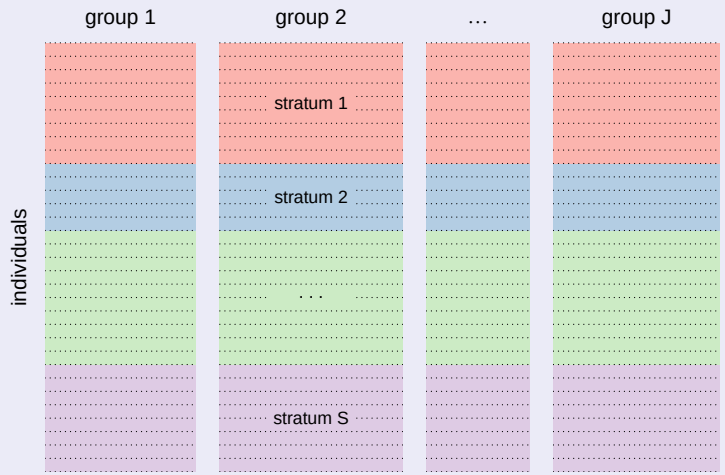
Gestion des individus manquantes en l'intégration de données omiques: imputation multiple dans le cadre de l'AFM

Valentin Voillet, Laurence Liaubet, Philippe Besse, Magali San Cristobal, Ignacio González

Toulouse, 03 décembre 2015

Intégration de données “omiques”

- étude des tableaux dans lesquels un même ensemble d’individus est décrit par plusieurs groupes de variables

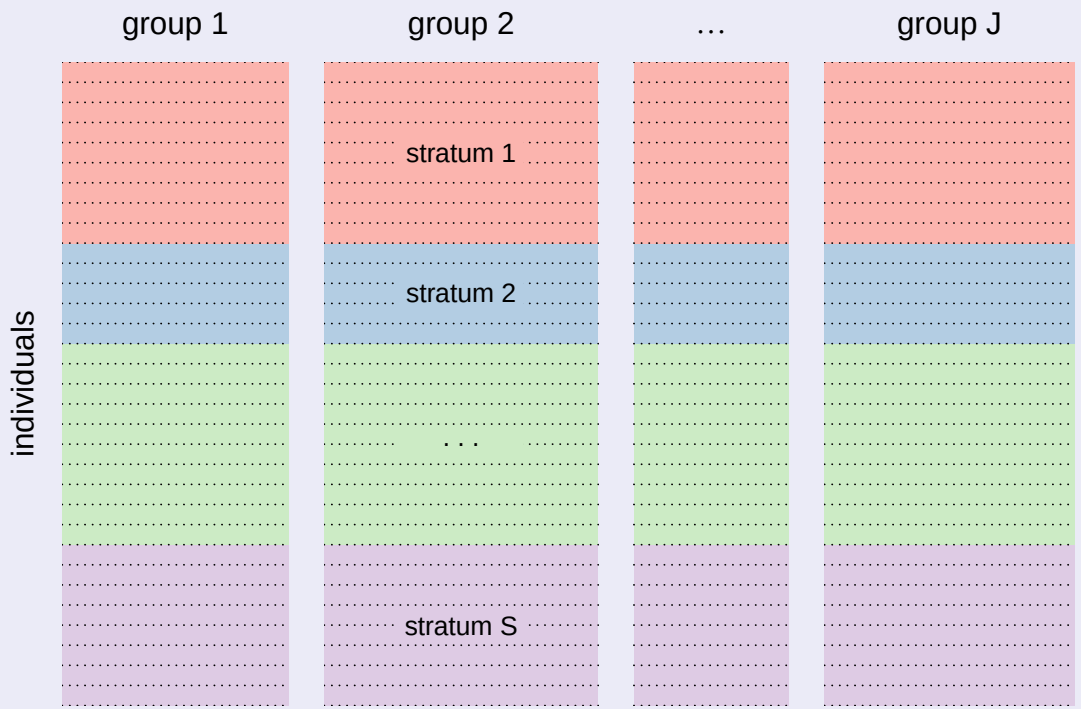


Le but :

- étudier les ressemblances entre des individus en termes de toutes les variables

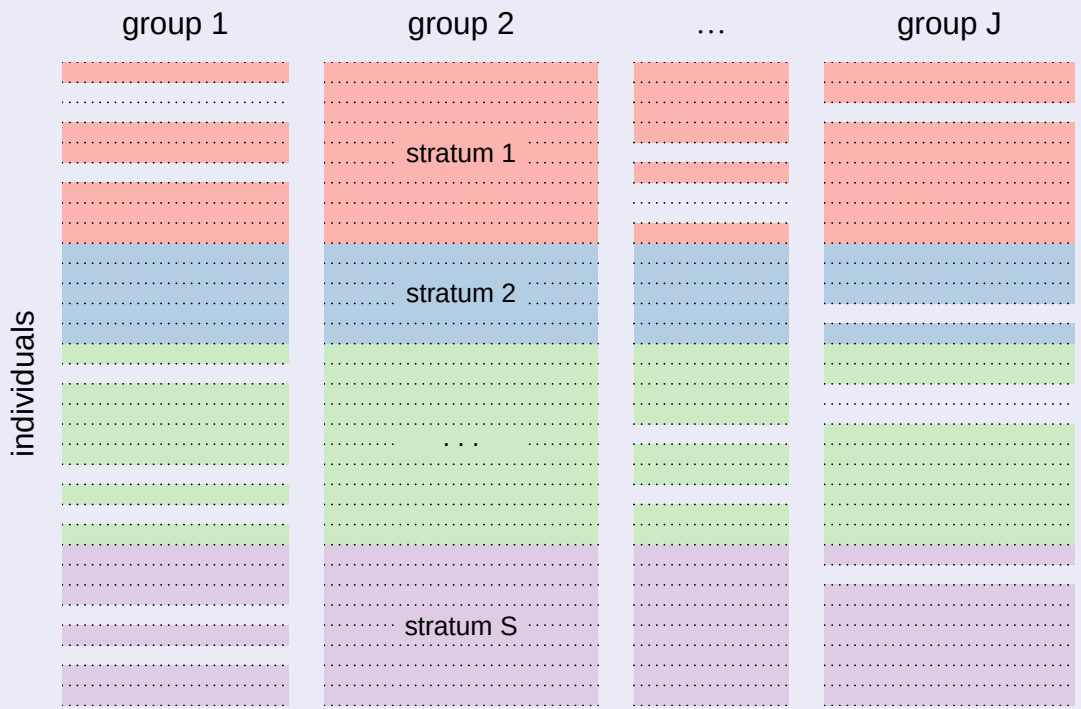
Problème

- Souvent, on est confronté au problème d'individus manquantes dans un ou plusieurs tableaux



Problème

- Souvent, on est confronté au problème d'individus manquantes dans un ou plusieurs tableaux



Motivation

Problématique des individus manquantes

- la plupart des méthodes statistiques ne peuvent pas être directement appliquées
- dans le cadre de l'analyse de multiple tableaux, peu d'approches sont proposées
- méthodes couramment disponibles pour analyser des données incomplètes :
  - suppression des individus ayant des valeurs manquantes
  - imputation simple, c'est-à-dire les compléter par des valeurs plausibles (par exemple les moyens des cas observés)

Motivation

Problèmes des méthodes courantes

- une grande proportion d'individus peut être supprimé
- imputation simple peut déformer les distributions et les relations entre variables
- ces méthodes souffrent de ne pas prendre en compte l'incertitude due aux données manquantes

Méthodes alternatives

- analyse canonique de corrélations généralisée
- AFM itérative régularisée (RI-MFA)
- imputation multiple en AFM (MI-MFA)

Motivation

Problèmes des méthodes courantes

- une grande proportion d'individus peut être supprimé
- imputation simple peut déformer les distributions et les relations entre variables
- ces méthodes souffrent de ne pas prendre en compte l'incertitude due aux données manquantes

Méthodes alternatives

- analyse canonique de corrélations généralisée
- AFM itérative régularisée (RI-MFA)
- imputation multiple en AFM (MI-MFA)

Motivation

Problèmes des méthodes courantes

- une grande proportion d'individus peut être supprimé
- imputation simple peut déformer les distributions et les relations entre variables
- ces méthodes souffrent de ne pas prendre en compte l'incertitude due aux données manquantes

Méthodes alternatives

- analyse canonique de corrélations généralisée
- AFM itérative régularisée (RI-MFA)
- imputation multiple en AFM (MI-MFA)

Motivation

Problèmes des méthodes courantes

- une grande proportion d'individus peut être supprimé
- imputation simple peut déformer les distributions et les relations entre variables
- ces méthodes souffrent de ne pas prendre en compte l'incertitude due aux données manquantes

Méthodes alternatives

- analyse canonique de corrélations généralisée
- AFM itérative régularisée (RI-MFA)
- imputation multiple en AFM (MI-MFA)

La méthode MI-MFA

Objectif de la MI-MFA

- estimer les coordonnées des individus sur les premiers composants MFA en présence d'individus manquantes dans un ou plusieurs tableaux
- prendre en compte l'incertitude induite par les individus manquantes sur les composants MFA

La méthode MI-MFA

Objectif de la MI-MFA

- estimer les coordonnées des individus sur les premiers composants MFA en présence d'individus manquantes dans un ou plusieurs tableaux
- prendre en compte l'incertitude induite par les individus manquantes sur les composants MFA

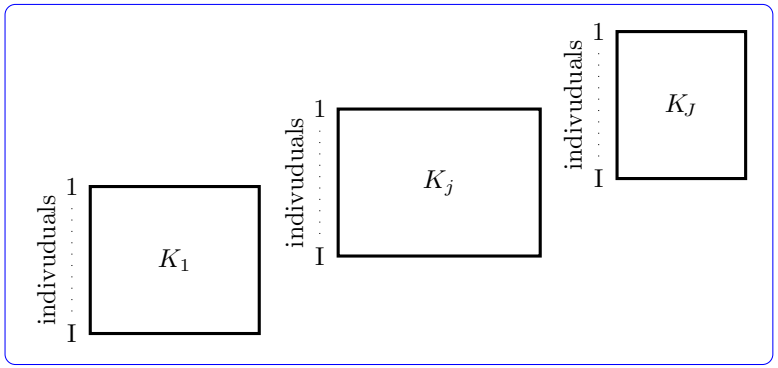
La méthode MI-MFA

Objectif de la MI-MFA

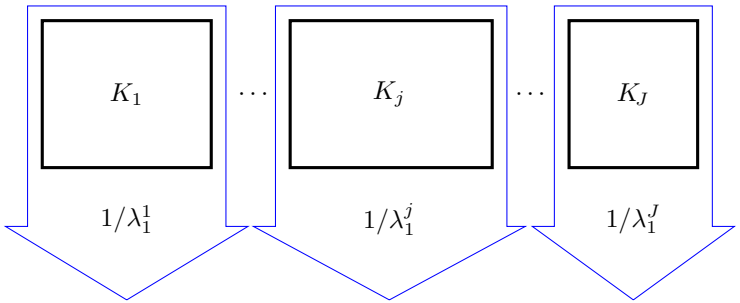
- estimer les coordonnées des individus sur les premiers composants MFA en présence d'individus manquantes dans un ou plusieurs tableaux
- prendre en compte l'incertitude induite par les individus manquantes sur les composants MFA

L'Analyse Factorielle Multiple (MFA)

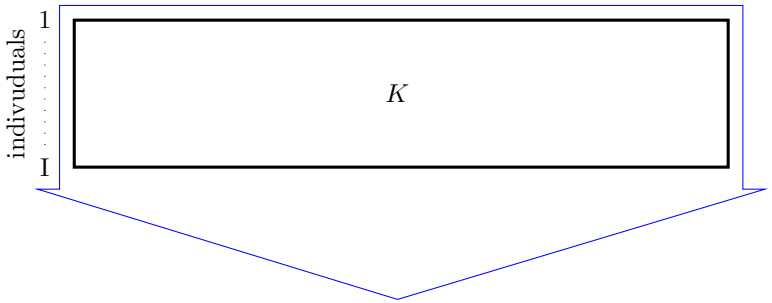
(i) J data tables



(ii) Separate PCAs and weighting



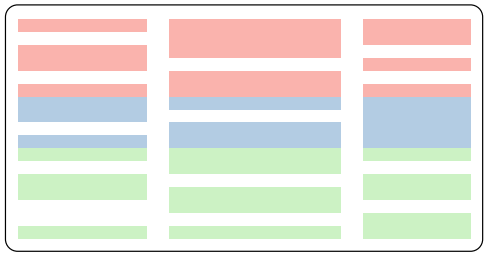
(iii) Merged and weighted matrix K



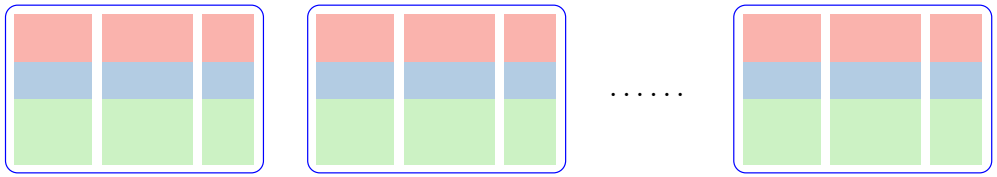
Global PCA

La méthode MI-MFA

Data with missing rows

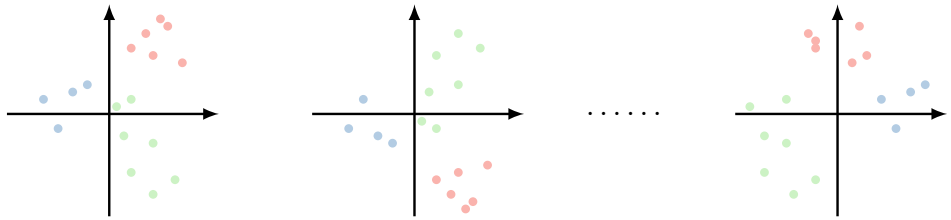


m imputed datasets



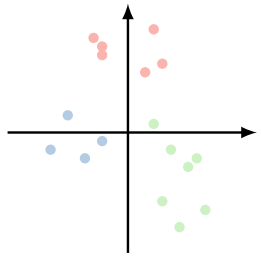
Analyse each dataset separately

MFA



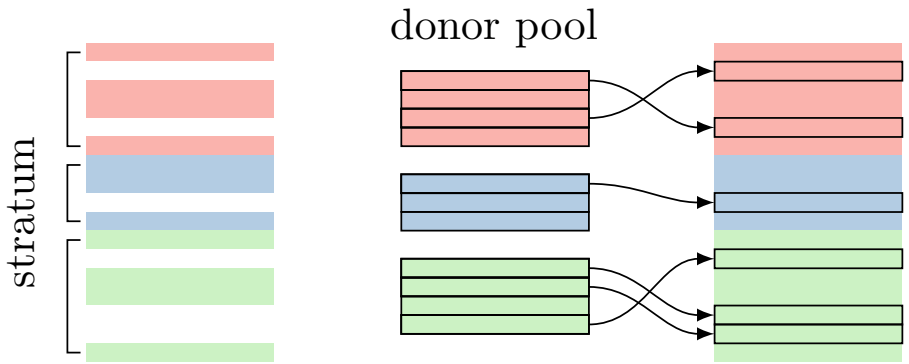
Combine m configurations into a compromise configuration

STATIS



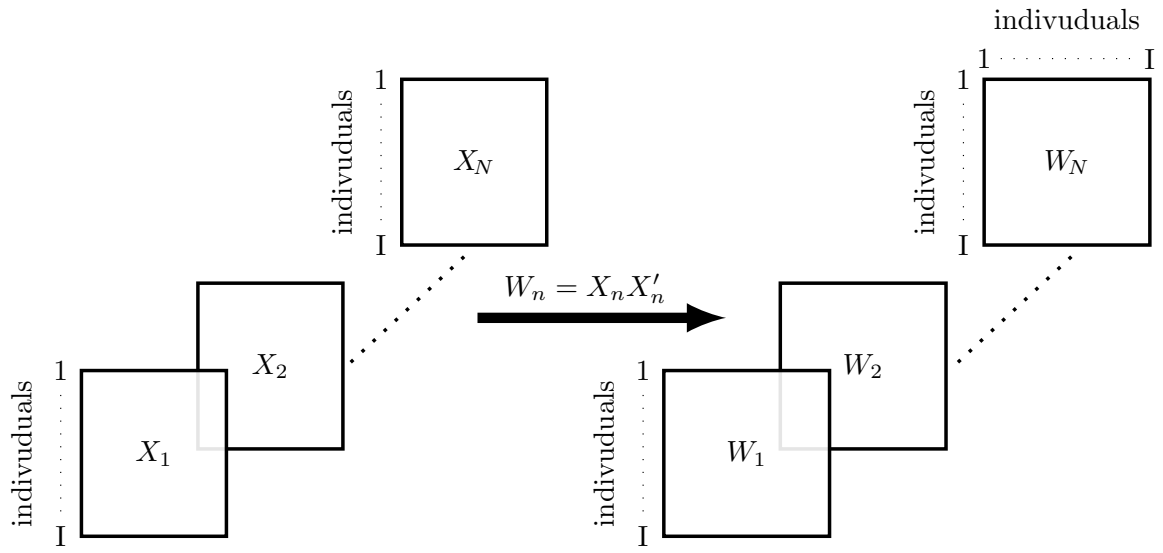
Comment imputer les individus manquantes ?

La méthode *hot-deck*

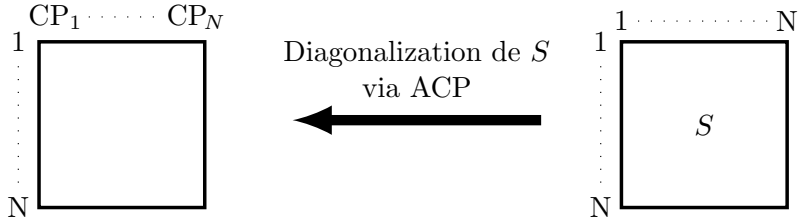


- il n'est pas nécessaire de définir un modèle pour la distribution des valeurs manquantes
- prend en compte les ressemblances entre individus et les liaisons entre variables
- peut être appliqué aux données de grands dimensions

La méthode STATIS (Structuration à Trois Indices de la Statistique)



$S_{jk} = \text{dist}(W_j, W_k)$



$\text{CP}_1 = [\alpha_1 \cdots \alpha_N]'$

Compromise space $W_{co} = \sum_{n=1}^{N} \alpha_n W_n$

$$S_{jk} = \frac{\text{trace}(W_j' W_k)}{\sqrt{\text{trace}(W_j' W_j) \cdot \text{trace}(W_k' W_k)}}$$

Les données *liver toxicity*

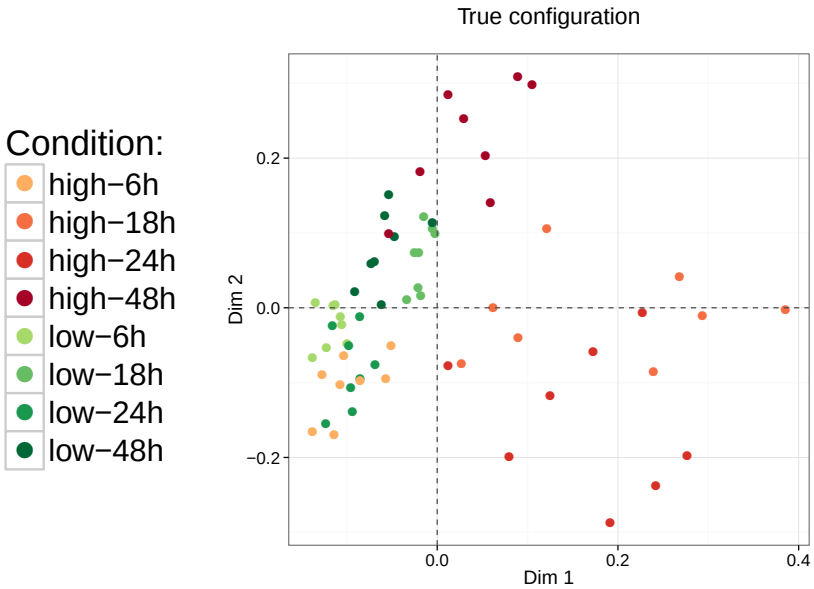
Elles proviennent d'une étude de toxicité du foie chez la souris

Les souris ont été exposés à doses de paracétamol dans une expérience contrôlée

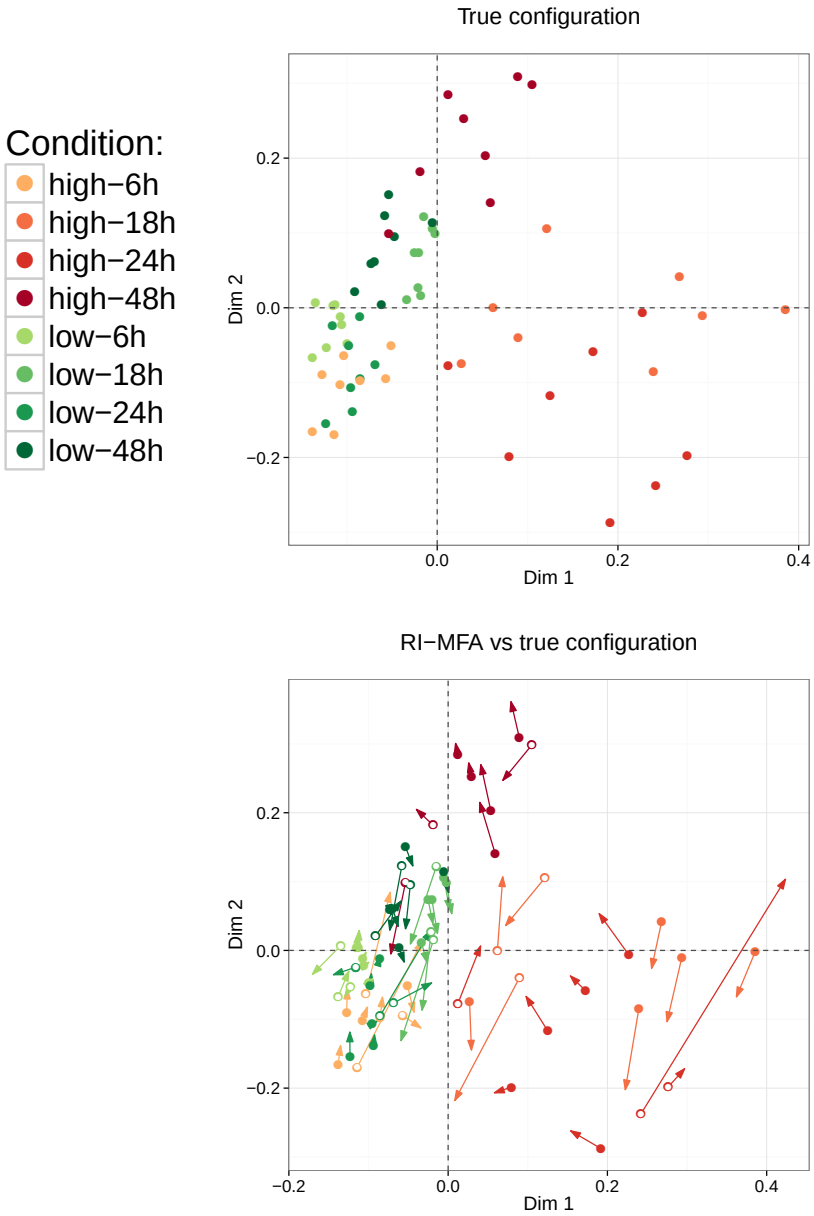
Pour 64 souris, nous disposons :

- des données d'expression de 3116 gènes
- des mesures de 10 variables cliniques contenant marqueurs de lésions du foie
- les 64 souris ont été croisées selon deux facteurs dans un plan à 8 répétitions :
  - la dose : fort (1500 mg/kg ou 2000 mg/kg) et faible (50 mg/kg ou 150 mg/kg)
  - le temps de nécropsie : à 6, 18, 24 et 48 heures

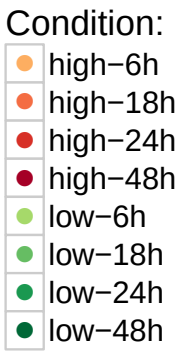
Les données *liver toxicity*

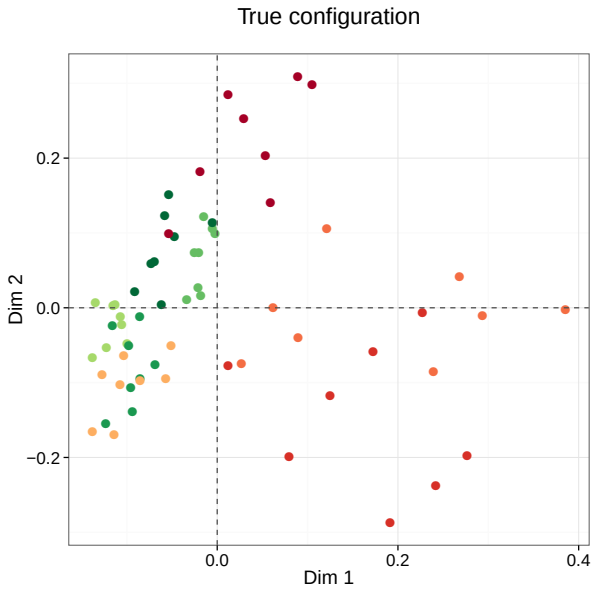


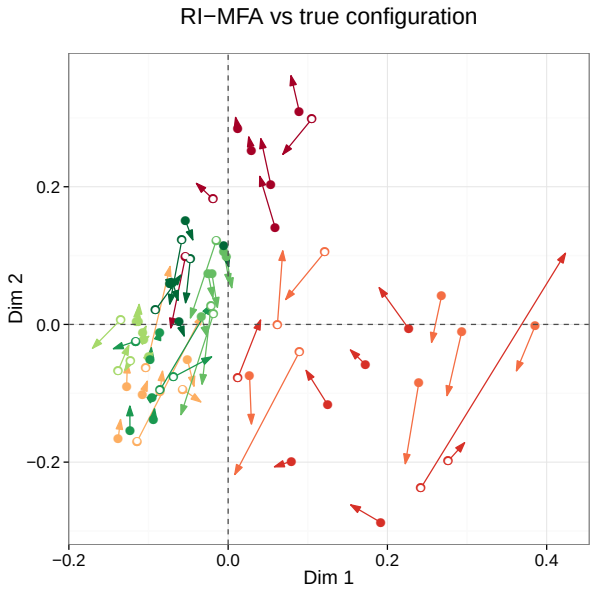
Les données *liver toxicity*

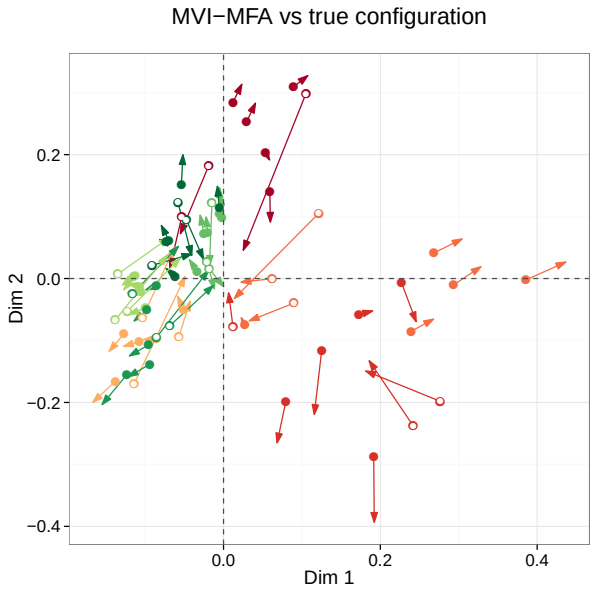


Les données *liver toxicity*

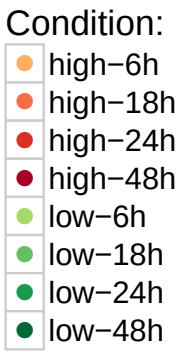


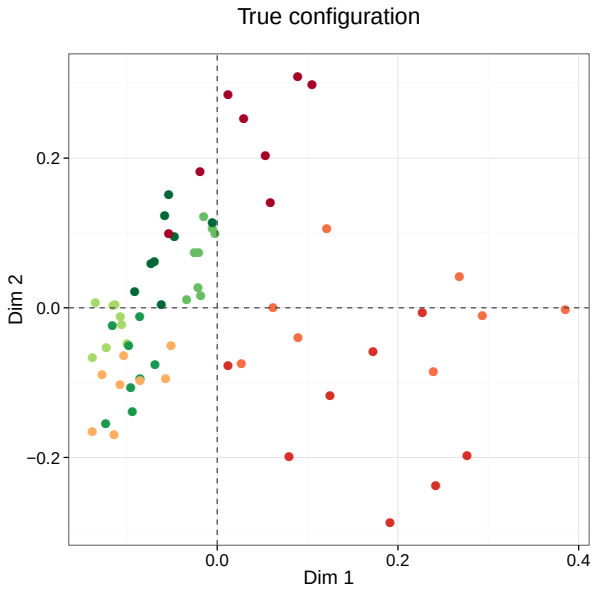


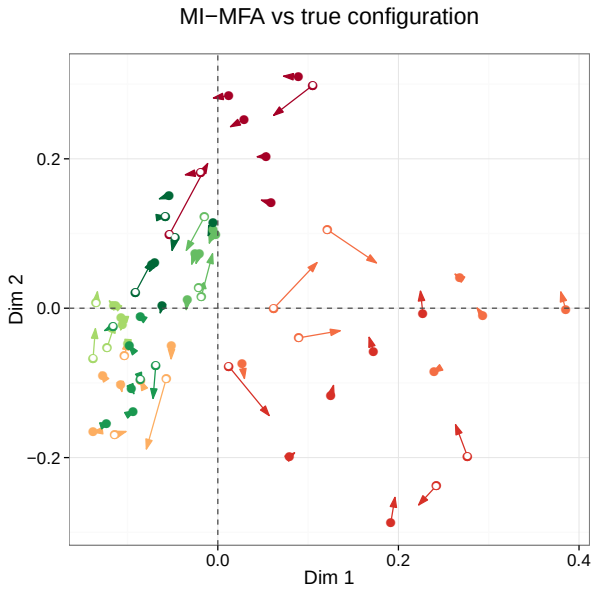


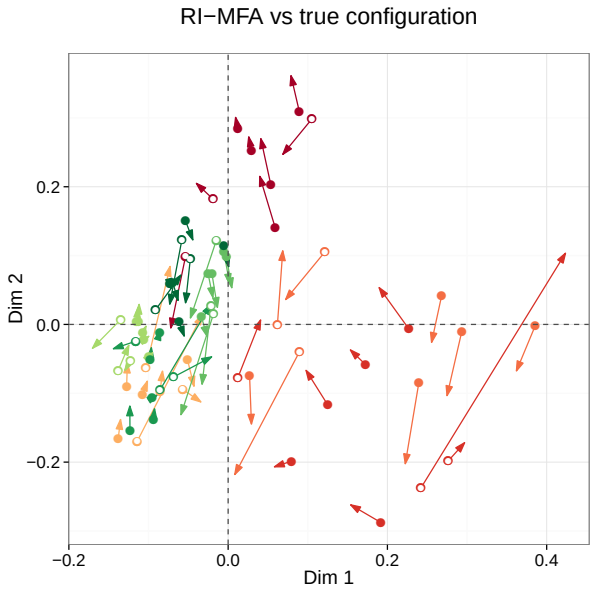


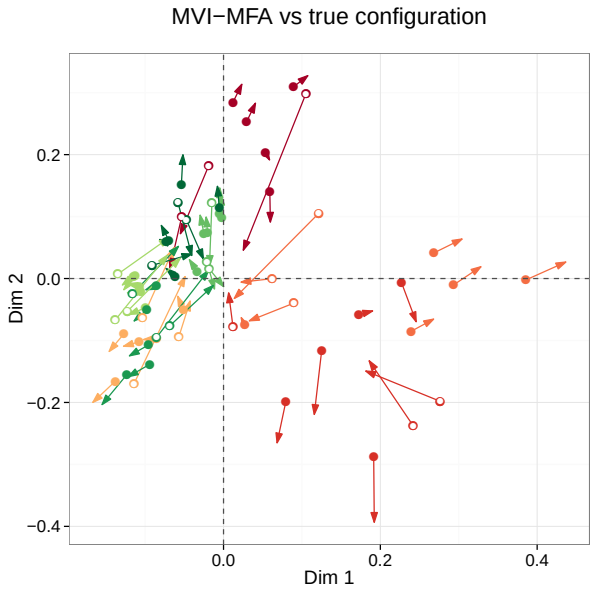
Les données *liver toxicity*











Les données liverTox

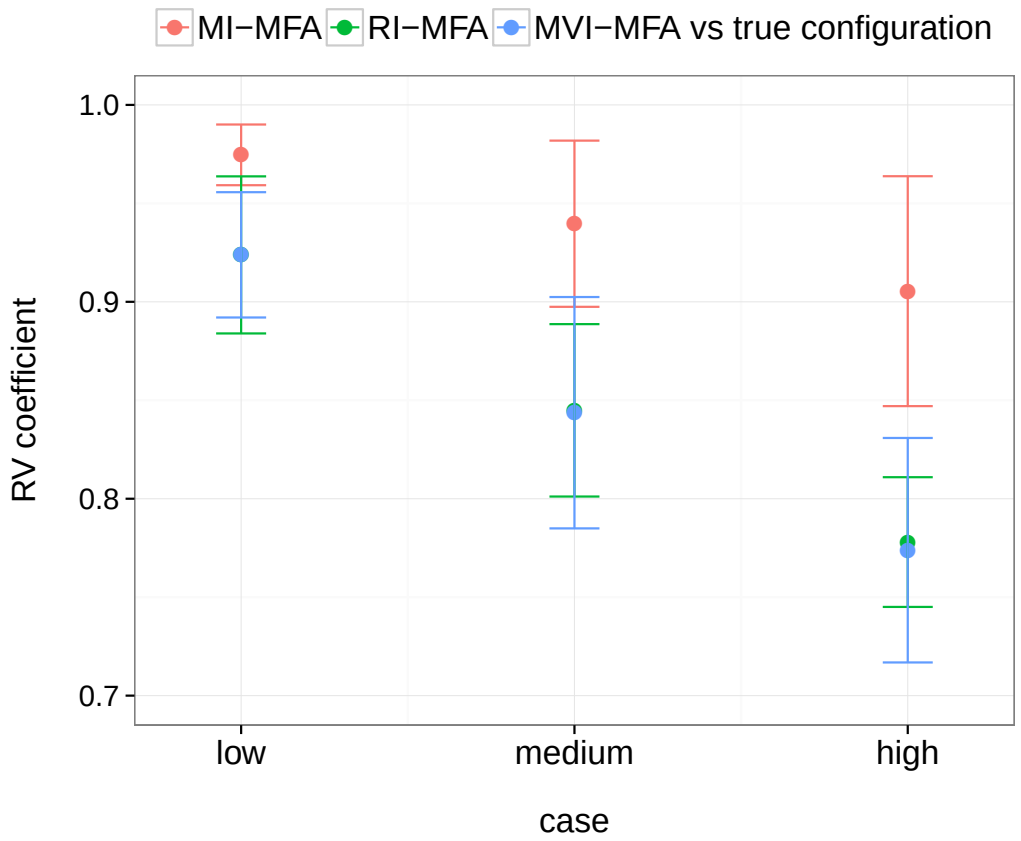
Le nombre d'individus supprimés dans chaque condition sur les données transcriptomiques a été :

$$\text{low} = 1 \text{ individu} \quad \text{medium} = 2 \text{ individus} \quad \text{high} = 3 \text{ individus}$$

Les données liverTox

Le nombre d'individus supprimés dans chaque condition sur les données transcriptomiques a été :

$$\text{low} = 1 \text{ individu} \quad \text{medium} = 2 \text{ individus} \quad \text{high} = 3 \text{ individus}$$



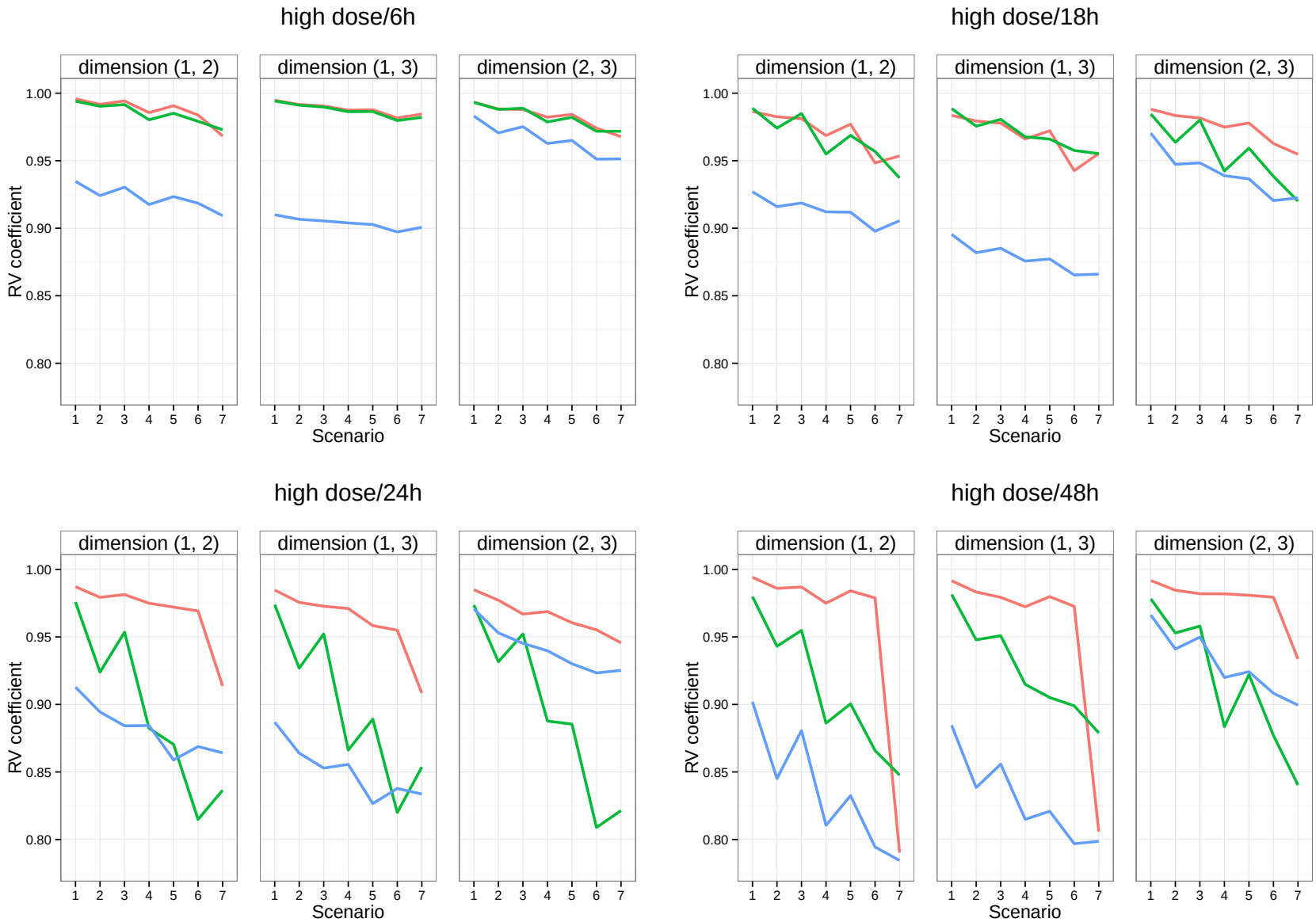
Moyenne des coefficients RV sur 50 configurations 2-dimensionnelles

Les données *liver toxicity*

Différentes scénarios avec des individus manquantes sur les données transcriptomiques et cliniques

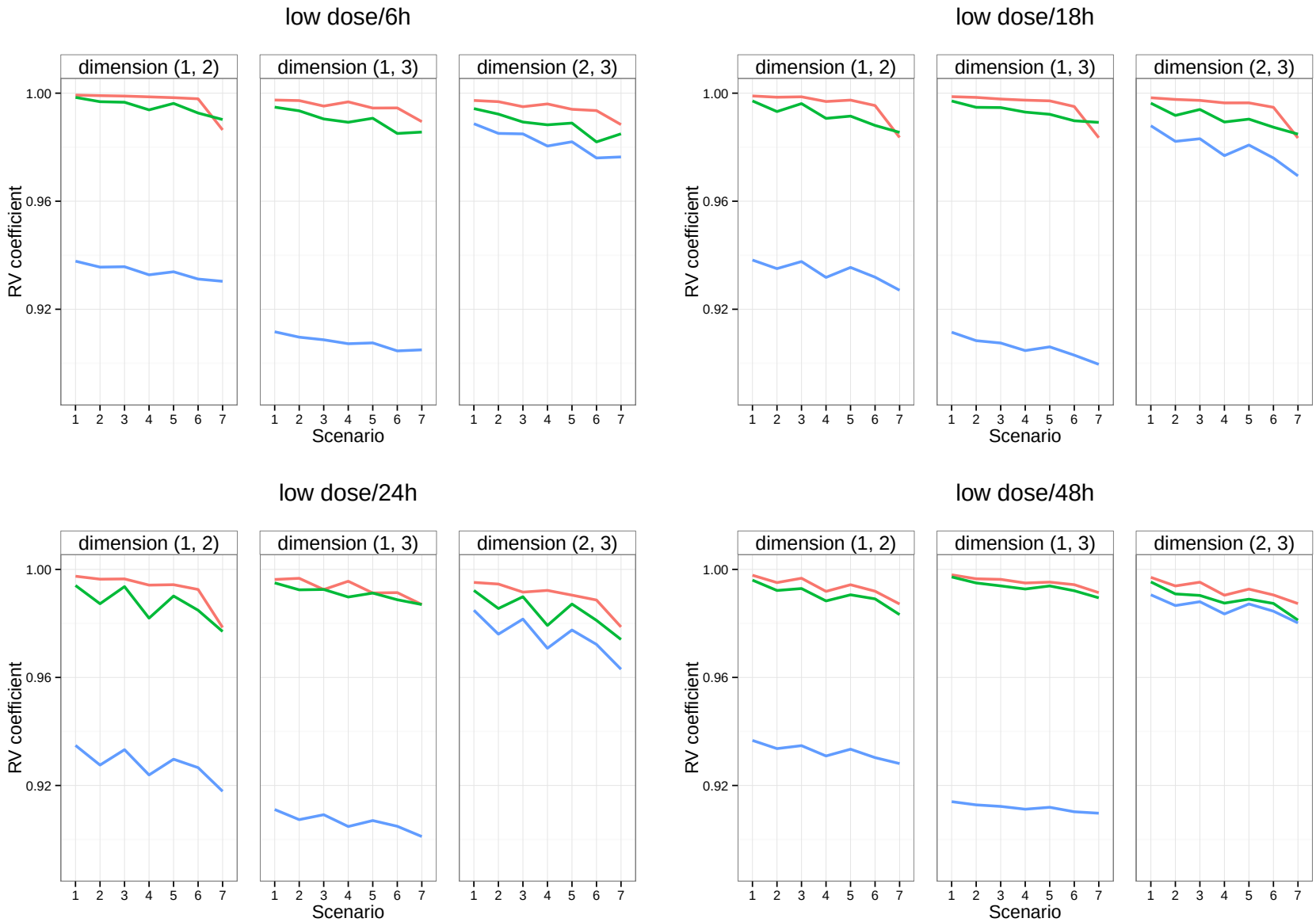
	Individus manquantes		
Scénario	Transcriptomique	Clinique	# cas
1	1	1	56
2	2	1	168
3	1	2	168
4	3	1	280
5	2	2	420
6	3	2	560
7	4	1	280

Les données liverTox



Moyenne des coefficients RV sur 50 configurations

Les données liverTox



Moyenne des coefficients RV sur 50 configurations

Les données *NCI-60*

NCI-60 contiennent données transcriptomiques et protéomiques d'une collection de 60 lignes cellulaires cancéreuses :

> côlon (7 lignes), rénal (8), ovarien (6), sein (8), prostate (2), poumon (9), système nerveux central (6), leucémies (6) et mélanomes (8)

Pour les 60 lignes cellulaires, nous disposons :

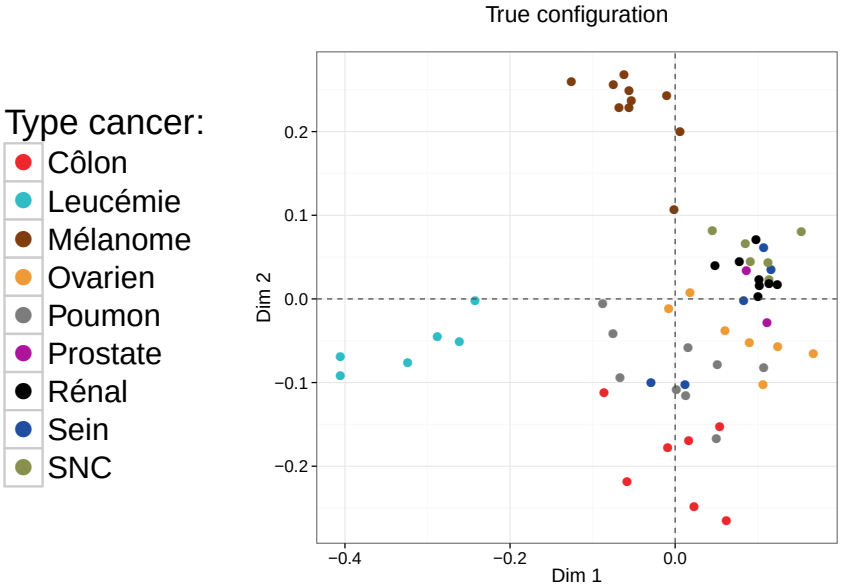
- des données d'expression de 1433 gènes
- des données protéomiques de 162 protéines

Les données *liver toxicity*

Différentes scénarios avec des individus manquantes sur les données transcriptomiques et protéomiques

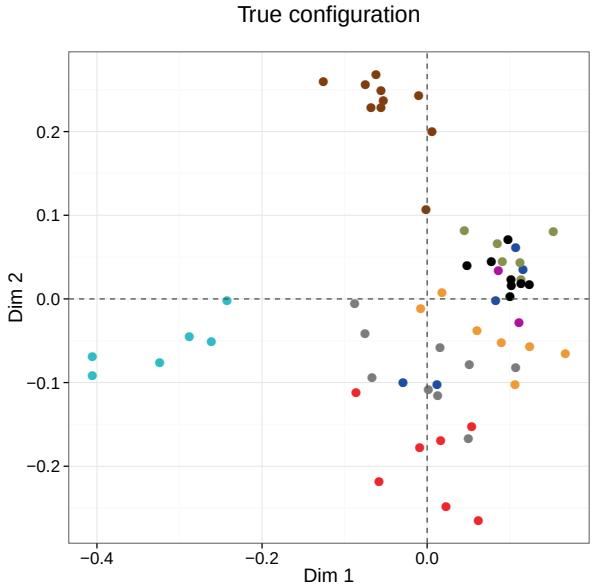
	Individus manquantes		
Ligne cellulaire	Transcriptomique	Protéomique	# cas
Sein	1	0	5
SNC	1	1	60
Côlon	2	0	42
Poumon	2	2	3024
Leucémie	1	1	60
Mélanome	2	2	5040
Ovarien	1	1	84
Prostate	0	0	//
Rénal	2	1	504

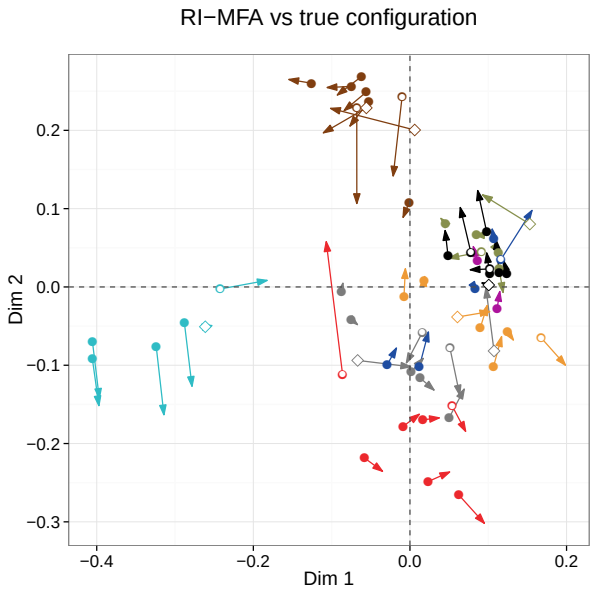
Les données *NCI-60*



Les données *NCI-60*

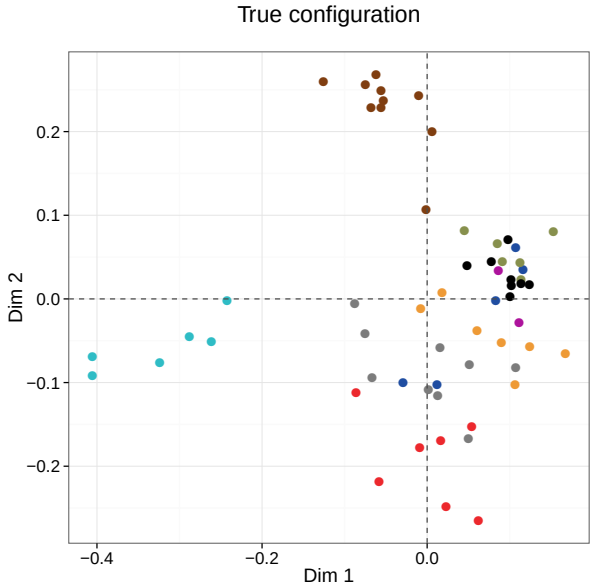
Type cancer:
- Côlon
- Leucémie
- Mélanome
- Ovarien
- Poumon
- Prostate
- Rénal
- Sein
- SNC

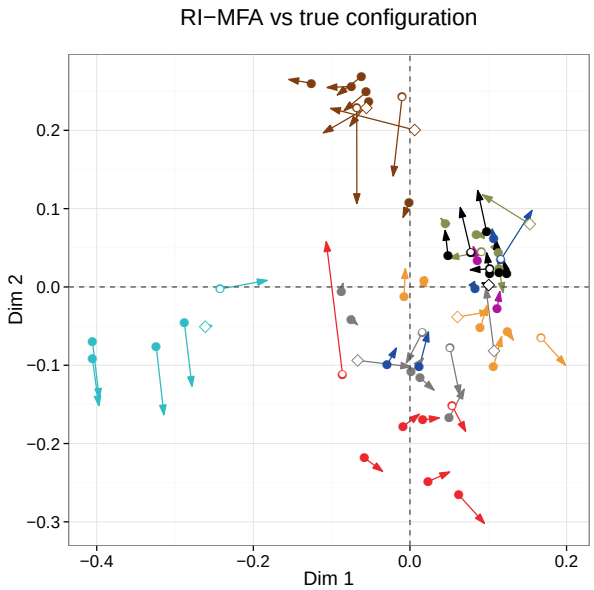


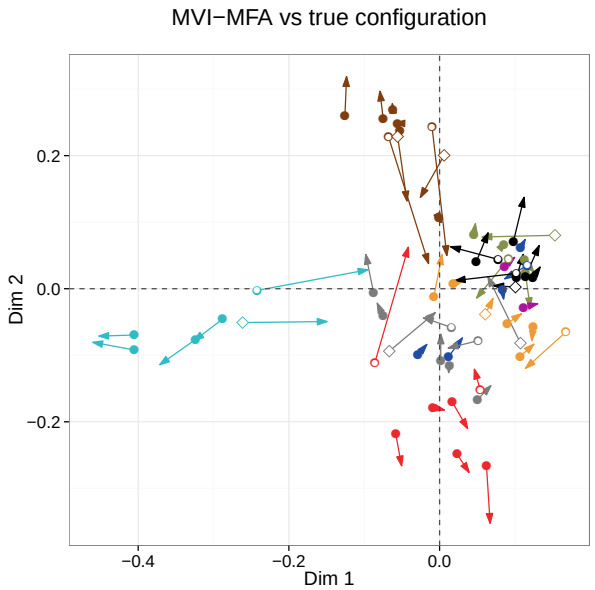


Les données *NCI-60*

Type cancer:
- Côlon
- Leucémie
- Mélanome
- Ovarien
- Poumon
- Prostate
- Rénal
- Sein
- SNC

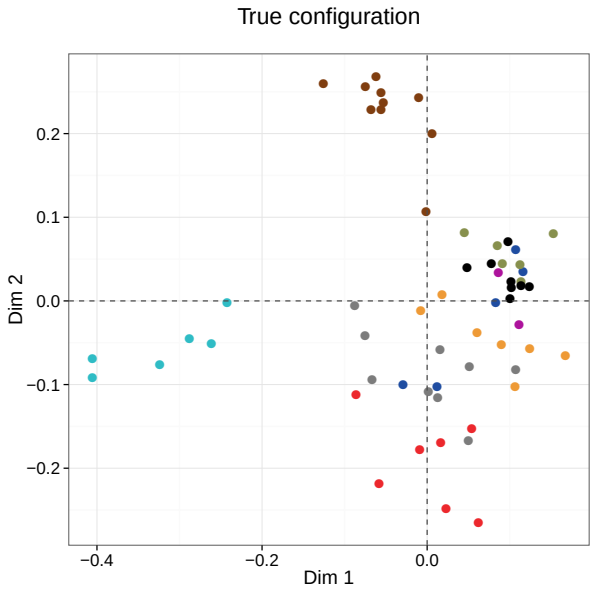


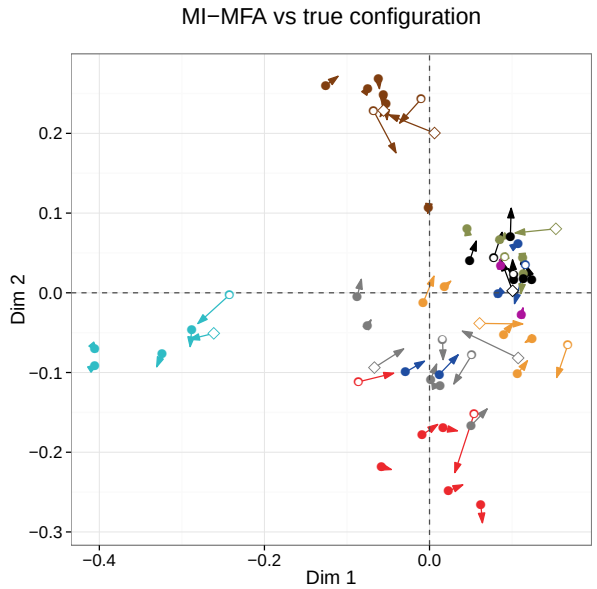


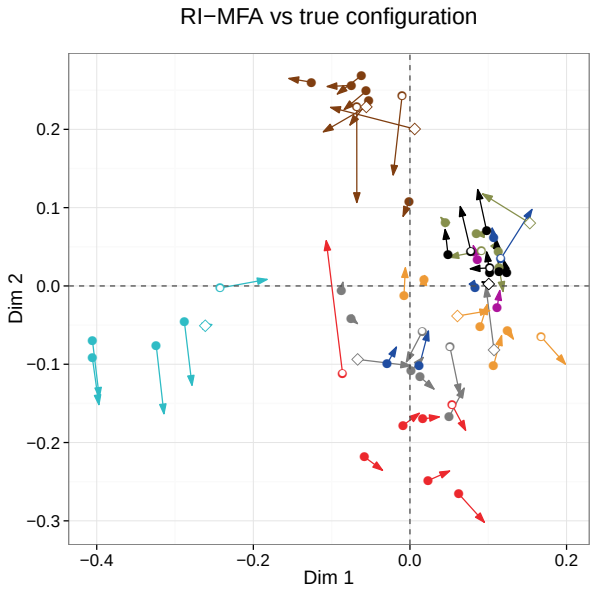


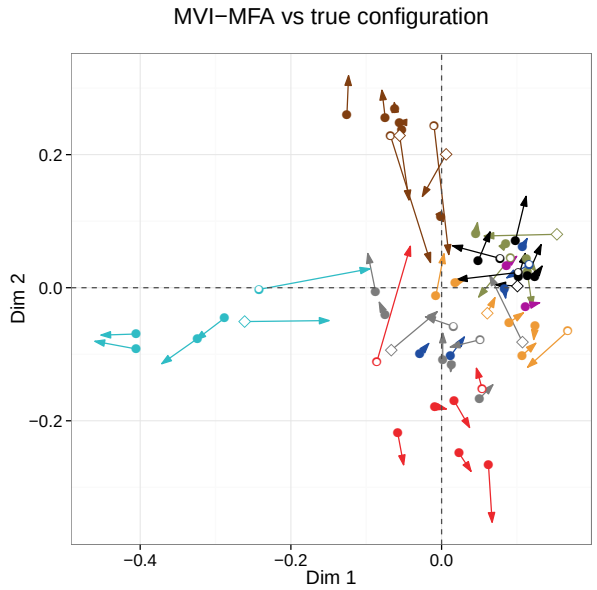
Les données *NCI-60*

Type cancer:
- Côlon
- Leucémie
- Mélanome
- Ovarien
- Poumon
- Prostate
- Rénal
- Sein
- SNC









Pourquoi est-il essentiel l'incertitude ?

Il est toujours possible d'obtenir une configuration d'individus de RI-MFA (ou MVI-MFA) en présence de individus manquantes

Mais, avec seulement cette configuration

- il est impossible savoir si les résultats obtenus sont plausibles
- et si l'utilisateur peut interpréter les résultats correctement
- une configuration unique ne peut pas refléter la variabilité de la estimation

Il est crucial d'avoir un outil permettant d'évaluer la fiabilité des résultats

Pourquoi est-il essentiel l'incertitude ?

Il est toujours possible d'obtenir une configuration d'individus de RI-MFA (ou MVI-MFA) en présence de individus manquantes

Mais, avec seulement cette configuration

- il est impossible savoir si les résultats obtenus sont plausibles
- et si l'utilisateur peut interpréter les résultats correctement
- une configuration unique ne peut pas refléter la variabilité de la estimation

Il est crucial d'avoir un outil permettant d'évaluer la fiabilité des résultats

Pourquoi est-il essentiel l'incertitude ?

Il est toujours possible d'obtenir une configuration d'individus de RI-MFA (ou MVI-MFA) en présence de individus manquantes

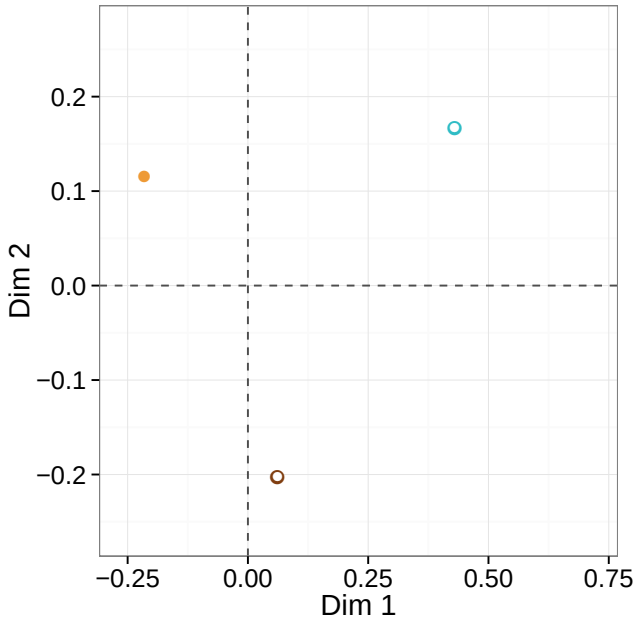
Mais, avec seulement cette configuration

- il est impossible savoir si les résultats obtenus sont plausibles
- et si l'utilisateur peut interpréter les résultats correctement
- une configuration unique ne peut pas refléter la variabilité de la estimation

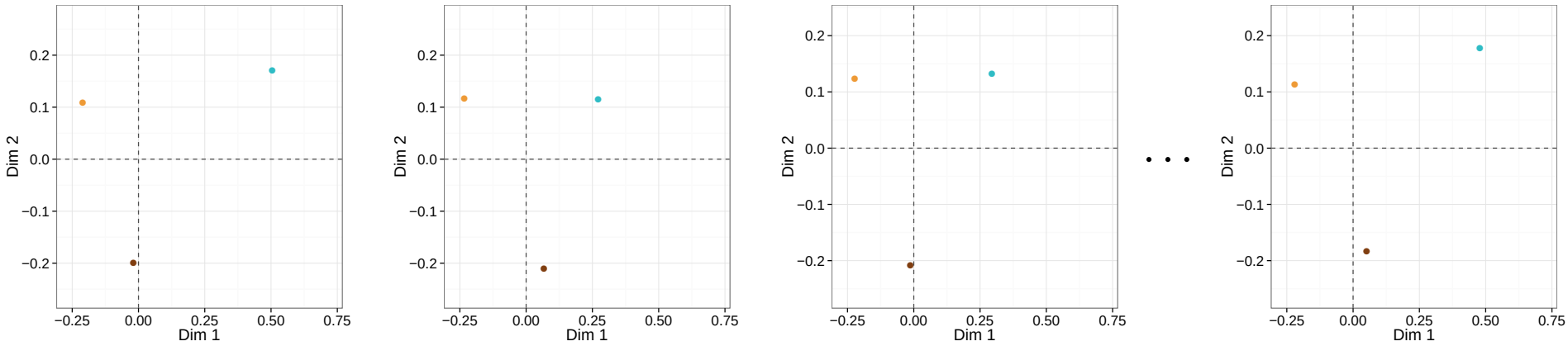
Il est crucial d'avoir un outil permettant d'évaluer la fiabilité des résultats

Pourquoi est-il essentiel l'incertitude ?

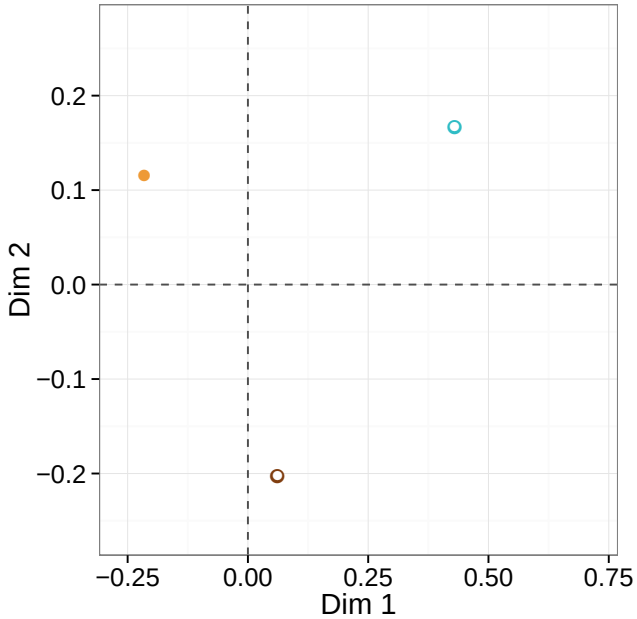
configuration compromise



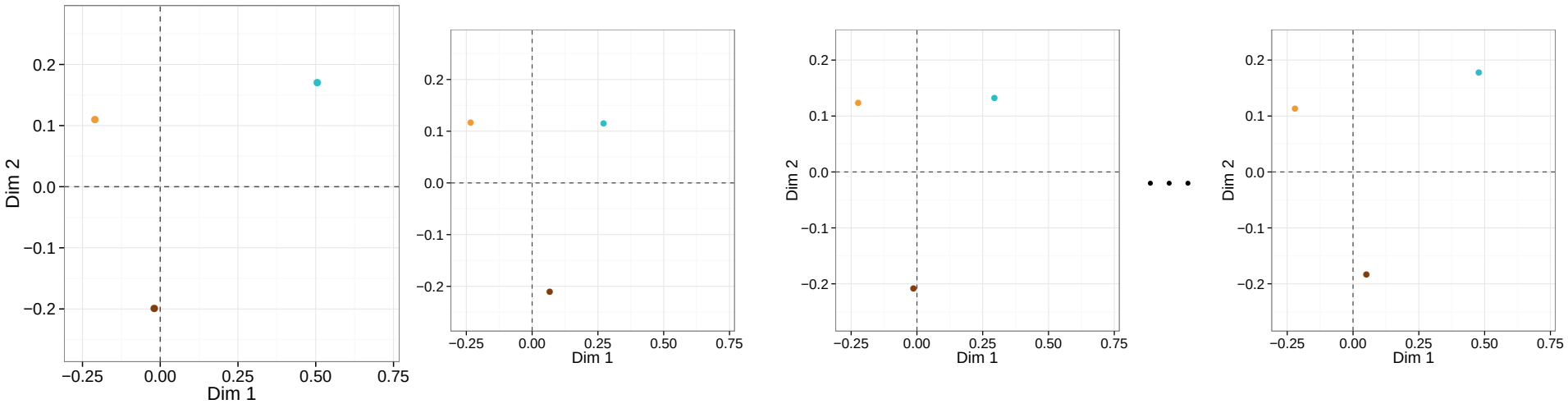
Pourquoi est-il essentiel l'incertitude ?



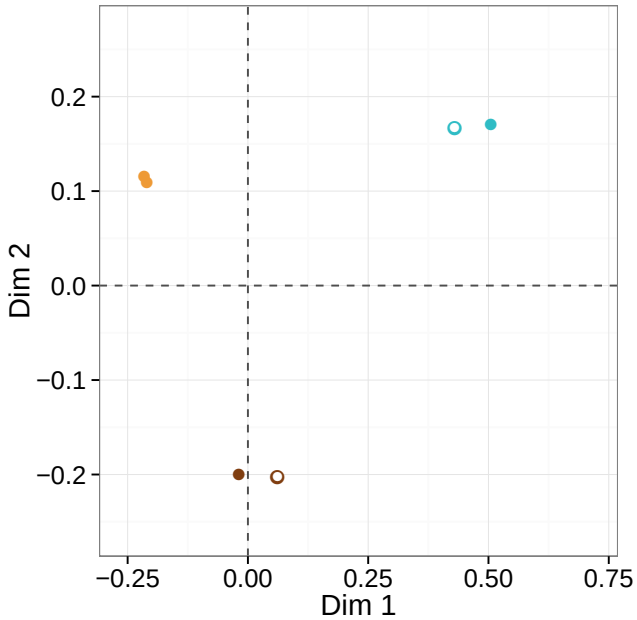
configuration compromise



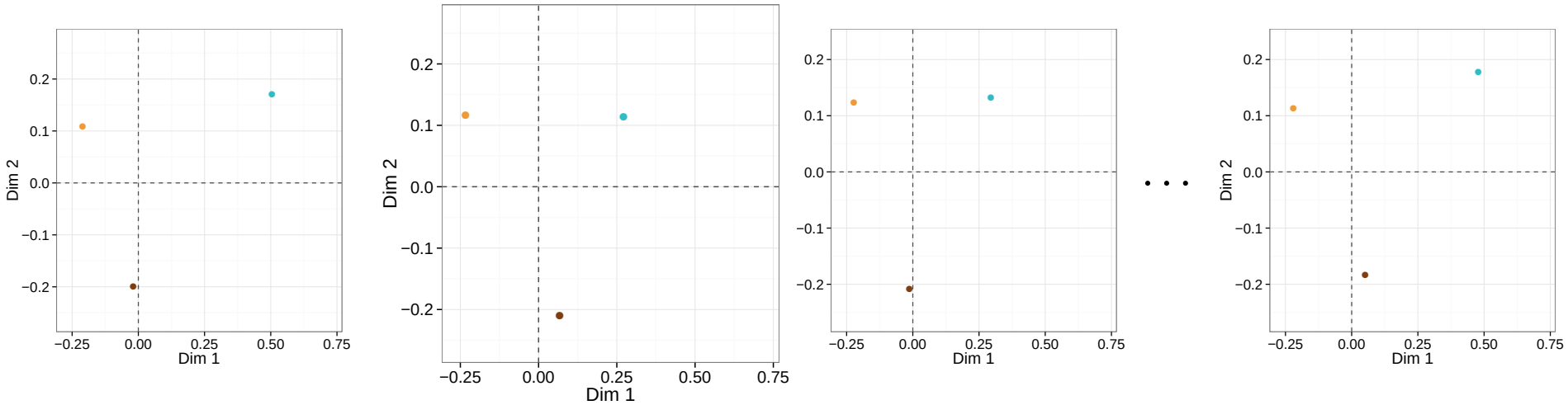
Pourquoi est-il essentiel l'incertitude ?



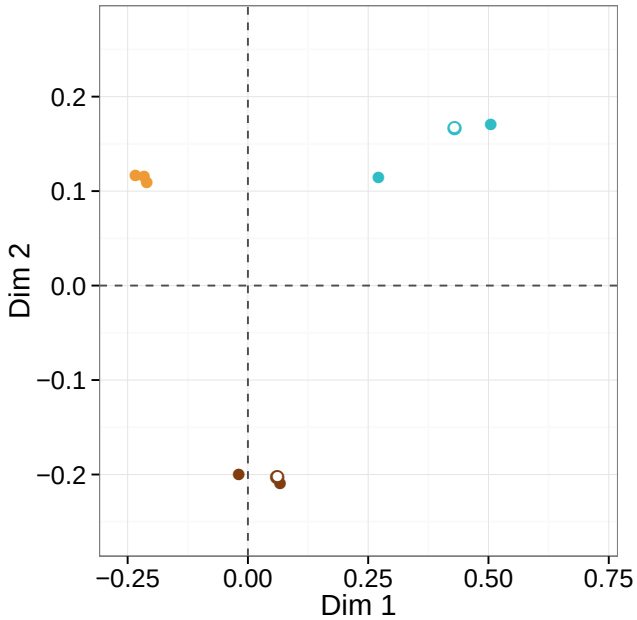
configuration compromise

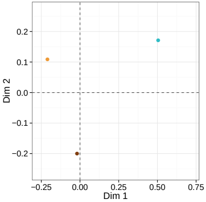


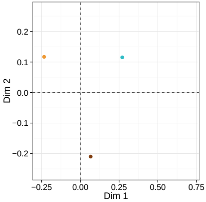
Pourquoi est-il essentiel l'incertitude ?

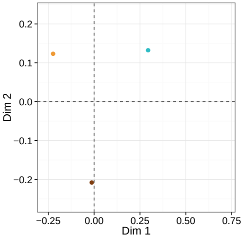


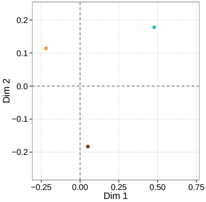
configuration compromise



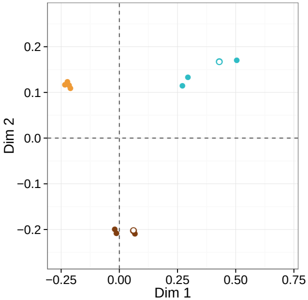




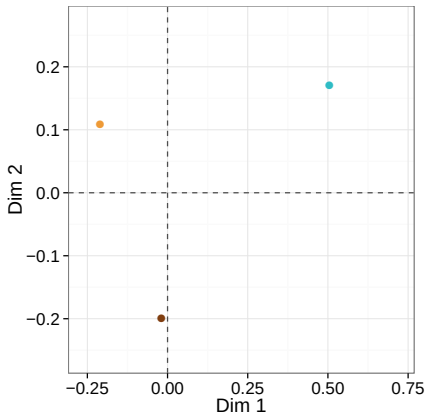


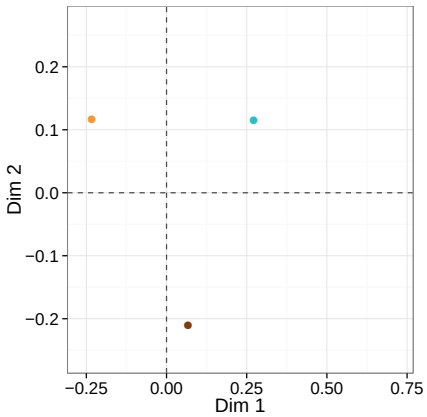


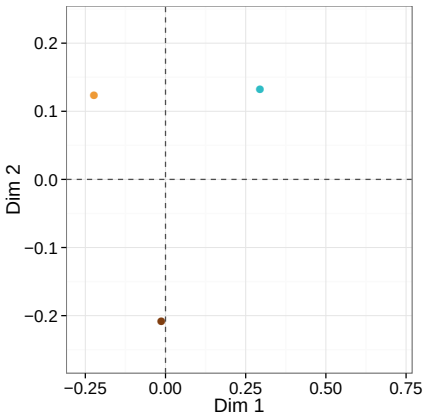
configuration compromise



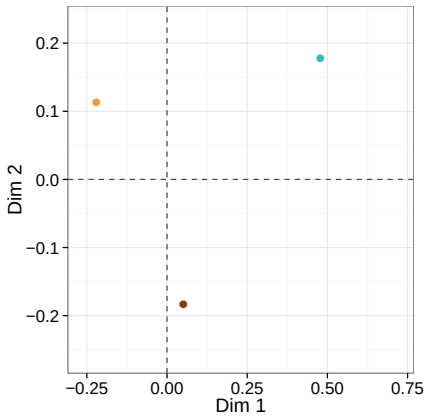
Pourquoi est-il essentiel l'incertitude ?



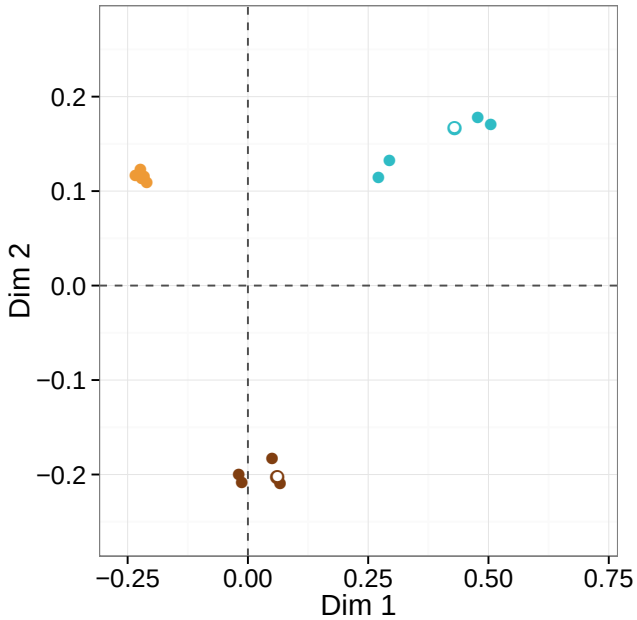




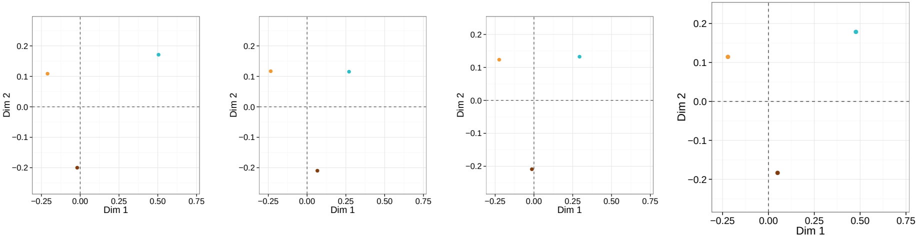
• • •



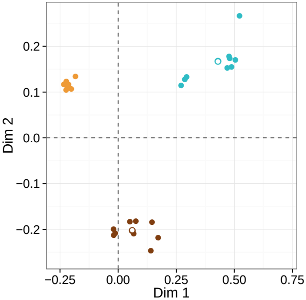
configuration compromise



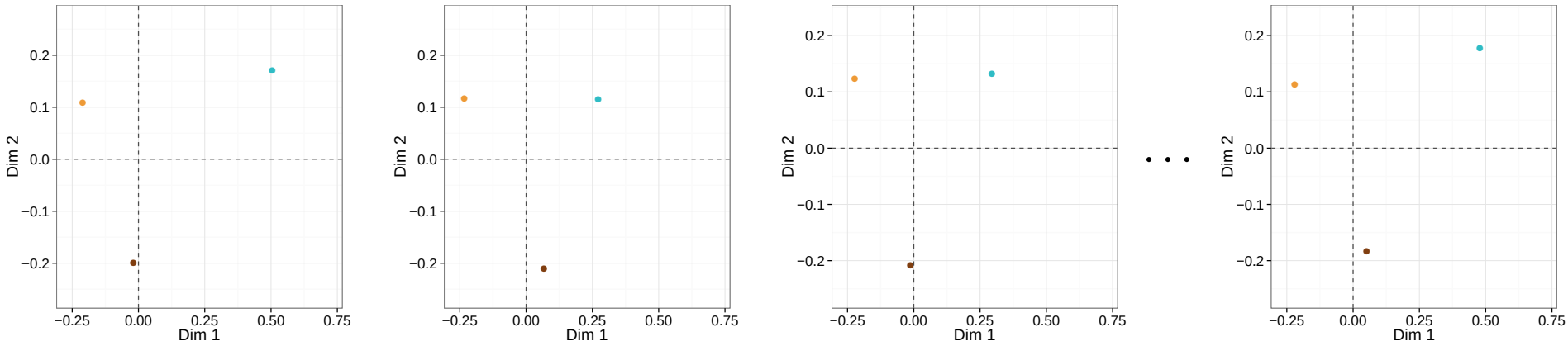
Pourquoi est-il essentiel l'incertitude ?



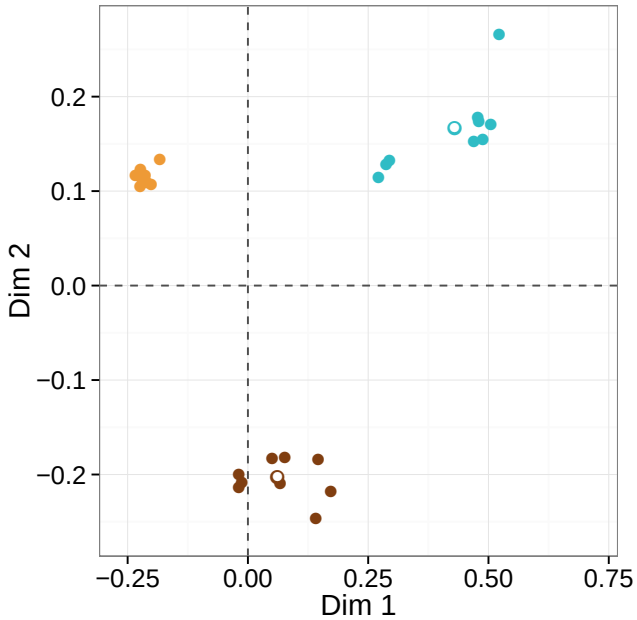
configuration
compromise



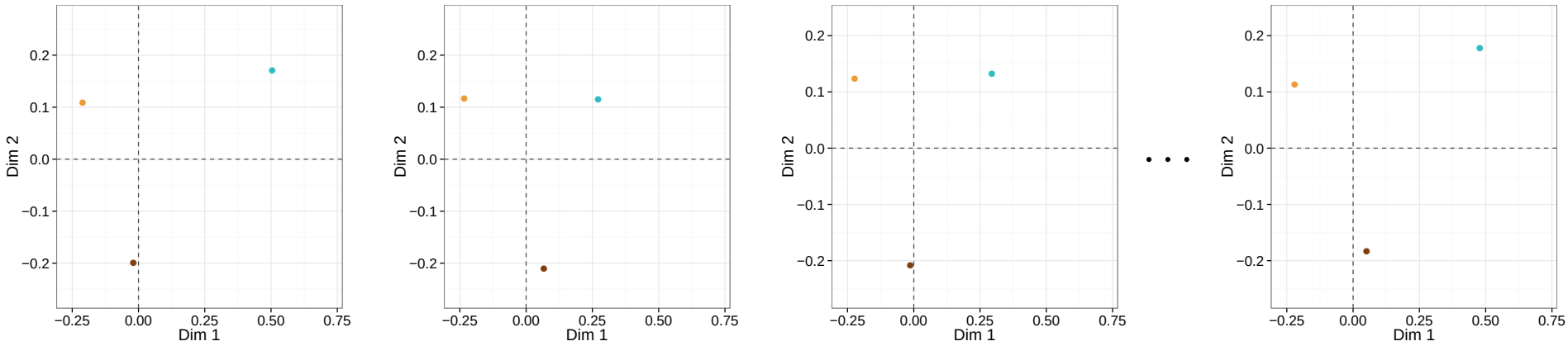
Pourquoi est-il essentiel l'incertitude ?



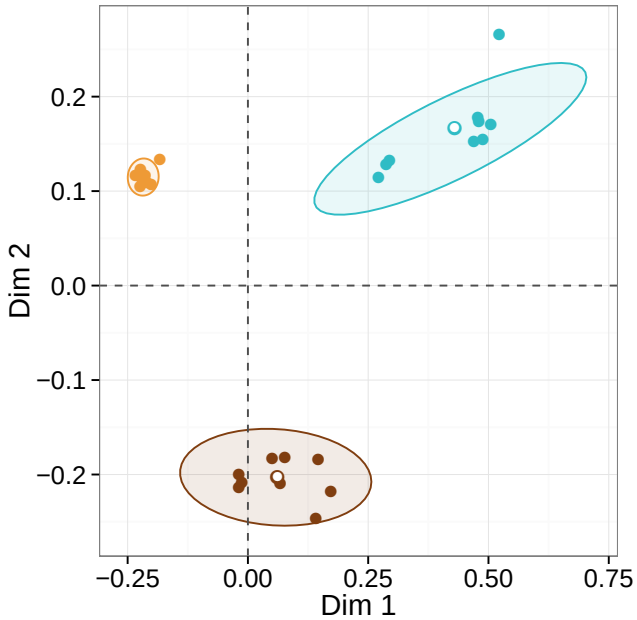
configuration compromise



Pourquoi est-il essentiel l'incertitude ?

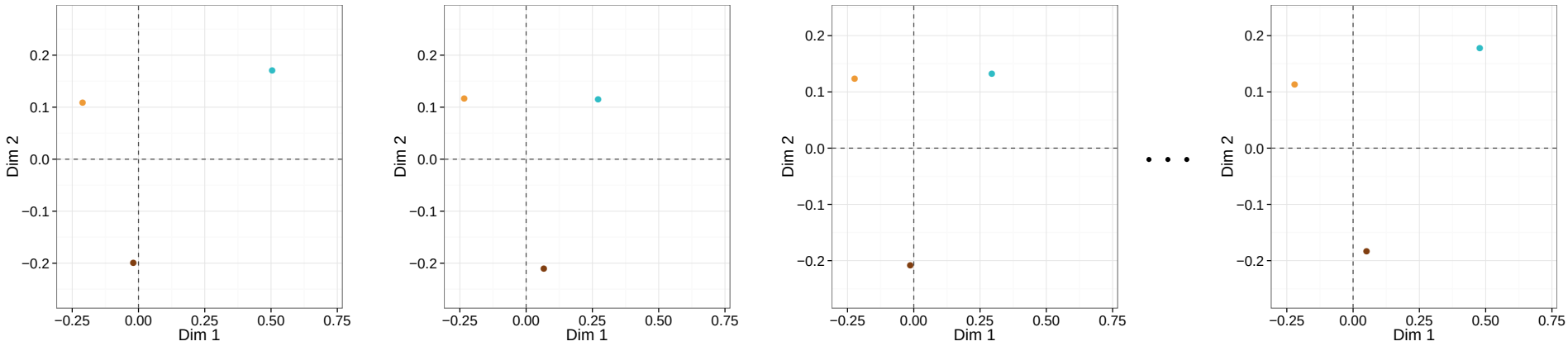


configuration compromise

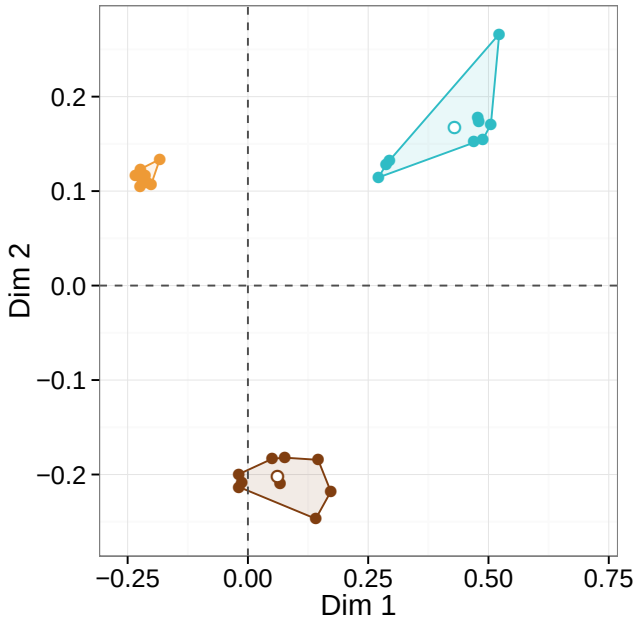


ellipses de confiance

Pourquoi est-il essentiel l'incertitude ?

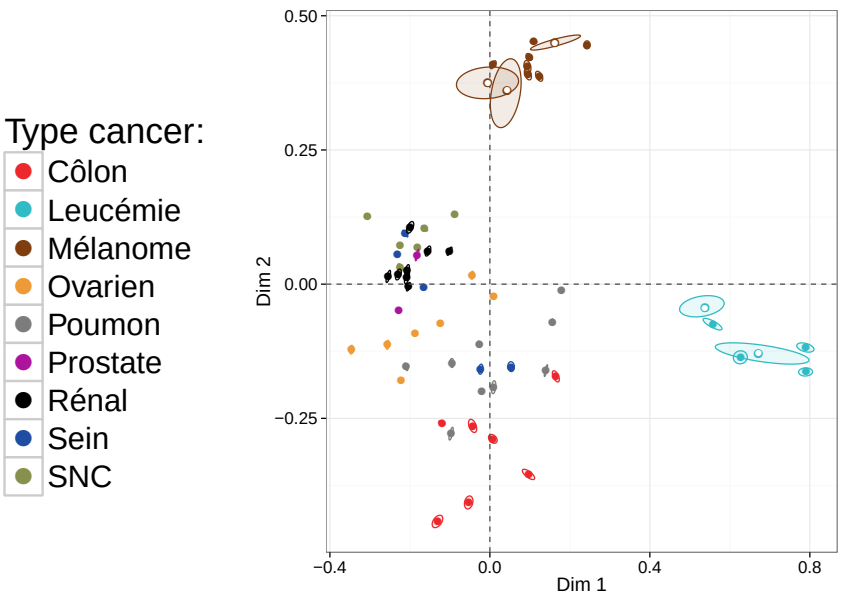


configuration compromise

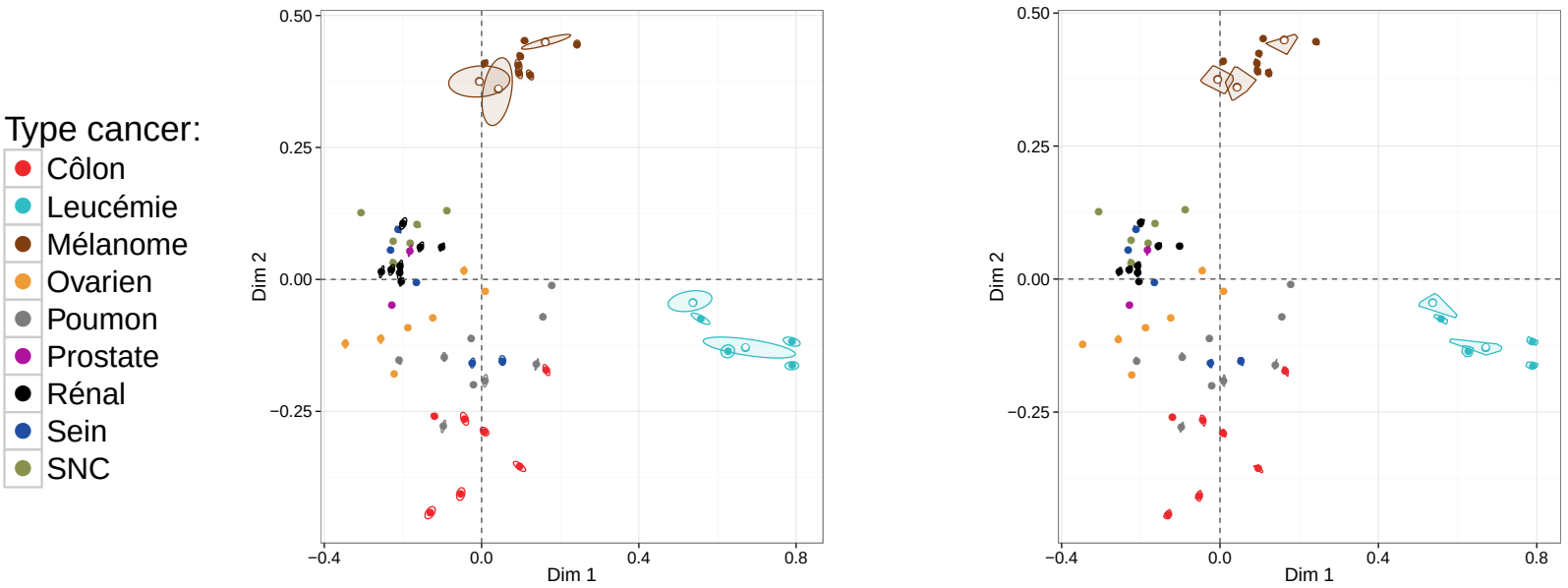


convex hulls

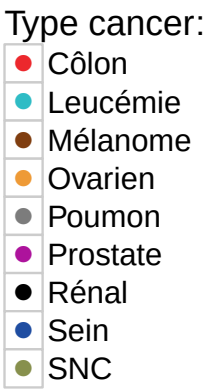
Pourquoi est-il essentiel l'incertitude ?

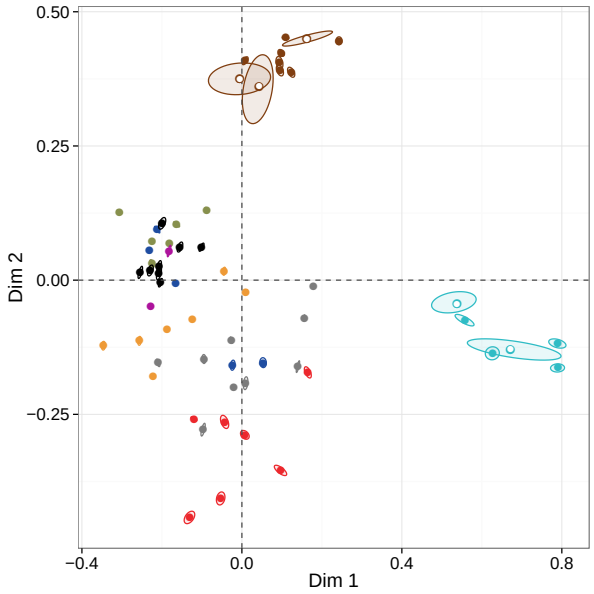


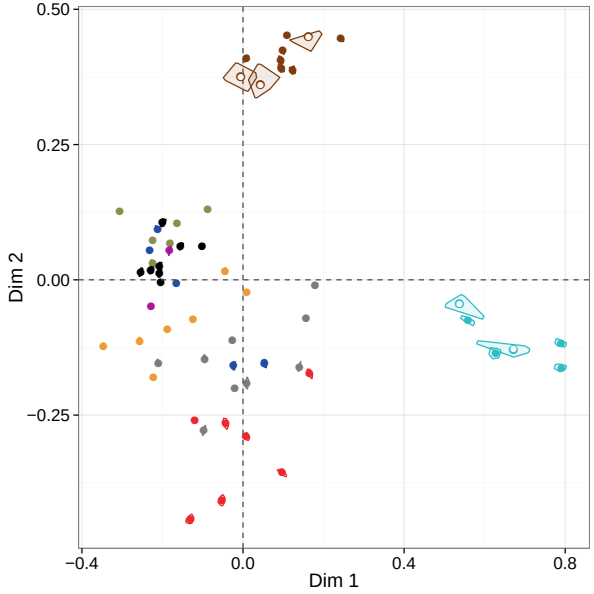
Pourquoi est-il essentiel l'incertitude ?

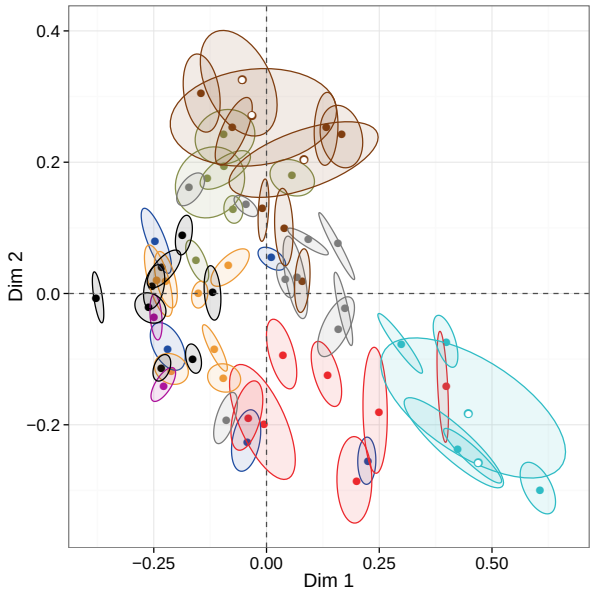


Pourquoi est-il essentiel l'incertitude ?









Pourquoi est-il essentiel l'incertitude ?