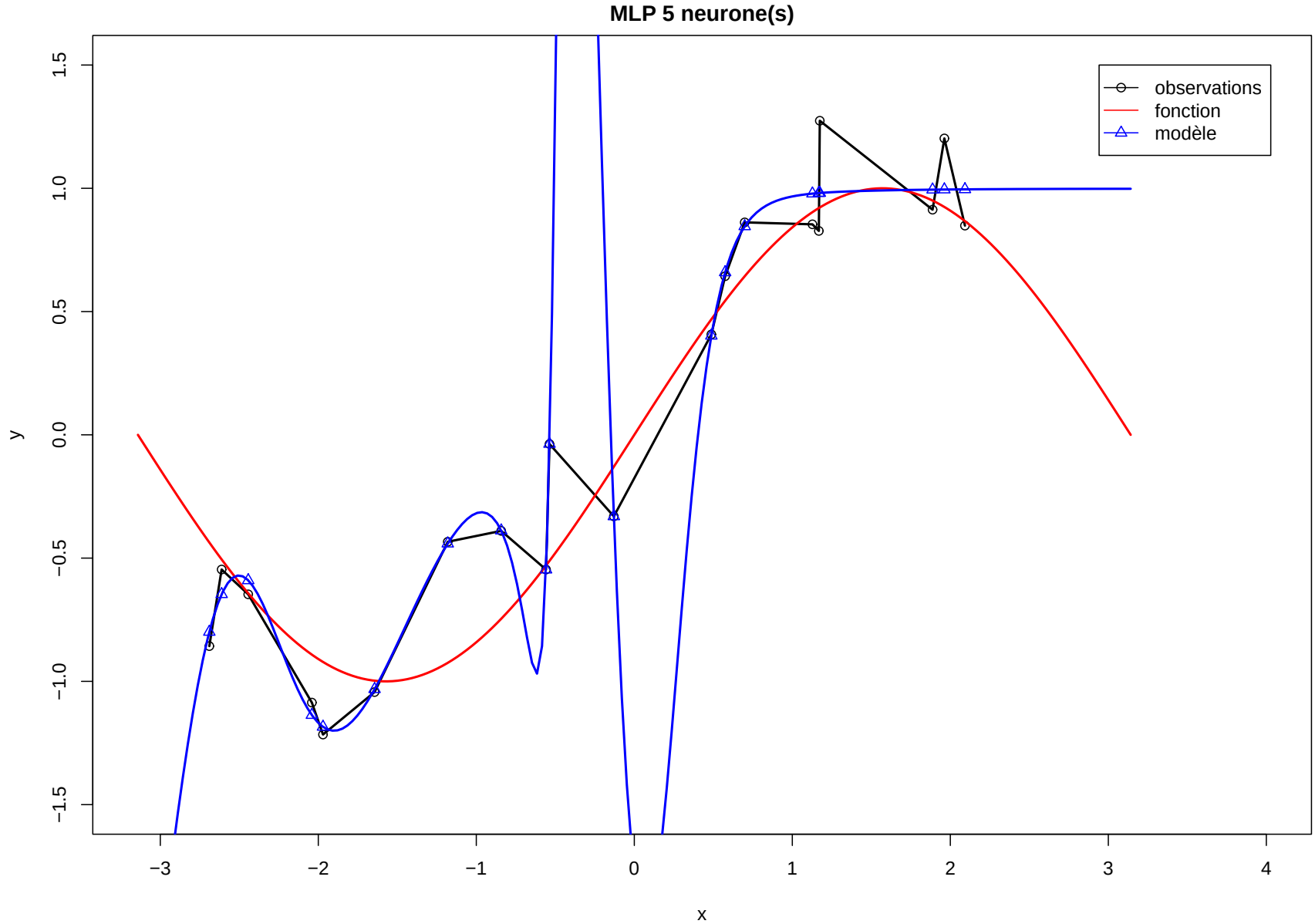
Comportement de divers modèles



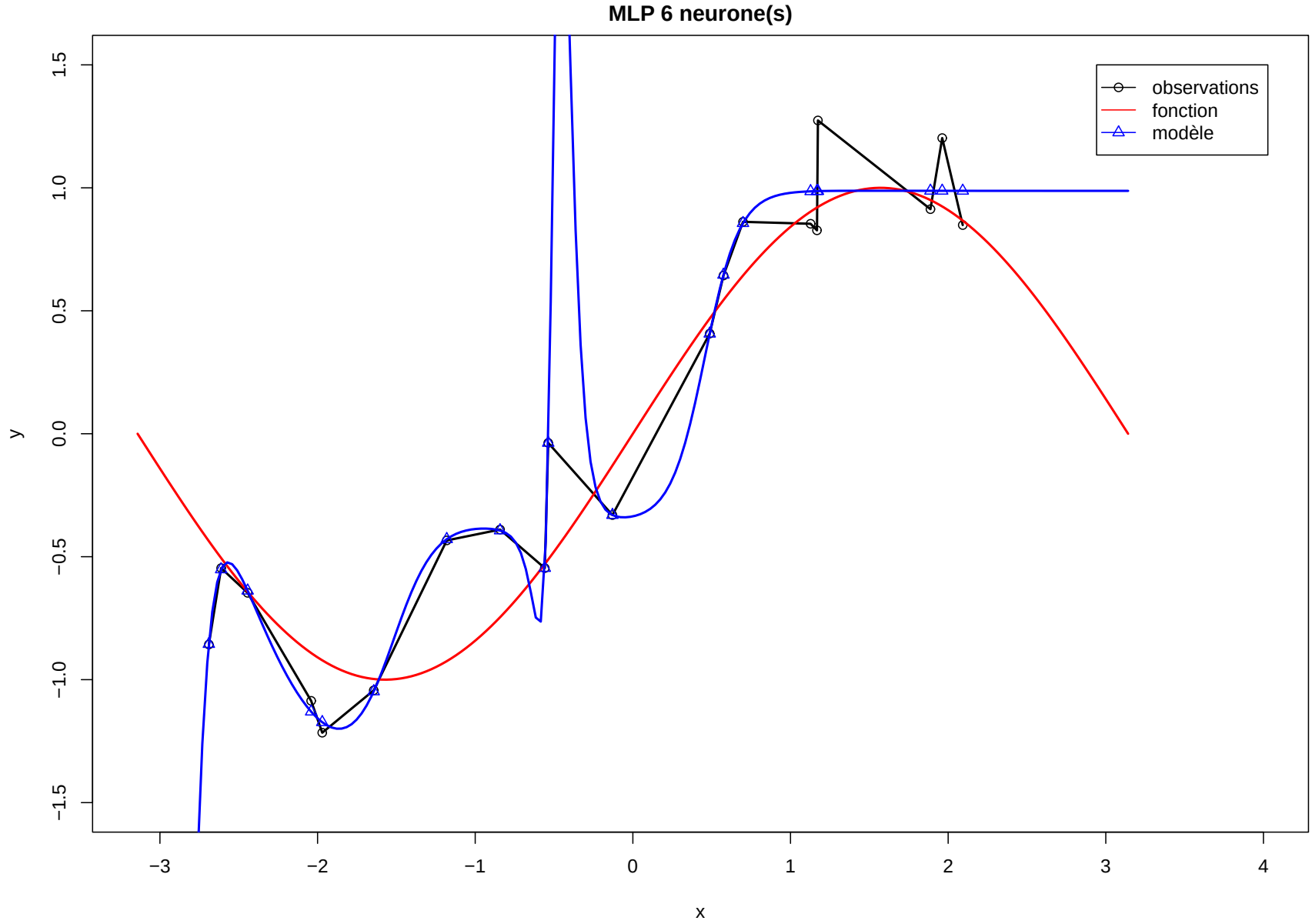
Comportement de divers modèles



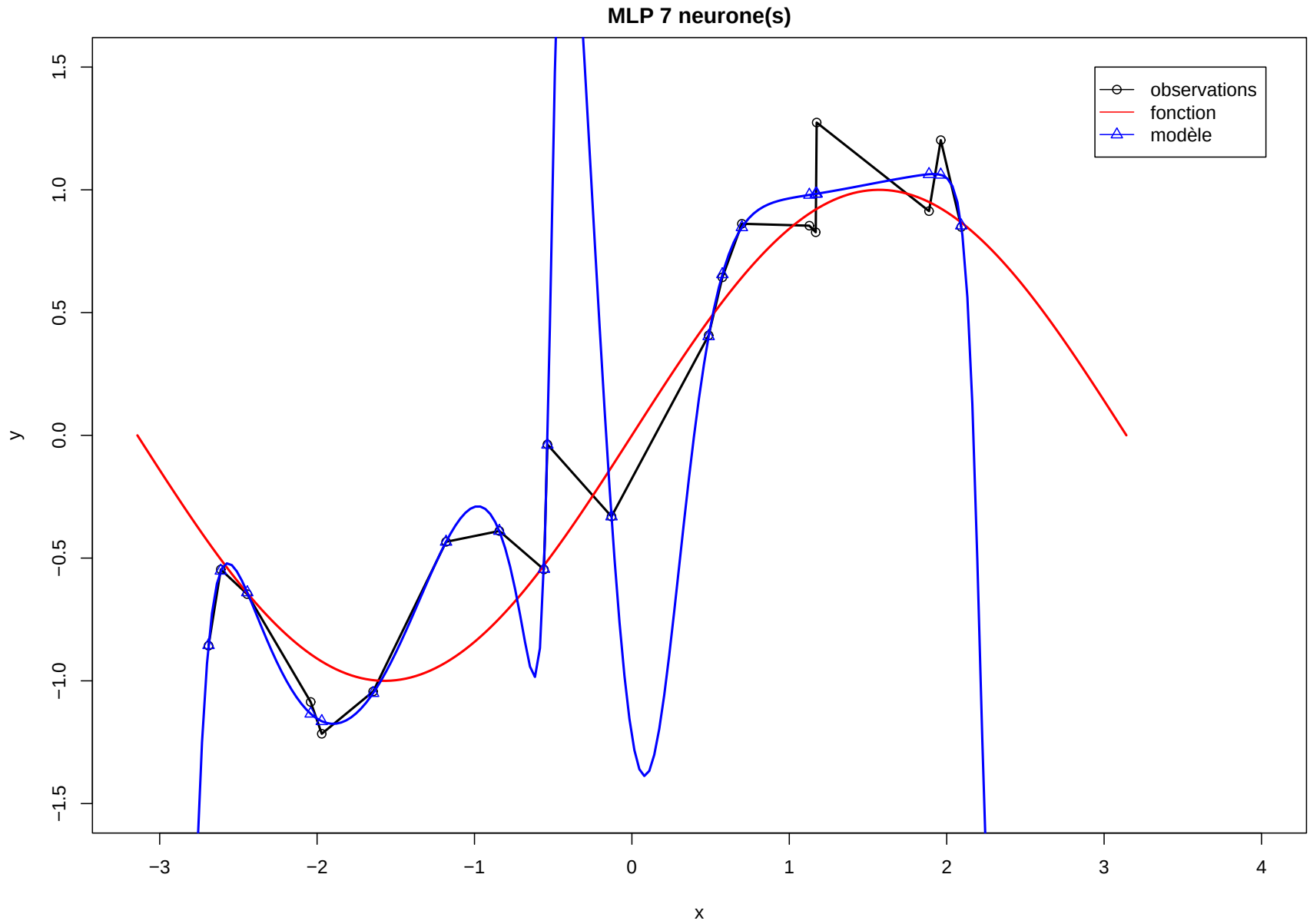
Comportement de divers modèles



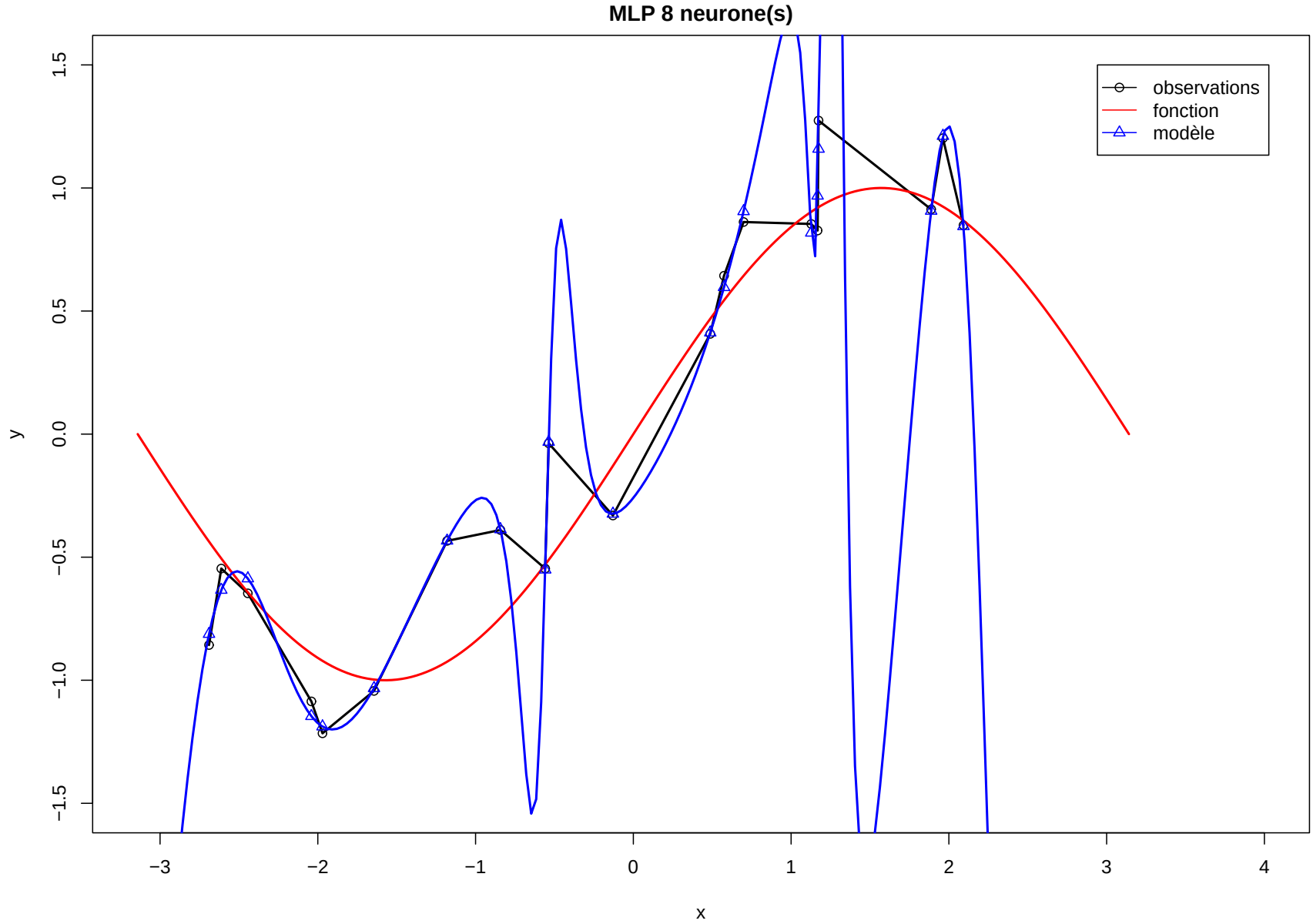
Comportement de divers modèles



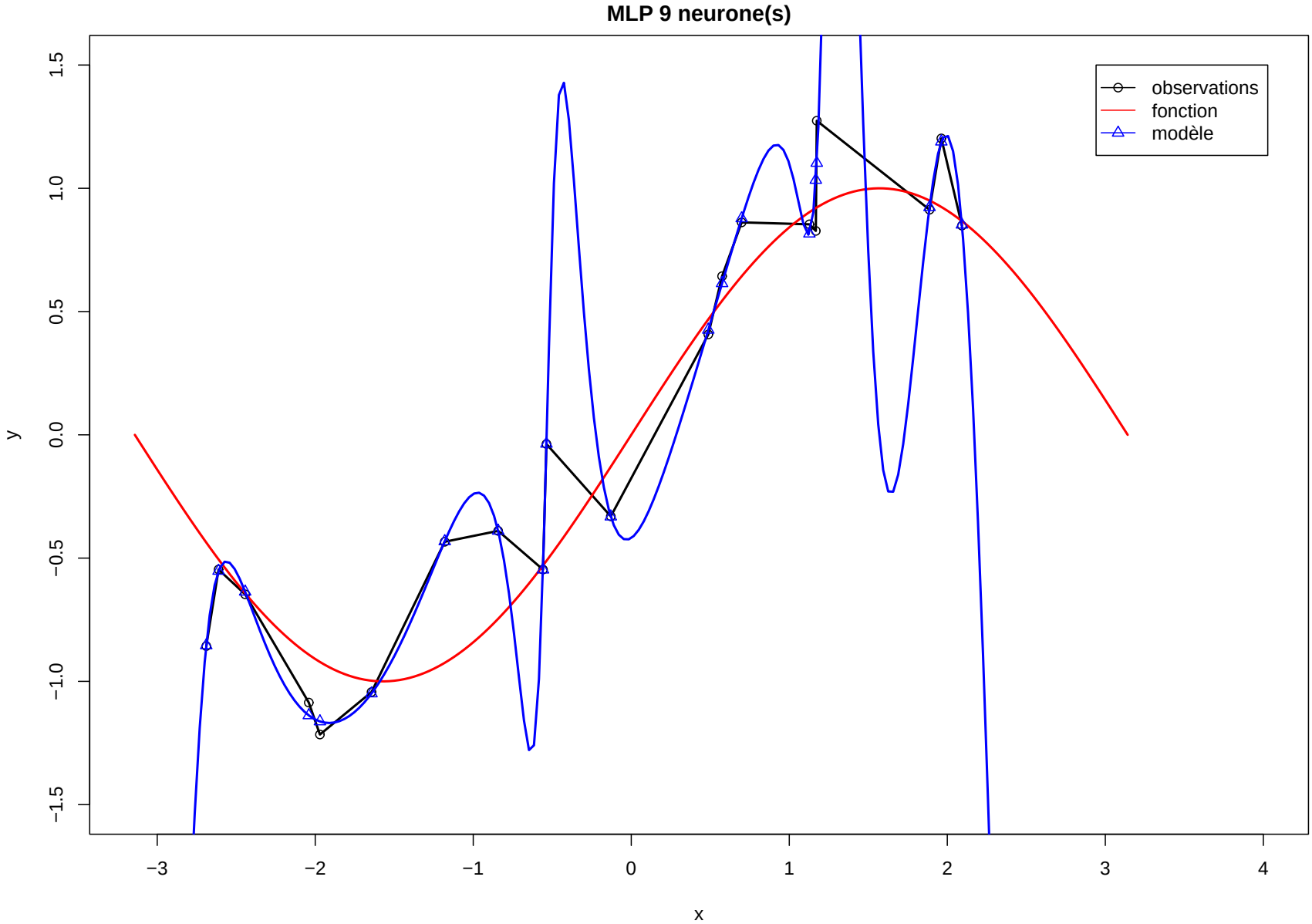
Comportement de divers modèles



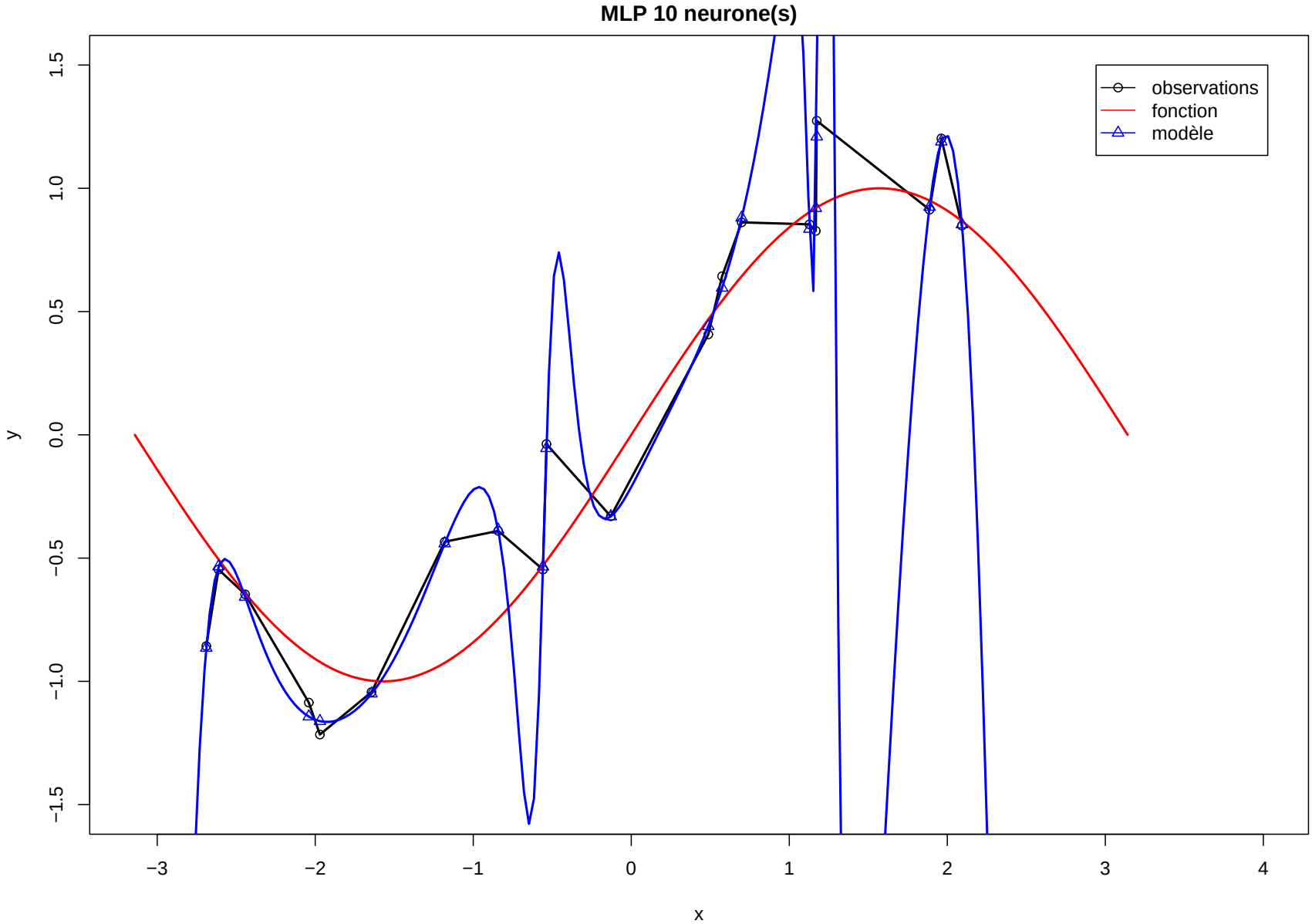
Comportement de divers modèles



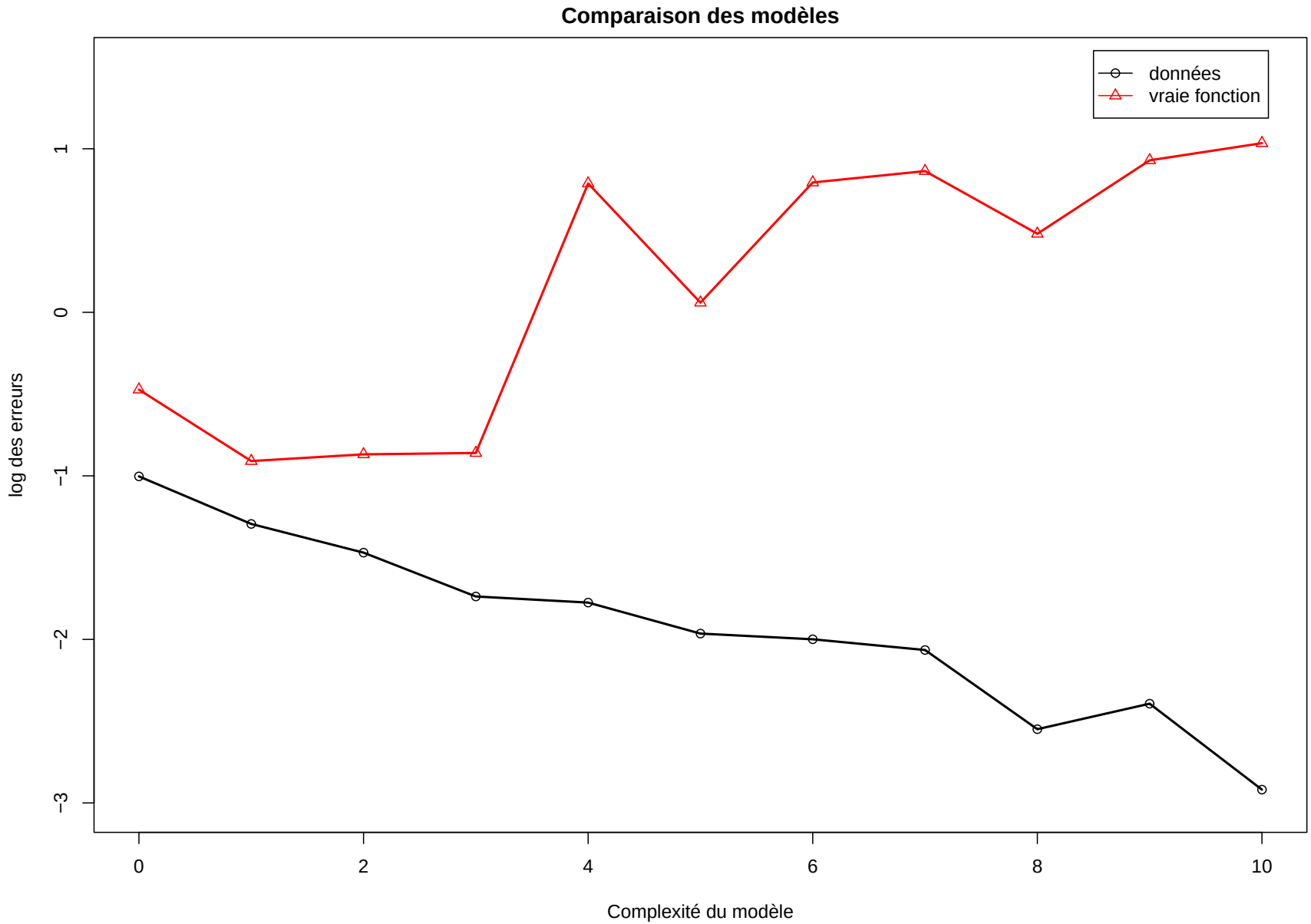
Comportement de divers modèles



Comportement de divers modèles



Evolution de l'erreur



Evolution de l'erreur (2)

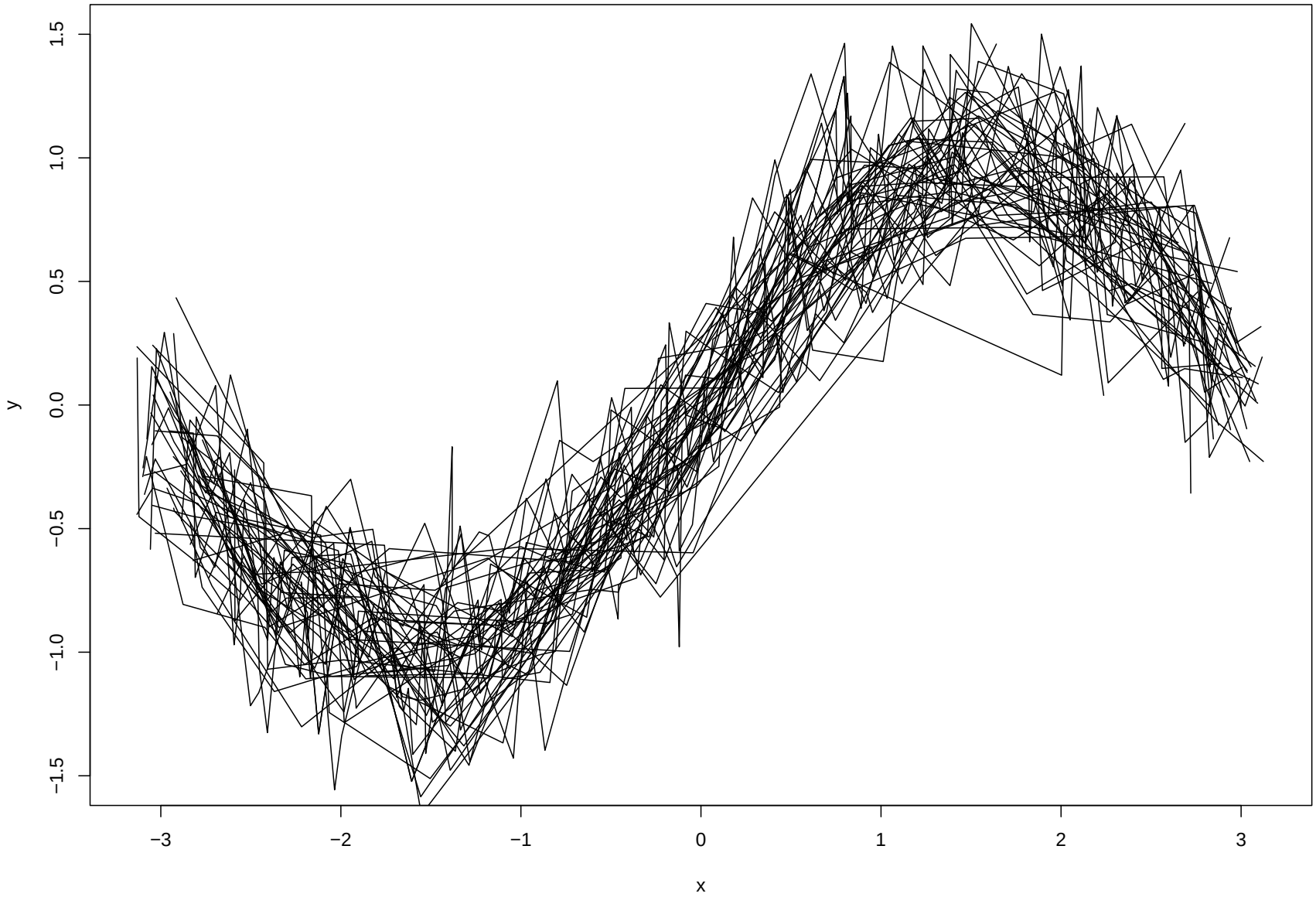
- Baisse de l'erreur sur les données : baisse du biais
- Augmentation de l'erreur par rapport à la vraie fonction : augmentation de la variance (sensibilité aux données, *overfitting*)
- Choix du modèle très difficile : l'erreur sur les données n'est pas un bon critère
- Pistes pour une solution :
  - contrôler finement la puissance du modèle (éviter les "vagues")
  - estimer la "vraie erreur" :
    - découpage des données
    - re-simulation

Sensibilité aux données

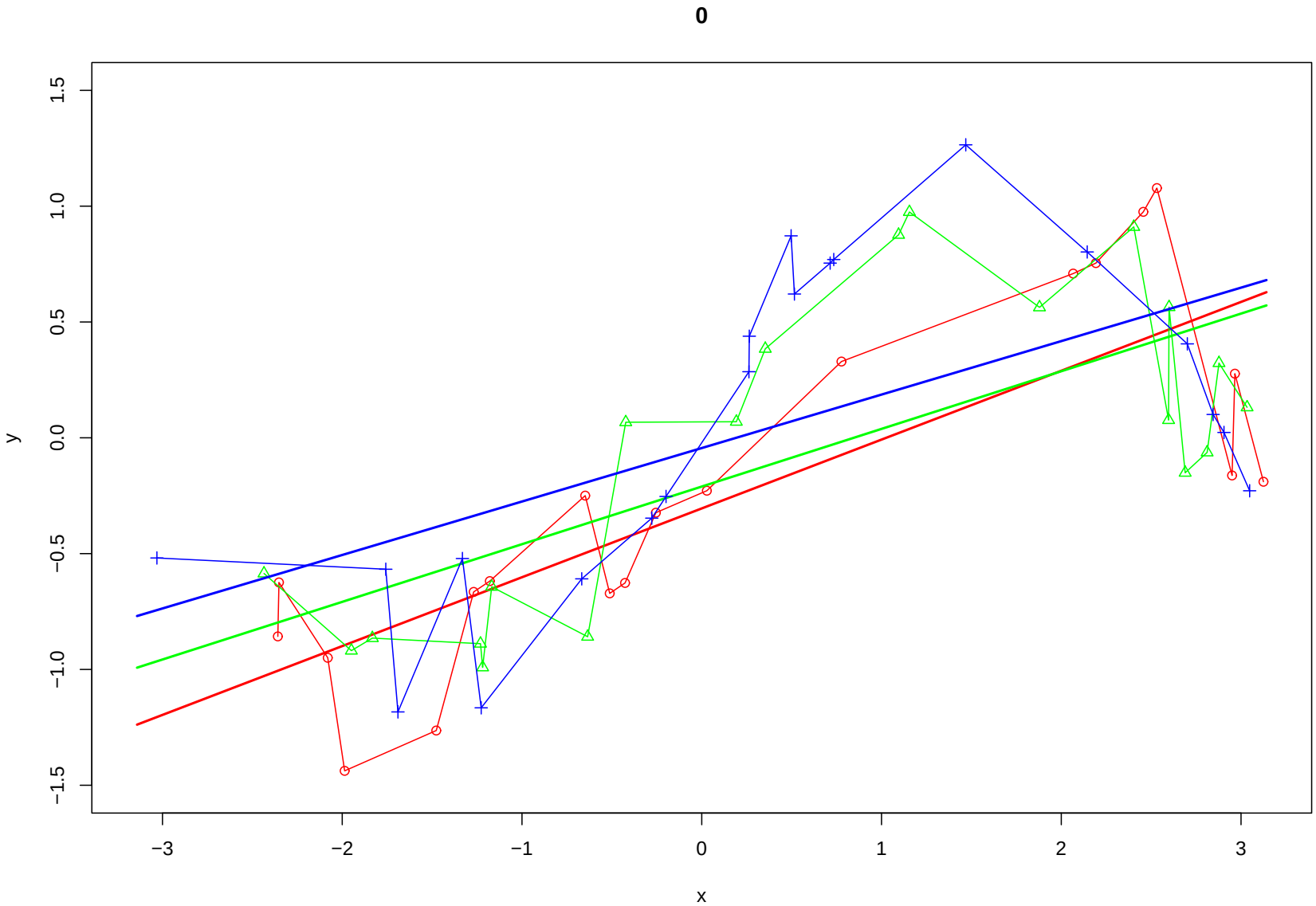
Illustration numérique de la variance :

- on engendre 50 jeux de données :
  - toujours sous la forme $y_i = f(x_i) + \epsilon_i$
  - les x_i sont choisis au hasard
- on estime le modèle pour chaque jeu de données
- on trace la réponse moyenne du modèle : la différence avec le vrai modèle est le **biais**
- on trace les valeurs extrêmes (moyenne $+/-$ 2 fois l'écart-type) : représentation de la **variance**
- sur cet exemple :
  - le biais devient rapidement nul (ou presque)
  - la variance explose

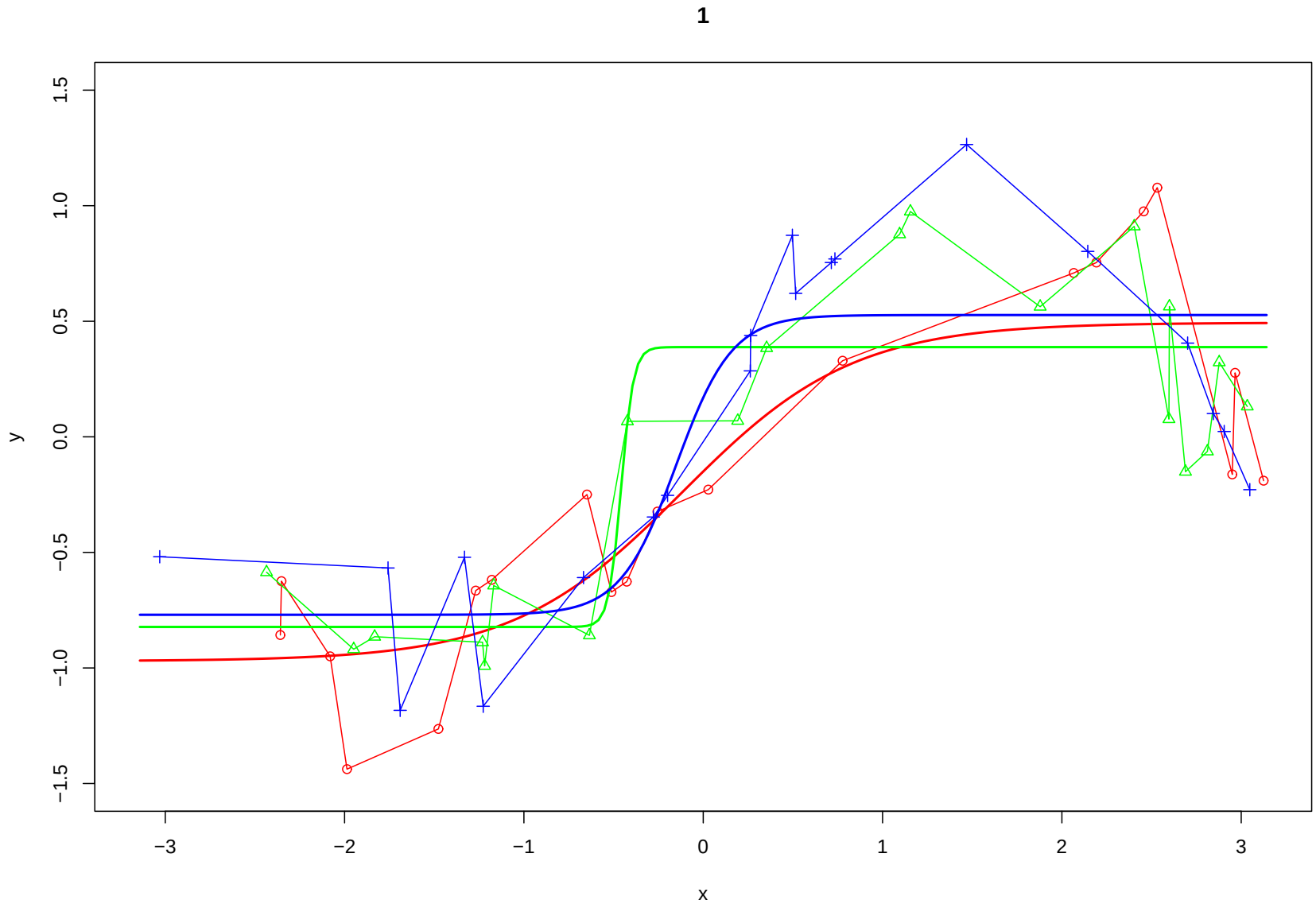
Sensibilité aux données : données



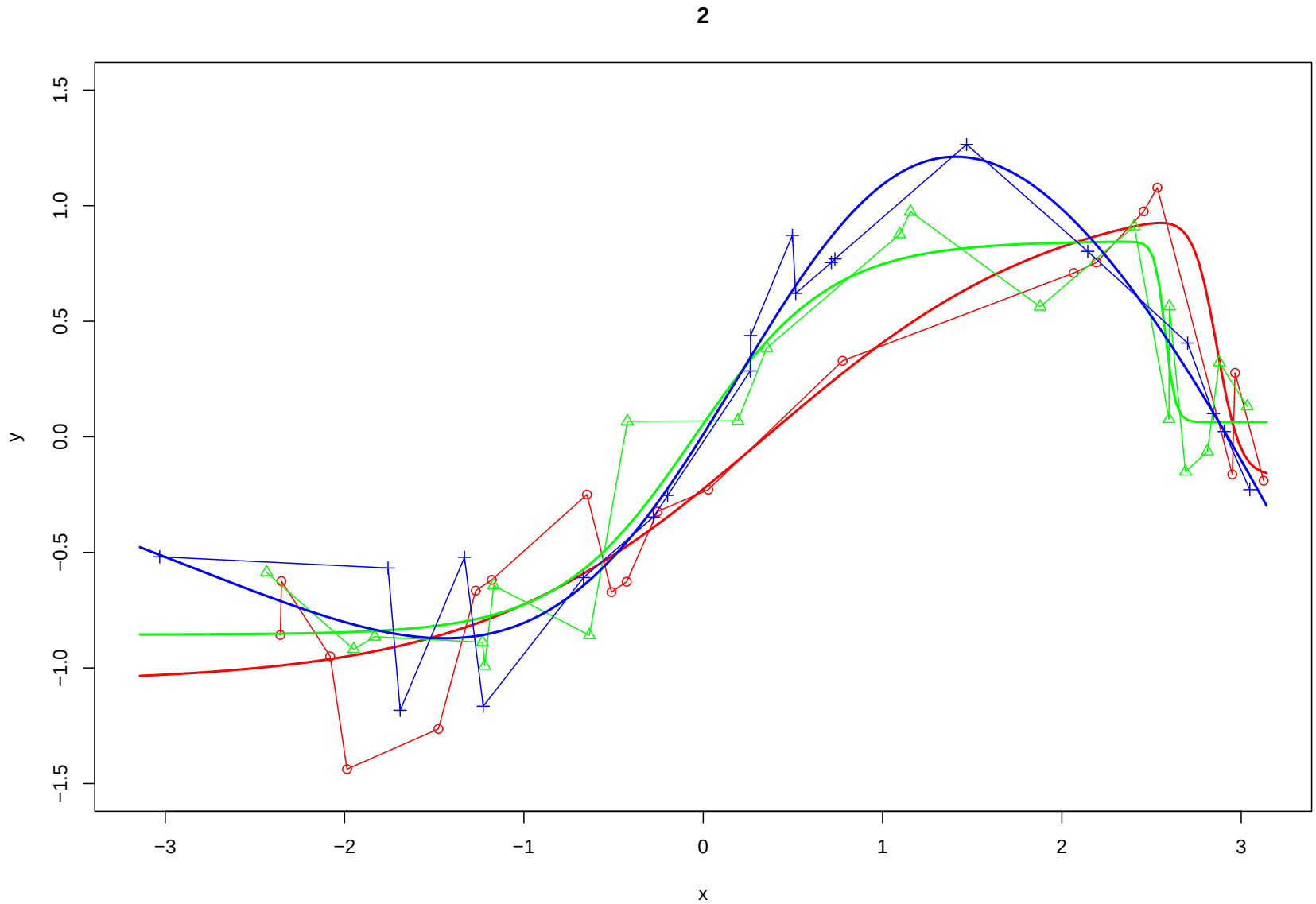
Sensibilité aux données : résultats sur 3 courbes



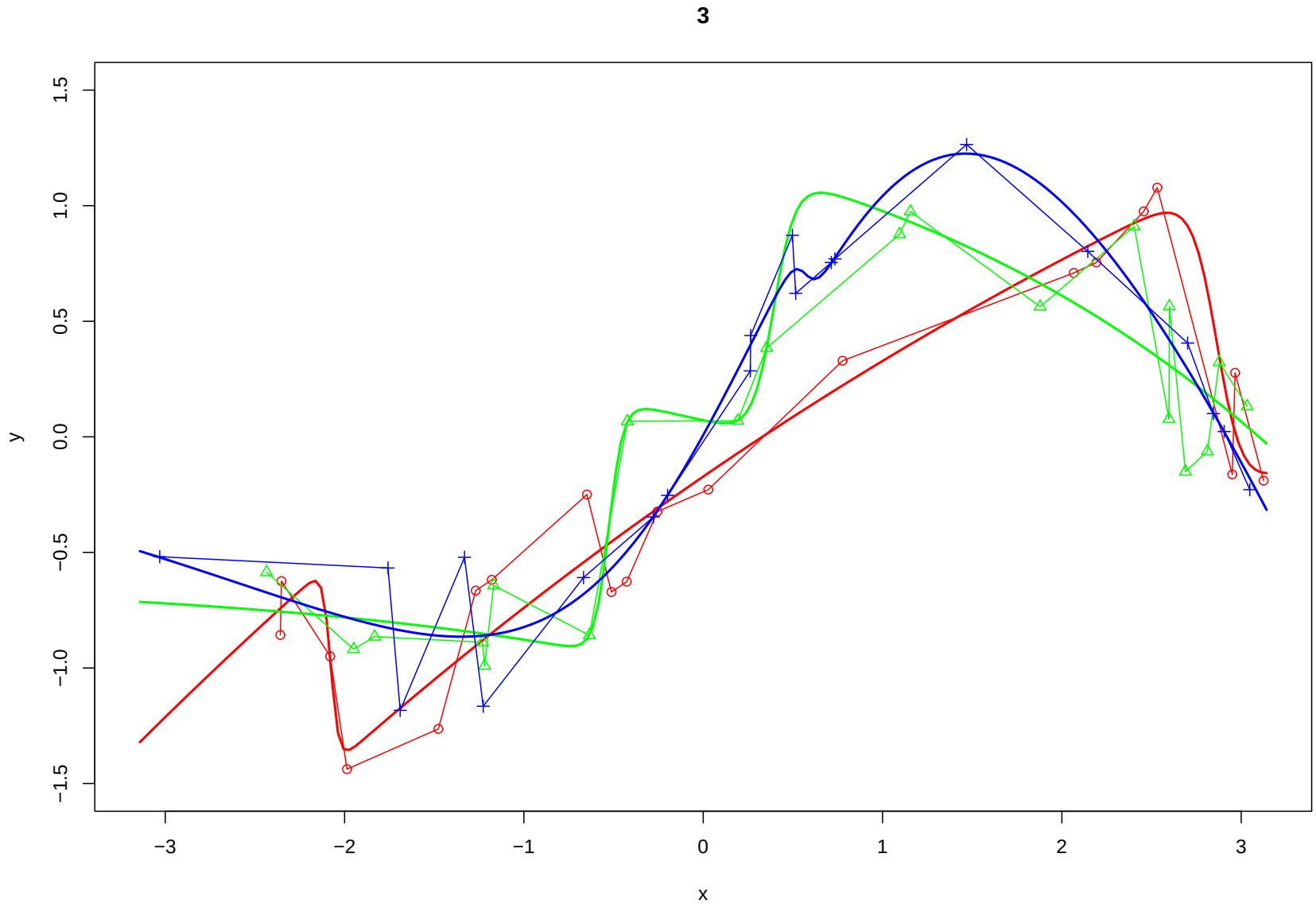
Sensibilité aux données : résultats sur 3 courbes



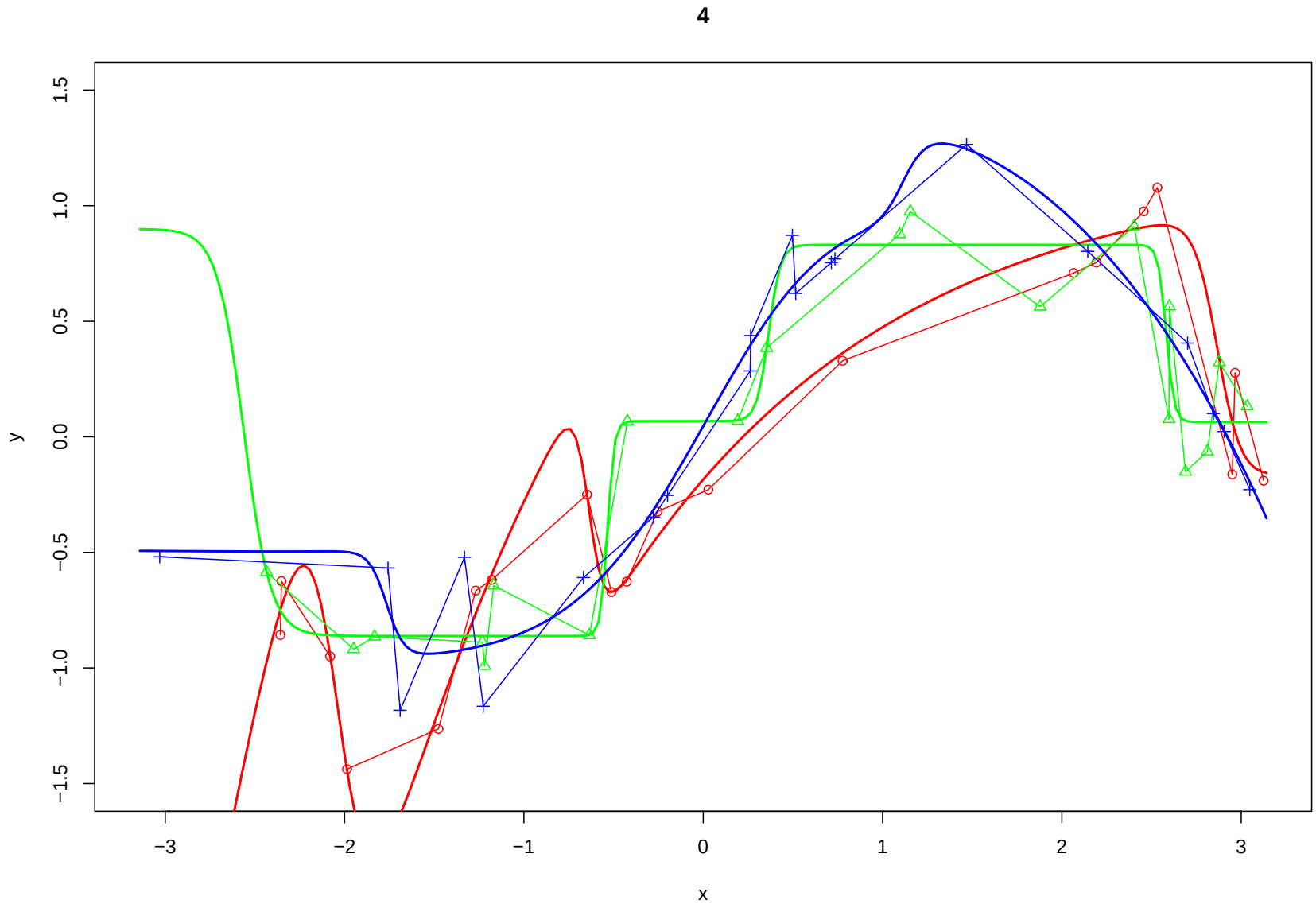
Sensibilité aux données : résultats sur 3 courbes



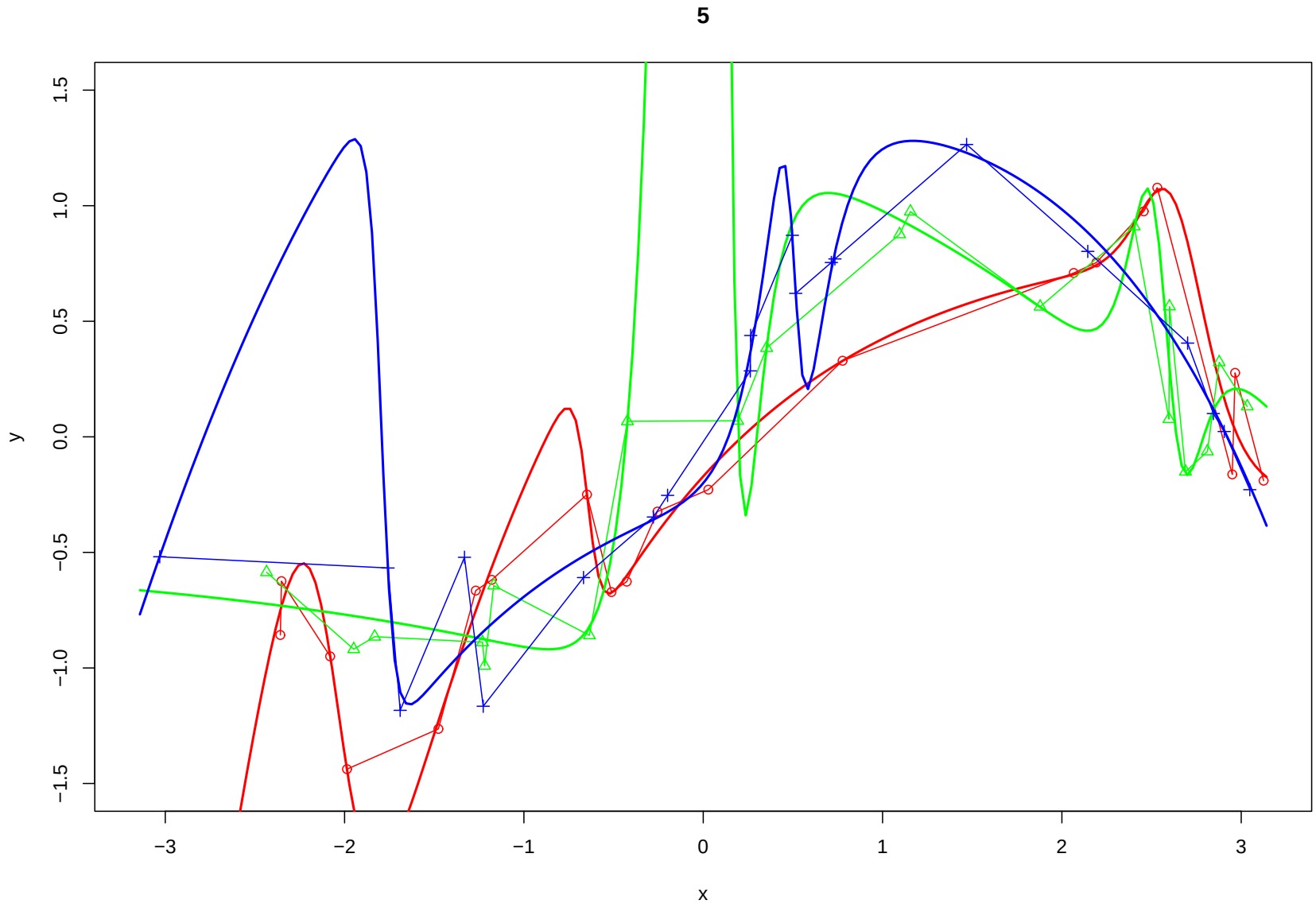
Sensibilité aux données : résultats sur 3 courbes



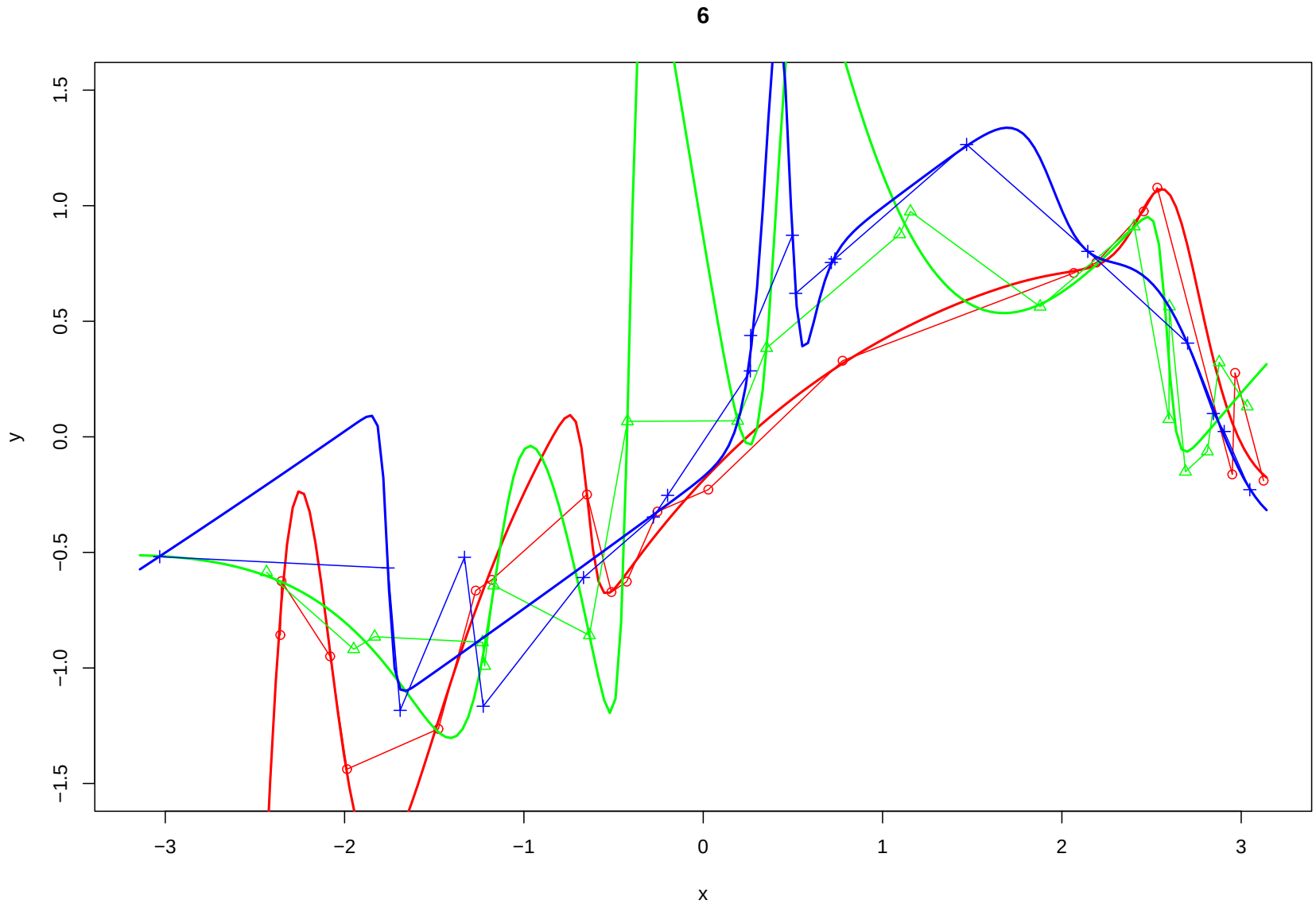
Sensibilité aux données : résultats sur 3 courbes



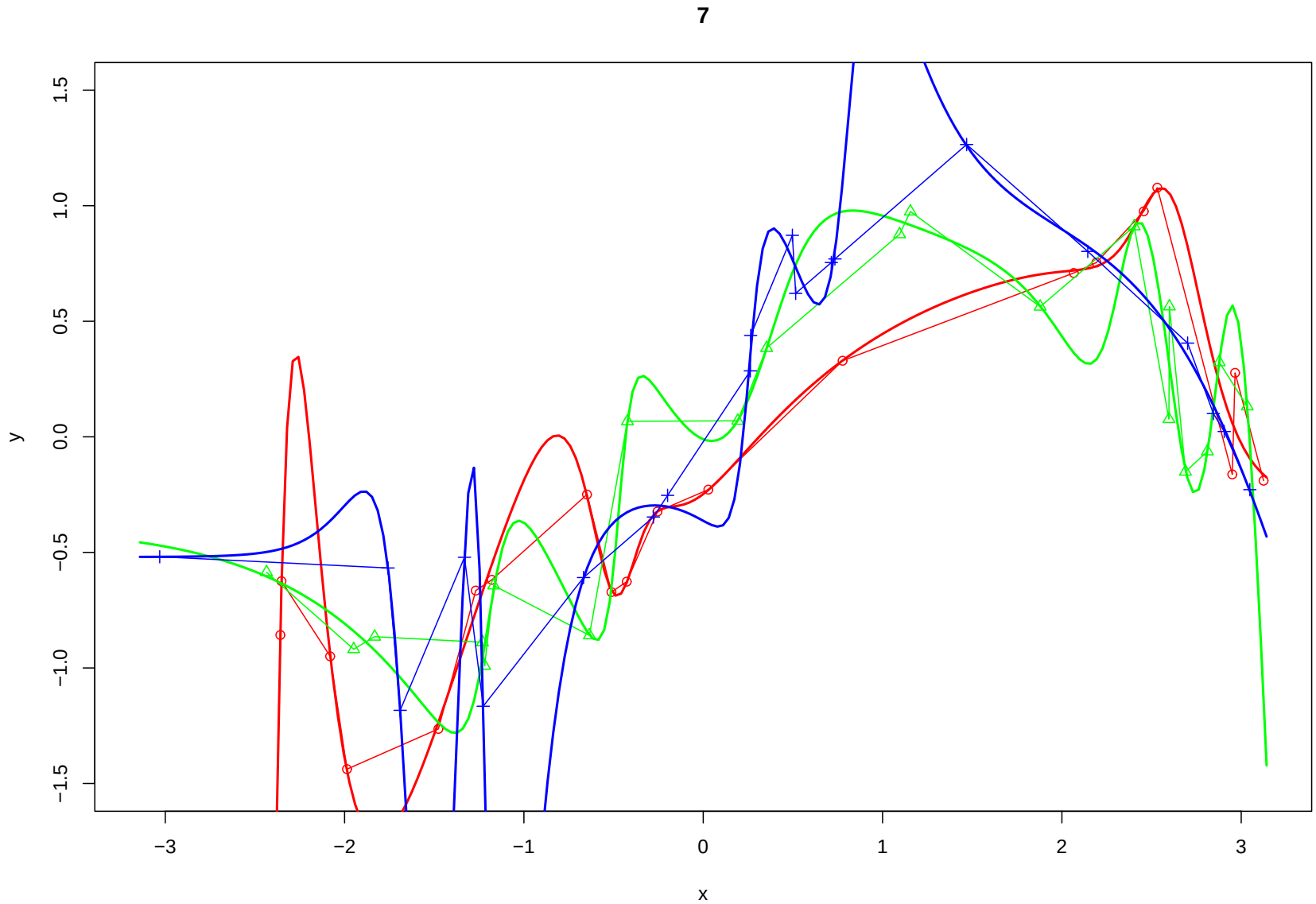
Sensibilité aux données : résultats sur 3 courbes



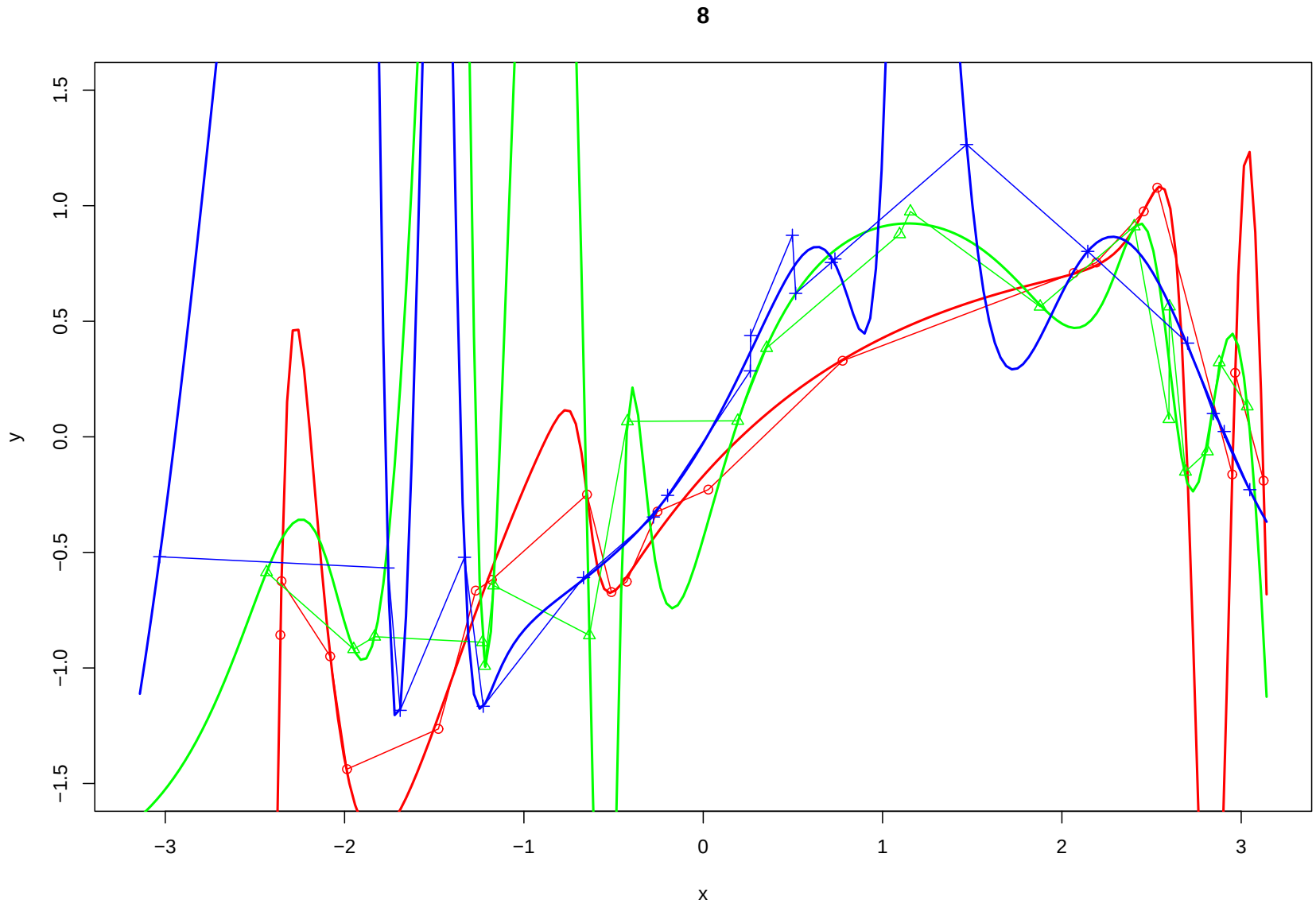
Sensibilité aux données : résultats sur 3 courbes



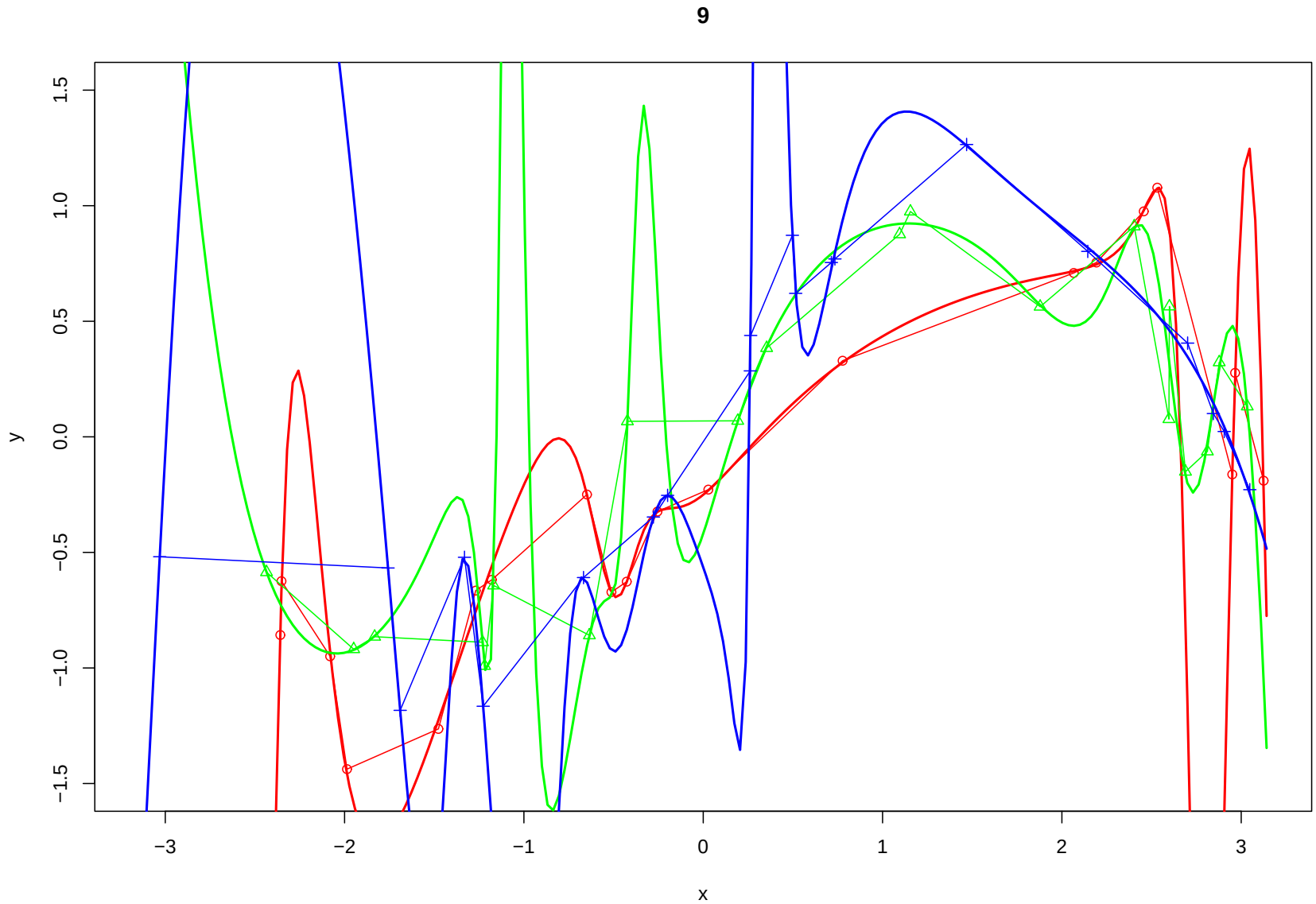
Sensibilité aux données : résultats sur 3 courbes



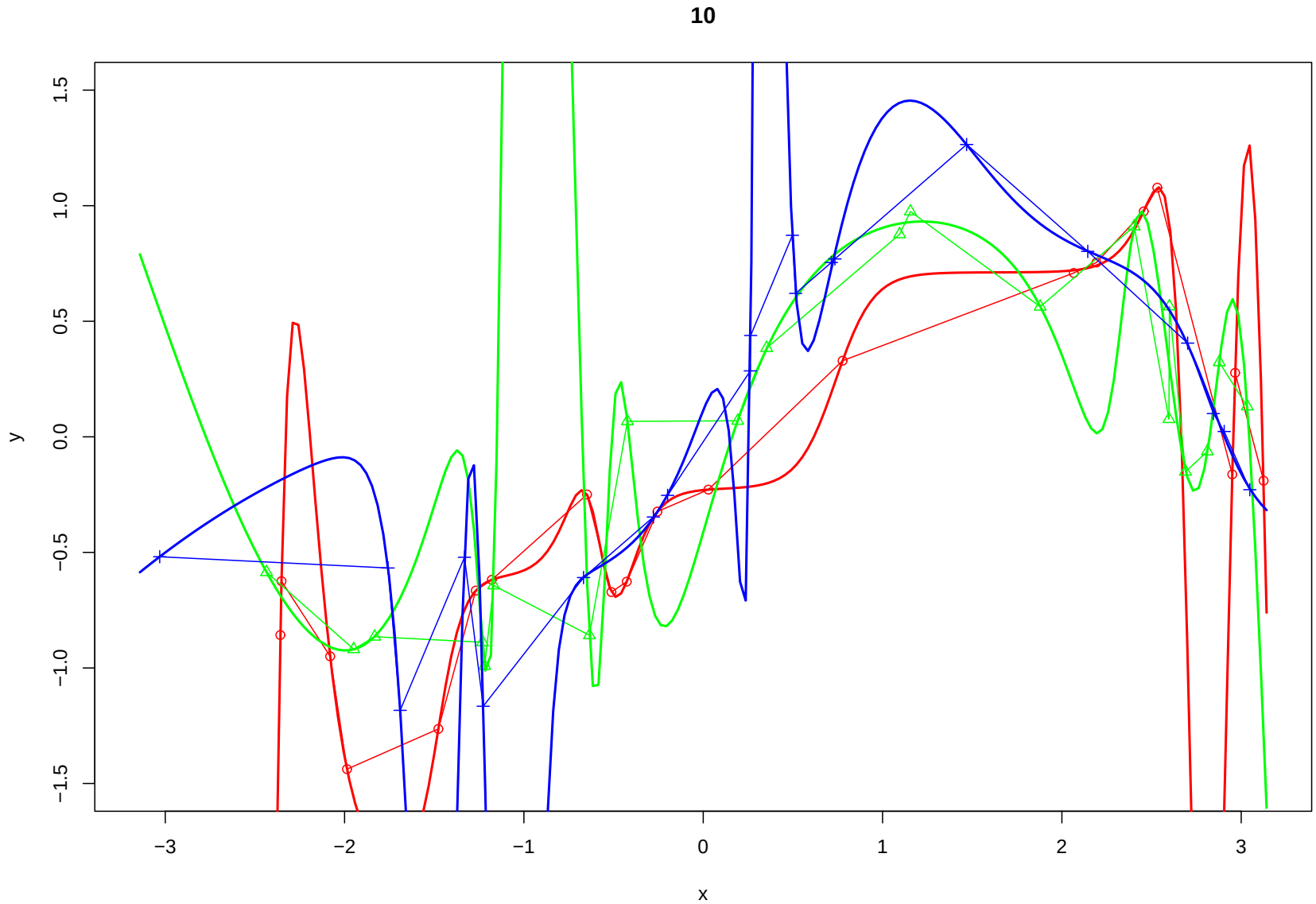
Sensibilité aux données : résultats sur 3 courbes



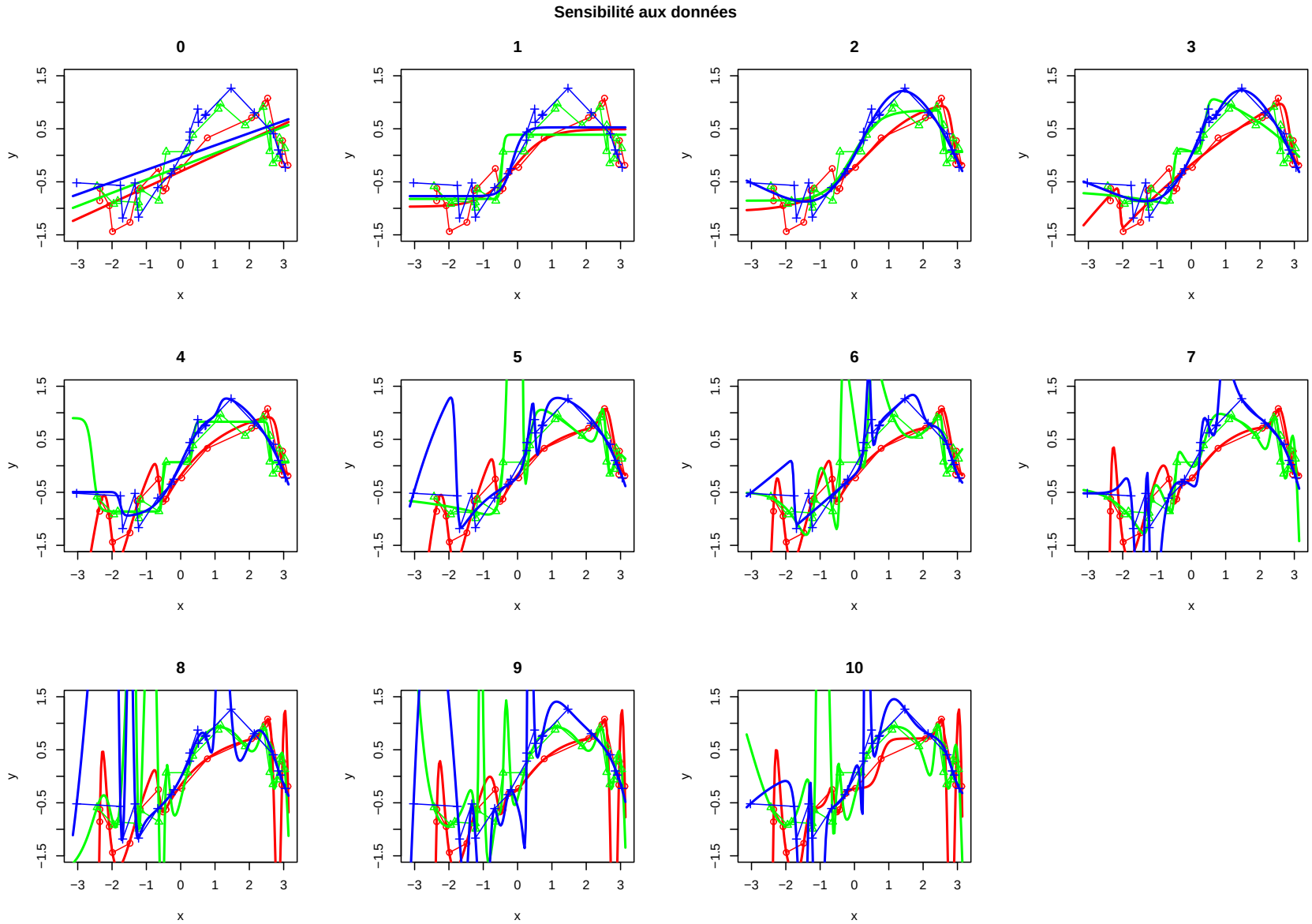
Sensibilité aux données : résultats sur 3 courbes



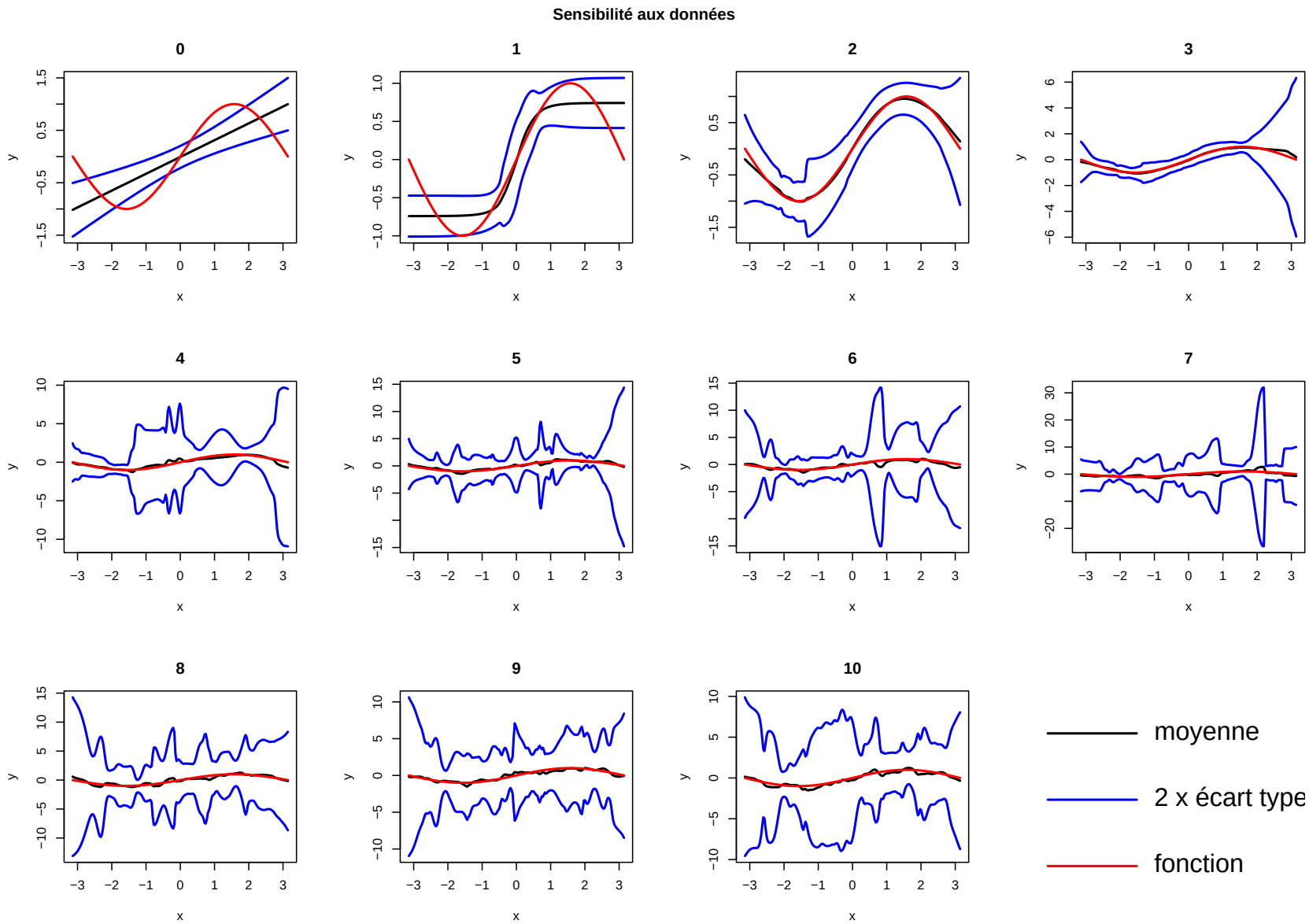
Sensibilité aux données : résultats sur 3 courbes



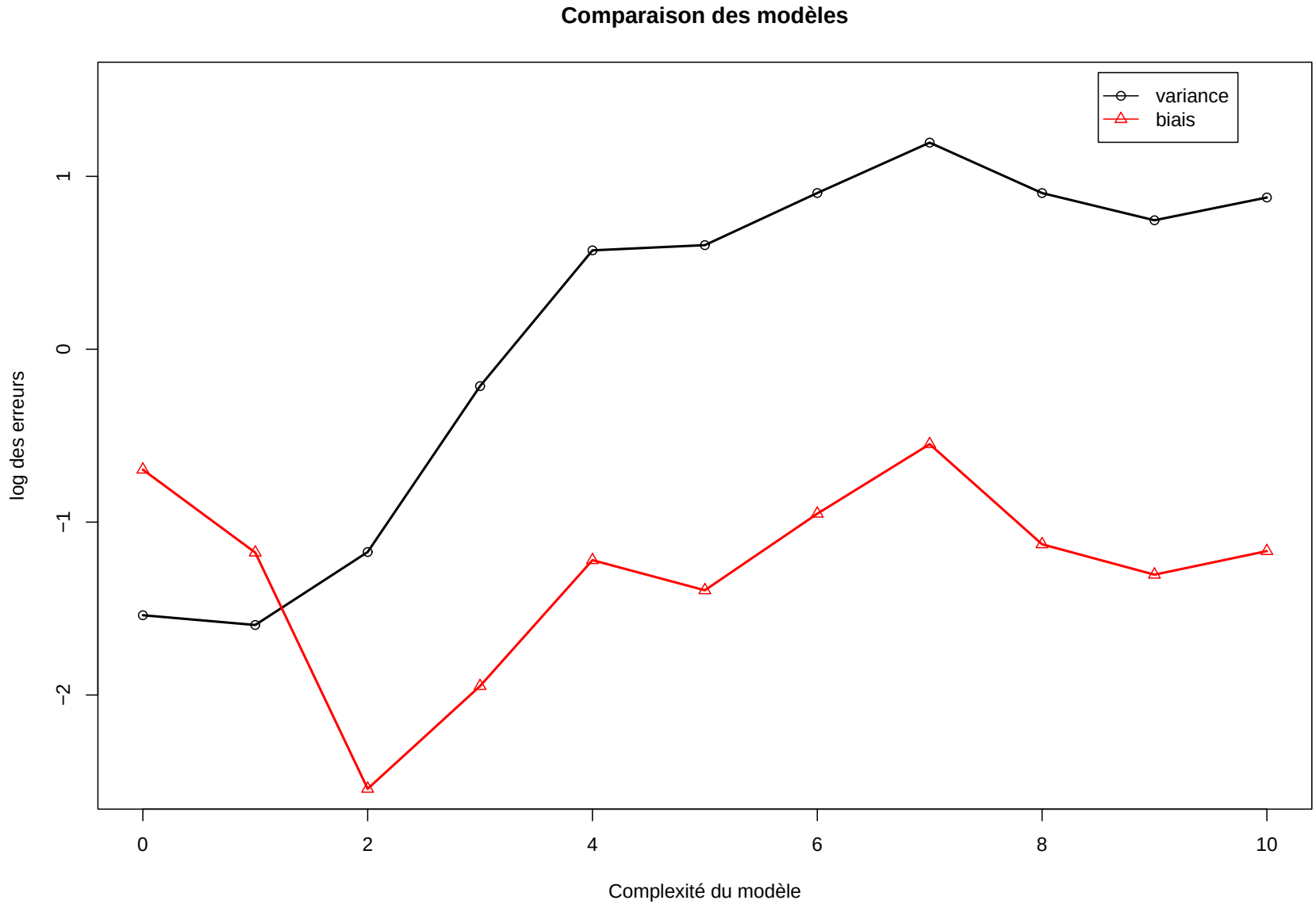
Sensibilité aux données : résultats sur 3 courbes



Sensibilité aux données : résultats



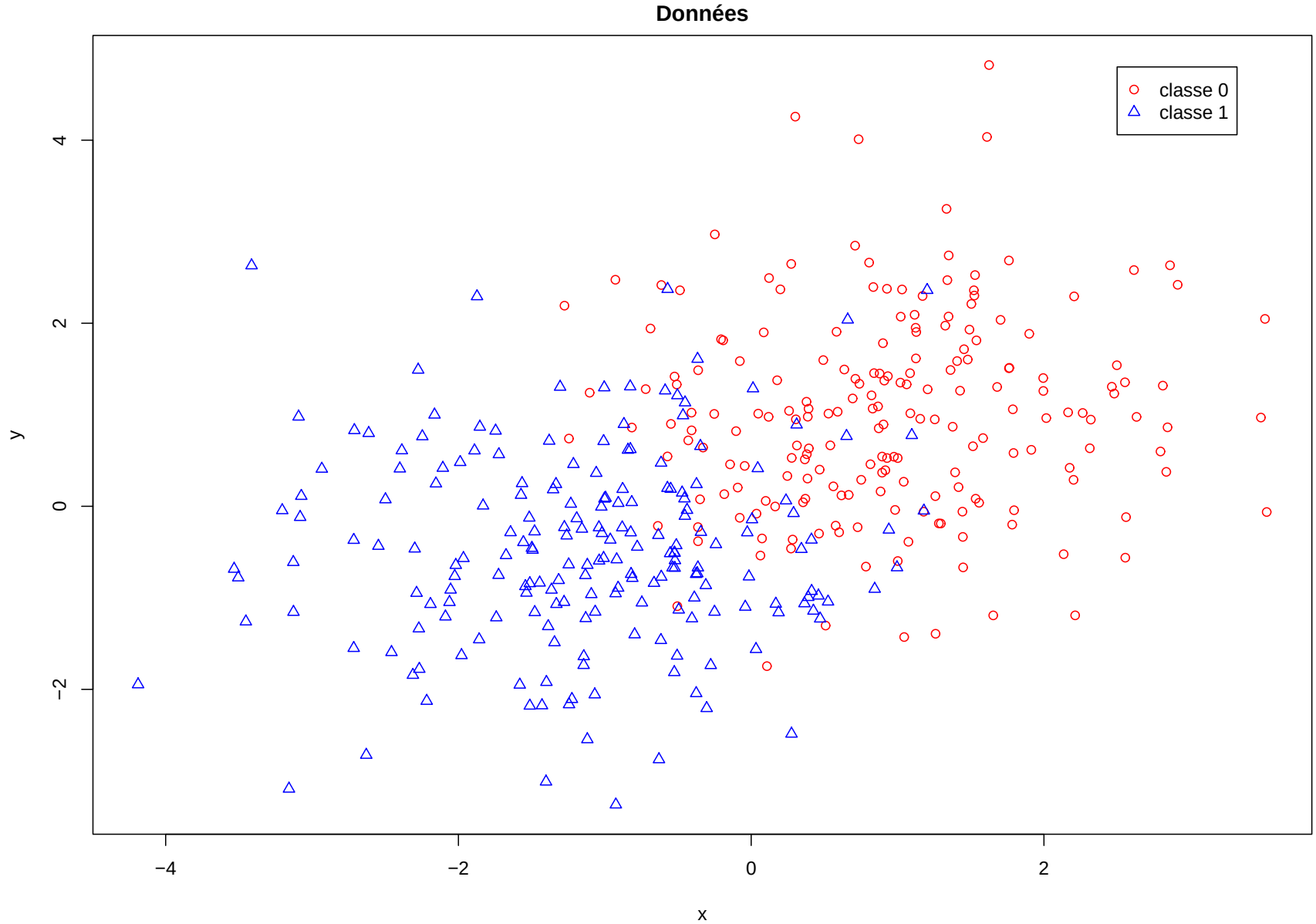
Décomposition biais+variance



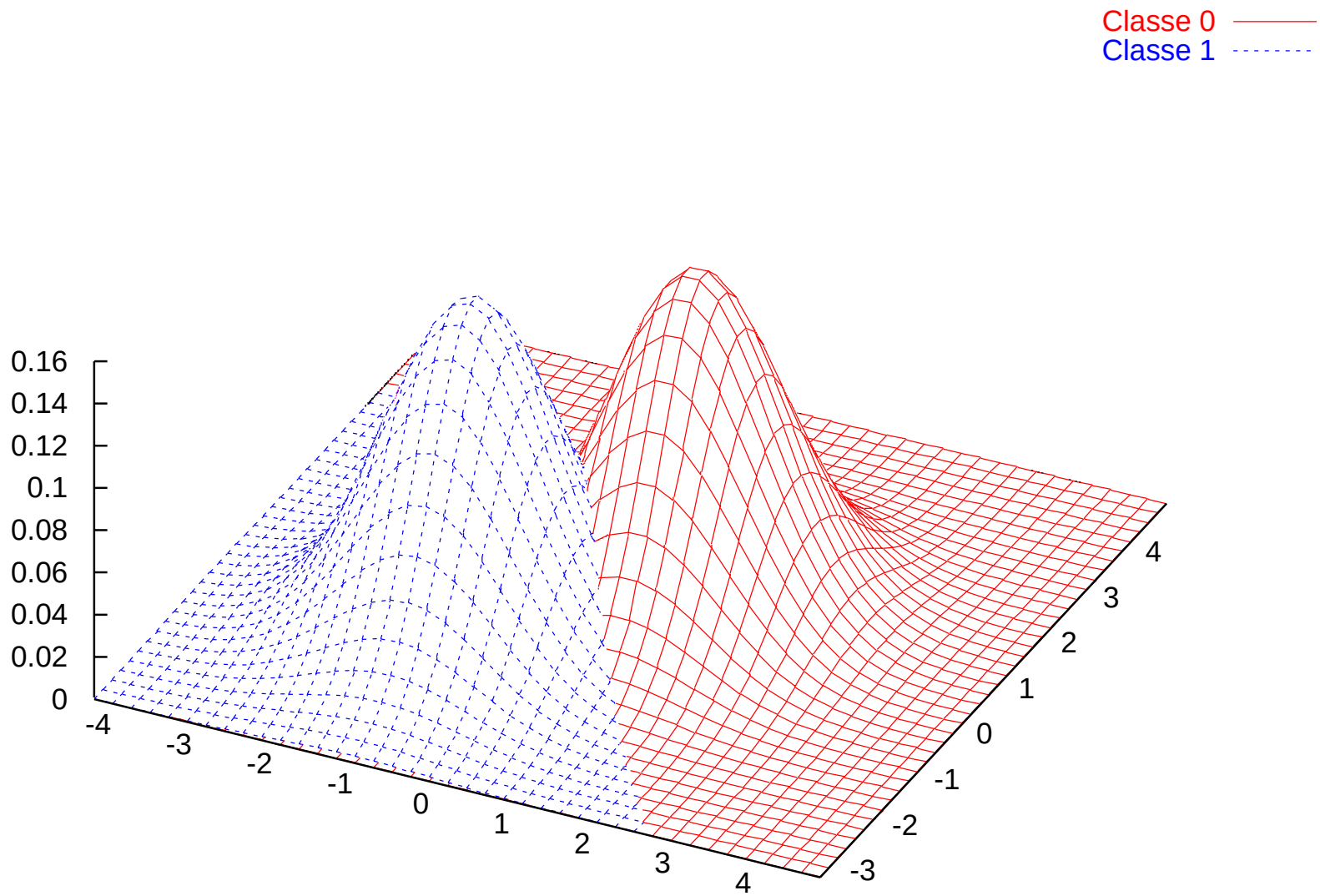
Un exemple de discrimination simple

- On possède une suite d'observations, les (x_i, y_i)
- Chaque observation (x, y) est placée dans une classe
- On cherche à expliquer modéliser la relation entre (x, y) et la classe d'appartenance
- Applications pratiques :
  - diagnostic médical ((x, y) : symptômes)
  - diagnostic informatique (détection d'intrusion, de virus, etc.)
  - reconnaissance des formes (écriture, parole, cibles, etc.)
- Le but principal est de pouvoir classer de nouvelles observations
- Mêmes problèmes que pour la régression

Un exemple de discrimination



Densités

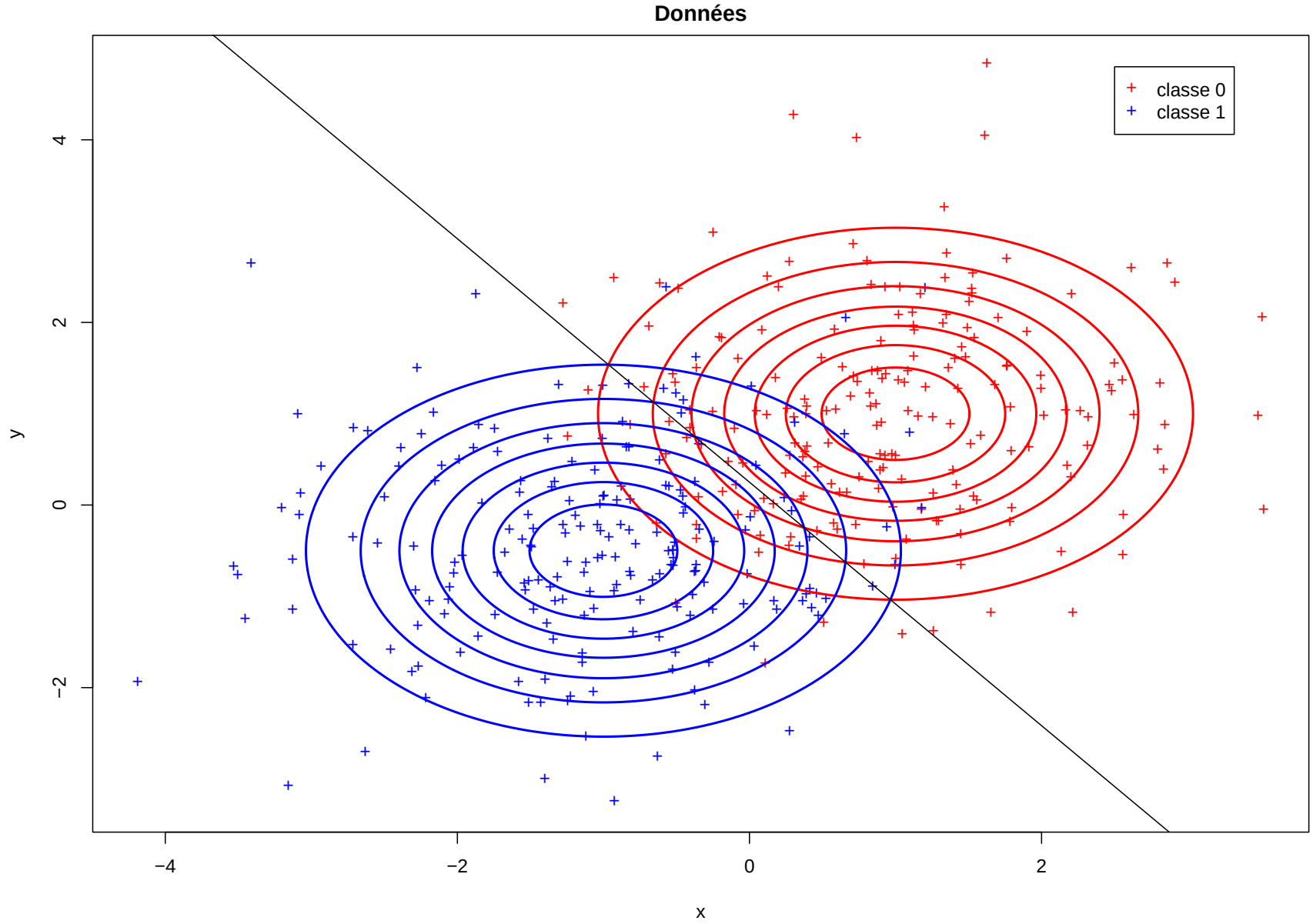

Classe 0
Classe 1
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-4
-3
-2
-1
0
1
2
3
4
-3
-2
-1
0
1
2
3
4

Connaissance imparfaite

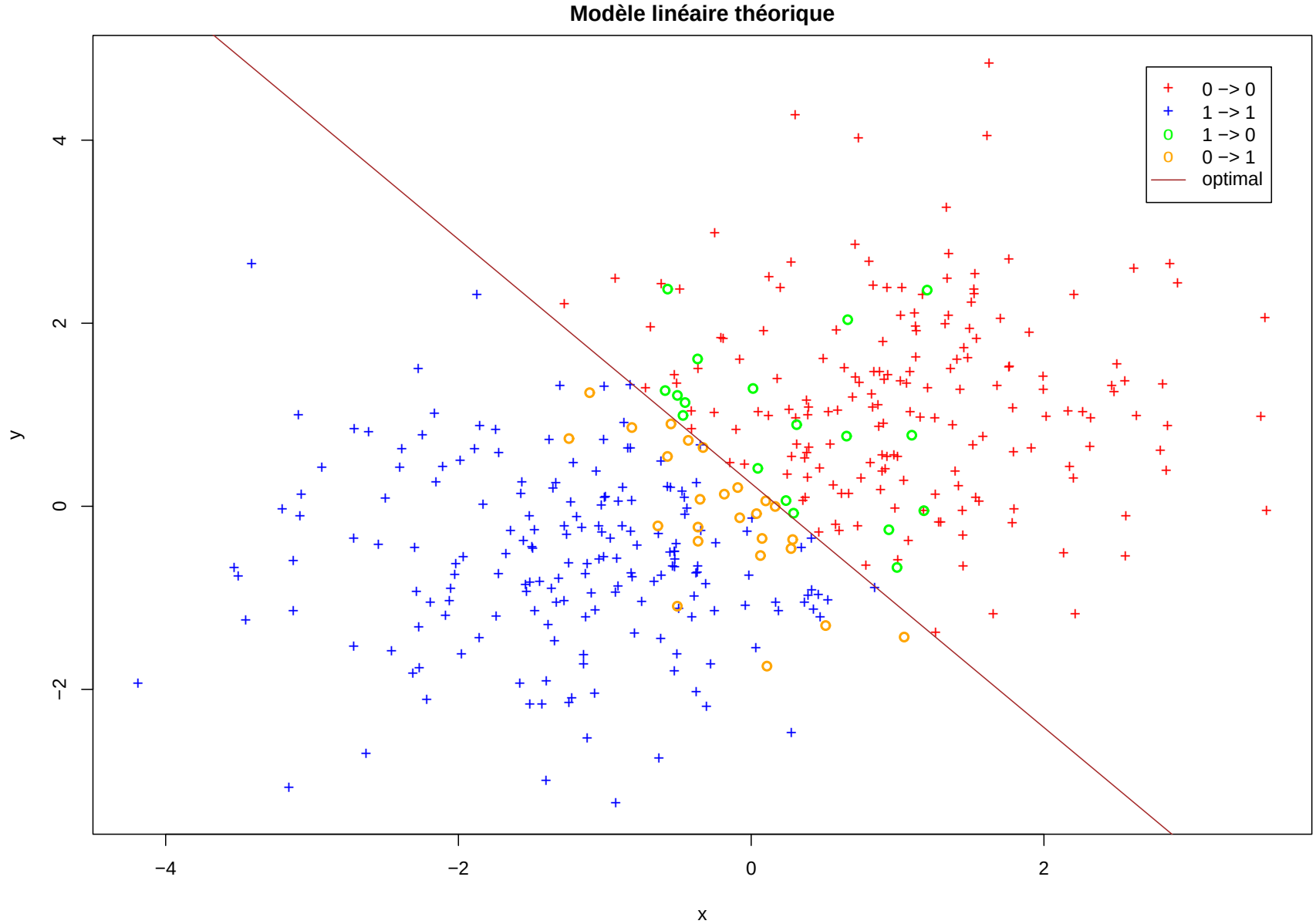
Les classes sont mélangées :

- connaissance imparfaite :
  - mesures fausses (diagnostic erroné, test manqué)
  - données manquantes (mesures impossibles à réaliser, dangereuses ou coûteuses)
- variabilité intrinsèque :
  - classe variable ou mal définie (consommateur en marketing par exemple)
  - évolution possible
  - variabilité naturelle (numération sanguine par exemple)
- le but **n'est pas** le classement parfait !
- estimer la probabilité d'appartenance à une classe
- sur l'exemple proposé, on peut démontrer qu'on ne peut pas descendre en dessous de 10.5% de mauvais classements

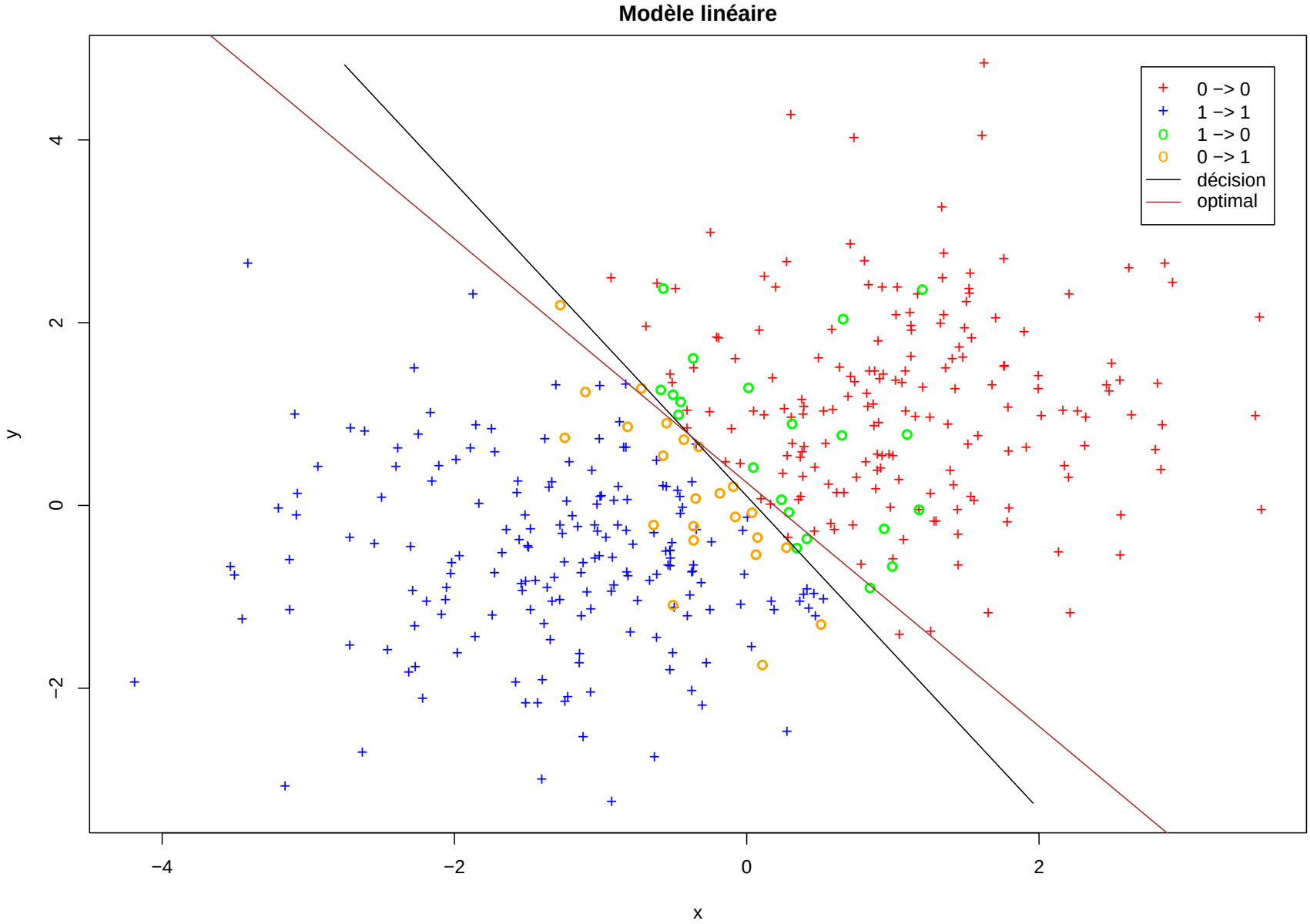
Modèle optimal (linéaire)



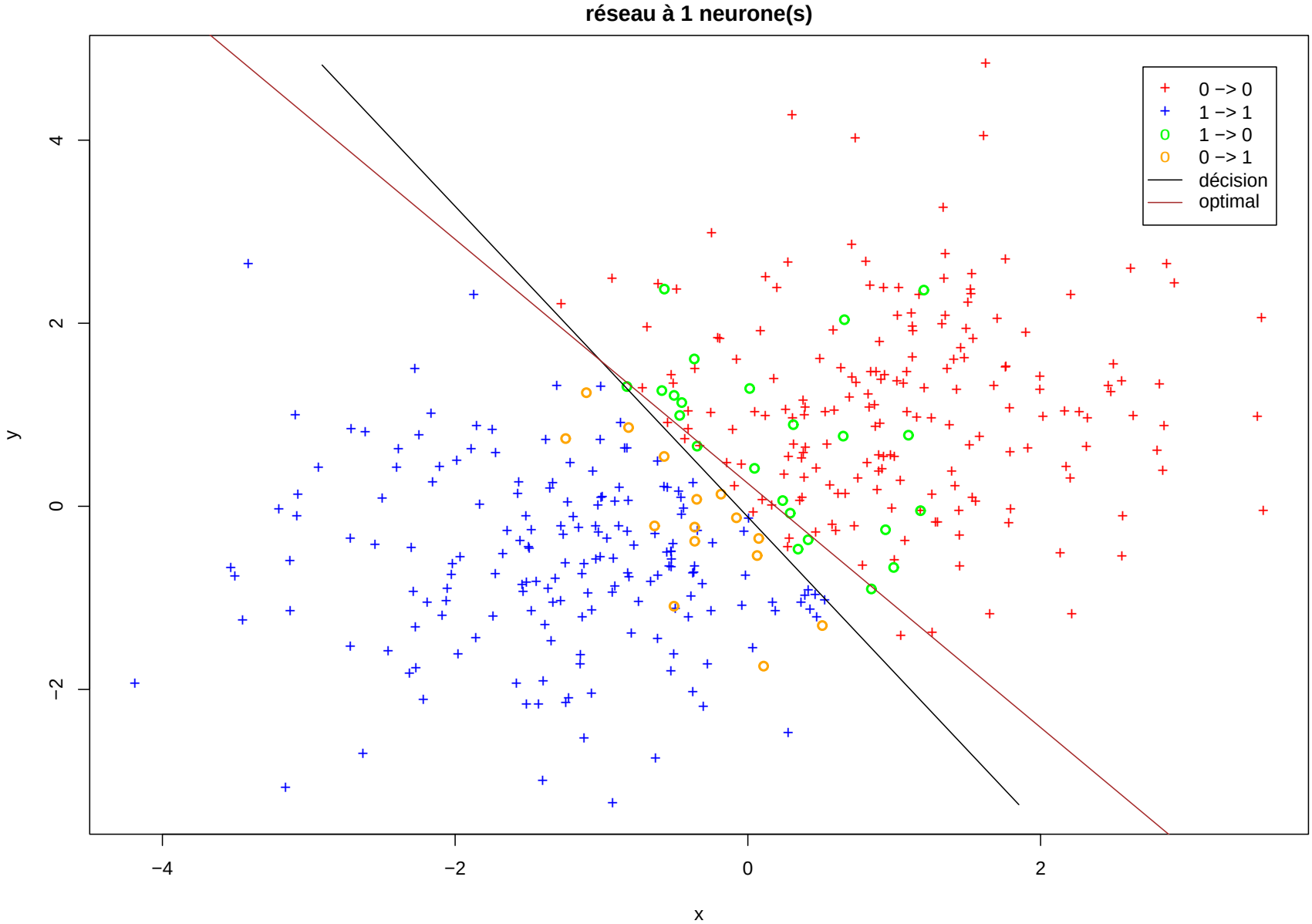
Le modèle optimal fait des erreurs (10.75%) !



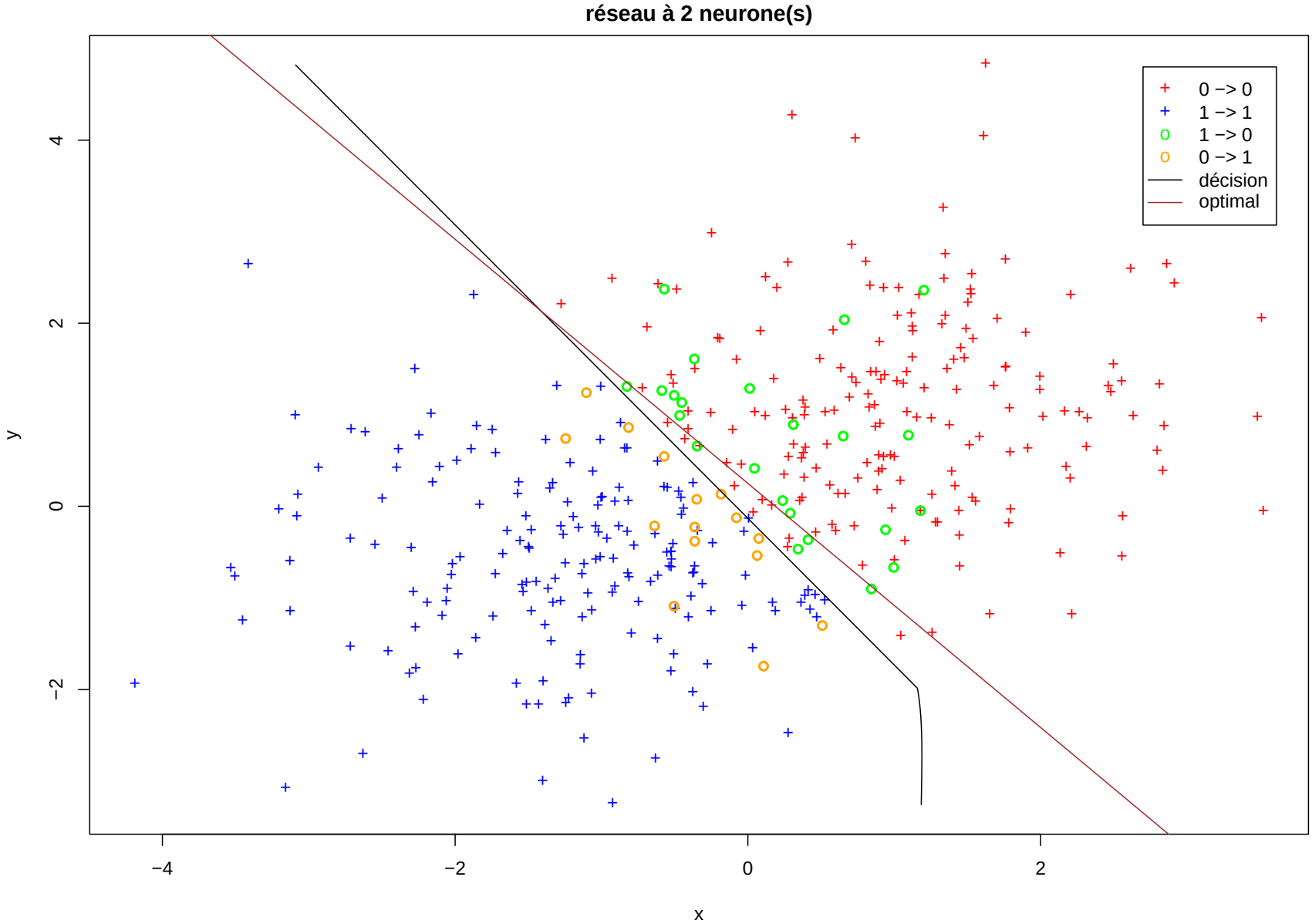
Comportement de divers modèles



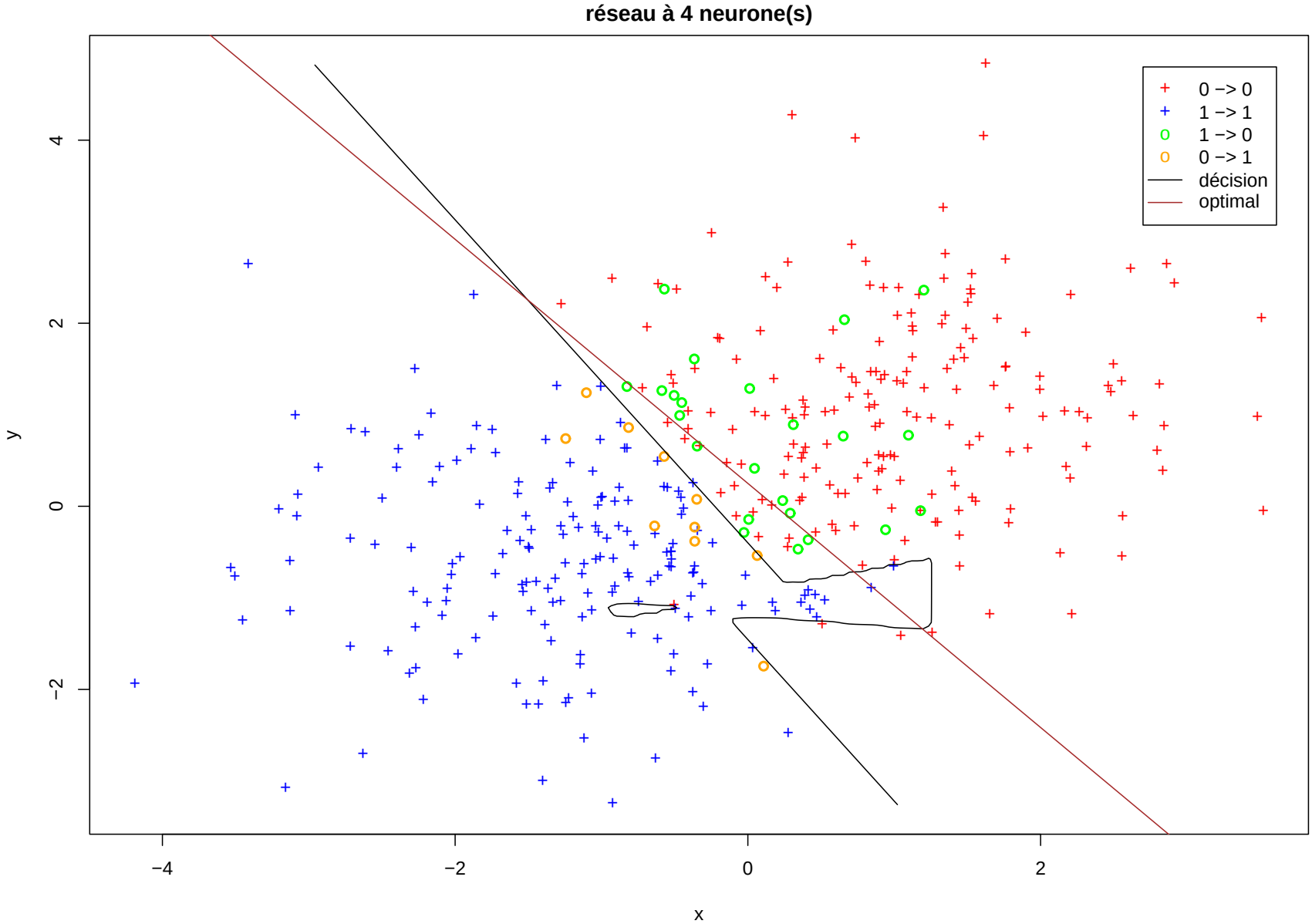
Comportement de divers modèles



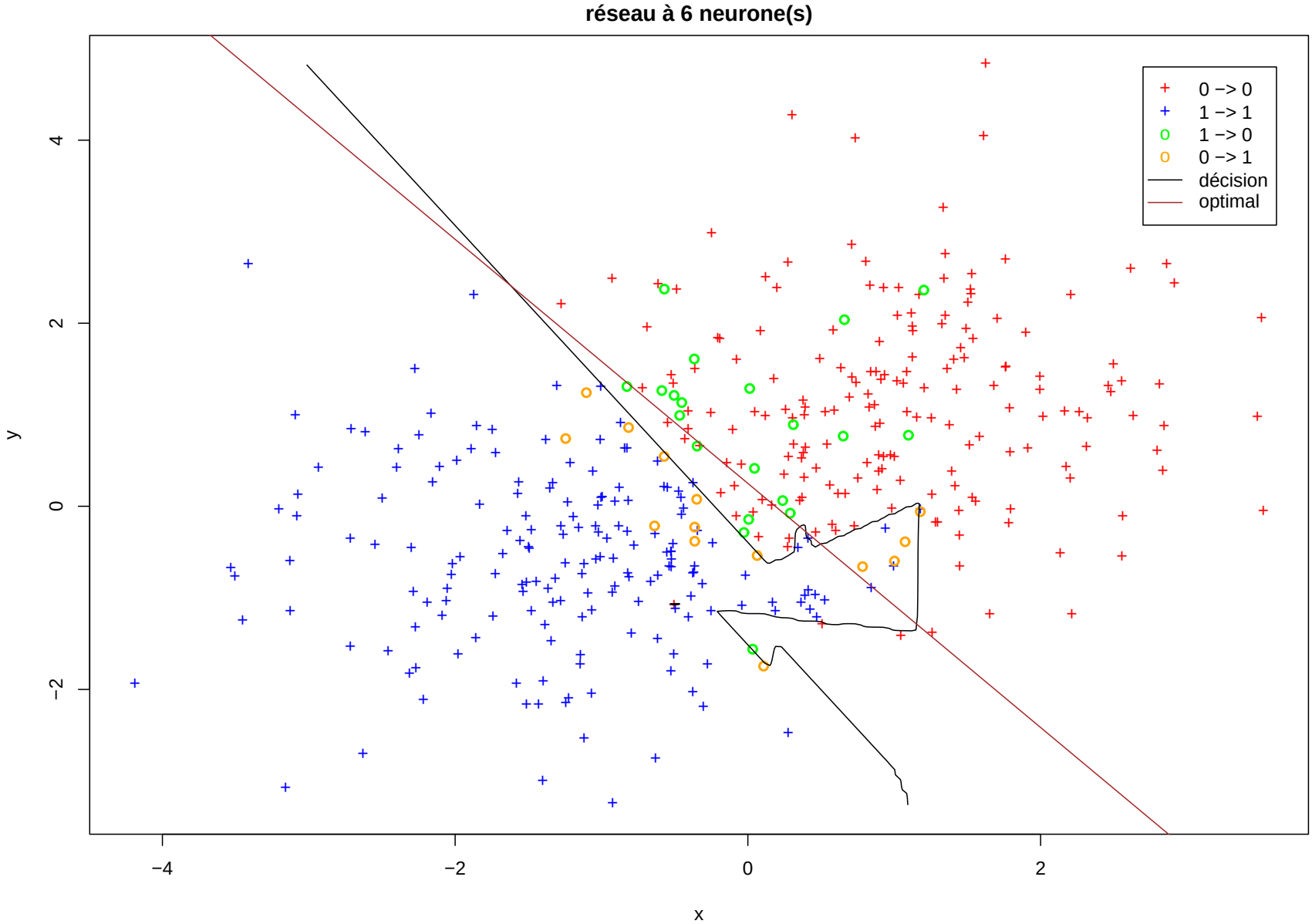
Comportement de divers modèles



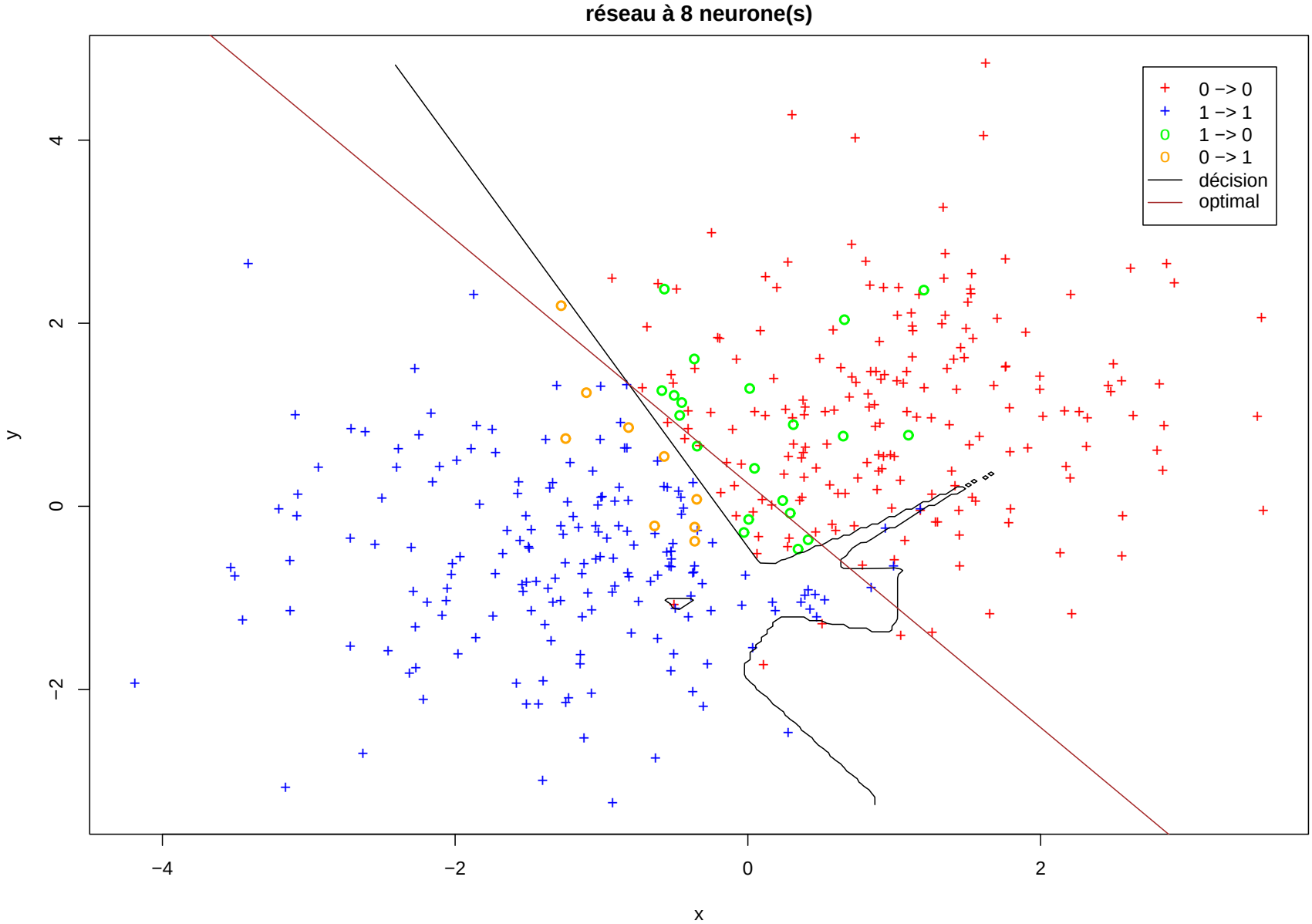
Comportement de divers modèles



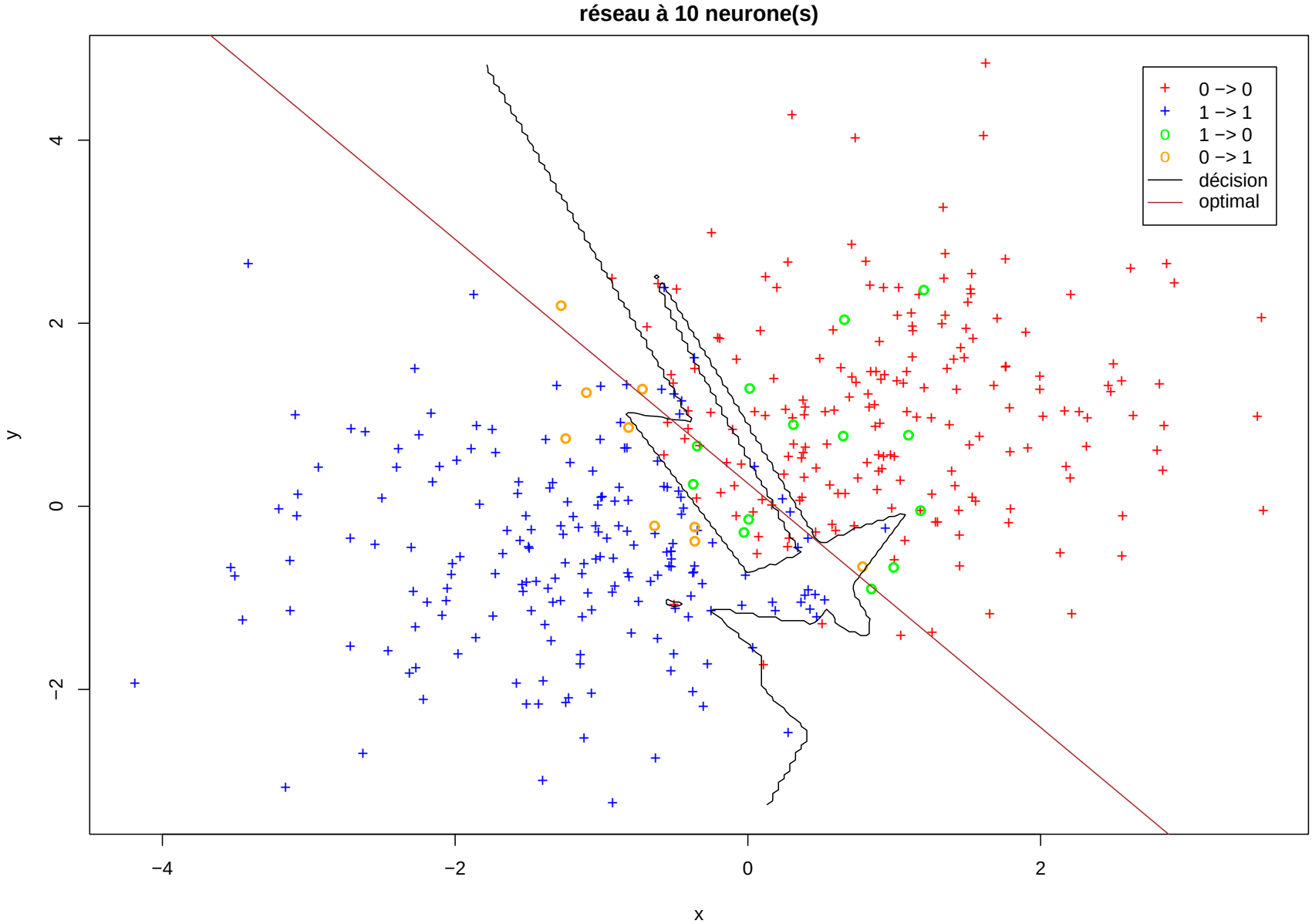
Comportement de divers modèles



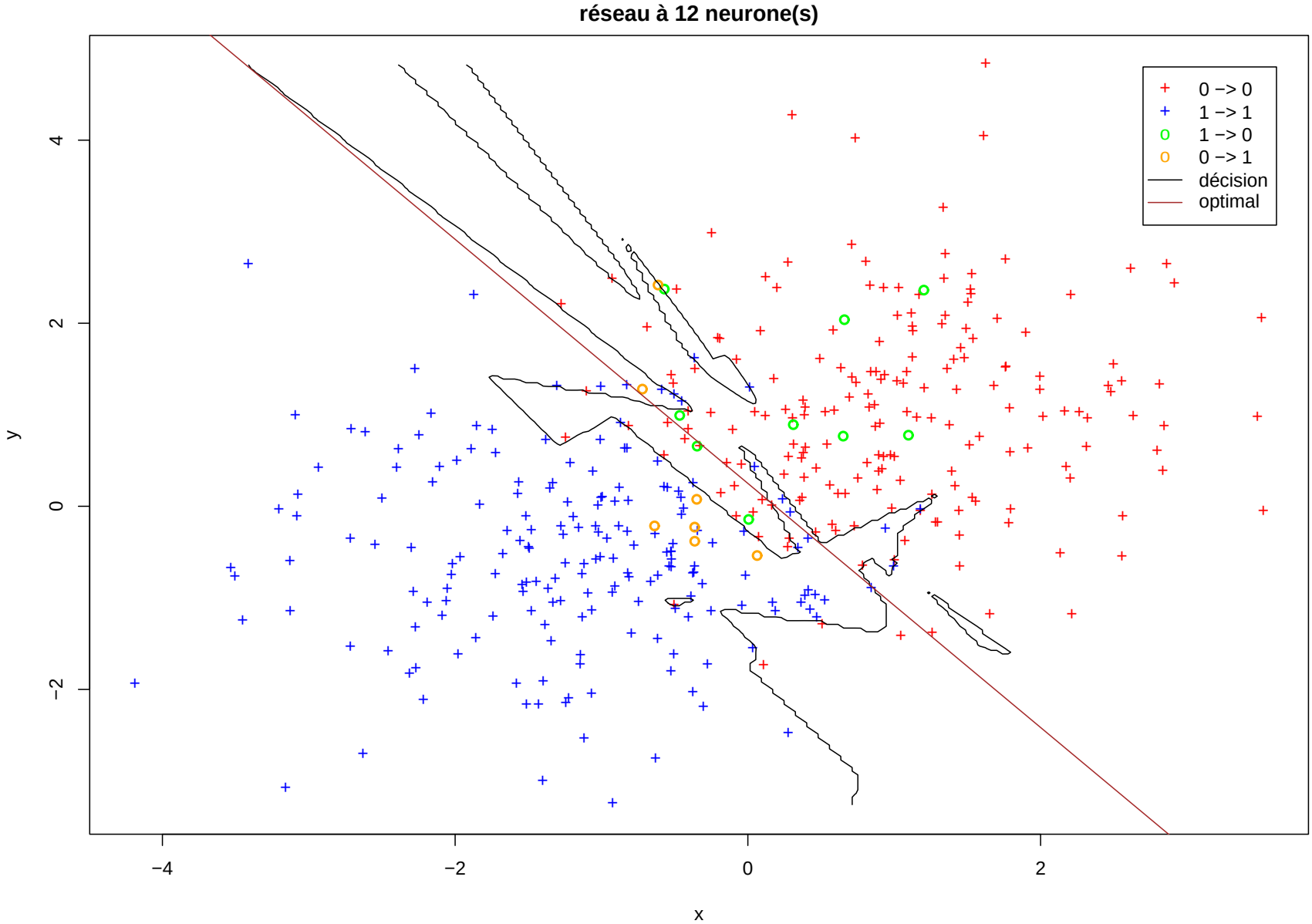
Comportement de divers modèles



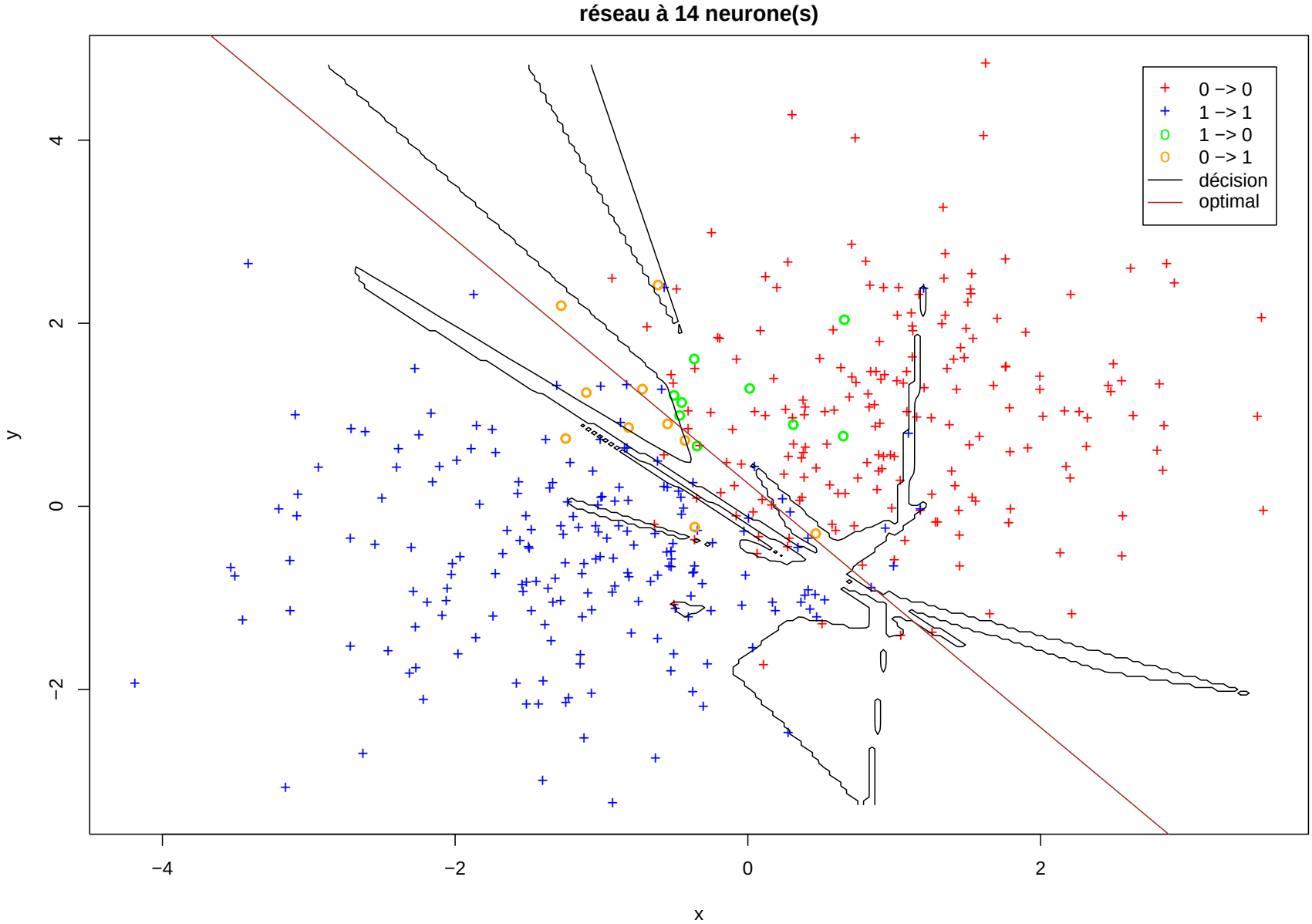
Comportement de divers modèles



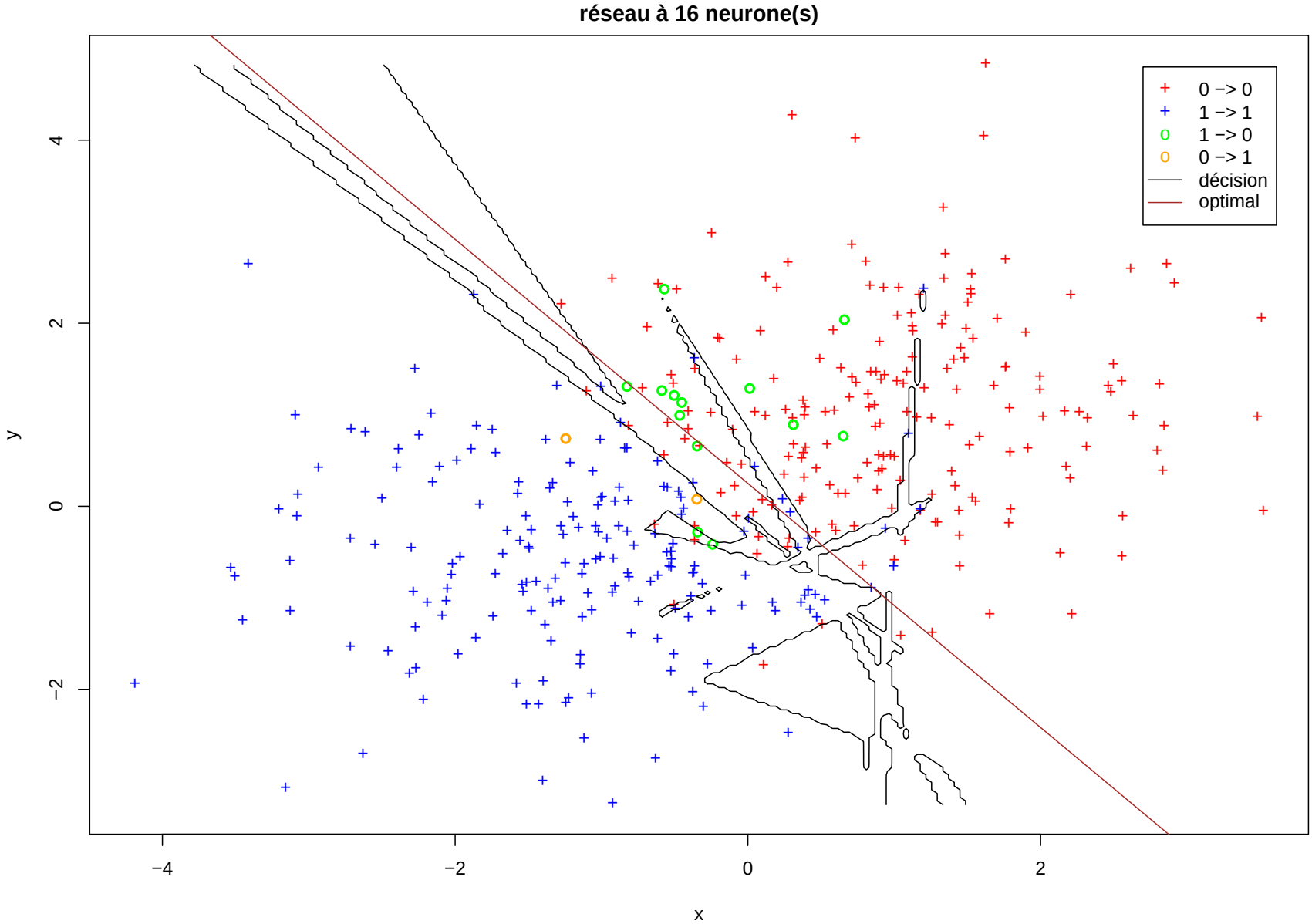
Comportement de divers modèles



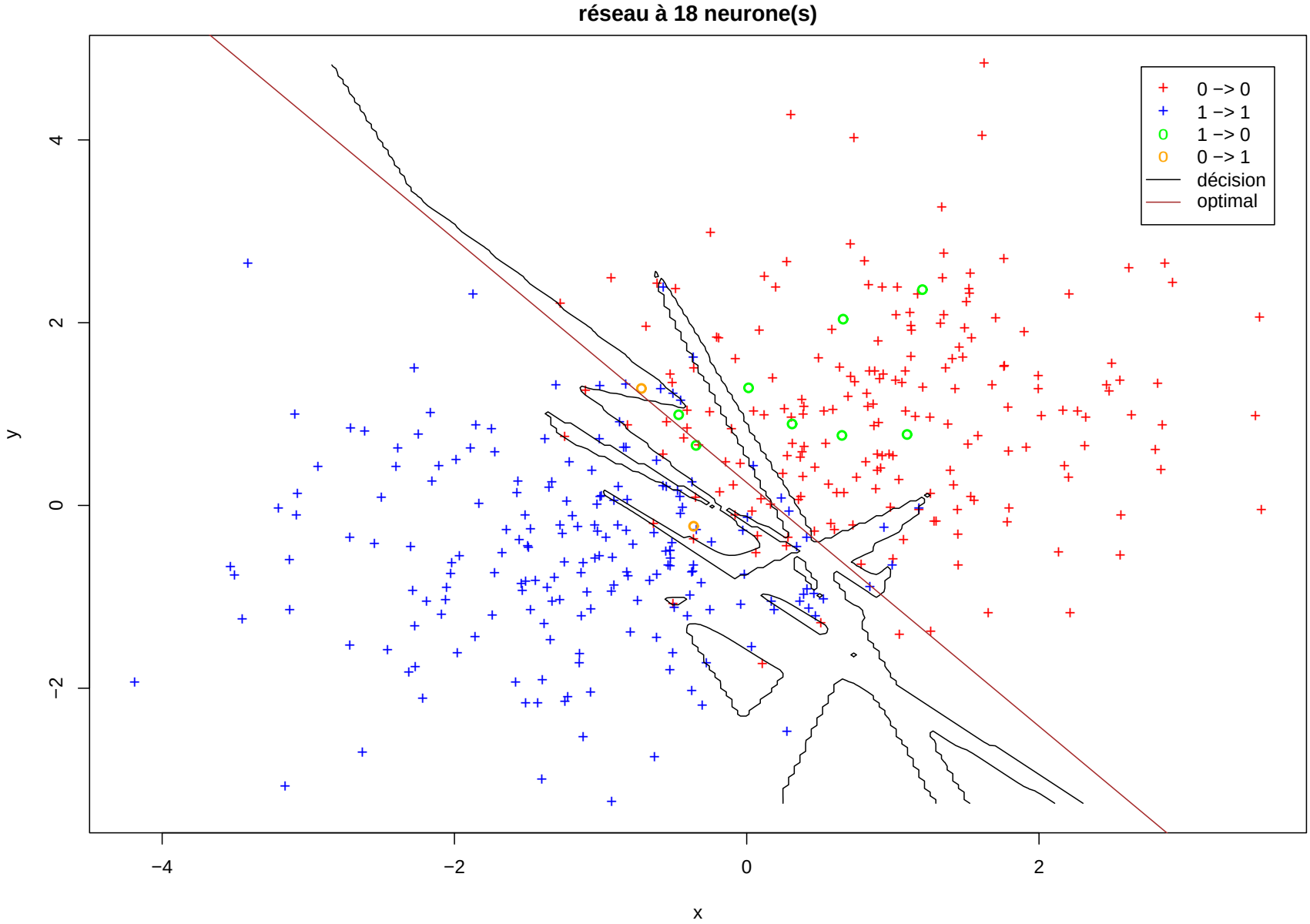
Comportement de divers modèles



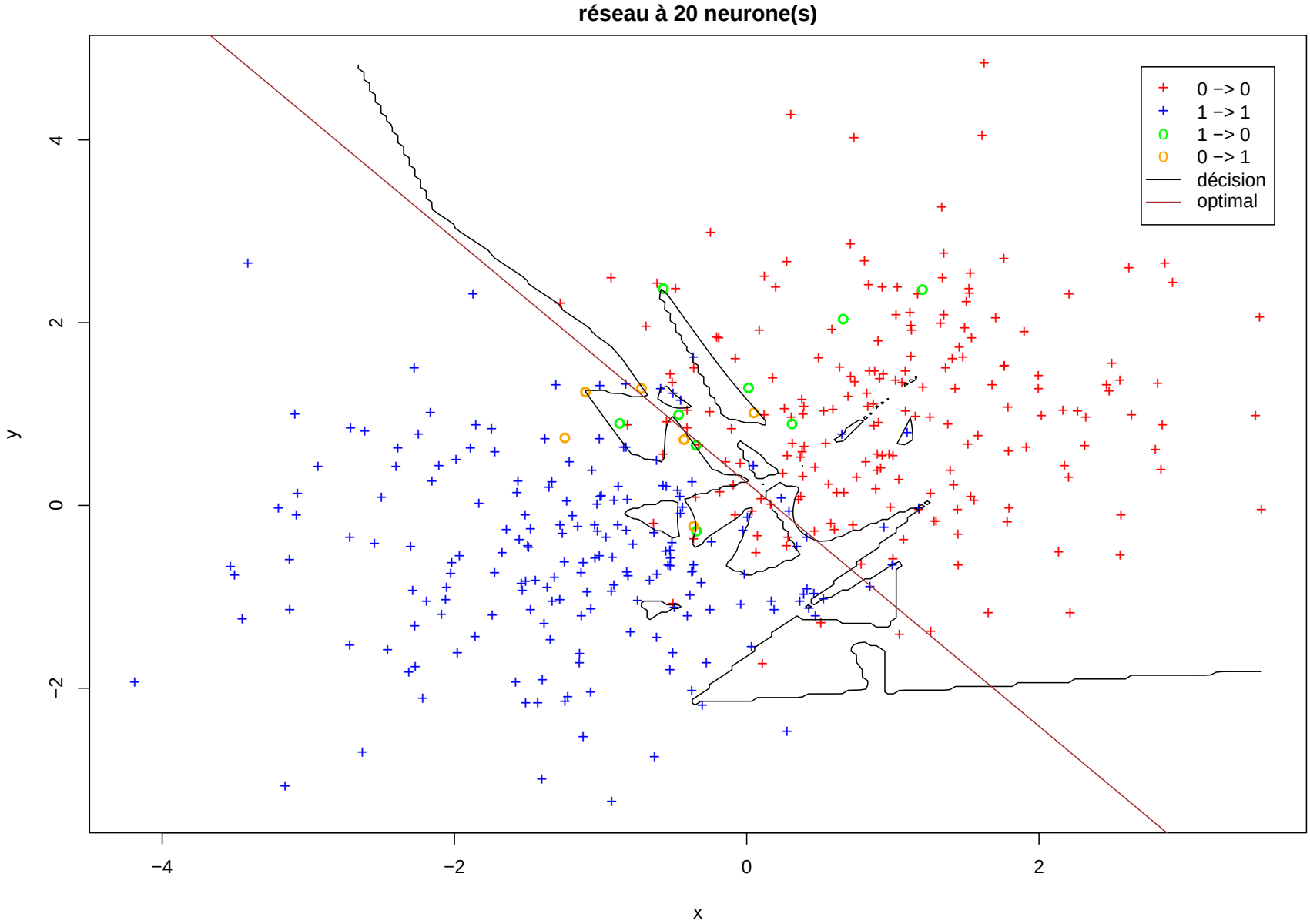
Comportement de divers modèles



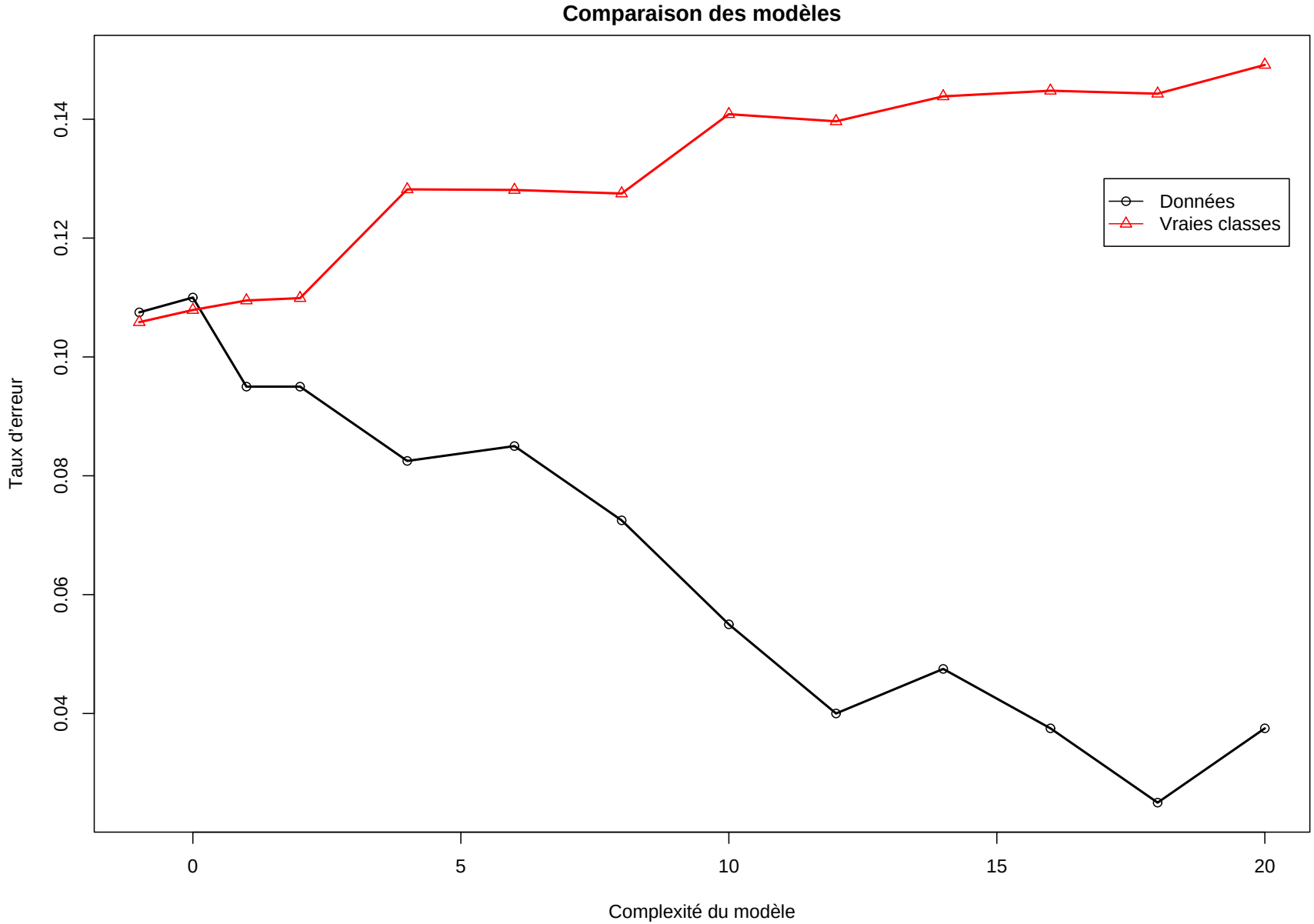
Comportement de divers modèles



Comportement de divers modèles



Evolution de l'erreur



Enseignement des exemples

- Une approche naïve est **impossible** :
  - les performances sur les données ne permettent pas de prédire le comportement du modèle sur de nouvelles données
  - les RN sont puissants : on peut presque toujours faire baisser l'erreur de modélisation en ajoutant des neurones
- Il faut **comprendre** le phénomène pour y remédier :
  - modélisation statistique
  - justification mathématique des résultats
- Les solutions (actuelles) sont coûteuses en temps de calcul

Cadre mathématique général

Les données à étudier sont décrites de la façon suivante :

- on dispose de N **individus** ou **exemples**
- chaque individu est décrit par n variables réelles, i.e. chaque exemple est un **vecteur** $x \in \mathbb{R}^n$
- dans le cas de la discrimination, chaque exemple est associé à une **classe** (un groupe)
- dans le cas de la régression, chaque exemple est associé à une **variable cible**, notée y (élément de $\mathbb{R}^p$).

On peut reformuler les trois problèmes de l'AD :

1. discrimination : trouver un lien entre x et sa classe
2. régression : trouver un lien entre x et y
3. classification : construire des classes en associant des étiquettes aux individus

Discrimination et statistique

En discrimination :

- pour chaque individu x^k, on connaît la classe C_j telle que $x^k \in C_j$
- on cherche à classer de nouveaux individus

On doit estimer :

$$P(C_j|x)$$

la probabilité que l'individu observé soit issu de la classe C_j sachant qu'il est décrit par le vecteur x. Intérêts :

- permet d'affecter un individu à une classe (la plus probable, en général)
- permet de mesurer l'erreur liée à cette affectation

Régression et statistique

En régression :

- chaque individu x^k est associé à une grandeur $y^k \in \mathbb{R}^p$
- on cherche à exprimer y^k comme une fonction de x^k

On doit estimer :

$$E(Y|X = x)$$

l'espérance conditionnelle de Y sachant $X = x$: c'est la meilleure approximation de Y par une fonction de X.

Classification et statistique

En classification :

- les individus ne sont associés à aucune donnée explicative
- on chercher à trouver des groupes (des classes)

On peut chercher à exprimer $p(x)$ comme un mélange :

$$p(x) = \sum_{j=1}^{M} p(x|C_j)P(C_j)$$

On peut ensuite affecter les individus aux classes, grâce à la règle de Bayes :

$$P(C_j|x) = \frac{p(x|C_j)P(C_j)}{p(x)}$$