Correction du bruit dans les microarrays

Sandra Plancade

Université de Tromsø - Norvège

7 mars 2011

Plan de l'exposé

Introduction : bruit dans les microarrays.

1. **Modèle de bruit additif.**
2. **Modèle normal-exponentiel.**
3. **Déconvolution non paramétrique pour Illumina.**
4. **Un nouveau modèle de bruit pour Illumina.**

Rappel sur les microarrays

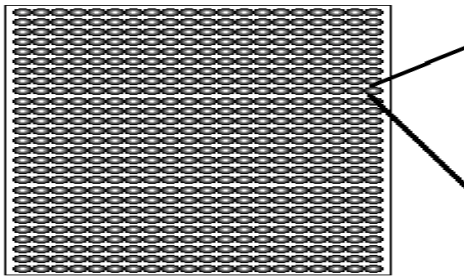
- Microarrays = dispositifs permettant de mesurer simultanément la concentration de milliers de gènes dans un échantillon (sang, tumeur...)

- Deux technologies les plus utilisées :
  - Illumina
  - Affymetrix

- Même principe général, mais analyse des données différente.

Principe général

- Plaque comprenant 100 000 - 1 000 000 de cellules.



Chaque cellule comporte une probe
(= sequence de nucléotide : A,T,C,G)
répetées 300 000 - 1 000 000 fois.

- Chaque probe est complémentaire d'une unique portion de gène :

  Gène ...ACTGGCGTAAGCTAG ...
  Probe CCGCATTCG

→ Théoriquement, une probe s'attache à un gène si il contient la séquence complémentaire, avec une probabilité proportionnelle à la concentration du gène.

- La microarray est éclairée par un laser et dans chaque cellule, les séquences hybridées emettent une intensité lumineuse

→ Estimation de la concentration du gène dans l'échantillon.

- Procédure :
  - Préparation de l'échantillon (étapes techniques).
  - Hybridation.
  - Elimination des probes non hybridées.
  - Mesure de l'intensité lumineuse émise par chaque cellule.

- Binding spécifique et non spécifique : une probe peut s'attacher
  - au gène qui lui correspond (specific binding)
  - à un autre gène (non-specific binding).

- L'intensité mesurée n'est pas exactement proportionnelle à la concentration du gène.
  - Erreurs techniques.
  - Binding non-spécifique.

Probes de contrôle

- Probes de contrôle : séquences qui ne sont complémentaires d'aucune portion de gène.
- Intensité des probes de contrôle = bruit.
- Construction différentes selon la plateforme :
  - Affymetrix : pour chaque probe régulière (PM = Perfect Match), il existe une probe de contrôle (MM = Mismatch) tel que PM et MM diffèrent d'un seul nucléotide.
  - Illumina : 759 probes negatives sur une array.

Modèle de bruit additif

Soit une probe p, et une array (= un échantillon) i. On considère que l'intensité mesurée $X_{i,p}$ s'écrit

$$X_{i,p} = S_{i,p} + B_{i,p} \quad \text{où}$$

- $S_{i,p}$: signal, proportionnel à la concentration du gène correspondant à p.
- $B_{i,p}$: terme de bruit (background noise).
- $S_{i,p} \perp B_{i,p}$
- $\{B_{i,1}, ..., B_{i,n_{probes}}\}$ i.i.d. de densité f_{B_i}.

On s'interesse à S
On mesure X
} On veut donc corriger l'effet de B.

Produit de convolution

$$X = S + B, \quad X \sim f_X,\ S \sim f_S,\ B \sim f_B \quad \text{et} \quad X \perp S.$$

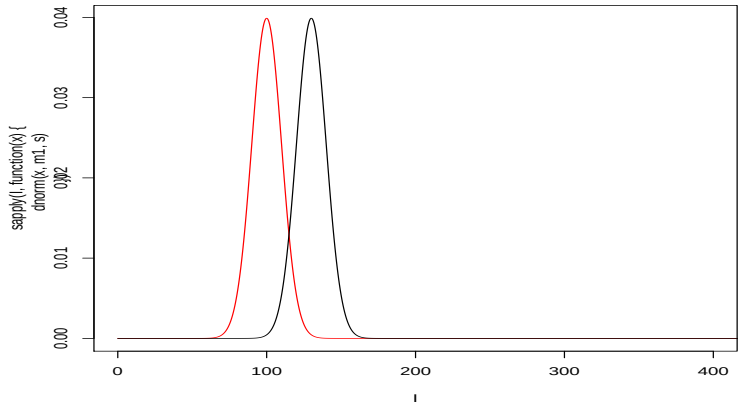
Alors

$$\begin{aligned} P[X = x] &= P[S + B = x] \\ &= \int P[S + B = x \cap S = s]ds \\ &= \int P[B = x - s \cap S = s]ds \\ &= \int P[B = x - s]P[S = s]ds \quad \text{(S et B indépendants)} \end{aligned}$$

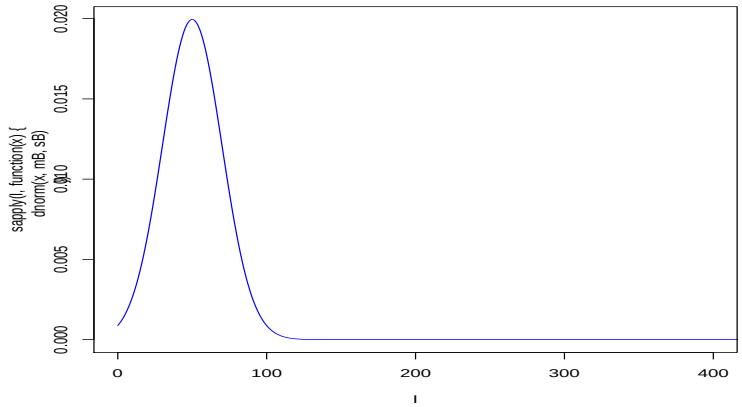
$$\Rightarrow f_X(x) = \int f_B(x - s)f_S(s)ds := f_B \star f_S(x) = f_S \star f_B.$$

Comparaison de l'intensité d'une probe p, parmi deux groupes d'individus.

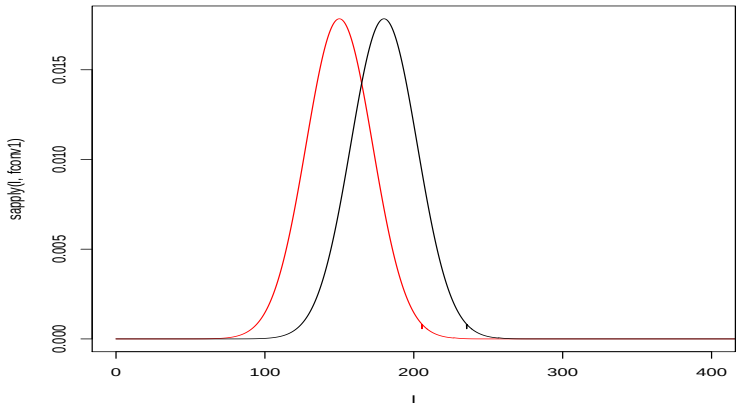
- Groupe 1 : signal S1, bruit B, intensité observée X1=S1+B
- Groupe 2 : signal S2, bruit B, intensité observée X2=S2+B



f_{S_1} (rouge) et f_{S_2} (noir)

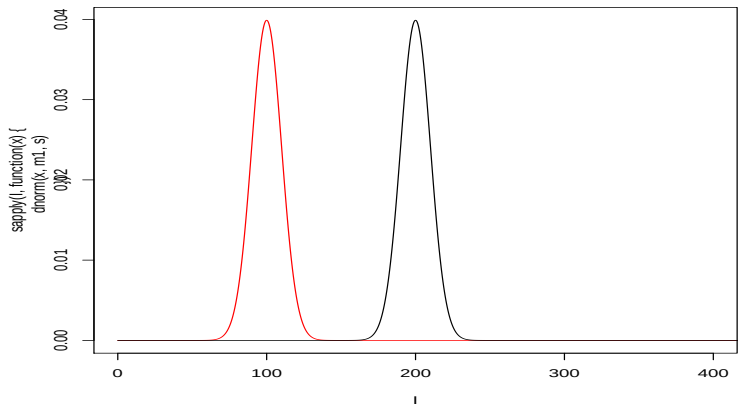


f_B

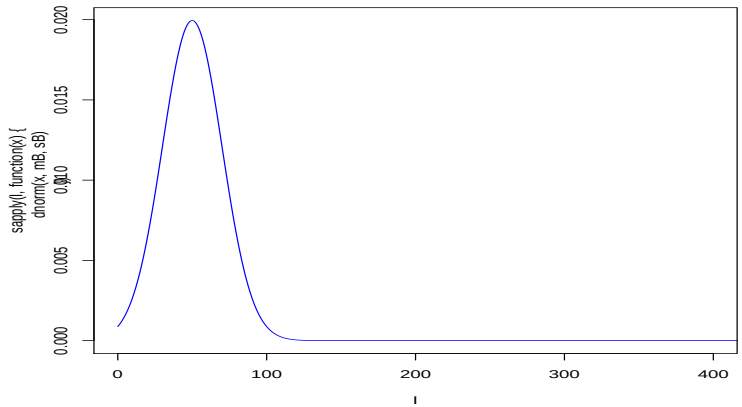


$f_{X_1} = f_{S_1} \star f_B$ (rouge)
et $f_{X_2} = f_{S_2} \star f_B$ (noir)

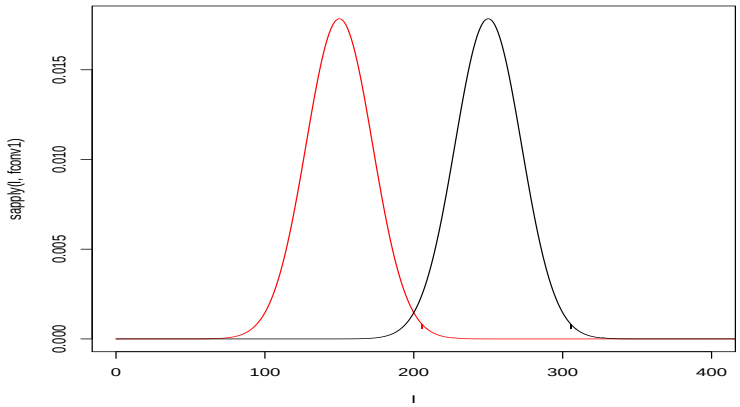
Remarque : L'effet du bruit est important pour des différences d'expression du même ordre que le bruit.



f_{S_1} (rouge) et f_{S_2} (noir)



f_B



$f_{X_1} = f_{S_1} \star f_B$ (rouge)
et $f_{X_2} = f_{S_2} \star f_B$ (noir)

Correction du bruit

- $X = S + B$,
  - On mesure X
  - On veut en déduire un estimateur $\widehat{S}$.
- Pour débruiter **une** mesure X_{i_0,j_0}, on utilise les informations provenant d'un **ensemble** de mesures $\{X_{i,j}, i, j\}$
  - Premier type d'information : probes négatives.
  - Second type d'information : modèle sur f_B et f_S.

Correction du bruit (2)

$X = S + B$. Supposons f_B connue (ou estimée)

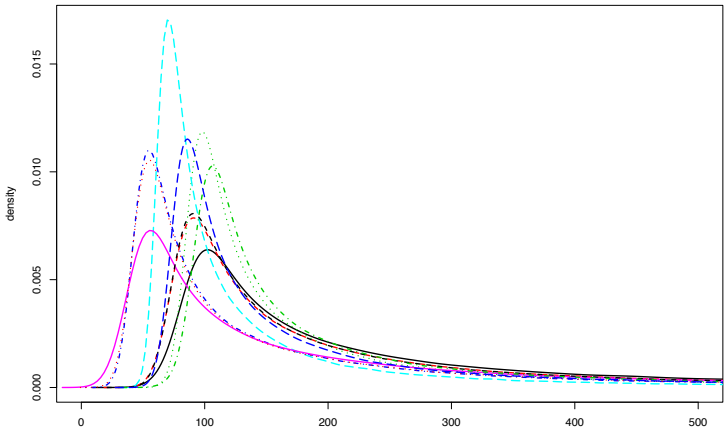
- Première approche : soustraction de la moyenne du bruit. $\widehat{S} = S - \mathbb{E}[B]$.
  ◇ Crée des valeurs négatives $\rightarrow$ Nécessité d'éliminer les valeurs $< \mathbb{E}[B]$.

- Deuxième approche : déconvolution. Supposons f_S connue (ou estimée),

$$\begin{aligned}
\mathbb{E}[S|X=x] &= \int sP[S=s|X=x]ds \\
&= \int sP[S=s \cap X=x]/P[X=x]ds \\
&= (1/P[X=x]) \int sP[S=s \cap S+B=x]ds \\
&= (1/P[X=x]) \int sP[S=s \cap B=x-s]ds \\
&= (1/P[X=x]) \int sP[S=s]P[B=x-s]ds
\end{aligned}$$

$$\Rightarrow \quad \widehat{S} = \frac{1}{(f_S \star f_B)(x)} \int s f_S(s) f_B(x-s) ds.$$

Le modèle normal-exponentiel



Distribution de l'ensemble des probes sur une array (une courbe = une array)

Soit i une array, et $p = 1, ..., n_{\text{probe}}$ l'ensemble des probes (non négatives). L'intensité mesurée $X_{i,p}$ se décompose en

$$X_{i,p} = S_{i,p} + B_{i,p} \quad \text{avec}$$

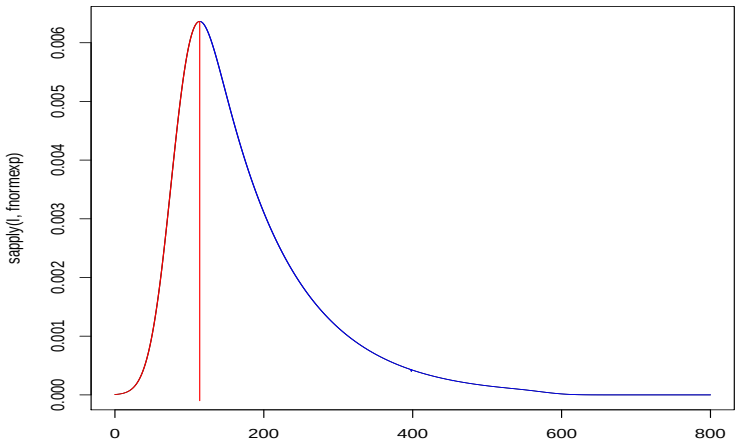
- $S_{i,p} \sim \mathcal{E}xp(\alpha_i)$
- $B_{i,p} \sim \mathcal{N}(\mu_i, \sigma_i^2)$.

Remarque : Implicitement, $(S_{i,1}, \ldots, S_{i,n_{\text{probe}}})$ sont considérés comme indépendants dans le calcul des paramètres.

- Faux d'un point de vue biologique.
- Approximation raisonnable car l'expression d'une majorité des gènes ne varie pas.

Affymetrix : Estimation de la densité du bruit et du signal

- $X_{i,p} = S_{i,p} + B_{i,p}$ où $X_{i,p}$ peut être la densité des probes PM ou la différence PM-MM.
- RMA (Robust Multi-array Average).



$\widehat{\mu_i}$: mode de la distribution
$\widehat{\sigma_i}$: SD de la distribution à gauche de μ_i
$\widehat{\alpha_i}$: on fitte la queue de distribution ($x \geq \widehat{\mu_i}$) par une exponentielle.

◇ Méthodes plus avancées pour estimer les paramètres.

- Estimateurs basés sur le maximum de vraisemblance (estimation conjointe de (α, μ, σ)).

Illumina : Estimation de la densité du bruit et du signal

- $X_{i,p} = S_{i,p} + B_{i,p}$ avec

$$S_{i,p} \sim \mathcal{E}(\alpha_i), \quad B_{i,p} \sim \mathcal{N}(\mu_i, \sigma_i^2).$$

- Estimation de f_{B_i} à partir des probes négatives.

$$\text{Intensité des probes négatives :} \quad X_{i,p}^{\mathsf{neg}} = B_{i,p}$$

.

  - $\widehat{\mu_i}$ moyenne des intensités des probes négatives.
  - $\widehat{\sigma_i}$: SD des intensités des probes negatives.

- Estimation de f_{S_i} :

$$\widehat{\alpha_i} = \frac{1}{n_{\mathsf{probe}}} \sum_{p=1}^{n_{\mathsf{probe}}} X_{i,p} - \widehat{\mu_i}$$

.

Déconvolution dans le modèle normal-exponentiel

Soit i une array, $\widehat{\alpha_i}$, $\widehat{\mu_i}$, $\widehat{\sigma_i}$ estimateurs des paramètres du modèles, et $x = X_{i,p}$ l'intensité mesurée pour la probe p. Alors,

$$\widehat{S}_{i,p} = \frac{1}{\int \widehat{f}_{S_i}(s)\widehat{f}_{B_i}(x-s)ds} \int s\widehat{f}_{S_i}(s)\widehat{f}_{B_i}(x-s)ds$$

avec

$$\widehat{f}_{S_i}(t) = \frac{1}{\widehat{\alpha_i}} \exp\left(-\frac{t}{\widehat{\alpha_i}}\right)$$

$$\widehat{f}_{B_i}(t) = \frac{1}{\sqrt{2\pi}\widehat{\sigma_i}} \exp\left(-\frac{(t-\widehat{\mu_i})^2}{2\widehat{\sigma_i}^2}\right).$$

Déconvolution non paramétrique

- Modèle additif sans hypothèse paramétrique : pour tout array i, probe j

$$X_{i,p} = S_{i,p} + B_{i,p}$$

avec $B_{i,p}$ de densité f_{B_i}, et $S_{i,p}$ de densité f_{S_i} sur l'array i.

- Estimation de f_{S_i}.

Proposition

Soit $\mathcal{F}$ la transformée de Fourier. Pour toutes fonctions f, g,

$$\mathcal{F}(f \star g) = \mathcal{F}(f)\mathcal{F}(g).$$

Conséquence : $\mathcal{F}(f_{S_i}) = \dfrac{\mathcal{F}(f_{X_i})}{\mathcal{F}(f_{B_i})}$.

- $\mathcal{F}(f_{X_i})$ estimé à partir des probes régulières :

$$\mathcal{F}(f_{X_i})(\lambda) = \int f_{X_i}(x) e^{i\lambda x} dx = \mathbb{E}[e^{i\lambda X_i}]$$

d'où

$$\widehat{\mathcal{F}(f_{X_i})}(\lambda) = \frac{1}{n_{\textsf{probe}}} \sum_{p=1}^{n_{\textsf{probe}}} e^{i\lambda X_{i,p}}$$

- $\mathcal{F}(f_B)$ estimé à partir des probes négatives :

$$\widehat{\mathcal{F}(f_{B_i})}(\lambda) = \frac{1}{n_{\textsf{neg}}} \sum_{p=1}^{n_{\textsf{neg}}} e^{i\lambda X_{i,p}^{\textsf{neg}}}$$

- Pour tout $\lambda \in \mathbb{R}$,

$$\widehat{\mathcal{F}(f_{S_i})}(\lambda) = \frac{\widehat{\mathcal{F}(f_{X_i})}(\lambda)}{\widehat{\mathcal{F}(f_{B_i})}(\lambda)}.$$

- Approximation : $\mathcal{F}(f_{S_i}) \approx \mathcal{F}(f_{S_i})\mathbb{1}_{[-\pi m,\pi m]}$, m à déterminer.
- Décomposition de $\mathcal{F}(f_{S_i})$ en série de Fourier :

$$\mathcal{F}(f_{S_i})\mathbb{1}_{[-\pi m,\pi m]}(\lambda) = \sum_{k=1}^{m} \left(\int \mathcal{F}(f_{S_i})(x)\overline{\psi}_{k,m}(x)dx \right) \psi_{k,m}(\lambda)$$

avec $\psi_{k,m}(u) = (1/\sqrt{m})e^{iku/m}\mathbb{1}_{[-\pi m,\pi m]}(u)$. De plus,

$$\psi_{k,m} = \mathcal{F}(\phi_{k,m}) \quad \text{où} \quad \phi_{k,m}(v) = \frac{\sin(\pi(mv-k))}{\pi(mv-k)}.$$

D'où

$$\begin{aligned} f_{S_i} &\approx \sum_{k=1}^{m} \left(\int \mathcal{F}(f_{S_i})(x)\overline{\psi}_{k,m}(x)dx \right) \phi_{k,m} \\ &\approx \sum_{k=1}^{m} \left[\frac{1}{N} \sum_{j=1}^{N} \left(\mathcal{F}(f_{S_i}) \times \overline{\psi}_{k,m} \right) \left(\pi m \left(-1 + \frac{2j}{N} \right) \right) \right] \phi_{k,m} \end{aligned}$$

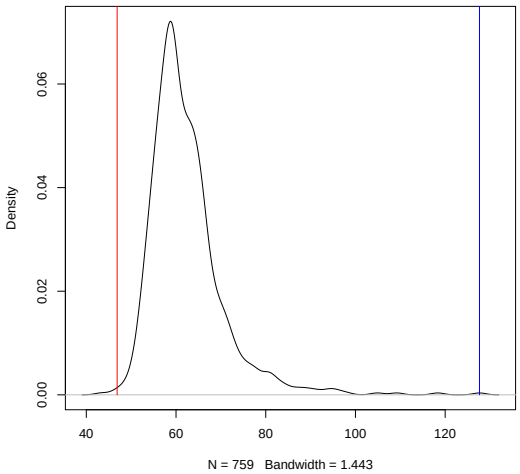
- Estimateur de f_{S_i} pour un m fixé.

$$\Rightarrow \quad \widehat{f_{S_i}} = \sum_{k=1}^{m} \left[\frac{1}{N} \sum_{j=1}^{N} \left(\widehat{\mathcal{F}(f_{S_i})} \times \overline{\psi}_{k,m} \right) \left(\pi m \left(-1 + \frac{2j}{N} \right) \right) \right] \phi_{k,m}$$

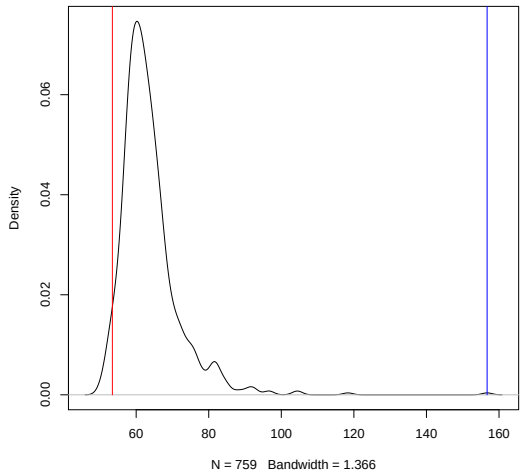
- Choix de m par sélection de modèle.
- Estimation de $S_{i,p}$:
  - $\widehat{f_{X_i}}$ estimateur de densité non paramétrique (noyau, histogramme...) construit à partir des probes régulières $\{X_{i,p}, p = 1, \ldots, n_{\text{probe}}\}$.
  - $\widehat{f_{B_i}}$ estimateur de densité non paramétrique construit à partir des probes négatives $\{X_{i,p}^{\text{neg}}, p = 1, \ldots, n_{\text{neg}}\}$.

$$\widehat{S}_{i,p} = \frac{1}{\widehat{f_{X_i}}} \int s \widehat{f_{S_i}}(s) \widehat{f_{B_i}}(x - s) ds.$$

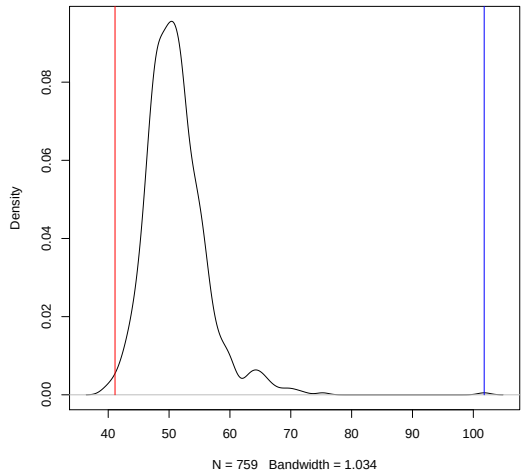
Densité des probes négatives pour une array (noir), intensité de la probe 514 (rouge) et 646 (bleu).



Array 1



Array 2



Array 3

- On observe un **effet probe**
- Les probes négatives ne représente pas un bruit i.i.d. sur l'array.

- Decomposition du bruit B en
  - Effets fixes.
  - Bruit technique résiduel non reproductible.
- Modèle :

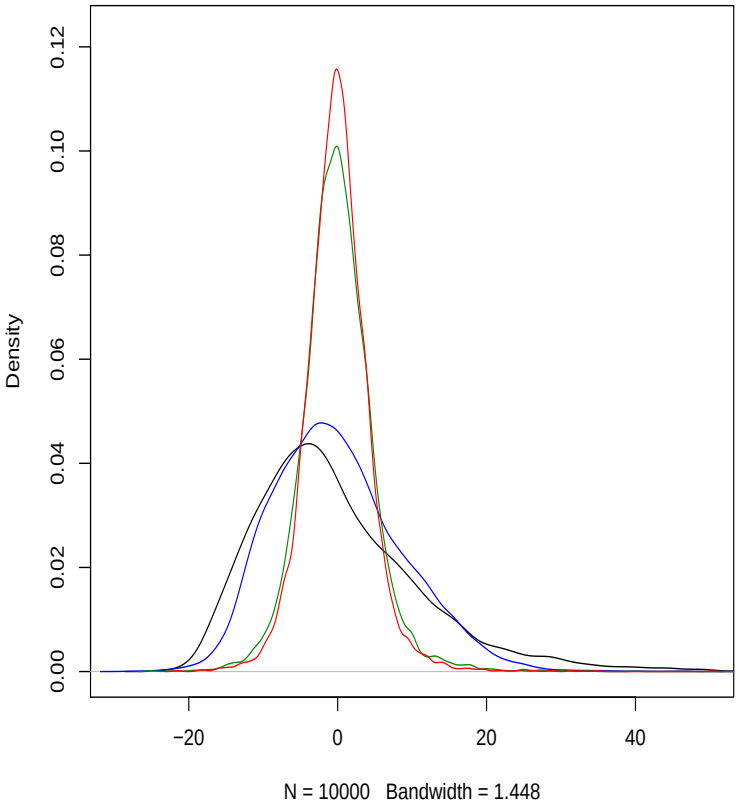
$$X_{i,p} = S_{i,p} + \underbrace{E_{i,p} + \tilde{B}_{i,p}}_{B_{i,p}}$$

  - $E_{i,p}$ = Effet fixe.
  - $\tilde{B}$ identiquement distribué sur **toutes les arrays** :
  $\{\tilde{B}_{i,p}, i = 1, \ldots, n_{array}, p = 1, \ldots, n_{probe}\}$ i.i.d. de densité $f_{\tilde{B}}$.
  - $\tilde{B}$ centré.
- Estimation des paramètres à partir des probes négatives :

$$X_{i,p}^{\text{neg}} = C_i + P_j + \tilde{B}_{i,p}$$

  - P_j et C_i coefficients du modèle linéaire
  - $f_{\tilde{B}}$: estimateur de densité des résidus du modèle linéaire.

Distribution du bruit résiduel $\tilde{B}$



- $\tilde{B}$ plus concentré que B.
- $\tilde{B}$ non gaussien.

noir : densité des probes negatives (recentrée)
bleu : densité de $\tilde{B}$ avec effet probe
rouge : densité de $\tilde{B}$ avec effet probe + array
vert : densité de $\tilde{B}$ avec effet probe + chip

Applications : différence d'expression entre deux groupes d'individus.

Considérons l'intensité d'une probe p pour deux groupes d'individus $\{i_1 = 1, ..., n_1\}$, $\{i_2 = 1, ..., n_2\}$:

$$\begin{aligned} \text{Groupe 1} \quad & X^{(1)}_{i_1,p} = S^{(1)}_{i_1,p} + C_{i_1} + P_p + \tilde{B}^{(1)}_{i_1,p} \\ \text{Groupe 2} \quad & X^{(2)}_{i_2,p} = S^{(2)}_{i_2,p} + C_{i_2} + P_p + \tilde{B}^{(2)}_{i_2,p} \end{aligned}$$

où C_i est un effet chip et P_p un effet probe.

- $f_{\tilde{B}}$ estimée à partir des résidus du modèle linéaire
- C_{i_1} et C_{i_2} estimés à partir des probes négatives (coefficients du modèle linéaire)

$$Y^{(j)}_{i,p} := X^{(j)}_{i,p} - C_i = S^{(j)}_{i,p} + P_p + \tilde{B}^{(j)}_{i,p} \quad j = 1, 2.$$

- Estimation de P_p impossible : on estime

$$\Sigma^{(j)}_{i,p} := S^{(j)}_{i,p} + P_p$$

$\rightarrow$ L'effet probe P_p est identique dans les deux groupes.

Déconvolution parmi les individus d'un groupe.

Pour $j = 1, 2$,

$$Y_{i,p}^{(j)} = \Sigma_{i,p}^{(j)} + \tilde{B}_{i,p}^{(j)}, \quad i = 1, dots, n_j$$

avec $Y_{i,p}^{(j)} = X_{i,p}^{(j)} - C_i$ et $\Sigma_{i,p}^{(j)} = S_{i,p}^{(j)} + P_p$.

- Pour $j = 1, 2$, $\{\Sigma_{i,p}^{(j)}, i = 1, \ldots, n_j\}$ i.i.d. de densité $f_{\Sigma_p^{(j)}}$.
- $\widehat{f}_{\Sigma_p^{(j)}}$ estimé par déconvolution non paramétrique.
- Pour tout $i = 1, \ldots, n_j$, $j = 1, 2$

$$\widehat{\Sigma_{i,p}^{(j)}} = \frac{1}{\widehat{f_{Y_p^{(j)}}}} \int s \widehat{f}_{\Sigma_p^{(j)}}(y - s) \widehat{f_{\tilde{B}}}(s) ds$$

avec
- $\widehat{f_{Y_p^{(j)}}}$ densité de $Y_{i,p}^{(j)}$, $i = 1, \ldots, n_j$.
- $y = Y_{i,p}^{(j)}$

Utilisations

- Comparaison de la distribution des signaux $f_{S_p^{(1)}}$ et $f_{S_p^{(2)}}$ parmi les deux groupes.
  $\rightarrow$ Comparaison des **distributions** $\widehat{f}_{\Sigma_p^{(1)}}$ et $\widehat{f}_{\Sigma_p^{(2)}}$.
- Dans le cas d'un design qui associe un individu du groupe 1 à un individu du groupe 2 (exemple : étude cas-contrôle),

$$S_{i_1,p}^{(1)} - S_{i_2,p}^{(2)} = \Sigma_{i_1,p}^{(1)} - \Sigma_{i_2,p}^{(2)}$$

est estimé par

$$\widehat{\Sigma_{i_1,p}^{(1)}} - \widehat{\Sigma_{i_2,p}^{(2)}}.$$

Conclusion

- Déconvolution : meilleure précision dans l'étude des différences d'intensité faibles.
- Modèle paramétrique : pas de réelle justification.
- Déconvolution non paramétrique.
- Modèle effets + bruits résiduel Illumina
  - Bruit plus concentré.
  - Estimation plus précise.
  - Déconvolution au sein d'un groupe.
- Extension à Affymetrix ?