

# Modélisation des variances pour l'analyse des données d'expression

F. Jaffrézic, G. Marot, J-L. Foulley  
Génétique Statistique  
INRA-SGQA

# Contexte

- **Objectif :**

Déterminer des gènes différentiellement exprimés entre deux ou plusieurs conditions.

**Exemples :**

Individus malades / sains.

Mode de reproduction (IA, FIV, clone).

# Analyse statistique

## Modèle

$$y_{ijk} = m_{ik} + e_{ijk}$$

$y_{ijk}$ : niveau d'expression du gène  $i$  ( $i=1, \dots, N$ ), réplication  $j$  ( $j=1, \dots, n_{ik}$ ) et condition  $k$  ( $k=1, \dots, K$ ).

$m_{ik}$ : effet du gène  $i$  dans la condition  $k$  (effet fixe).

$e_{ijk}$ : résidus supposés indépendants et normalement distribués :

$$e_{ijk} \sim N(0, \sigma_{ik}^2)$$

Variances  $\sigma_{ik}^2$  peuvent varier par gène  $i$  et par condition  $k$ .

# Analyse statistique

## Modélisation des variances $\sigma_{ik}^2$ :

### Une variance commune pour tous les gènes :

=> Statistique de test surestimée pour beaucoup de gènes.

=> Trop de gènes détectés.

=> Fort taux de faux positifs.

### Une variance pour chaque gène :

=> Manque de puissance car pas assez d'information.

=> Trop peu de gènes détectés.

**Nécessité pour une modélisation plus fine.**

# Analyse statistique

## Modélisation des variances

### SAM t-test (Tusher et al., 2001):

On ajoute une constante aux variances gène à gène afin de stabiliser les variances trop petites. Utilisation de méthodes Bayésiennes empiriques pour estimer cette constante.

### VarMixt (Delmar et al., 2005):

Utilisation d'un modèle de mélange sur les distributions des variances pour déterminer des groupes de variances égales.

**VM2** : Affectation stricte ; **VM** : Affectation souple.

## Modèle proposé :

**Modèle mixte structural sur les variances** (effet condition fixe et effet gène aléatoire).

# Présentation du modèle

- **Modèle mixte structural sur les variances** (Foulley et al., 1992):

$$\ln(\sigma_{ik}^2) = \mu_k + \delta_{ik}$$

$\mu_k$ : effet condition (fixe)

$\delta_{ik}$ : effet gène dans la condition k (aléatoire)

Effets  $\delta_{ik}$  supposés indépendants et identiquement distribués :

$$\delta_{ik} \sim N(0, \tau_k^2)$$

=> **Flexibilité** : 1 variance pour chaque gène dans chaque condition.

=> **Modèle parcimonieux** (Seulement 2K paramètres à estimer).

# Modèle de variance

## Estimation des paramètres

Deux approches possibles :

**1) Estimation par les méthodes MCMC** de type Gibbs sampling.

Par exemple, estimation par le logiciel winBUGS.

**2) Estimation par une méthode approchée** en travaillant sur les variances empiriques.

Méthode beaucoup **plus rapide** en temps de calcul.

# Estimation des paramètres

**2) Méthode approchée** : Modèle structural sur le log des variances empiriques.

$$\ln(s_{ik}^2) = \mu_k + \delta_{ik} + \varepsilon_{ik}$$

$$s_{ik}^2 = \frac{1}{n_{ik} - 1} \sum_{j=1}^{n_{ik}} (y_{ijk} - y_{ik.})^2$$

$\mu_k$ : effet condition (fixe)

$\delta_{ik}$ : effet gène (aléatoire)

$$\delta_{ik} \sim N(0, \tau_k^2)$$

$\varepsilon_{ik}$ : erreur d'échantillonnage due à l'estimation des variances vraies  $\sigma_{ik}^2$  par les variances empiriques  $s_{ik}^2$ .

Supposés indépendants et normalement distribués :

$$\varepsilon_{ik} \sim N(0, \omega_{ik}^2)$$

Variance d'échantillonnage  $\omega_{ik}^2$  peut être estimée par  $2/d_{ik}$  (théorie asymptotique).

$d_{ik}$  = degrés de liberté du gène  $i$  dans la condition  $k$  ( $d_{ik} = n_{ik} - 1$ ).

# Estimation des paramètres

Du fait de l'utilisation de lois normales conjuguées :

$$\ln s_{ik}^2 \mid \ln \sigma_{ik}^2 \sim N(\ln \sigma_{ik}^2, \omega_{ik}^2)$$

$$\ln \sigma_{ik}^2 \sim N(\mu_k, \tau_k^2)$$

Le meilleur prédicteur de  $\ln \sigma_{ik}^2$  est :

$$\hat{\ln \sigma_{ik}^2} = \mu_k + \lambda_{ik} (\ln s_{ik}^2 - \mu_k)$$

Où

$$\lambda_{ik} = \tau_k^2 / (\tau_k^2 + \omega_{ik}^2)$$

Est un **facteur de shrinkage** de  $\ln \sigma_{ik}^2$  vers la moyenne  $\mu_k$

# Estimation des paramètres

**Estimation « shrinkée » des paramètres de variances :**

$$\hat{\ln \sigma_{ik}^2} = \mu_k + \lambda_{ik} (\ln s_{ik}^2 - \mu_k)$$

**Facteur de shrinkage:**  $\lambda_{ik} = \tau_k^2 / (\tau_k^2 + \omega_{ik}^2)$

Quand  $\tau_k^2 = 0$  : une variance commune pour tous les gènes égale à  $\mu_k$ .

Quand  $\tau_k^2 \rightarrow +\infty$ , le facteur de shrinkage devient 1.

Une variance estimée pour chaque gène dans chaque condition :

$$\hat{\ln \sigma_{ik}^2} = \ln s_{ik}^2$$

# Modèle de variance

Statistique de test :

$$t_{i,kl} = \frac{m_{ik} - m_{il}}{\sqrt{\hat{\sigma}_{ik}^2 / n_{ik} + \hat{\sigma}_{il}^2 / n_{il}}}$$

Si variances supposées connues => approximation normale.

**Modèle structural** : Statistique de Welch

=> Suit approximativement une loi de Student à  $v_i$  degrés de liberté.

Calcul des degrés de liberté par la **méthode de Satterthwaite**.

# Présentation des données

## Etude de génomique fonctionnelle sur les embryons bovins avant implantation.

Protocole expérimental réalisé par S. Degrelle (2006)

### Trois modes de reproduction :

- Insémination Artificielle (IA).
- Fécondation in vitro (FIV).
- Clonage : 3 lignées.

10 embryons Holstein disponibles pour IA, FIV et chacune des trois lignées de clones.  
(50 embryons au total).

10214 ADNc unique spottés sur des membranes de Nylon N+ pour chaque embryon.

### Normalisation des données

Passage au  $\log_2$  des données d'expression des gènes.

Données centrées par membrane et par gène.

# Estimation des paramètres

## Modèle structural

### Estimation MCMC :

Gibbs sampling par winBUGS (Spiegelhalter et al., 2002)

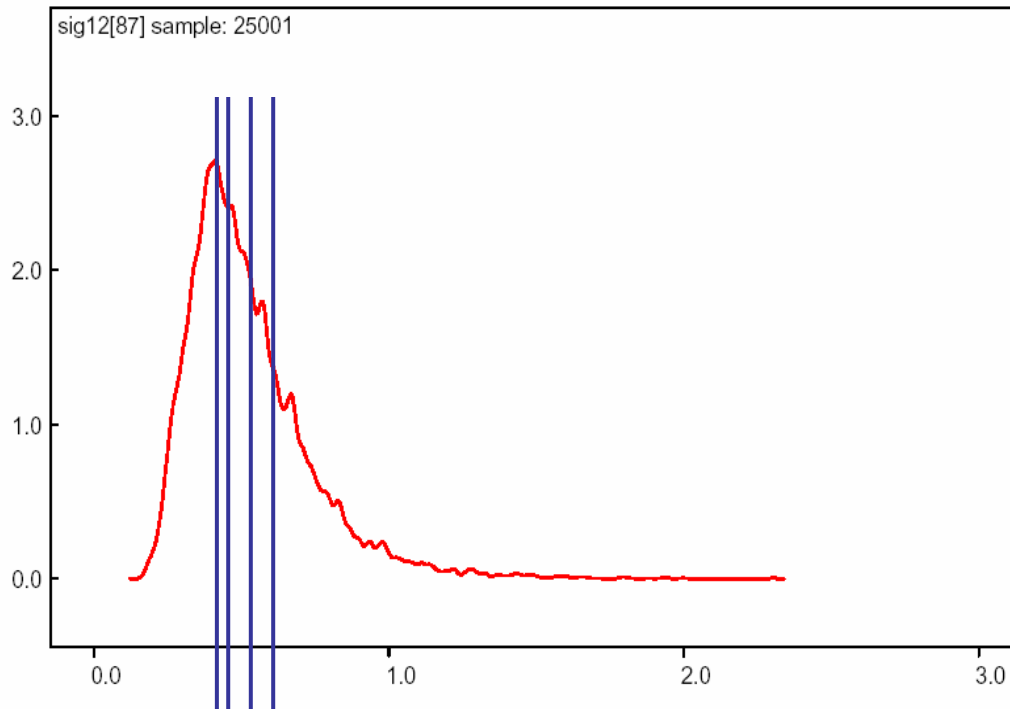
- A priori Uniforme sur l'écart-type  $\tau_k$  (Gelman, 2006) ( $U(0,10)$ )
- A priori non informatif pour  $\mu_k$  (a priori  $N(0,10^3)$ )
- 30000 itérations avec 5000 de burn-in.

=> Grand temps de calcul.

# Estimation des paramètres

Estimation des paramètres  $\sigma_{ik}^2$  : Distribution a posteriori très asymétrique

Mode < Méthode approchée << Médiane



Mode	Approximated	Median	Mean
0.42	0.44	0.48	0.53

# Résultats

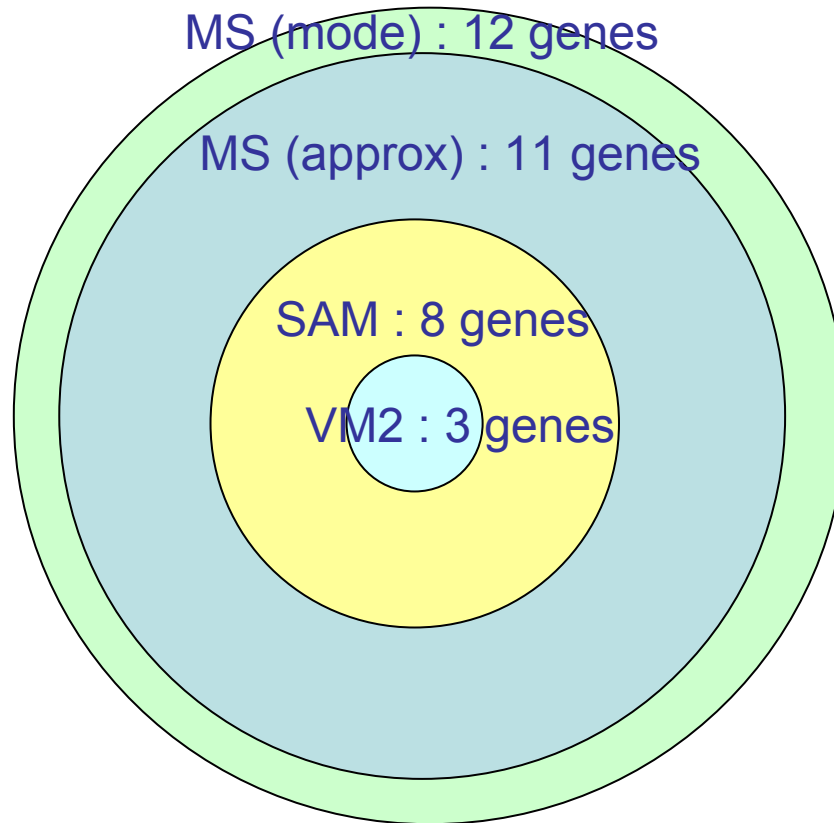
## Gènes trouvés différentiellement exprimés (BH 10%) :

- Avec le modèle à variance homogène : 364 gènes
- Avec le modèle gène à gène : aucun gène
- Avec le modèle structural (méthode approchée): 11 gènes

Gene	Cond1	Cond2	Cond2-Cond1	Adj.p-value
AW461513	3	5	-	0.07902557
BF041564	3	5	-	0.07902557
CN433942	4	5	+	0.07902557
CR551977	4	5	+	0.07902557
CR553385	4	5	+	0.07902557
AW461689	3	5	-	0.08822997
CR552847	4	5	+	0.08822997
CN434085	4	5	+	0.08822997
CN435027	4	5	+	0.08822997
BF046406	1	4	-	0.08963992
CN432421	4	5	+	0.08963992

# Résultats

**Gènes détectés différemment exprimés.**  
Benjamini Hochberg à 10 %



**MS (médiane) :**  
**aucun gène**

# Degrés de liberté

## Calcul des degrés de liberté :

Liste de gènes trouvés avec le modèle structural

**Méthode de Satterthwaite** : 11 gènes.

Nombre de degrés de liberté estimés en moyenne = 24.5  
(variant entre 23 et 27).

**Approximation normale** : 246 gènes.

# Simulations

## 100 simulations pour chaque jeu de données

### Jeux de données

- ✓ **Spleendata** (package Varmixt dans R)  
données du CEA  
comparaison de souris irradiées in vivo avec des souris normales  
4300 gènes  
→ données appariées
- ✓ **Extrait du jeu de données précédent:**  
seulement les conditions 1,2 et 5 (IA, FIV, 3ème lignée de clones)  
5000 gènes: 100 premiers gènes trouvés par MS,  
4900 aléatoires parmi les autres.  
→ données non appariées

# Simulations

## Spleendata

Même méthode de simulation que dans l'article de Delmar et al. (2005)

Chaque gène suit une loi normale

- ✓ 43 gènes simulés différentiellement exprimés avec un ratio moyen variant uniformément entre 1.2 et 2.5.
- ✓ 4317 gènes simulés avec une log intensité égale dans chaque condition à la moyenne des deux conditions
- ✓ Variances distribuées aléatoirement parmi tous les gènes.
- ✓ BH 5%

# Simulations

Number of replicates*	4	6	8	12
<b>Number of true positives</b>				
Structural Model	27.7 (2.9)	35.2 (1.8)	38.0 (1.8)	40.5 (1.6)
VM	27.9 (2.8)	34.6 (1.7)	37.5 (2.0)	40.1 (1.7)
VM2	26.2 (2.9)	33.3 (2.0)	36.8 (2.0)	39.9 (1.7)
SAM	25.6 (2.8)	35.0 (1.8)	37.6 (2.0)	40.4 (1.6)
Gene specific	18.9 (3.8)	31.9 (2.3)	36.1 (2.3)	39.6 (1.7)
Homoskedastic	34.0 (1.9)	36.7 (1.3)	39.9 (1.3)	41.5 (1.0)
VMpaired	22.3 (2.8)	32.3 (2.0)	35.9 (2.0)	39.5 (1.7)
VM2paired	20.7 (3.2)	31.0 (2.2)	35.0 (2.2)	39.1 (1.9)
SAMpaired	0.0 (0.0)	29.4 (2.2)	35.5 (2.2)	39.5 (1.8)
<b>Number of false positives</b>				
Structural Model	2.4 (1.5)	2.9 (2.0)	3.1 (1.8)	3.1 (2.0)
VM	1.8 (1.4)	2.2 (1.6)	2.5 (1.5)	2.8 (1.8)
VM2	1.6 (1.3)	2.0 (1.5)	2.3 (1.5)	2.6 (1.9)
SAM	0.8 (0.9)	2.2 (1.7)	2.7 (1.6)	3.0 (1.9)
Gene specific	1.2 (1.1)	2.5 (1.6)	2.5 (1.6)	2.6 (1.6)
Homoskedastic	40.7 (7.8)	41.5 (7.8)	41.6 (7.3)	42.6 (8.1)
VMpaired	1.5 (1.3)	2.0 (1.3)	2.1 (1.5)	2.6 (1.6)
VM2paired	1.6 (1.4)	2.0 (1.4)	2.3 (1.7)	2.4 (1.7)
SAMpaired	0.0 (0.0)	0.6 (0.8)	1.4 (1.3)	2.2 (1.5)

\*Replicates correspond to the number of measurements for each gene within each condition

## Résultats

- Plus de 6 réplifications:
  - ✓ SAM, VM et MS aussi bons
  - ✓ VH mauvais
- 4 réplifications :
  - ✓ SAM apparié mauvais
- Modèles non appariés meilleurs que les modèles appariés

# Simulations

## Repro125

Chaque réplication de gène est simulée suivant une loi normale.

- ✓ 100 gènes simulés différentiellement exprimés avec une différence de log intensité simulée avec une loi gamma (10.8,0.07).
- ✓ 4900 gènes simulés avec une log intensité égale dans chaque condition à la moyenne des trois conditions
- ✓ Variances distribuées aléatoirement parmi les gènes différentiellement exprimés
- ✓ Modèles de variance → récupération des p-values brutes
- ✓ Ajustement BH 10%

# Simulations

Number of replicates*	5	10
<b>Number of true positives</b>		
Structural Model	24.6 (6.2)	62.4 (3.6)
VM	15.6 (6.1)	57.5 (3.8)
VM2	10.3 (5.7)	55.1 (4.1)
SAM	16.5 (5.6)	60.7 (3.8)
Gene specific	7.3 (5.0)	51.8 (4.0)
Homoskedastic	42.7 (4.5)	69.6 (3.9)
<b>Number of false positives</b>		
Structural Model	8.2 (4.6)	15.3 (4.8)
VM	3.0 (2.4)	11.0 (3.7)
VM2	1.9 (2.1)	9.5 (3.6)
SAM	3.3 (2.6)	12.7 (4.0)
Gene specific	1.7 (1.8)	10.7 (4.3)
Homoskedastic	93.3 (12.4)	103.2 (11.7)

\*Replicates correspond to the number of measurements for each gene within each condition

**Plus de puissance avec le modèle structural surtout dans le cas de faibles réplifications**

# Discussion

## **Simulations : comparaison de méthodes.**

**Grande sensibilité des listes de gènes** trouvés suivant le modèle de variance choisi (même quand peu de différences dans l'estimation des paramètres).

**Modèle à variance homogène** très mauvais. Beaucoup trop de faux positifs.

**Modèle gène à gène** : pas assez de puissance. Trop peu de gènes détectés.

**Modèle structural** : se comporte bien comparé aux autres modèles surtout quand comparaisons entre plus de deux conditions.

**Problème avec le SAM apparié** quand peu de répliquions.

**=> Modélisation des variances = point clef de l'analyse différentielle.**

# Perspectives

## **Modèle structural :**

=> Possibilité d'inclure **d'autres facteurs** (sexe, âge, fonctions du temps, etc.)  
Problème dans le package nlme de R pour fixer la variance résiduelle

## **=> Méthode d'estimation exacte :**

**Estimation des paramètres par le SAEM** (cf. Thèse Mylène Duval).

## **Hypothèse couramment faite : gènes supposés indépendants.**

=> Prise en compte des **corrélations** entre les gènes

=> Prise en compte des corrélations quand mesures sur les gènes faites au cours du temps.