Analyse statistique des données issues des biopuces à ADN

Julie Peyre

LMC-IMAG, UJF, Institut Curie

Thèse dirigée par Anestis Antoniadis et Marie Dutreix

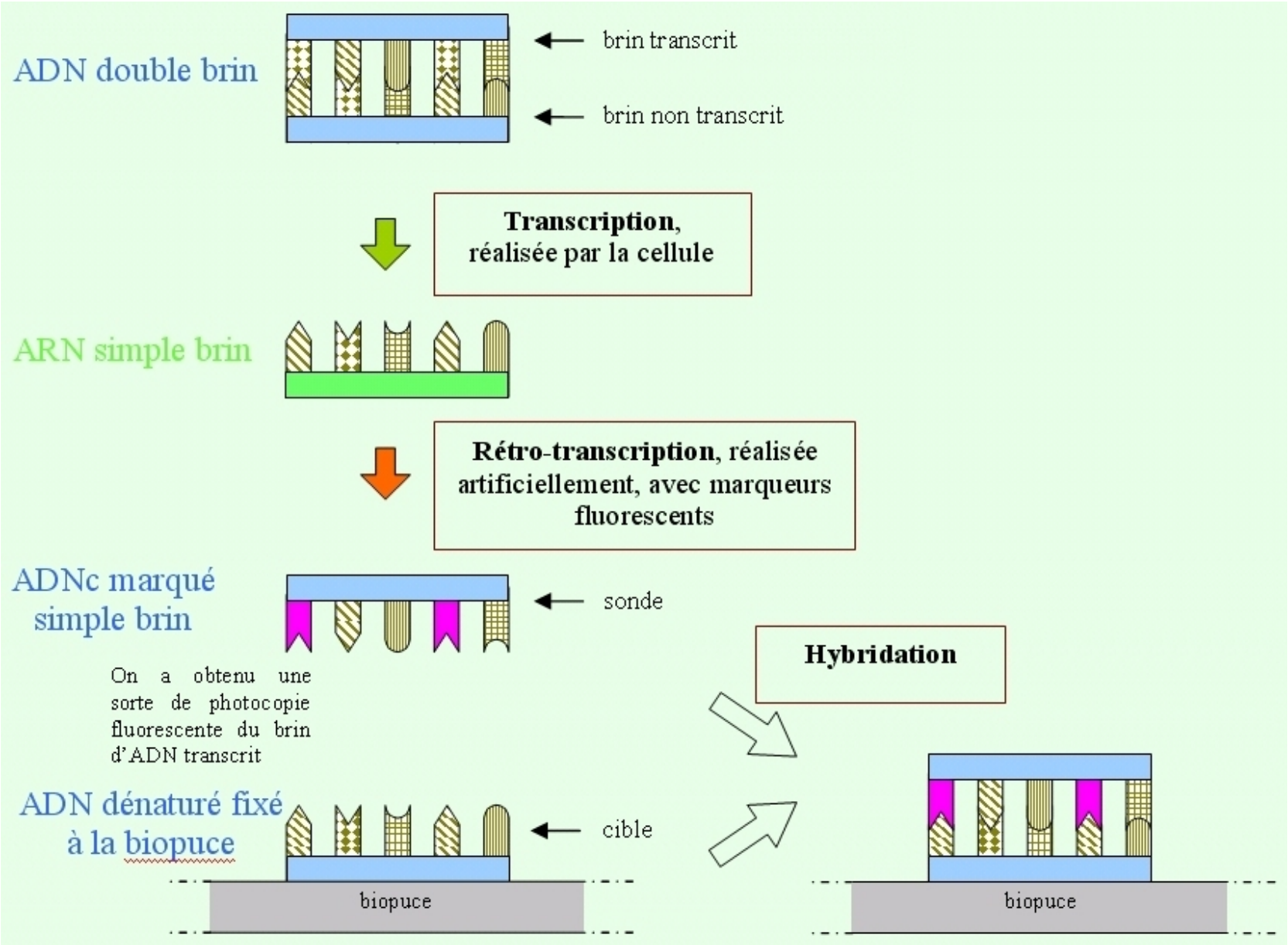
Introduction au contexte biologique

Le dogme central de la biologie moléculaire

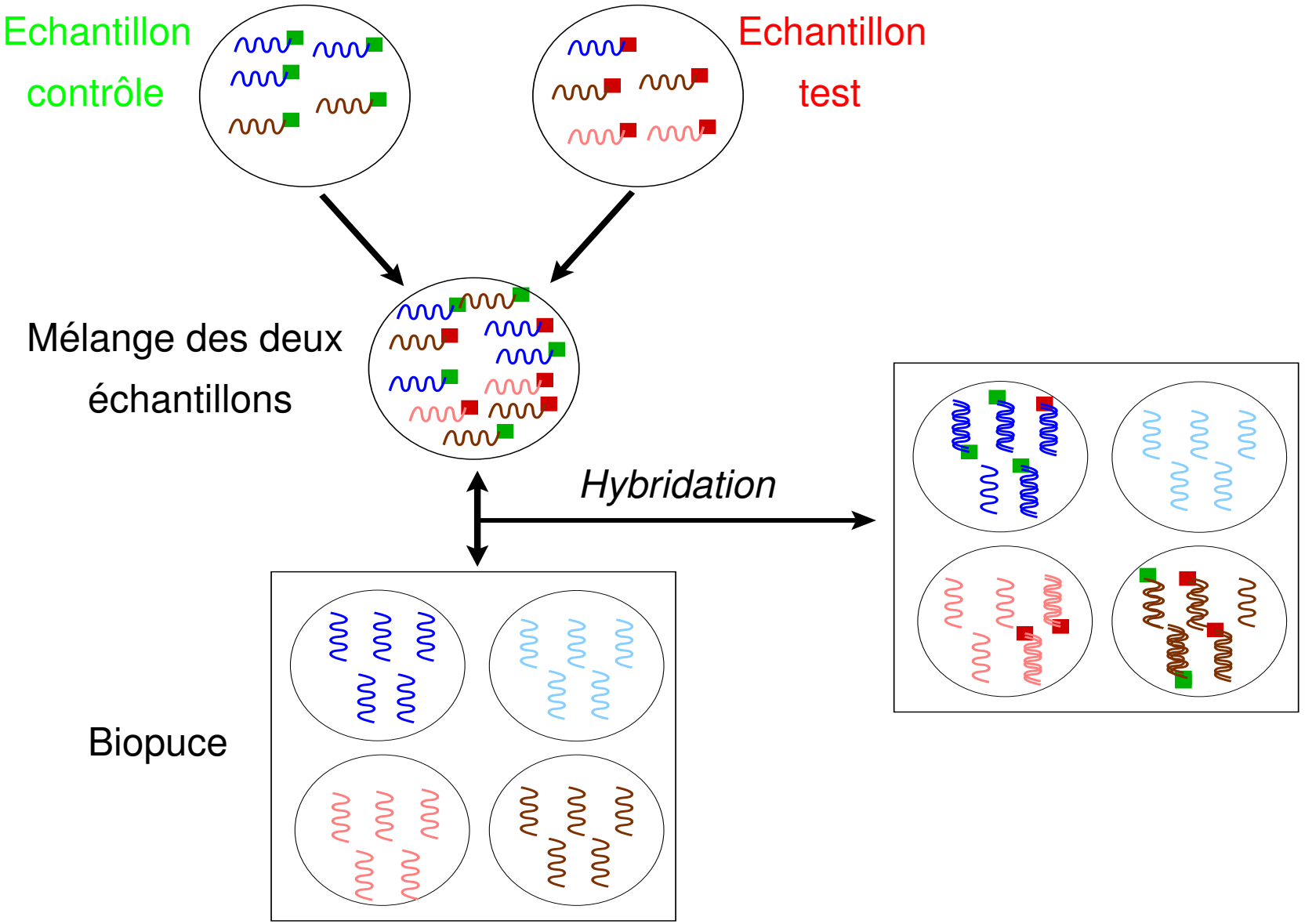
Protéines : main d'œuvre de la cellule.

Principe biopuces : étudier la composition en ARN pour caractériser l'activité d'une cellule.

Le principe des biopuces

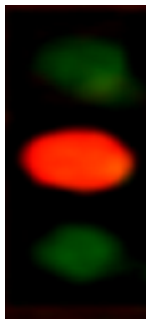


Expériences à double marquage



Un exemple de résultat

Expression différente entre les 2 populations

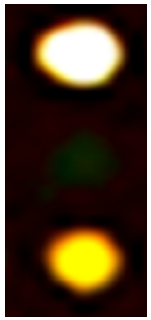


réprimé

induit

réprimé

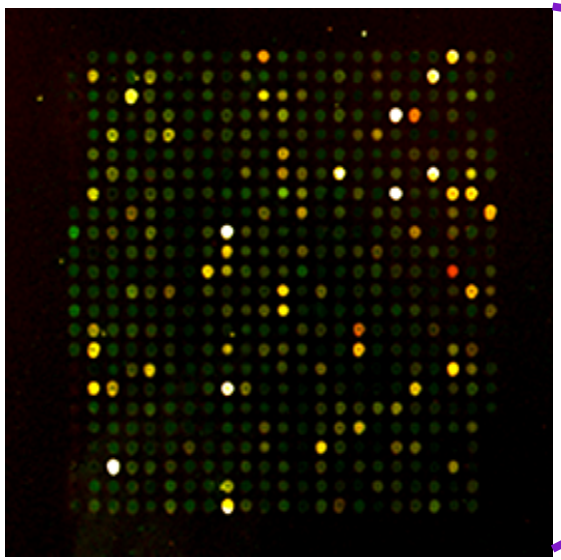
Même niveau d'expression

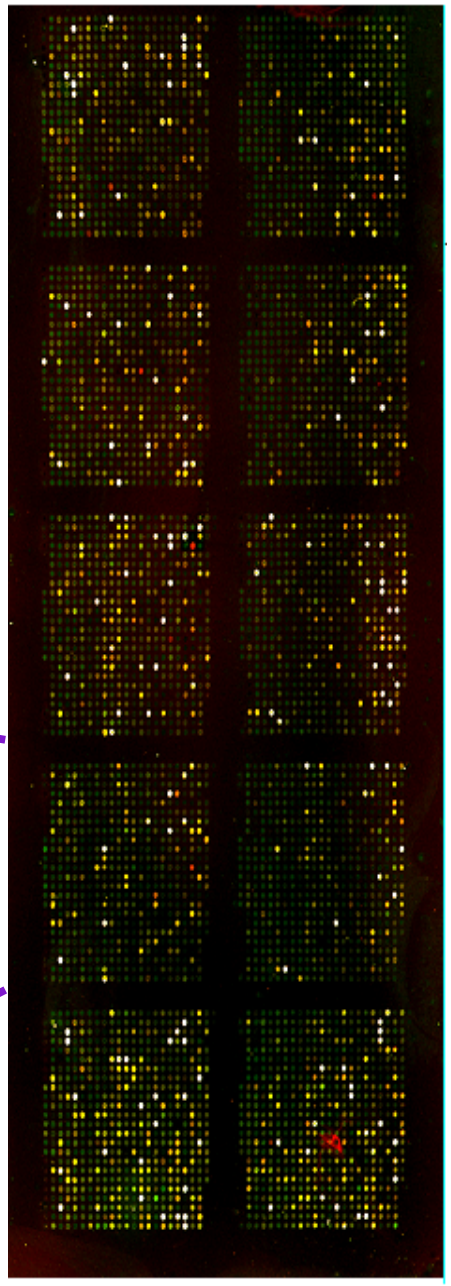


abondant

rare

intermédiaire





Une technique qui appelle des méthodes statistiques

Trois problèmes envisagés ici :

- Normalisation des données
- Détection des gènes différentiellement exprimés
- Réduction de dimension et classification supervisée

Normalisation et simulation des données de biopuces

Une technologie soumise à de nombreuses variations

- Matériel biologique
- Fabrication des puces
- Incorporation des fluorochromes
- Paramètres expérimentaux
- Analyse d'image

Objectif et principe de la normalisation

- But : éliminer les variations dues à la technique pour ne conserver que les différences liées au phénomène biologique

- Hypothèse fondamentale : faible pourcentage de gènes différentiellement exprimés

Transformation préalable des données

Notations :

p désigne le nombre total de gènes sur la biopuce
$G = (G_i)_{i=1...p}$ intensités mesurées sur l'échantillon contrôle
$R = (R_i)_{i=1...p}$ intensités mesurées sur l'échantillon test

On fait une transformation logarithmique :

Log-intensité totale

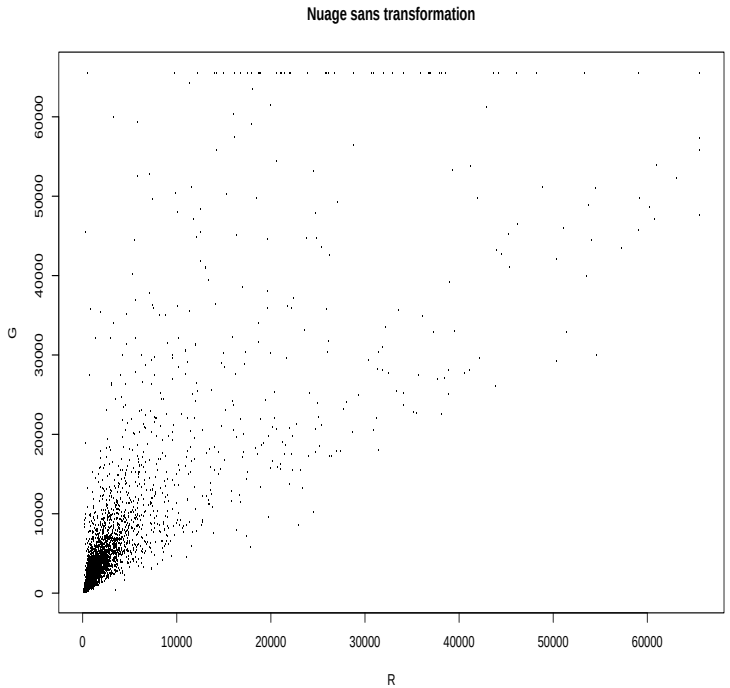
$$A = log_2(\sqrt{R \times G}) = \frac{log_2(R) + log_2(G)}{2}$$

Log-ratio d'expression

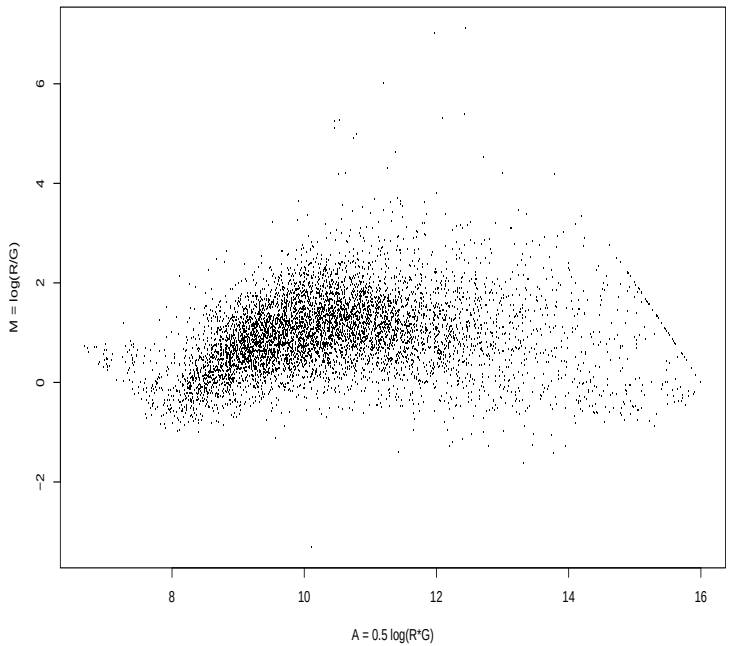
$$M = log_2\left(\frac{R}{G}\right) = log_2(R) - log_2(G)$$

Transformation des données : résultats

Avant transformation



Après transformation



Une première méthode : normalisation par la médiane

Modèle :

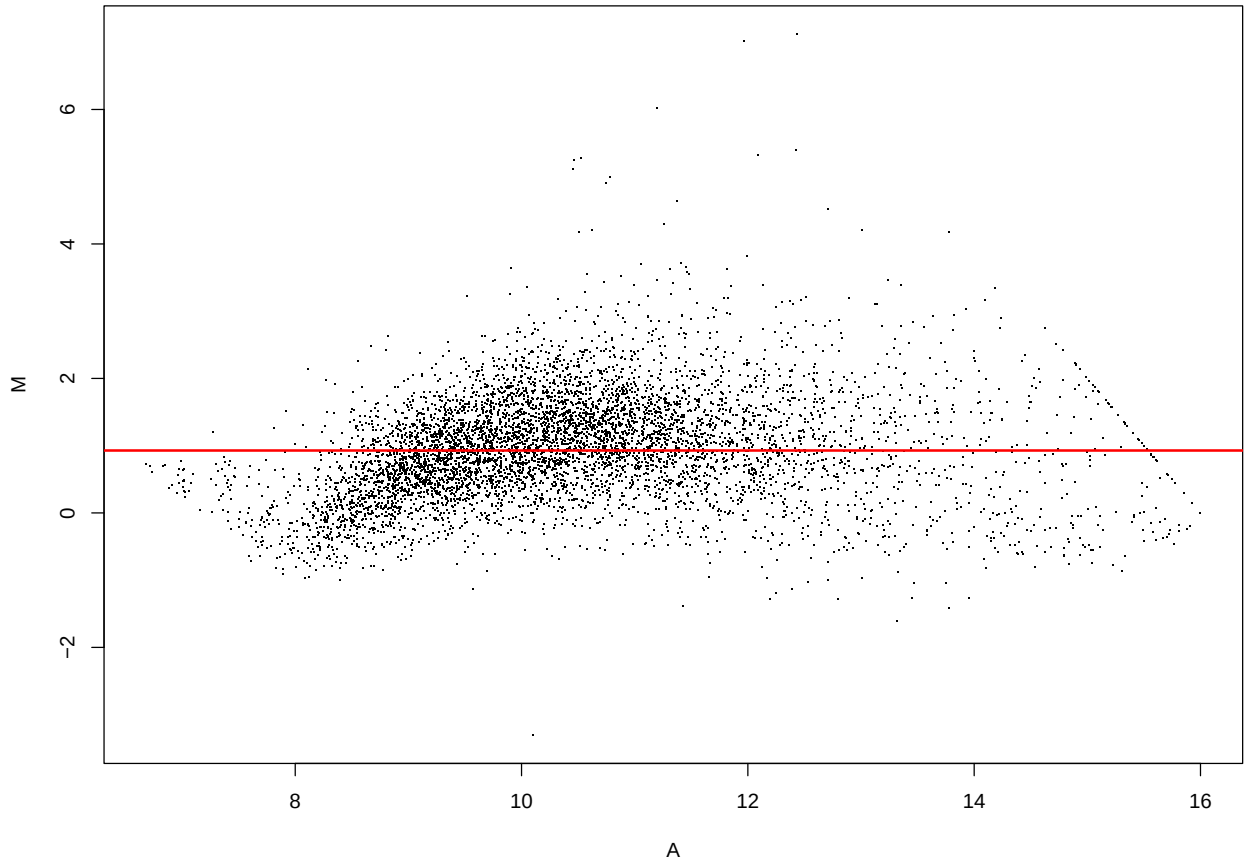
$$R_i = c \times G_i$$

où $i = 1 \ldots p$

$$\Updownarrow$$

$$M_i = K$$

où $i = 1 \ldots p$



Normalisation lowess *(Speed et al. 2001)*

Modèle :

$$R_i = c_i \times G_i$$

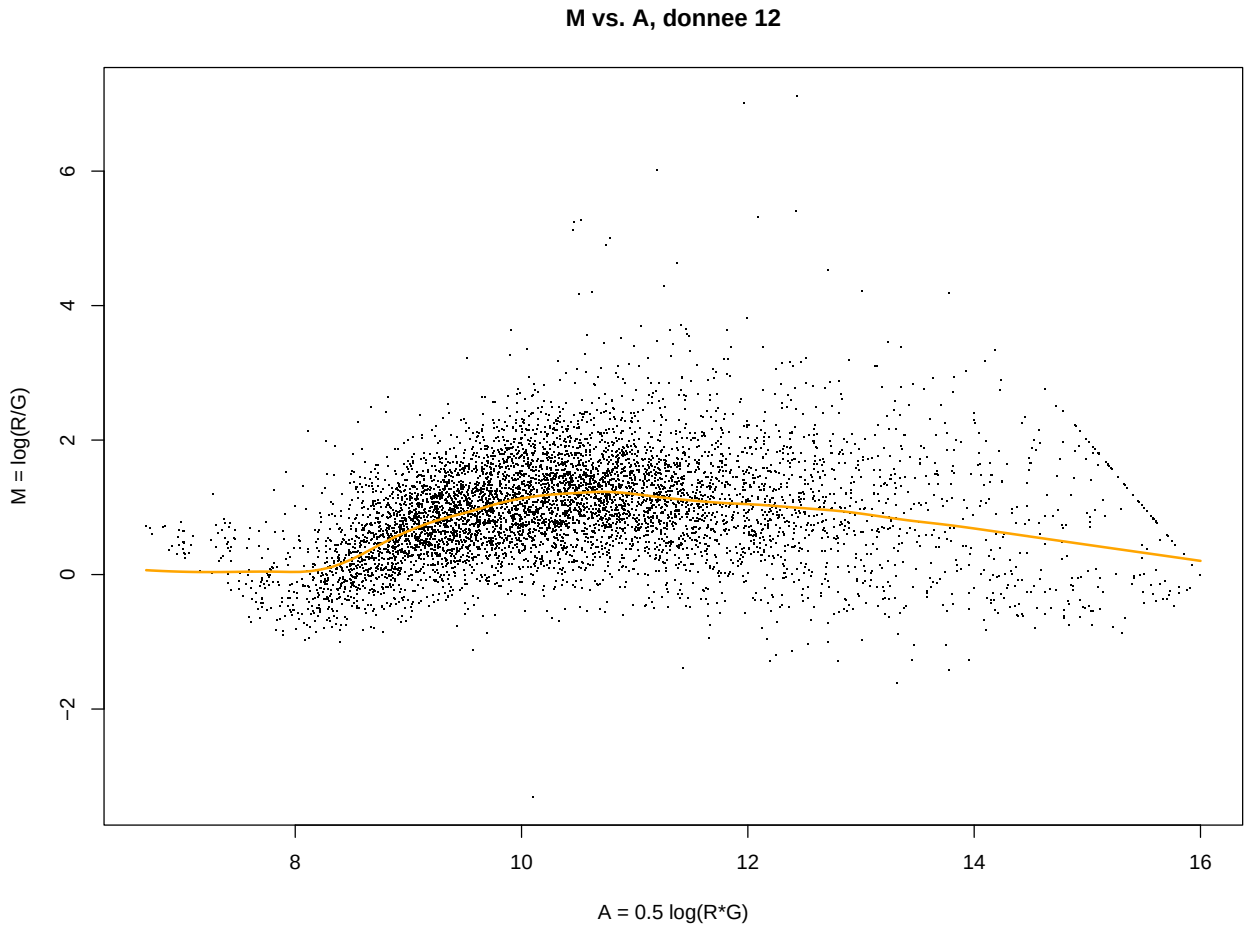
où $i = 1 \ldots p$

$$M_i = K_i$$

où $i = 1 \ldots p$

On estime une fonction lisse telle que :

$$M_i = \rho(A_i)$$

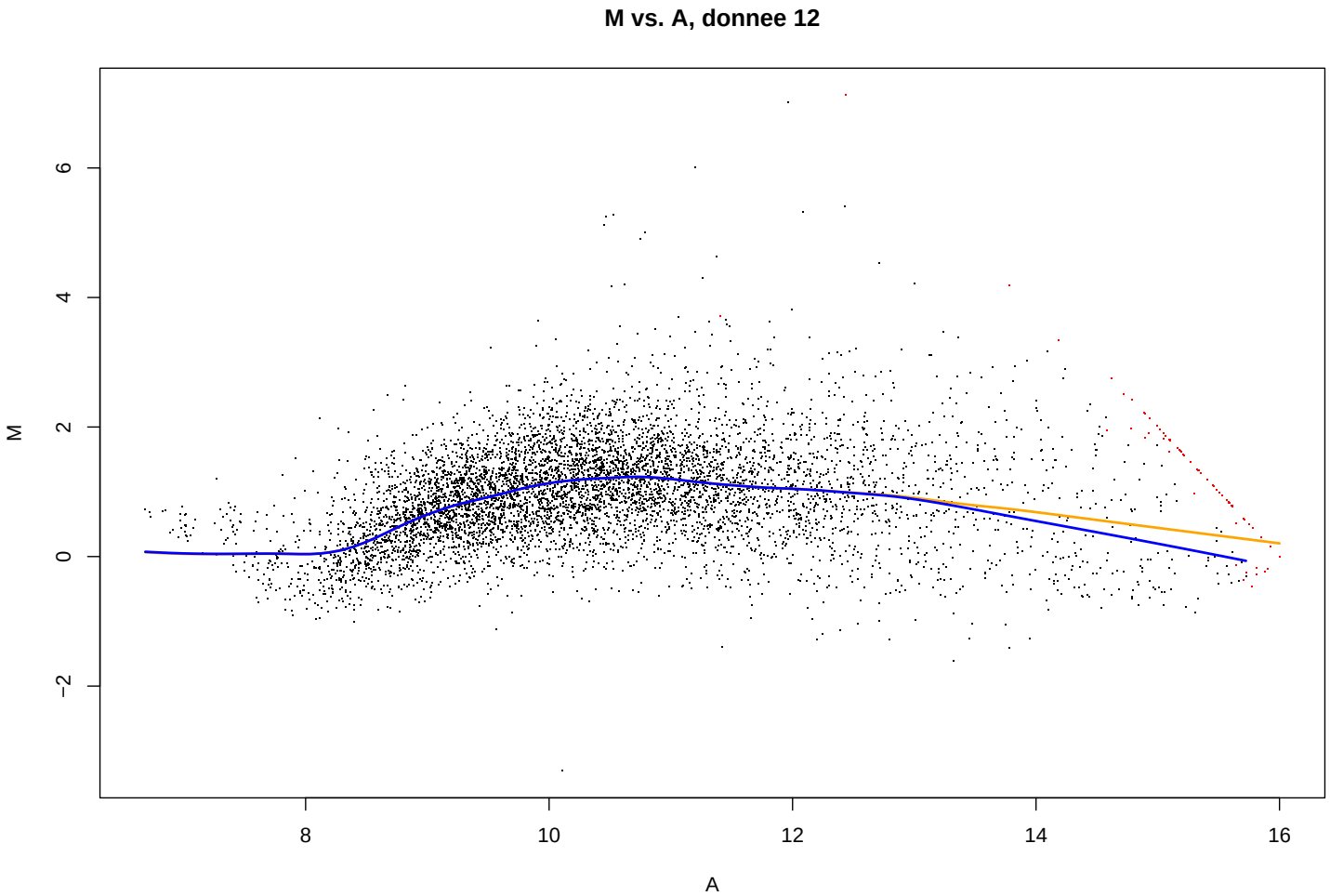


ρ est estimée par LOWESS

LOcally WEighted Scatterplot Smoothing

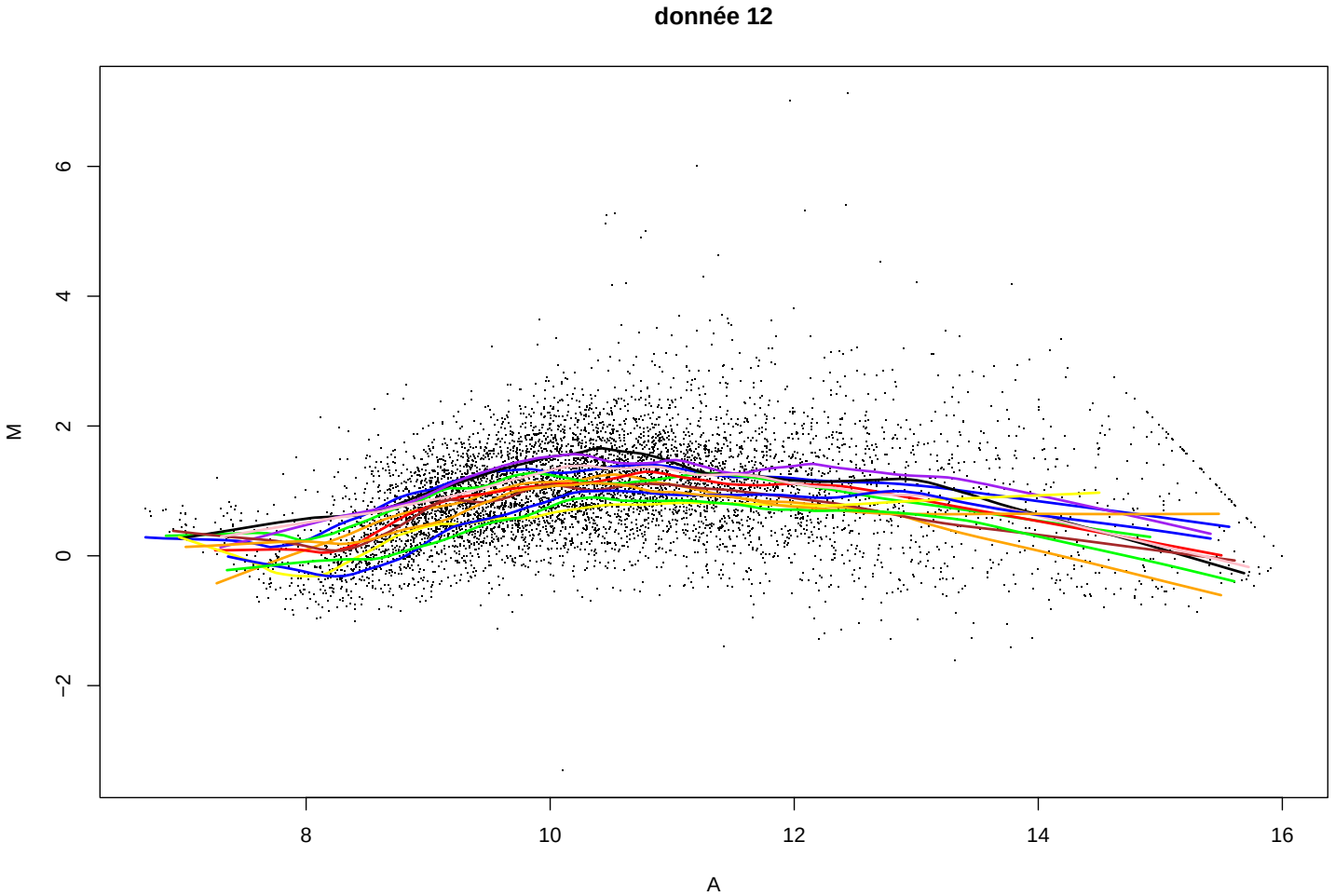
Améliorations proposées

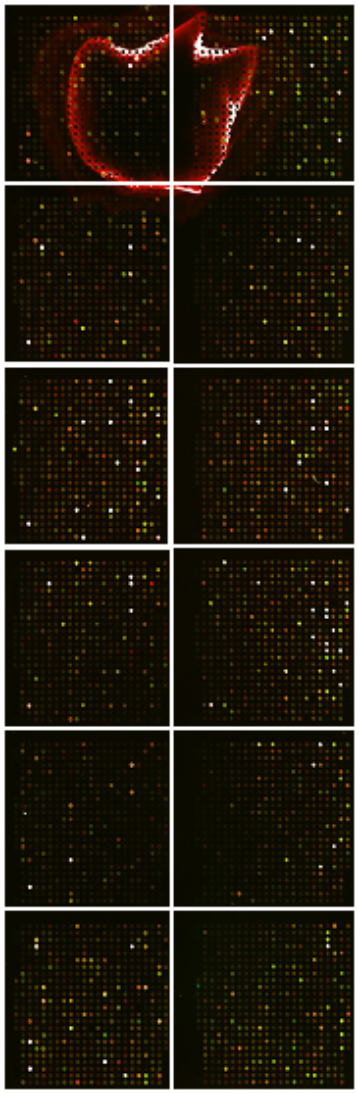
Eviter les saturants



Améliorations proposées

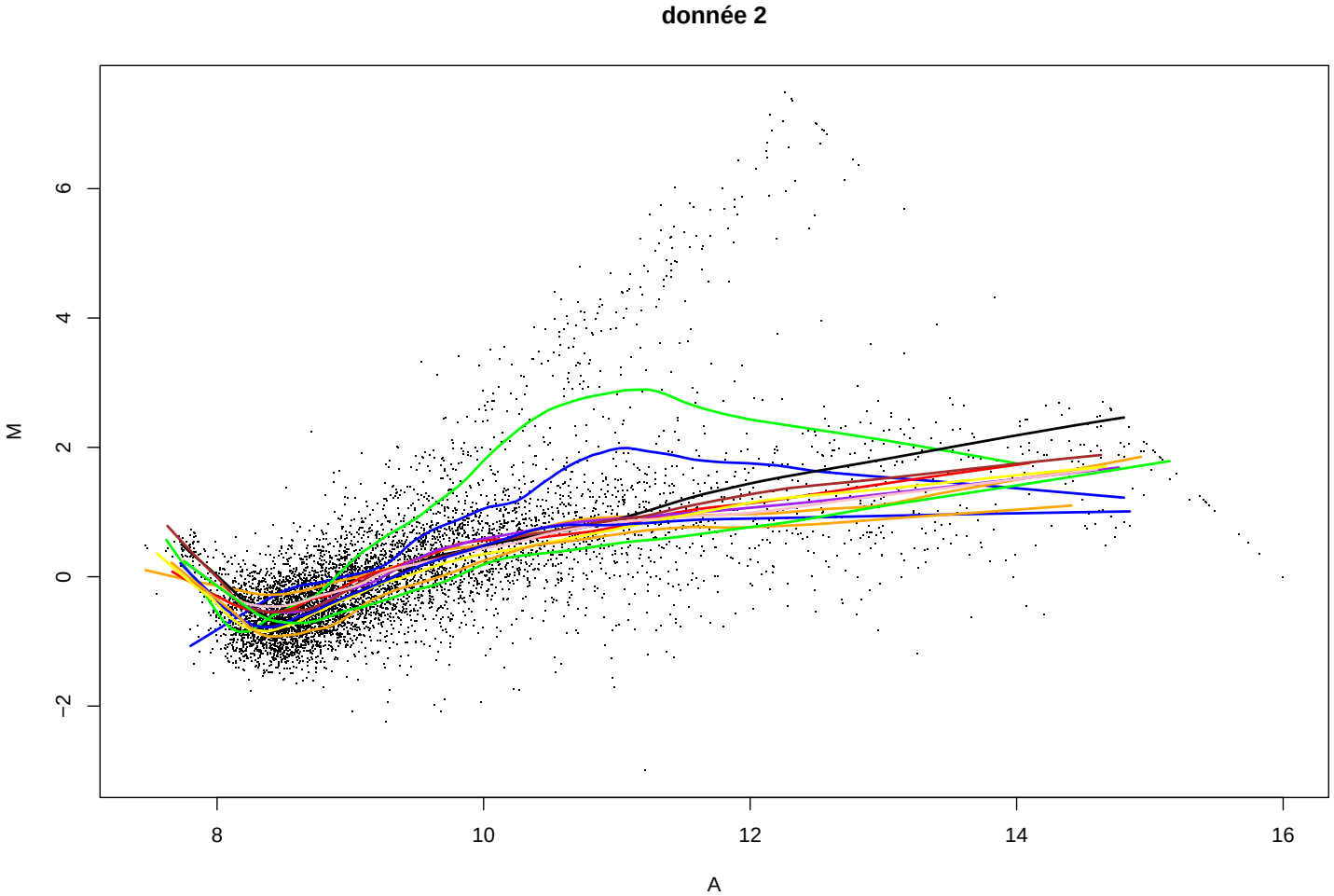
Normalisation par bloc

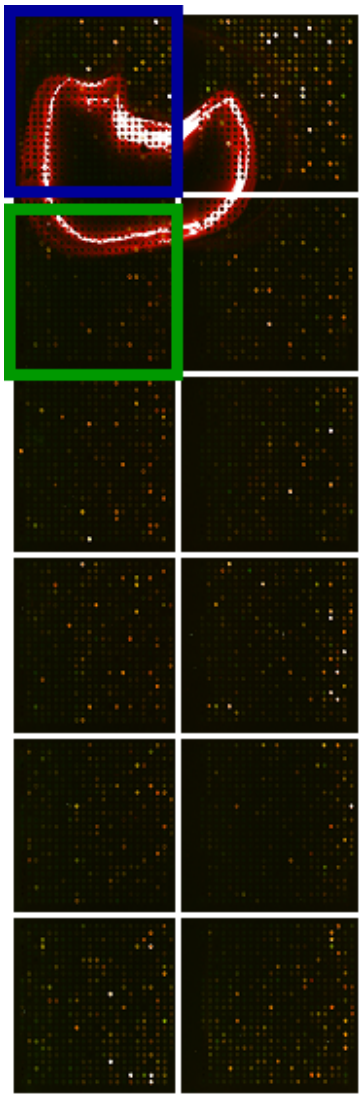




Améliorations proposées

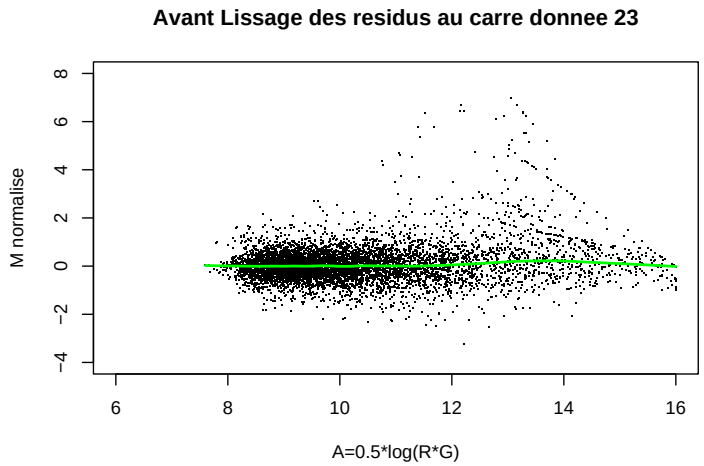
Normalisation par bloc

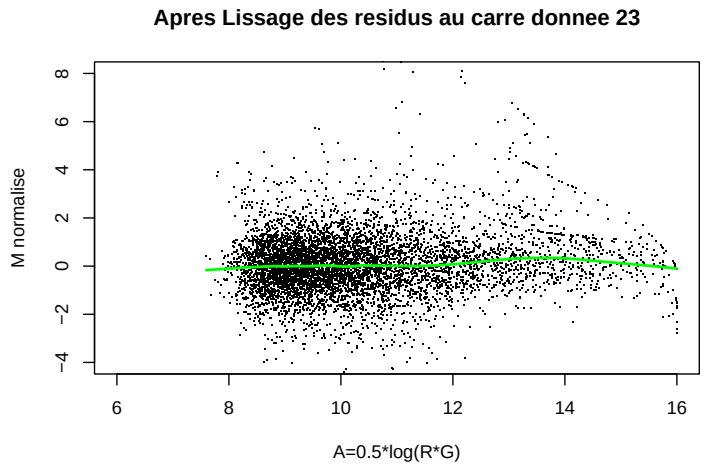


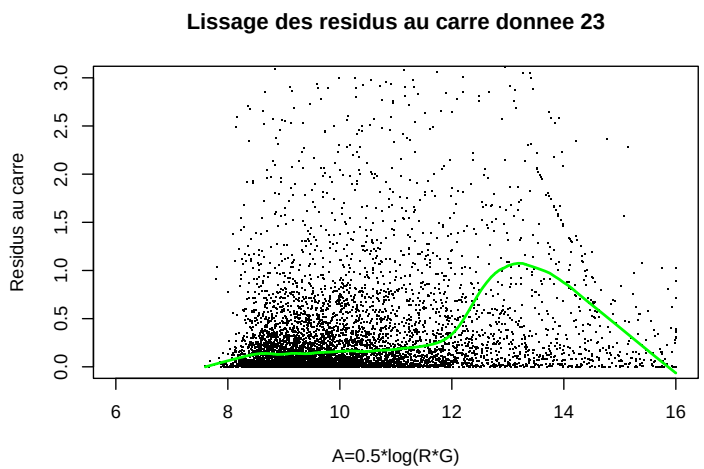


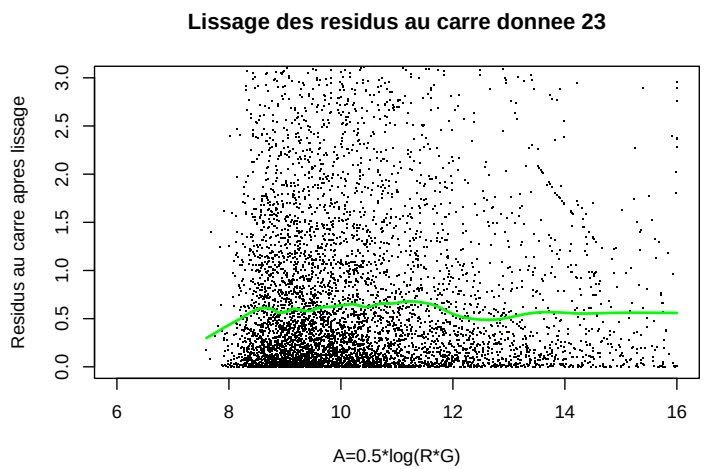
Améliorations proposées

Lissage des résidus









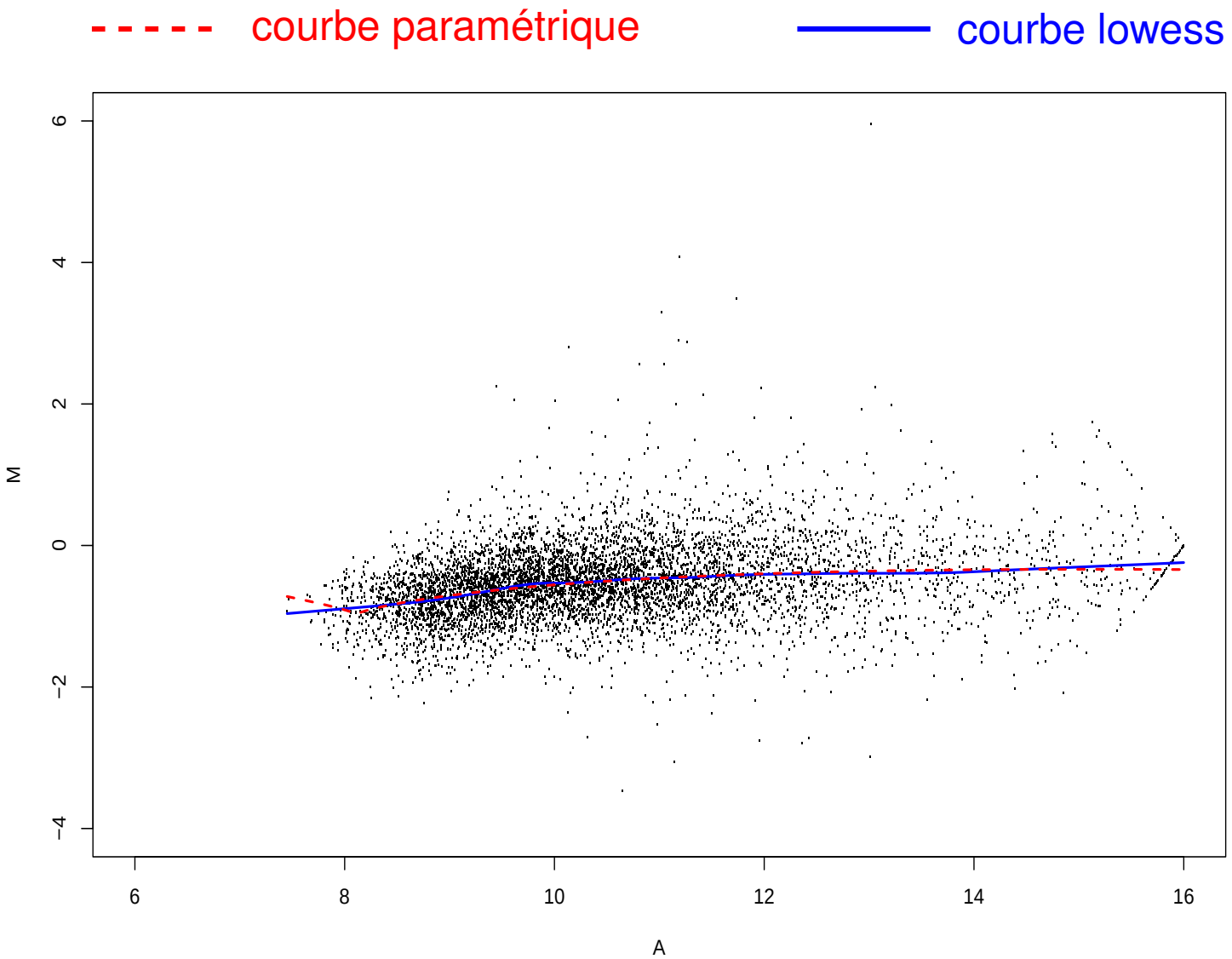
Modèle de simulation de données

- Modélisation des formes de nuage par fonction “baignoire”
  - Rappel : lowess pour estimer ρ telle que $M = \rho(A)$.
  - Ici, représentation paramétrique par fonction “baignoire” :

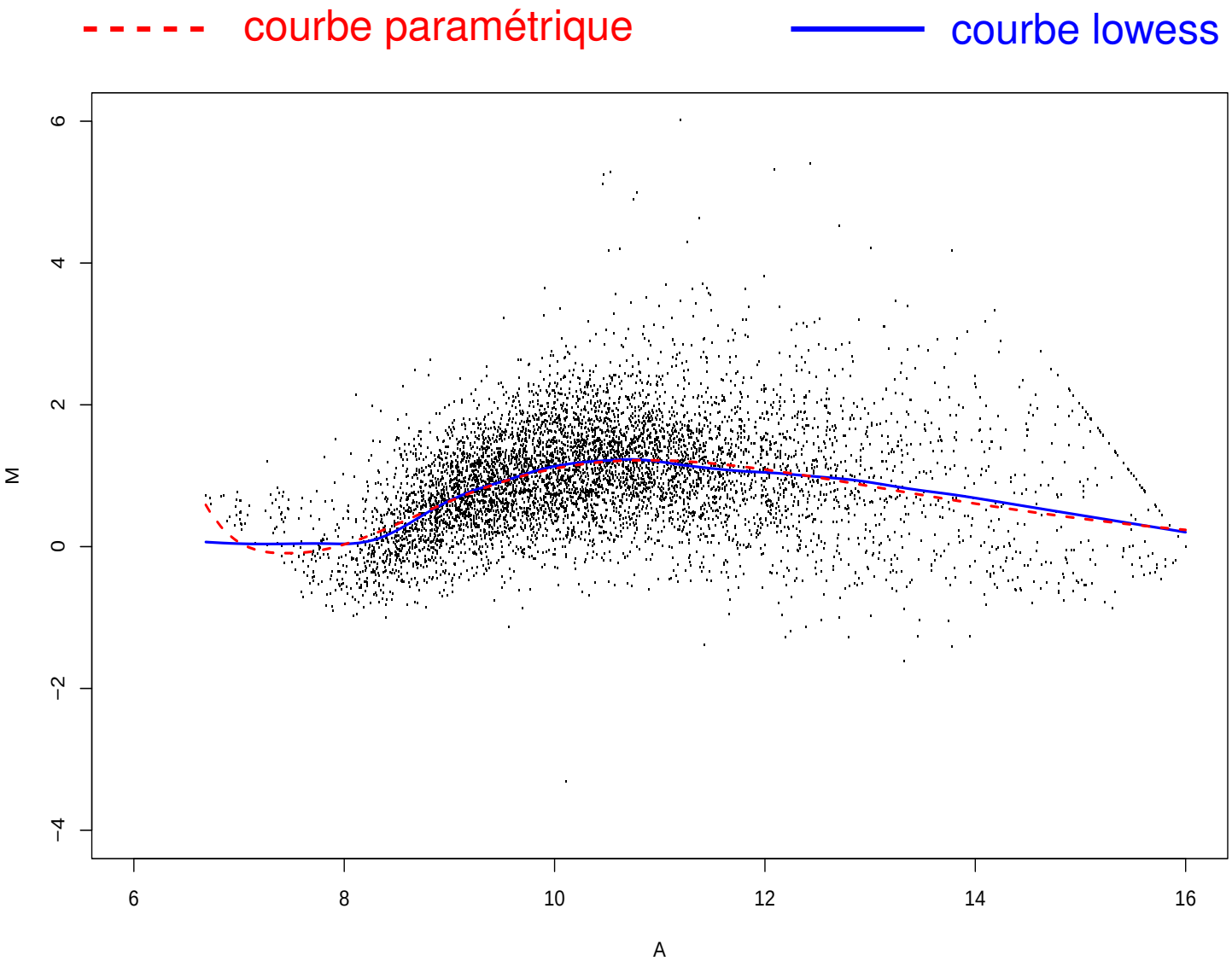
$$\rho(A, \theta) = (A - x_0)^{\alpha} e^{-\beta(A - x_1)} - k$$

    où $\theta = (x_0,\ \alpha,\ \beta,\ x_1,\ k)$ représente le vecteur des cinq paramètres de la fonction.
  - Estimation de θ avec une procédure de moindres carrés non linéaires.
  - Apprentissage de la loi de θ pour simuler ensuite de nouvelles courbes.
  - Courbes simulées dans le repère (A,M), on repasse en (G,R).

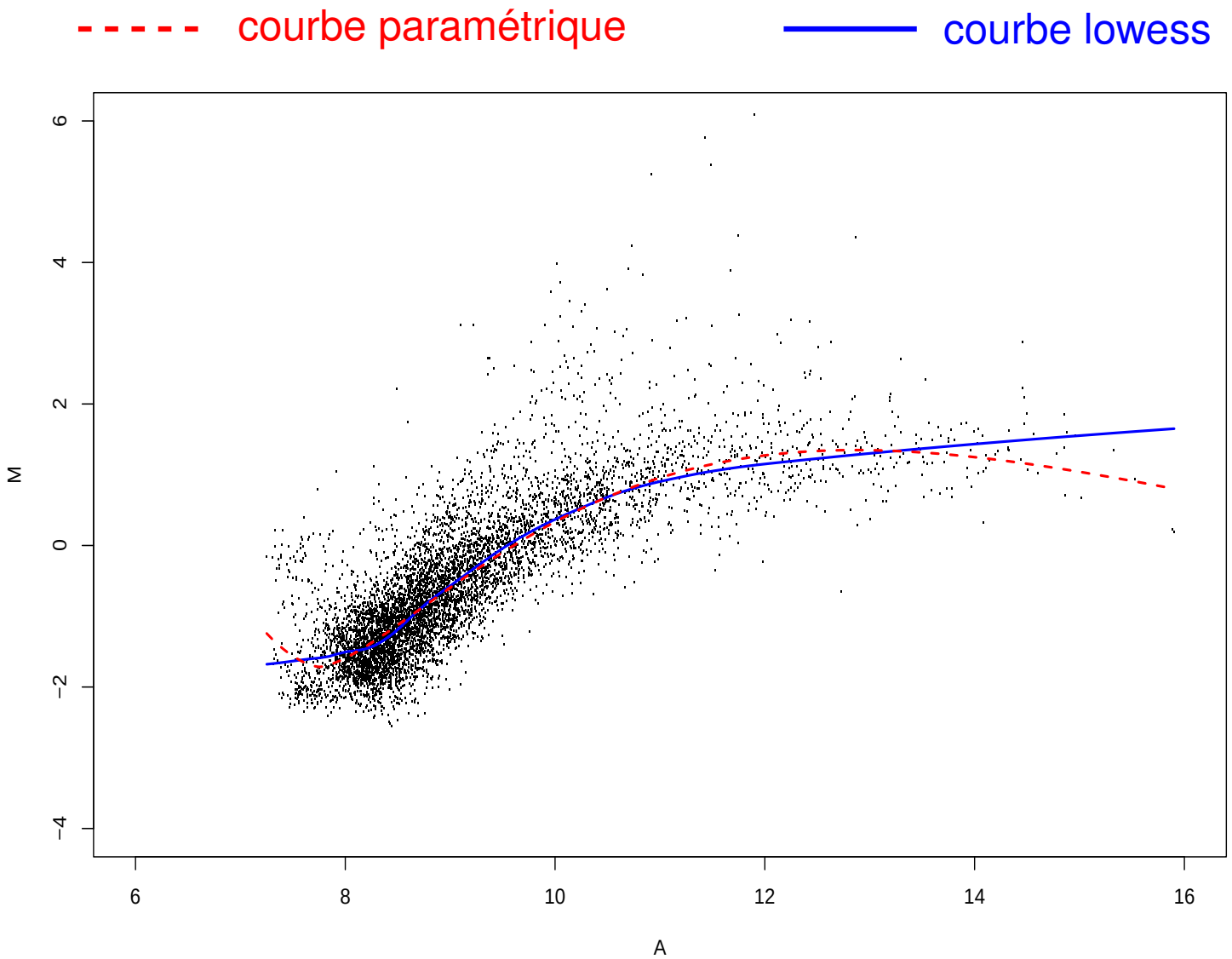
Exemples de nuages réels



Exemples de nuages réels



Exemples de nuages réels



Modèle de simulation de données (suite)

- Modélisation de l'aléa
  - On repasse en écriture (G, R) au lieu de (A, M).
  - Trois facteurs d'aléa : l'aléa dû à l'échantillon contrôle G, l'aléa dû à l'échantillon test R et celui commun aux deux échantillons, effet de l'expérience noté S.
  - *Modèle :*

$$\begin{aligned} G_i &= \alpha_G + \mu_{G,i} e^{\eta_{S,i}+\eta_{G,i}} + \epsilon_{S,i} + \epsilon_{G,i} \\ R_i &= \alpha_R + \mu_{R,i} e^{\eta_{S,i}+\eta_{R,i}} + \epsilon_{S,i} + \epsilon_{R,i} \end{aligned}$$

Pour un indice $k \in \{G, R, S\}$, on a :

$$\begin{aligned} \eta_{k,i} &\sim N(0, \sigma^2_{\eta_k}) \\ \epsilon_{k,i} &\sim N(0, \sigma^2_{\epsilon_k}) \end{aligned}$$

Modèle de simulation de données (suite)

- Modélisation de l'aléa
  - On repasse en écriture (G, R) au lieu de (A, M).
  - Trois facteurs d'aléa : l'aléa dû à l'échantillon contrôle G, l'aléa dû à l'échantillon test R et celui commun aux deux échantillons, effet de l'expérience noté S.
  - *Modèle :*

$$\begin{aligned} G_i &= \alpha_G + \mu_{G,i} e^{\eta_{S,i}+\eta_{G,i}} + \epsilon_{S,i} + \epsilon_{G,i} \\ R_i &= \alpha_R + \mu_{R,i} e^{\eta_{S,i}+\eta_{R,i}} + \epsilon_{S,i} + \epsilon_{R,i} \end{aligned}$$

Pour un indice $k \in \{G, R, S\}$, on a :

$$\begin{aligned} \eta_{k,i} &\sim N(0, \sigma^2_{\eta_k}) \\ \epsilon_{k,i} &\sim N(0, \sigma^2_{\epsilon_k}) \end{aligned}$$

Modèle de simulation de données (suite)

- Modélisation de l'aléa
  - On repasse en écriture (G, R) au lieu de (A, M).
  - Trois facteurs d'aléa : l'aléa dû à l'échantillon contrôle G, l'aléa dû à l'échantillon test R et celui commun aux deux échantillons, effet de l'expérience noté S.
  - *Modèle :*

$$\begin{aligned} G_i &= \alpha_G + \mu_{G,i} e^{\eta_{S,i}+\eta_{G,i}} + \epsilon_{S,i} + \epsilon_{G,i} \\ R_i &= \alpha_R + \mu_{R,i} e^{\eta_{S,i}+\eta_{R,i}} + \epsilon_{S,i} + \epsilon_{R,i} \end{aligned}$$

Pour un indice $k \in \{G, R, S\}$, on a :

$$\begin{aligned} \eta_{k,i} &\sim N(0, \sigma^2_{\eta_k}) \\ \epsilon_{k,i} &\sim N(0, \sigma^2_{\epsilon_k}) \end{aligned}$$

Modèle de simulation de données (suite)

- Modélisation de l'aléa
  - On repasse en écriture (G, R) au lieu de (A, M).
  - Trois facteurs d'aléa : l'aléa dû à l'échantillon contrôle G, l'aléa dû à l'échantillon test R et celui commun aux deux échantillons, effet de l'expérience noté S.
  - *Modèle :*

$$\begin{aligned} G_i &= \alpha_G + \mu_{G,i} e^{\eta_{S,i}+\eta_{G,i}} + \epsilon_{S,i} + \epsilon_{G,i} \\ R_i &= \alpha_R + \mu_{R,i} e^{\eta_{S,i}+\eta_{R,i}} + \epsilon_{S,i} + \epsilon_{R,i} \end{aligned}$$

Pour un indice $k \in \{G, R, S\}$, on a :

$$\begin{aligned} \eta_{k,i} &\sim N(0, \sigma^2_{\eta_k}) \\ \epsilon_{k,i} &\sim N(0, \sigma^2_{\epsilon_k}) \end{aligned}$$

Modèle de simulation de données (suite)

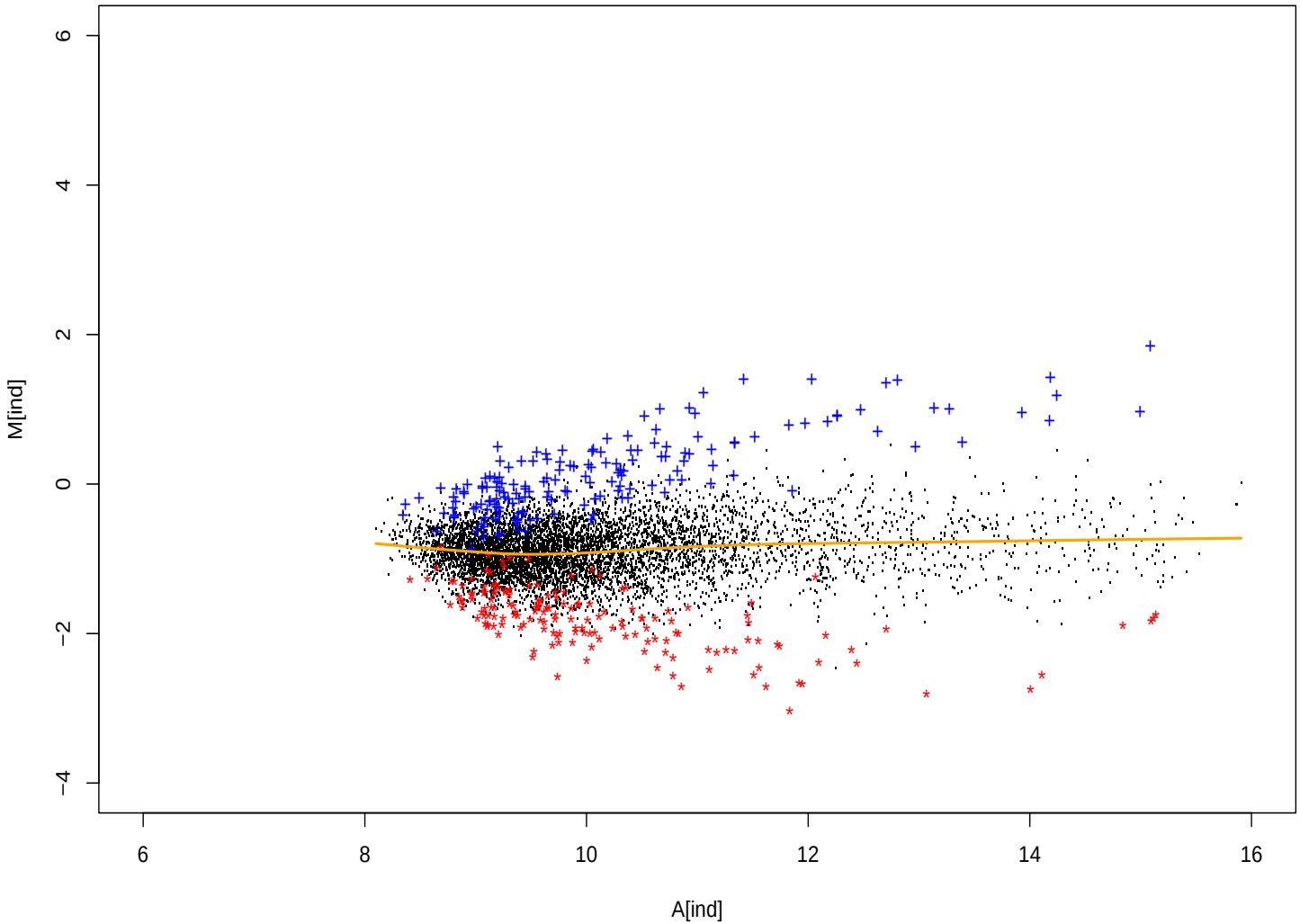
- Modélisation de l'aléa
  - On repasse en écriture (G, R) au lieu de (A, M).
  - Trois facteurs d'aléa : l'aléa dû à l'échantillon contrôle G, l'aléa dû à l'échantillon test R et celui commun aux deux échantillons, effet de l'expérience noté S.
  - *Modèle :*

$$\begin{aligned} G_i &= \alpha_G + \mu_{G,i} e^{\eta_{S,i}+\eta_{G,i}} + \epsilon_{S,i} + \epsilon_{G,i} \\ R_i &= \alpha_R + \mu_{R,i} e^{\eta_{S,i}+\eta_{R,i}} + \epsilon_{S,i} + \epsilon_{R,i} \end{aligned}$$

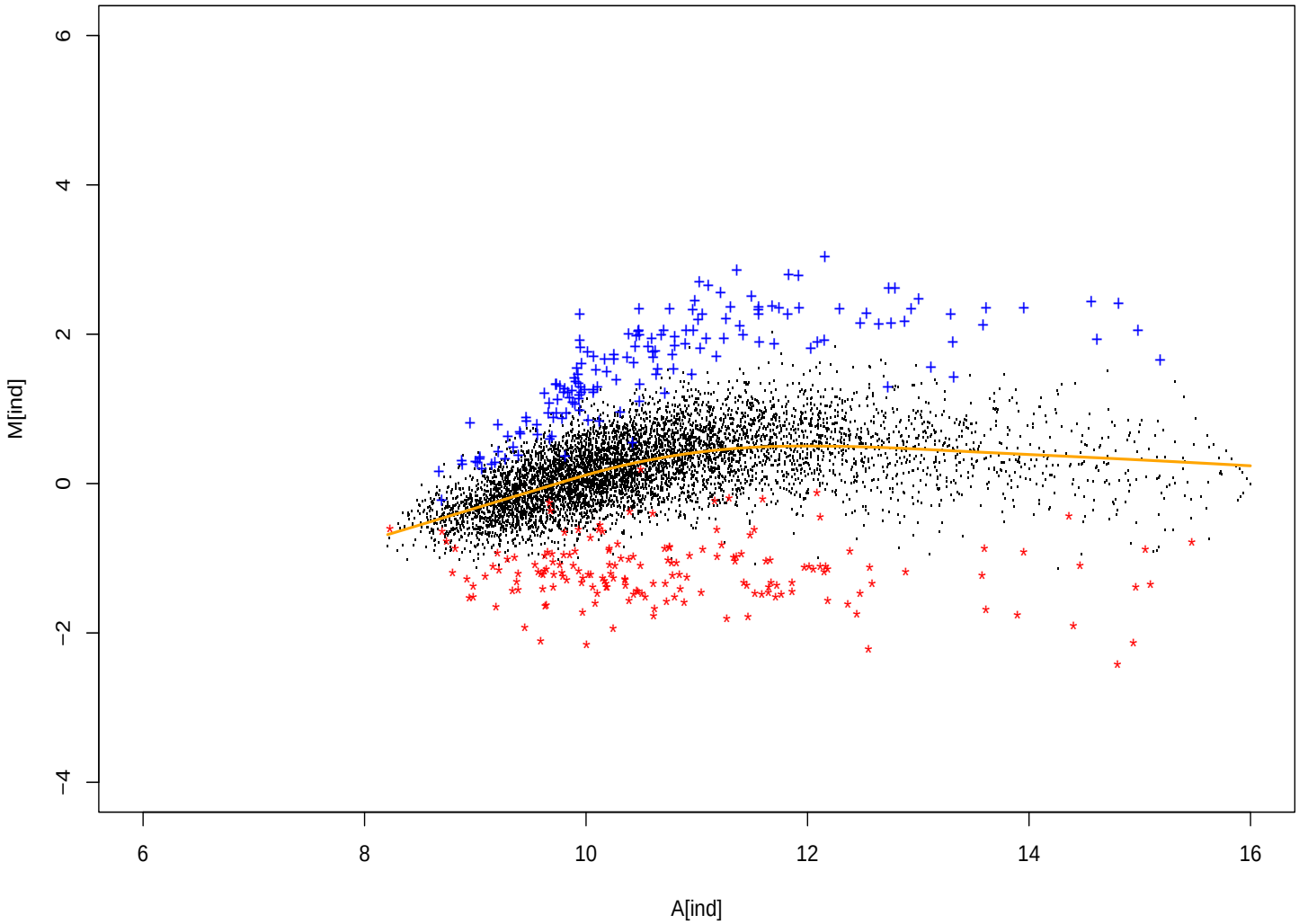
Pour un indice $k \in \{G, R, S\}$, on a :

$$\begin{aligned} \eta_{k,i} &\sim N(0, \sigma^2_{\eta_k}) \\ \epsilon_{k,i} &\sim N(0, \sigma^2_{\epsilon_k}) \end{aligned}$$

Des données simulées



Des données simulées



Résultat pour le choix de la méthode de normalisation

- Sur les données simulées, on connaît les gènes exprimés, on peut donc mesurer la qualité d'une normalisation.

- Les simulations effectuées nous amènent à choisir une normalisation lowess.

Détection des gènes différentiellement exprimés

Problématique

- But : identifier les gènes qui présentent une différence statistique significative entre deux conditions expérimentales.
- Notations : p gènes, n_1 biopuces condition 1, n_2 biopuces condition 2.
- Test d'hypothèses multiples : pour chaque gène i, on teste

H_{0i} : « le gène i n'est pas différentiellement exprimé entre les populations 1 et 2 »

contre

H_{1i} : « le gène i est différentiellement exprimé entre les populations 1 et 2 »

Deux éléments importants dans ce problème

- Choix de la statistique de test notée Z_i, i indice du gène
- Règle de décision / Seuil du test
  Critère : *FDR* (False Discovery Rate)

$$Q = \begin{cases} \dfrac{\text{nb faux positifs}}{\text{nb découvertes}} & \\ 0 & \text{si aucun gène découvert} \end{cases}$$

On définit $FDR = \mathbb{E}(Q)$

Pour ces deux points plusieurs approches envisagées.

Deux éléments importants dans ce problème

- Choix de la statistique de test notée Z_i, i indice du gène
- Règle de décision / Seuil du test
  Critère : *FDR* (False Discovery Rate)

$$Q = \begin{cases} \dfrac{\text{nb faux positifs}}{\text{nb découvertes}} & \\ 0 & \text{si aucun gène découvert} \end{cases}$$

On définit $FDR = \mathbb{E}(Q)$

Pour ces deux points plusieurs approches envisagées.

Règles de décision envisagées

- Sélection de modèle par procédure FDR
- Sélection de modèle et pénalisation
  (Abramovich et al. *2000 et Golubev 2002)*
- FDR-ondelettes
- Seuillage bayésien
  (Johnstone et Silverman 2002)

Règles de décision envisagées

- Sélection de modèle par procédure FDR
- Sélection de modèle et pénalisation
  (Abramovich et al. *2000 et Golubev 2002)*
- FDR-ondelettes
- Seuillage bayésien
  (Johnstone et Silverman 2002)

Règles de décision envisagées

- Sélection de modèle par procédure FDR
- Sélection de modèle et pénalisation
  (Abramovich et al. *2000 et Golubev 2002)*
- FDR-ondelettes
- Seuillage bayésien
  (Johnstone et Silverman 2002)

Sélection de modèle par procédure FDR (1)

Le modèle que l'on cherche à sélectionner est celui qui nous donne la liste des gènes différentiellement exprimés. On estime l'ensemble I_1 des indices de ces gènes par un ensemble $\hat{I}$.

- On suppose qu'on a les p-valeurs π_i où i désigne l'indice du gène.
- On les range par ordre croissant

$$\pi_{(1)} \leq \ldots \leq \pi_{(p)}$$

- On calcule $k = \max\{i \mid \pi_{(i)} \leq \frac{i}{p} q\}$ ($q =$ seuil)
- On rejette les hypothèses $H_{0(j)}$ pour tout $j \in \{1, \ldots, k\}$.
  Si un tel k n'existe pas, on ne rejette aucune hypothèse *i.e.* aucun gène n'est détecté.

Consistance de la procédure

Hypothèses (HT) sur la statistique de test Z_i :

$$Z_i - \mu_i \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

pour une suite μ_i, dépendante de n, telle que :

(a) $\limsup_{n\to+\infty} |\mu_i| = 0 \quad \forall i \notin I_1$

(b) $\liminf_{n\to+\infty} \frac{|\mu_i|}{b_n} > 1 \quad \forall i \in I_1$ pour une suite b_n telle que $b_n \xrightarrow[n\to+\infty]{} +\infty$

Proposition - Consistance de la procédure FDR

Si *(HT)* est vraie et si $q_n \xrightarrow[n\to+\infty]{} 0$ satisfait $\forall j \in I_1$, $\lim_{n\to+\infty} \frac{1-\Phi(|\mu_j|)}{q_n} = 0$

Alors :

$$\lim_{n\to+\infty} \mathbb{P}(\hat{I} = I_1) = 1$$

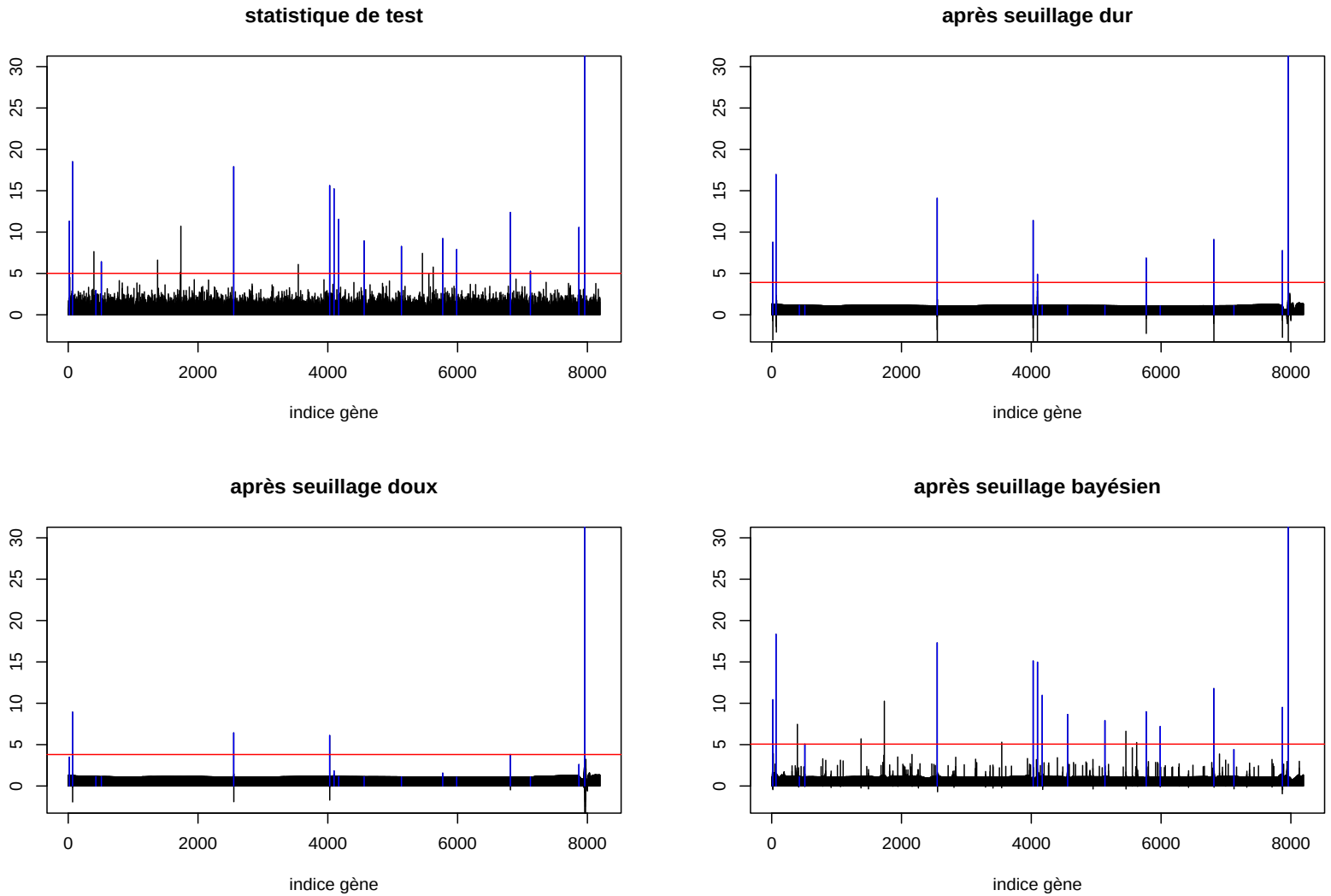
FDR-ondelettes (1)

- Idée : tenir compte du caractère creux des données et s'approcher de l'hypothèse d'indépendance pour rendre la procédure FDR efficace.

- Principe :
  - utiliser une transformation en ondelettes sur la statistique de test pour décorréler les données
  - faire un seuillage des coefficients d'ondelettes pour réduire le nombre d'hypothèses à tester (p^* hypothèses)
  - appliquer une méthode FDR sur les p^* statistiques de test restantes

- Inconvénient : il faut avoir p proche d'une puissance de 2.

FDR-ondelettes (2)



Eset 12 - Reconstruction de la statistique de test (en val. abs.) selon le type de seuillage des coefficients d'ondelettes utilisé

FDR-ondelettes (3)

Eset 12 : jeu de données “artificiel” où 16 gènes connus sont différentiellement exprimés

$$p = 12\,626, \quad n_1 = n_2 = 12$$

Remarque : on se limite à $p' = 8192 = 2^{14}$

Seuillage ondelettes	Nb détectés	Nb vrai diff.	FDR obs.
Aucun seuillage	23	15	0.35
Seuillage dur	10	8	0.20
Seuillage doux	5	4	0.20
Seuillage bayésien	20	14	0.30

Résultats sur *Eset12*

$$p = 12\,626, \quad n_1 = n_2 = 12$$

Méthode	Nb détectés	Nb vrai diff.	FDR obs.
FDR BH	23	15	0.348
FDR BY	19	14	0.263
Pen. Abramovich 1	30	15	0.500
Pen. Cons. Golubev	20	14	0.300
Subst. Golubev	23	15	0.348
Ebayes Cauchy	25	15	0.400

Résultats sur des données de l'Institut Curie

Fortes doses d'irradiation (200 Gy) contre faibles doses (entre 7.5 μGy et 36.2 Gy). $p = 6\,327,\ n_1 = 24,\ n_2 = 10$

Méthode	Nb gènes détectés
FDR BH	2307
FDR BY	1128
Pen. Abramovich 1	2379
Pen. Cons. Golubev	6327
Subst. Golubev	2082
Ebayes Cauchy	6327
Ebayes Laplace	6327

Réduction de dimension et classification supervisée

avec Sophie Lambert-Lacroix

Problématique

- On dispose de n couples indépendants (Y_i, X_i) de même loi que (Y, X)
  - Y_i variable réponse binaire. Par exemple : cancer ou pas cancer
  - X_i variable p-dimensionnelle. X_i correspond au vecteur du profil d'expression génétique de l'individu i.

- Objectif : faire de la classification supervisée
  i.e. pouvoir prédire la classe Y_i en fonction de X_i.

- La difficulté : $n << p$

  On s'intéresse aux méthodes de réduction de dimension.

Le cadre : les Modèles Linéaires Généralisés (GLM)

La loi conditionnelle de Y sachant X est supposée appartenir à la famille des lois exponentielles, de densité conditionnelle donnée par

$$\exp\left(\frac{y^T\eta(x) - b(\eta(x))}{\phi} + c(y,\phi)\right). \quad (1)$$

Les fonctions b, c et le paramètre de dispersion $\phi > 0$ sont connus.
Soit $g : \mathbb{R} \to \mathbb{R}$ une fonction inversible, spécifiant la relation entre le prédicteur η et la fonction de régression μ (*i.e.* l'espérance conditionnelle $\mu(x) = \mathbb{E}[Y|X = x]$) selon la relation

$$\eta = g(\mu).$$

g est appelée la fonction de lien canonique.

$$\text{On a : } Y = \mu(X) + \epsilon.$$

Modèle de réduction de dimension (1)

- Pour réduire la dimension, on projette X sur un espace vectoriel engendré par les covariables et on ajuste une courbe non paramétrique de ces combinaisons linéaires

$$Y = \tilde{\mu}(B^T X) + \epsilon, \quad \mathbb{E}[\epsilon|X] = 0 \ p.s.$$

où $B \in \mathcal{M}(p, \kappa)$ avec κ le nombre de combinaisons linéaires conservées

Modèle de réduction de dimension (1)

- Pour réduire la dimension, on projette X sur un espace vectoriel engendré par les covariables et on ajuste une courbe non paramétrique de ces combinaisons linéaires

$$Y = \tilde{\mu}(B^T X) + \epsilon, \quad \mathbb{E}[\epsilon|X] = 0 \ p.s.$$

où $B \in \mathcal{M}(p, \kappa)$ avec κ le nombre de combinaisons linéaires conservées

- Modèles en indice simple : $\kappa = 1$, on note $B = \beta$

Modèle de réduction de dimension (2)

- Par définition des modèles linéaires généralisés, on avait

$$Y = \mu(X) + \epsilon.$$

Modèles en indice simple : $\exists \beta \in \mathbb{R}^p$ et $\tilde{\mu} : \mathbb{R} \mapsto \mathbb{R}$ tel que $\mu(X) = \tilde{\mu}(\beta^T X)$.

- β défini à un facteur d'échelle et un facteur de signe près, on pose $\beta = \mathbb{E}[\nabla \mu(X)]$, en effet on a :

$$\mathbb{E}[\nabla \mu(X)] = \mathbb{E}\left[\nabla\{\tilde{\mu}(\beta^T X)\}\right] = \mathbb{E}\left[\nabla \tilde{\mu}(\beta^T X)\right] \beta = c\beta.$$

Approche GSIM

- Objectif : estimer β (tel qu'on l'a fixé par convention) et $\tilde{\eta} = g(\tilde{\mu})$.
- Remarque :

$$\left.\begin{array}{l} \beta = \mathbb{E}[\nabla \mu(X)] \\ \mu = g^{-1}(\eta) \end{array}\right\} \Longrightarrow \beta = \mathbb{E}\left[(g^{-1})'(\eta(X))\ \nabla \eta(X)\right].$$

- Idée : estimer η et $\nabla\eta$ par polynômes locaux à partir de la fonction de vraisemblance conditionnelle (1).
  On pose alors :

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} (g^{-1})'(\hat{\eta}(X_i))\widehat{\nabla \eta}(X_i).$$

On régresse Y_i sur $\hat{\beta}^T X_i$ par polynômes locaux pour obtenir $\widehat{\tilde{\eta}}$ et $\hat{\mu}(x) = g^{-1}(\widehat{\tilde{\eta}}(\hat{\beta}^T x))$.

GSIM : algorithme

Étape A : Pour $j = 1, \ldots, n$, calculer

$$\hat{\eta}(X_j) = \widehat{a}_j \in \mathbb{R}, \quad \widehat{\nabla\eta}(X_j) = \widehat{\underline{b}_j} \in \mathbb{R}^p,$$

obtenus en maximisant

$$\sum_{i=1}^{n} \mathcal{L}\left(a_j + \underline{b}_j(X_i - X_j), Y_i\right) K_H^p(X_i - X_j),$$

comme fonction de a_j et $\underline{b}_j$. Poser

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} (g^{-1})'(\hat{\eta}(X_i)) \widehat{\nabla\eta}(X_i).$$

Étape B : Déterminer $\widehat{\eta}(x) = \widehat{a}_0$ en maximisant

$$\sum_{i=1}^{n} \mathcal{L}\left(a_0 + b_0 \hat{\beta}^T (X_i - x), Y_i\right) K_{h_B}^1\left(\hat{\beta}^T (X_i - x)\right),$$

comme fonction de a_0 et b_0.

GSIM : problèmes de résolution

IRLS : Iterative Reweighted Least Squares. Dans le cadre des modèles de Bernoulli, on a

$$gradient\ critère = Z^T W_k [Y - \Pi] \approx 0.$$

Z de rang plein en colonnes $\rightarrow$ $Y = \Pi = e^\eta/(1 + e^\eta)$ et donc

$$\eta(X_i) = ln(Y_i) - ln(1 - Y_i) = +/ - \infty \text{ !!!}$$

Pénalité de type Ridge :

$$\sum_{i=1}^{n} \mathcal{L}\left(a_j + \underline{b}_j(X_i - X_j), Y_i\right) K_H^p(X_i - X_j) - \frac{1}{2}\lambda \underline{b}_j^T \Sigma^2 \underline{b}_j$$

où $\Sigma^2 =$ matrice de variance empirique de X.

GSIM : problèmes de résolution

IRLS : Iterative Reweighted Least Squares. Dans le cadre des modèles de Bernoulli, on a

$$gradient\ critère = Z^T W_k [Y - \Pi] \approx 0.$$

Z de rang plein en colonnes $\rightarrow$ $Y = \Pi = e^\eta/(1 + e^\eta)$ et donc

$$\eta(X_i) = ln(Y_i) - ln(1 - Y_i) = +/ - \infty \text{ !!!}$$

Pénalité de type Ridge :

$$\sum_{i=1}^{n} \mathcal{L}\left(a_j + \underline{b}_j(X_i - X_j), Y_i\right) K_H^p(X_i - X_j) - \frac{1}{2}\lambda \underline{b}_j^T \Sigma^2 \underline{b}_j$$

où $\Sigma^2 =$ matrice de variance empirique de X.

Inconvénient : choisir le paramètre de régularisation λ.

GSIM$_\lambda$: paramètres et choix d'implantation

- Etape A :
  - choix des noyaux

  $\longrightarrow$ noyau produit gaussien

  - matrice H = $h_A \mathrm{Id}_p$ (stand. en colonnes, $\Sigma^2 = \mathrm{Id}_p$)
  - paramètre de régularisation λ

  $\longrightarrow$ choisis simultanément par validation croisée

- Etape B :
  - noyau unidimensionnel gaussien
  - fenêtre de lissage h_B choisie par "plug-in"
    (Fan et Gijbels 1996)

Méthodes de comparaison

- Méthodes simples : *kNN*, *DLDA* et *DQDA*
- Méthode *rOPG* *(Xia* et al. *2002)* : méthode d'estimation d'un espace de réduction de dimension.
  - Correspond à l'étape A de *GSIM* mais avec un critère de moindres carrés à la place d'un critère de vraisemblance.
  - Comme pour *GSIM*, on ajoute une pénalité de type Ridge.
  - Appliquée dans les mêmes conditions que $GSIM_\lambda$: fenêtre à l'étape A et paramètre de régularisation λ choisis par validation croisée et étape B identique à $GSIM_\lambda$.

Application aux données de biopuces : *Colon*

- 62 profils d'expression dont 40 tissus tumoraux et 22 tissus sains. On a $p = 2000$ gènes
- Méthode de test : 100 subdivisions aléatoires en
  - un ensemble d'apprentissage
    41 individus soit 2/3 de la population
  - un ensemble de test
    21 individus soit 1/3 de la population
- Prétraitement préalable nous amène à ≈ 1200 gènes

	DLDA	DQDA	KNN	rOPG$_\lambda$	GSIM$_\lambda$
moy	0.144	0.154	0.204	0.153	0.148
std	0.057	0.064	0.071	0.060	0.056

Colon. Etude resampling : moyenne et écart-type du taux d'erreur dans le fichier test.

Conclusion

- un modèle de simulation de données qui semble raisonnable
- des méthodes originales de détection de gènes différentiellement exprimés
- une méthode de réduction de dimension intéressante au niveau pratique et théorique

Perspectives

- développer le modèle de simulation pour le rendre plus général
- combiner plusieurs classifieurs
- extension au cas multiclasse
- utiliser $\hat{\beta}$ pour faire de la sélection de gènes

Merci pour votre attention.

Application aux données de biopuces : *Leukemia multiclasse*

- 72 profils d'expression dont
  - 47 tissus leucémie ALL (38 type B, 9 type T)
  - 25 tissus leucémie AML

- 7129 gènes avant prétraitement

	DLDA	DQDA	KNN	GSIM$_\lambda$
moy	0.039	0.046	0.055	0.025
std	0.037	0.039	0.036	0.026