Mise à jour MixtComp été 2017

Vincent KUBICKI - InriaTech

Inria Lille - Nord Europe

17 Octobre 2017

MixtComp - Généralités

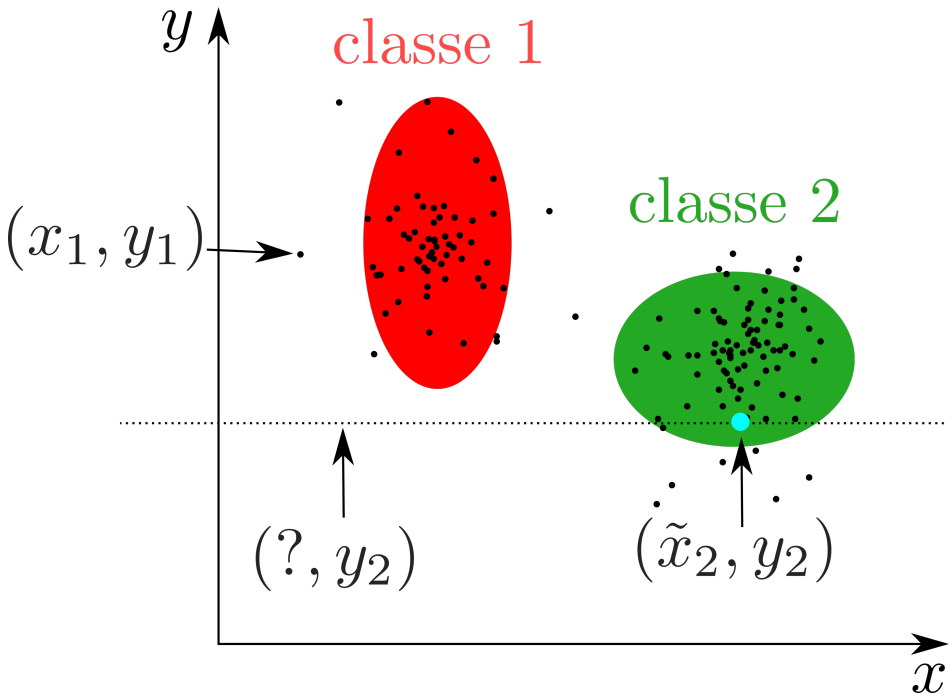

Figure 1 – Modes de MixtComp

- Classification de données hétérogènes et partiellement observées
- Prédiction basée sur la classification
- Utilisation de modèles statistiques : $f(x; \hat{\theta})$
- Estimation par maximisation de vraisemblance

MixtComp - Structure et problèmes

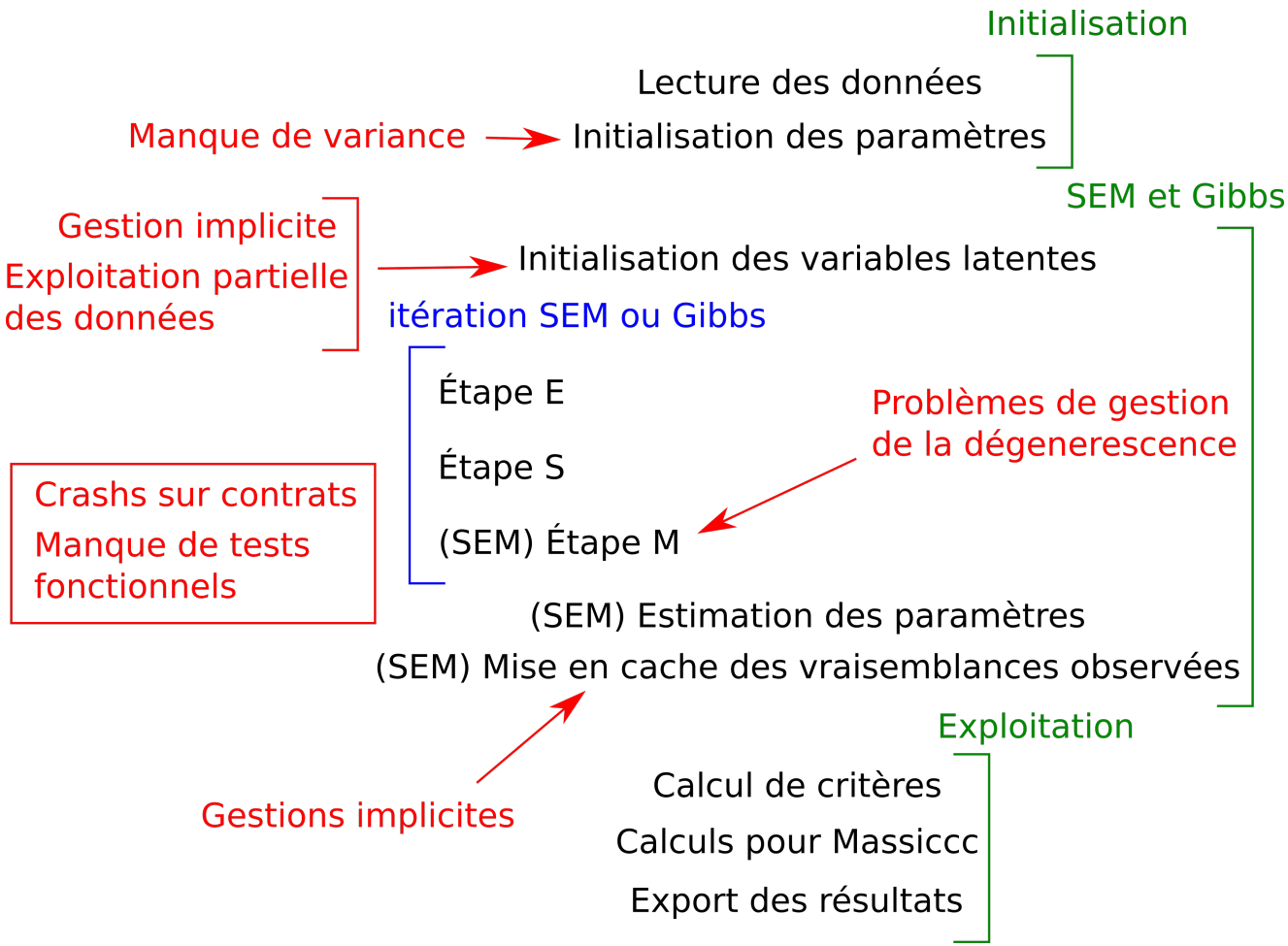

Figure 2 – Vue synthétique de l'algorithme

Manque de variance dans les initialisations

Estimation par individu représentant

Ancienne méthode

- Partitionnement des individus
- Estimation par maximum de vraisemblance

Nouvelle méthode

- Tirage d'un individu par classe
- Déduction des valeurs de paramètres, modèle par modèle

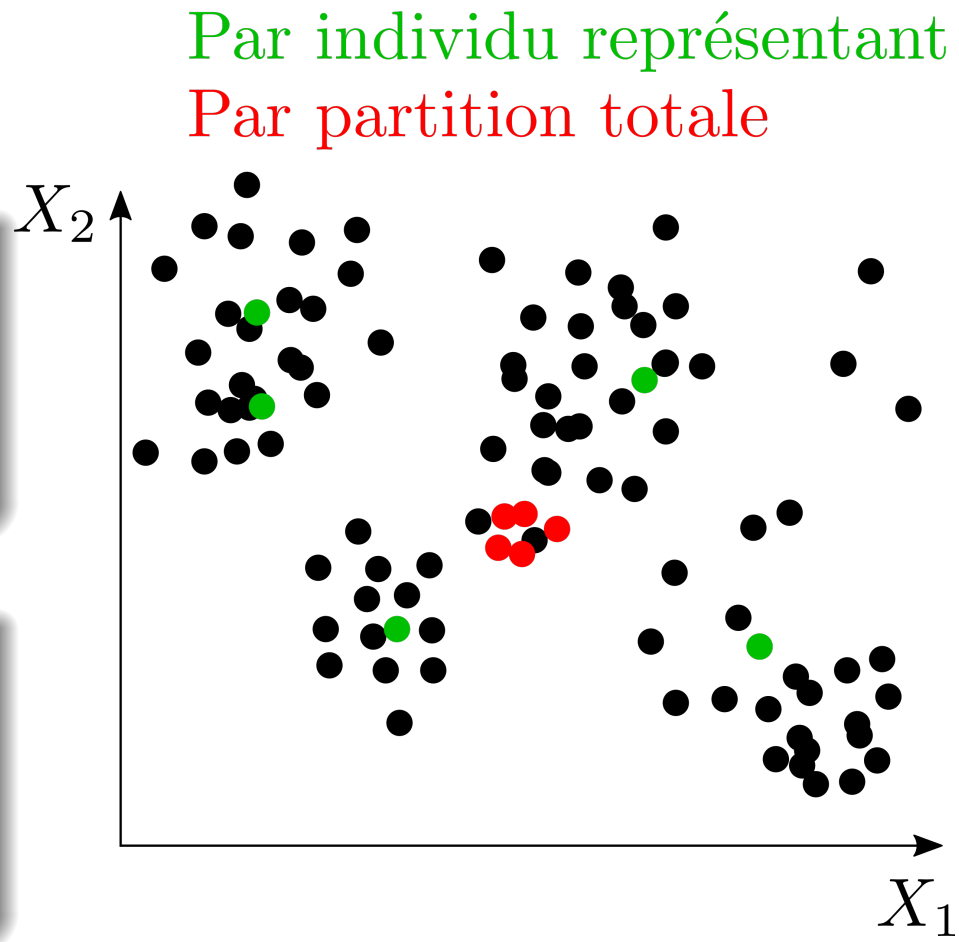

Figure 3 – Comparaison des initialisations

Gestion complexe des dégénérescences

Lancements multiples

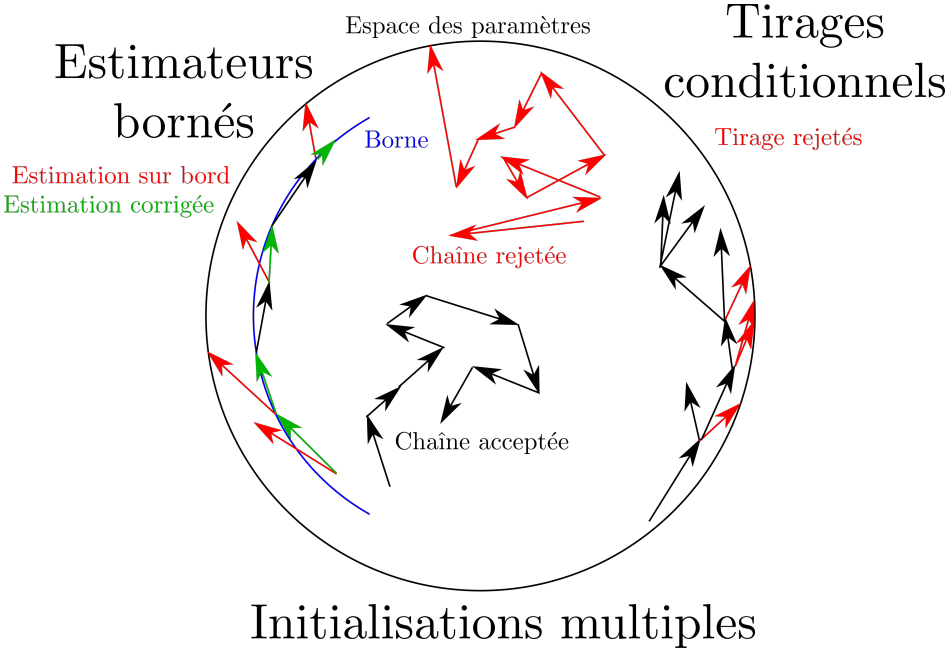

Figure 4 – Comparaison des méthodes d'initialisation

Dégénérescence avant

- Gibbs : lente
- Borne : ad-hoc et crash

⇒ Code complexe et lent : déclenchements et tirages

Dégénérescence maintenant

- Relance si détection de dégénérescence
- Descriptif erreur pour utilisateur

Utilisation sous-optimale des données lors de l'initialisation

Utilisation de la vraisemblance observée

Ancienne méthode

Initialisation uniforme des variables latentes

$\Downarrow$

Initialisation éloignée de la distribution finale

Nouvelle méthode

- Utilisation de la vraisemblance observée
- Limitation calculatoire

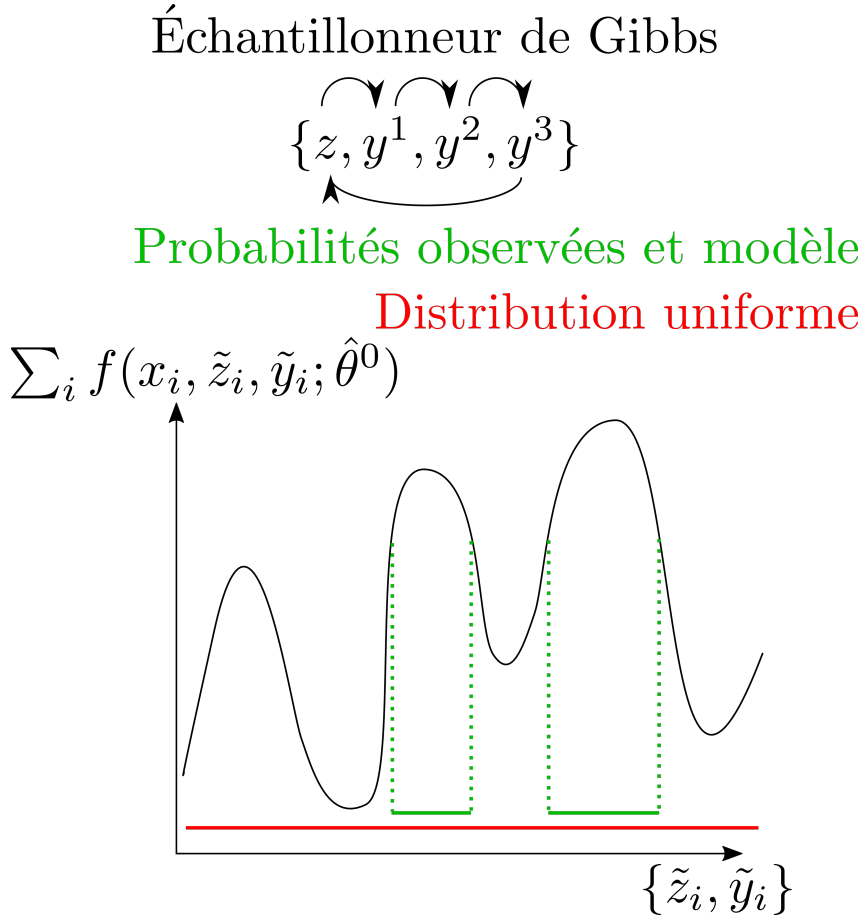

Figure 5 – Tirages avec modèle

Gestion implicite des initialisations, mises en cache et tests

Génie logiciel

Ancienne architecture d'initialisation

- Héritée des modèles simples
- Suppose que le modèle connaît l'algo appelant
- Peu lisible

Ajout de nouvelles méthodes

- Calcul de distribution observée empirique $f(x, y, z; \hat{\theta}) \rightarrow f(x; \hat{\theta})$
- Mise en cache des probabilités observées
- Initialisation de chaîne de Markov

Ajout de tests

- Un test fonctionnel par modèle
- Tests unitaires mis à jour suite aux autres modifications

Résumé des modifications

Figure 6 – Connexion entre les principales modifications

Performances

- Tests à effectuer sur données réelles
- Création de sorties pour l'analyse de performance données générées
- Le modèle ordinal fonctionne mieux

Développements possibles

- Méthodes alternatives d'initialisation possibles : multiples individus, et étape de maximisation de vraisemblance
- Méthodes alternatives plus faciles à implémenter
- Affinement des conditions de détection de dégénérescence
- Ré-introduction partielle des estimateurs bornés ?