Ecole des Ponts Paristech

2010–2011

Rapport de stage scientifique

Bensalah Antoine

élèves ingénieurs, première année

Modélisation de l'anisotropie d'un réseau de discontinuités 3D par mélanges de lois de probabilités

Stage réalisé au sein du LEESU
LEESU, Ecole des Ponts ParisTech, 6-8 avenue Blaise Pascal, Cité Descartes, Champs-sur-Marne, 77455 Marne-la-Vallée Cedex 2 (France)
mai-juillet

Maître de stage : Olivier Fouché

Fiche de synthèse

- Type de stage : stage scientifique
- Année académique : 2010-2011
- Auteur : Bensalah Antoine
- Formation : élève ingénieur, première année Ecole des Ponts ParisTech
- Titre du rapport : Modélisation de l'anisotropie d'un réseau de discontinuités 3D par mélanges de lois de probabilités
- Organisme d'accueil : LEESU
- Pays d'accueil : France
- Maître de stage : Olivier Fouché
- Tuteur de stage : Olivier Fouché
- Mots-clefs : Algorithmes EM-SEM, réseaux de discontinuités, tests statistiques, distance de Hausdorff, classification automatique
- Thème Ecole : Mathématiques appliquées, Géologie-Géophysique, Recherche

Remerciements

Je tiens à remercier dans un premier temps l'équipe pédagogique de l'Ecole des Ponts ParisTech, ainsi que les intervenants du laboratoire LEESU de m'avoir permis d'effectuer un stage scientifique très enrichissant dans les meilleurs conditions. Ce stage a été financé par le projet ANCRES du programme GESSOL (Ademe et MEDDTL) coordonné par Olivier Fouché. Ainsi je souhaite remercier l'Ademe et MEDDTL pour leur soutient financié et tout particulièrement Olivier Fouché pour sa patience et pour sa pédagogie, ainsi que Jean Diebolt, pour sa patience également, et pour ses explications précieuses qui ont su me guider et me faire découvrir un domaine fascinant. J'ai beaucoup apprécié travailler avec ces deux chercheurs, comme au sein d'une équipe, je les en remercie.

Résumé

Ce rapport présente une approche probabiliste de la modélisation de l'anisotropie d'un réseau de discontinuités (RD) et conclue un stage financé par le projet ANCRES du programme GESSOL (Ademe et MEDDTL). Seule la première étape de la modélisation est abordée, nous avons décrit et discuté seulement ici des méthodes de classification automatique de données d'orientation planaire ou vectorielle 3D dont les applications ne se cantonnent pas aux RD. Ces méthodes de classification en familles des orientations des discontinuités utilisent des estimations de paramètres de mélanges de lois de probabilités ; sur la sphère dans le cas vectoriel 3D. Ces estimations peuvent alors servir pour modéliser le réseau de discontinuités. Nous nous sommes dans un premier temps attachés à fixer les conventions et les notations qui sont prises tout au long du rapport. Ensuite, nous avons utilisé différentes lois de probabilités sur la sphère unité pouvant être utilisées dans la classification en familles des discontinuités. Si la loi de Fisher nous avait paru la plus simple, pour des raisons de symétrie, nous avons retenu pour le cas de la 3D des lois de probabilités formées par des projections de lois normales sur la sphère. Nous avions déjà une méthode non probabiliste de classification des RD qui nous permettait également de décider a posteriori du nombre de familles que nous voulions former. Les algorithmes probabilistes que nous avons utilisés sont EM et SEM, ce dernier ayant été mis de côté car il ne nous satisfaisait pas vraiment. La description du fonctionnement de ces algorithmes est enrichie de nombreux exemples présentés dans ce rapport qui analysent également leurs faiblesses et leurs limites. A l'issue de cette description, nous avons mis en place un test statistique de validation des paramètres estimés basé sur des méthodes utilisées habituellement pour la reconnaissance de formes et utilisant notamment la distance de Hausdorff. Ce critère constitue donc une validation a posteriori du fonctionnement des algorithmes. Nous avons donc mené une étude statistique de ce critère afin de déterminer le risque de première espèce et de borner l'erreur faite sur les paramètres estimés lorsque le test est positif. Nous sommes arrivés à limiter le risque de première espèce à 5%. Cependant, lors de l'étude de sa senbilité, nous avons observé que les erreurs permises sur la matrice de covariances pour les projections de lois multi-normales du plan tangent sur la sphère sont de l'ordre de 100%. Si nous n'arrivons pas à contrôler l'erreur sur la matrice de covariances, l'erreur sur le vecteur moyen est au maximum d'une dizaine de degrés, ce qui est plutôt satisfaisant. Nous avons alors mis en pratique le fonctionnement de ces algorithmes pour une application à la géologie, où la modélisation des fractures dans des massifs rocheux est recherchée car les mesures directes n'y sont pas précises. Nous

avons dans cette optique commencé une étude de l'influence des erreurs de mesures dues aux relèvement manuels du positionnement des fractures sur les carottes. Puis nous avons fini par l'étude d'un cas particulier de forage, où sont présents deux principaux types de discontinuités, que nous avons cherché à regrouper en familles. Si les résultats de la modélisation sont exploitables, il reste néanmoins quelques problèmes dans le fonctionnement des algorithmes, qui ne donnent pas entière satisfaction quant aux paramètres estimés.

Mots-clés : algorithmes EM/SEM, simulation réseaux de discontinuités, distance de Hausdorff, test statistique

Abstract

In this report, we shall present a probabilistic model of a discontinuties network's anisotropy (RD). This is the conclusion of a training course financied by ANCRES project from the GESSOL program (Ademe & MEDDTL). We present here the first step of the modelisation, we have just described and discussed about automatic classification methods of plannar or 3D vectoriel orientation data which applications are not bounded to RD. These methods of classification in discontinuties orientation groups, are using estimations of parameters of mixes probabilities distribution, and in the 3D vectoriel case, on the unit sphere. Those estimations may later be used to model the discontinuties network.

In the first place we will detail all the conventions and notations that we used in this report. Afterwards, we will describe several probabilities distributions on the unit sphere that may be used to classify in different classification groups. If the Fisher distribution appeared to be the most simple, for some symetric reasons, we would used, in the 3D case, probabilities distributions created by projections of normal distributions on the sphere. We will present another method, a non-probabilistic classification of the RD which allows us to decide later on the number of groups that we want to form. The probabilistics algorithms that we have used are the EM and SEM algorithms. The SEM's one is not used because it doesn t give good results. The description of functioning algorithms is enriched by a lot of exemple presented in this report which analysed the strengths and weaknesses of those algorithms.

At the term of this description, we will set up a statistic test in order to validate the estimated parameters based on methods which ordinary are used for form recognitions, like the Hausdorff distance. So this criterion is a post-validation of functioning algorithms. Thereforme, we will make a statistical study of this criterion, in order to calculate the error of the first kind and bound up the error made on the estimated parameters, when the test is positive. We conclude that we can bound the error of the first kind up to 5%. However, during the sensitivity test, we observe that the errors on the covariance matrix for multi-normal distributions projections of the tangent plan on the sphere are around 100%. If we didn t succeded to control the error on the covariance matrix, the erreur on the average vector would reach its peak at 10 degrees, which is rather satisfying.

Then we will study the functioning of the algorithms for an application in geology, which is the modelisation of fractures in a rock massif. At the beginning of this part, we study the influence of measure errors because of the manual determination of the orientation of the fractures of the samples taken from a bore. Finaly we will end this report with a particular case of drilling, where we can find 2 principal types of discontinuities, which we

will exploit. Even if the results of modelisation are exploitable, it remains some issues in the functioning of algorithms, which are not given the entire satisfaction in the estimated parameters.

Key words : EM/SEM algorithms, simulation of discontinuted networks, Hausdorff distance. Statistical test.

Contents

List of Figures

Présentation de l'organisme d'accueil et du maître de stage

J'ai été accueilli dans le cadre de mon stage scientifique de première année à l'Ecole des Ponts ParisTech au laboratoire LEESU (Laboratoire Eau, Environnement et Systèmes Urbains). C'est un laboratoire commun à l'Université Paris-Est Créteil, l'Université Paris-Est Marne-la-Vallée, Agro ParisTech et à l'Ecole des Ponts ParisTech. Ses recherches s'organisent autour de trois principaux thèmes que sont le cycle de l'eau et des contaminants, les extrêmes hydrologiques ainsi que les techniques multi-échelles de modélisation des milieux complexes et de leur climat instationnaire, comme c'est le cas pour le milieu urbain. Ce stage a pour visée la découverte du monde de la recherche par l'exécution d'un travail de recherche au sein d'une équipe d'un laboratoire. Pour ma part, j'ai travaillé en liaison avec mon maître de stage Olivier Fouché, chercheur au sein du LEESU, également maître de conférence de la Chaire de Géotechnique du Conservatoire national des arts et métiers (CNAM), dont un des principaux sujet de recherche actuel est la modélisation multi-échelles de la géométrie des milieux discontinus et des écoulements dans les aquifères fissurés. J'ai également beaucoup travaillé avec Jean Diebolt, directeur de recherches au CNRS en mathématiques appliquées attaché au Laboratoire d'Analyse et de mathématiques appliquées de l'Université de Marne-la-Vallée. J'ai donc eu accès à un bureau au sein du laboratoire à l'Ecole des Ponts ParisTech, avec un ordinateur à ma disposition. J'ai souvent travaillé directement en contact avec Jean Diebolt, qui à supervisé l'avancement des travaux, essentiellement en mathématiques, tandis que je m'accordais avec Olivier Fouché pour la partie géologie et sont rapport avec les mathématiques.

1 Introduction

1.1 Contexte

Ce stage a été financé par le projet ANCRES du programme GESSOL (Ademe et MEDDTL). Dans le cadre du projet ANCRES, nous nous intéressons à l'étude des hétérogénéités et discontinuités du terrain naturel dans les premiers mètres, du sol à la roche mère, voire jusqu'à la nappe. Macropores dus à l'activité biologique dans le sol, fissures dues à la sécheresse ou au gel, fractures d'origine tectonique, itinéraires de contournement des lentilles argileuses, toutes ces hétérogénéités sont des chemins préférentiels d'écoulement depuis la surface vers la nappe. En facilitant le transfert vers la profondeur, ces chemins empêchent le matériau naturel en place d'exercer sa fonction de filtration et d'épuration. Il en résulte un transfert de polluants vers la nappe. En terme de mesure physique des propriétés du terrain, leur présence modifie grandement les résultats des essais d'infiltration, dont les difficultés d'interprétation sont alors accrues. Dans ce stage, on a pris à titre d'illustration une population de fractures tectoniques d'un massif rocheux, mesurées sur le terrain. Mais le sujet théorique est beaucoup plus général. En effet, il s'agit d'étudier un moyen de classification automatique des orientations de réseaux de discontinuité (RD) en familles, pour ensuite procéder à leur simulation. Nous pouvons remarquer que puisque les orientations des RD sont des données vectorielles, l'étude peut s'appliquer à toute classification de d'orientations vectorielles en 3D.

1.2 Objectifs

En nous appuyant sur cette constatation, nous souhaitons utiliser des algorithmes de regroupement de points (sur une sphère) en familles. Le choix des points sur une sphère s'explique par la représentation choisie pour les orientations des discontinuités.

Nous souhaitons mettre en place un algorithme de classification des orientations de RD grâce à des algorithmes probabilistes de regroupement de points de la sphère unité représentant les orientations possibles des discontinuités. Pour cela, nous allons dans un premier temps définir les conventions et les notations prises tout au long de ce rapport. Puis nous introduirons les différentes lois de probabilités que nous avons étudiés pour reconnaître les familles de points. Nous présenterons alors un premier algorithme de classification que nous possédions déjà et qui permet une décision a posteriori du nombre de familles à chercher dans l'échantillon. Nous pourrons alors décrire l'algorithme EM et nous illustrerons cette description de nombreux exemples afin d'en étudier les capacités et les limites. L'algorithme SEM sera défini et étudier en comparaison avec l'algorithme EM. Nous définirons un critère basé sur la distance de Hausdorff dans le but d'avoir un critère de validation des paramètres estimés par nos algorithmes. Nous souhaitons

alors avoir une étude de ce critère la plus précise possible afin de contrôler le risque de première espèce et sa sensibilité. Nous terminerons ce rapport par une application à la géologie conformément aux attentes du stage. Dans ce cadre, nous étudierons l'influence des erreurs de mesures lors du relèvement manuel des orientations des fractures. Enfin, nous exécuterons l'algorithme EM sur un échantillon de données issues d'un forage. Ce sera l'occasion de mettre en pratique nos algorithmes et nos programmes sur des données réelles.

1.3 Langages de programmation utilisés

Nous avons utilisé essentiellement le logiciel Scilab, pour sa simplicité et la concisions de son langage de programmation. Cependant, pour certaines fonctions, le temps d'exécution s'est révélé beaucoup trop conséquent. Nous avons donc décidés d'écrire quelques fonctions en C++. Le gain de temps fut alors au niveau de nos attentes.

2 Conventions et notations

2.1 Représentation de l'orientation des discontinuités

Nous attribuons à chaque fracture un vecteur normal unitaire. Pour qu'il y ait un unique choix possible, nous prenons la convention "normale montante" (ou bien "normale descendante"). On choisit alors le vecteur normal unitaire montant (ou descendant).

Dans un repère orthonormé de référence $(O, \overrightarrow{e_x}, \overrightarrow{e_y}, \overrightarrow{e_z})$, un vecteur $\overrightarrow{u}$ est dit montant (resp. descendant) si $\overrightarrow{u}.\overrightarrow{e_z} > 0$ (resp. $\overrightarrow{u}.\overrightarrow{e_z} < 0$). Dans notre étude, le repère orthonormé de référence $(O, \overrightarrow{e_x}, \overrightarrow{e_y}, \overrightarrow{e_z})$ sera tel que l'axe Oz soit la direction de la carotte considérée.

2.2 Repère et coordonnées choisis

Nous choisissons alors les coordonnées sphériques ainsi que les coordonnées cartésiennes pour représenter les vecteurs normaux des discontinuités. Nous confondrons par la suite les vecteurs unitaires et les points sur la sphère unité, essentiellement du point de vue des coordonnées. Comme il y ambiguïté à propos de la définition des coordonnées sphériques, nous suivrons les définitions habituellement utilisées en physique :

Considérons un point $M \in \mathbb{R}^3$ repéré initialement dans le repère cartésien $(O, \overrightarrow{e_x}, \overrightarrow{e_y}, \overrightarrow{e_z})$ par les coordonnées (x, y, z). Les coordonnées sphériques de M notées (r, θ, φ) associées à la base mobile $(\overrightarrow{e_r}, \overrightarrow{e_\theta}, \overrightarrow{e_\varphi})$ sont définies par :

Soient H le projeté orthogonal de M sur le plan Oxy, θ l'angle $(\overrightarrow{e_z}, \overrightarrow{OM})$ pris modulo π et φ l'angle $(\overrightarrow{e_x}, \overrightarrow{OH})$ pris modulo 2π.

Alors

$$\overrightarrow{e_r} = \frac{\overrightarrow{OM}}{\left\|\overrightarrow{OM}\right\|}$$

$$\overrightarrow{e_\theta} = r\cos(\varphi)\sin(\theta)\overrightarrow{e_x} + r\sin(\varphi)\sin(\theta)\overrightarrow{e_y} + r\cos(\theta)\overrightarrow{e_z}$$

$$\overrightarrow{e_\theta} = -r\sin(\varphi)\sin(\theta)\overrightarrow{e_x} + r\cos(\varphi)\sin(\theta)\overrightarrow{e_y}$$

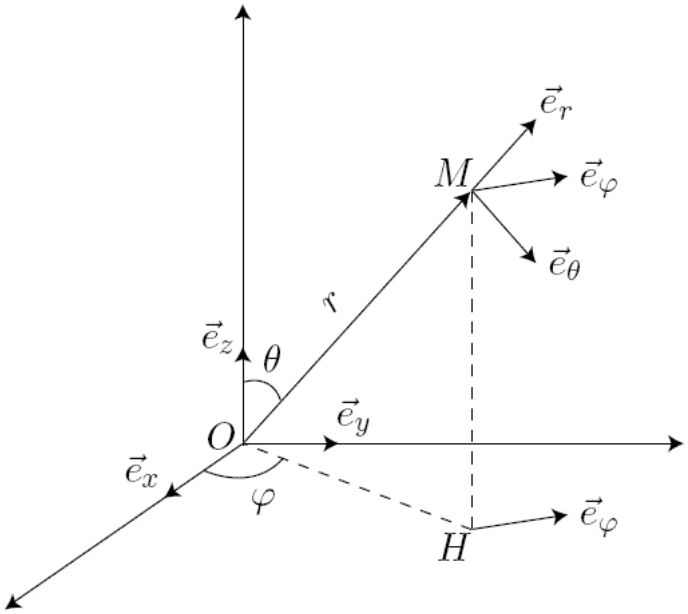

Figure 1: Conventions pour les coordonnées sphériques

Rappelons les passages de l'un à l'autre des systèmes de coordonnés :

$$\begin{cases} x = r\sin(\theta)\cos(\varphi) \\ y = r\sin(\theta)\sin(\varphi) \\ z = r\cos(\theta) \end{cases} \qquad \begin{cases} r = x^2 + y^2 + z^2 \\ \theta = \arccos(\frac{z}{\sqrt{x^2+y^2+z^2}}) \\ Si\ y > 0 \qquad \varphi = \arccos(\frac{x}{\sqrt{x^2+y^2}}) \\ Si\ y \leq 0 \quad \varphi = 2\pi - \arccos(\frac{x}{\sqrt{x^2+y^2}}) \end{cases}$$

2.3 Visualisation des familles de discontinuités, diagramme de Wulff

Il est commode et habituel d'utiliser les diagrammes de Wulff pour représenter les points sur la demi-sphère supérieure ou inférieure (convention normale montante ou descendante). Nous utiliserons donc aussi cette représentation par la suite. Nous indiquerons par un symbole $\cap$ si c'est la convention normale montante qui à été choisie, par le symbole $\cup$ si c'est la convention

normale descendante. Tracer un diagramme de Wulff consiste en une projection stéréographique sur le plan équatorial des points de la demi-sphère supérieure par rapport au pôle sud, ou des points de la demi-sphère inférieure par rapport au pôle nord.

La projection stéréographique :

Étudions le cas de la convention normale montante, le cas de la convention normale descendante est similaire.

On note $\mathbb{S} = \{(x, y, z) \in \mathbb{R}^3 \ tels\ que\ x^2 + y^2 + z^2 = 1\}$ la sphère unité. Soit $S = (0, 0, -1)$ le pôle sud et $\mathbb{S}_+ = \mathbb{S} \cap \{(x, y, z) \in \mathbb{R}^3 \ tels\ que\ z \geq 0\}$ la demi sphère surpérieure.

La projection stéréographique de $\mathbb{S}_+$ sur le plan équatorial $\mathbb{R}^2$ est l'application qui à $M \in \mathbb{S}_+$ associe l'unique point d'intersection entre la droite (MS) et le plan équatorial. On obtient aisément :

$$\begin{array}{rccc} \pi : & \mathbb{S}_+ & \longrightarrow & \mathbb{R}^2 \\ & (x, y, z) & \longmapsto & (\frac{x}{1+z}, \frac{y}{1+z}) \end{array} \tag{1}$$

π est alors un difféomorphisme de $\mathbb{S}_+$ sur le disque unité de $\mathbb{R}^2$.

Visualisation graphique de la projection stéréographique Dans la figure 2, le choix de la convention est inversée : c'est la demi-sphère inférieure qui est projetée sur le plan équatorial.

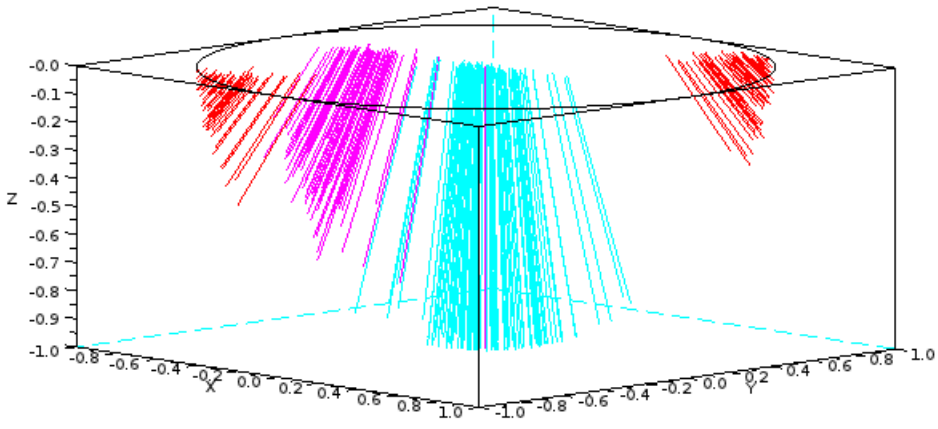

Figure 2: Projection stéréographique convention normale descendante

La figure 3 présente le diagramme de Wulff associé.

∪

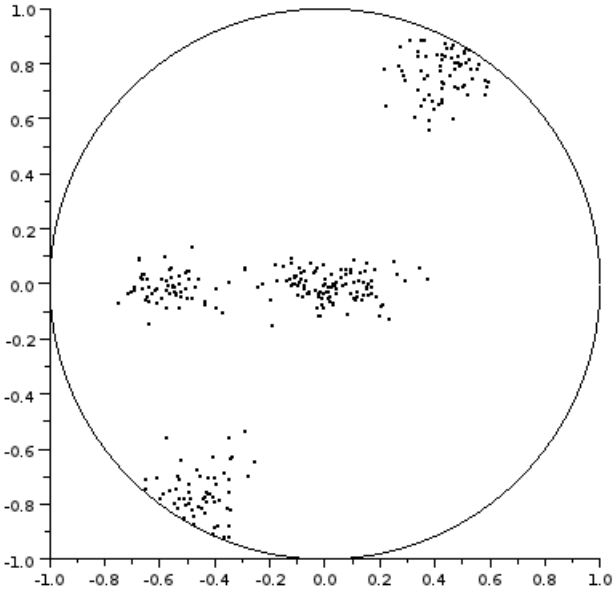

Figure 3: Diagramme de Wulff convention normale descendante

Remarques :

- Les couleurs représentent les différentes familles de points identifiées par l'algorithme, nous y reviendrons.

- Les droites ont pour extrémités un point de la sphère et le point sur le plan équatorial image par la projection stéréographique.

Remarques :

- Les points situés au milieu du disque représentent des vecteurs normaux proches du vecteur $\overrightarrow{e_z}$, ce sont donc des discontinuités qui sont quasiment horizontales.

- Les points situés proches du bord du disque représentent des discontinuités presque verticales.

- Des points très proches du bord et diamétralement opposés représentent des discontinuités ayant quasiment les mêmes orientations.

Cette dernière remarque est très importante et est symptomatique de toute convention où la normale est un vecteur orienté et non un axe (droite

orthogonale) coupant la sphère en deux points antipodaux. Elle justifie le paragraphe suivant.

2.4 Se passer de la convention normale montante dans le cadre des algorithmes présentés

La convention normale montante peut poser quelques problèmes, si l'on considère une famille de discontinuités d'orientations proches telles que leur vecteur normal unitaire montant est proche de l'équateur. Alors certains points seront éloignés sur la sphère alors qu'ils appartiennent à la même famille. Cela gêne nos algorithmes de regroupement en familles car nous cherchons à regrouper dans une même famille des points proches sur la sphère. Nous avons donc décidé d'abandonner les conventions normale montante et descendante. On représente alors chaque discontinuité par ses vecteurs normaux unitaires montant et descendant. Les points obtenus sont situés sur toute la sphère unité et cet ensemble de points possède alors une symétrie centrale par rapport au point 0. On cherche toujours à regrouper tous ces points en familles, seulement nous gardons à l'esprit que les familles de points seront à regrouper par deux, pour trouver finalement des familles de couples de points. Remarquons qu'il est plus satisfaisant d'un point de vue mathématique de représenter des directions de plan par des couples de points antipodaux car ces derniers représentent les directions des droites orthogonales.

L'inconvénient de cette méthode de représentation réside dans l'augmentation de la taille des données d'entrée pour les algorithmes qui auront à les traiter. Cela double naturellement le nombre de points à regrouper en familles. Cela peut être gênant si les algorithmes utilisés ont des complexités exponentielles.

3 Différentes lois de probabilités sur la sphère unité

Nous utiliserons essentiellement des algorithmes d'estimation de mélanges de lois de probabilités pour le regroupement en familles. Chaque famille correspondra à une composante du mélange de la loi envisagée. C'est pourquoi nous commençons par un bref rappel sur les mélanges de lois de probabilités.

3.1 Introduction aux mélanges de lois de probabilités

Soit $(\Omega, \mathcal{T})$ un ensemble mesurable. Soit $n \in \mathbb{N}$ et $\mu_1, ..., \mu_n$ des mesures de probabilités sur $(\Omega, \mathcal{T})$.

Définition : On appelle mélange des lois de probabilités $\mu_1, ..., \mu_n$ selon les proportions $p_1, ..., p_n \in \mathbb{R}_+$, telles que $\sum_i p_i = 1$ la mesure de probabilités

μ définie par :

$$\mu = \sum_{i=1}^{n} p_i \mu_i \tag{2}$$

Un algorithme d'estimation de mélange de probabilités à partir d'un échantillon peut alors nous permettre de regrouper les points de la sphère en familles, chaque famille correspond alors à une composante du mélange selon une certaine proportion.

Il faut cependant définir des mélanges de lois de probabilités sur la sphère unité qui soient susceptibles de reconnaître les formes voulues. A ce titre, définissons deux lois de probabilités sur la sphère unité.

3.2 Loi de Fisher

A strictement parler, la loi de Fihser n'est pas une loi de probabilité sur la sphère unité, mais sur $[0, \pi[\times[0, 2\pi[$ que nous identifierons à $\mathbb{S}\backslash\{S\}$ grâce au difféomorphisme des coordonnées sphériques déjà défini :

$$\mathbb{S}\backslash\{S\} = \left\{ \begin{pmatrix} \sin(\theta)\cos(\varphi) \\ \sin(\theta)\sin(\varphi) \\ \cos(\theta) \end{pmatrix}, (\theta, \varphi) \in [0, \pi[\times[0, 2\pi[\right\}$$

La loi de Fisher de concentration $K \in \mathbb{R}_+^*$ possède une densité par rapport à la mesure de Lebesgue sur $[0, \pi[\times[0, 2\pi[$, donnée par :

$$(\theta, \varphi) \longmapsto C \sin(\theta) e^{-K\cos(\theta)} \tag{3}$$

C est la constante de normalisation et vaut : $C = \frac{K}{2\pi(e^K - e^{-K})}$

Les figures 4 et 5 présentent, sous forme d'histogrammes, la répartition empirique des points sur la sphère selon la loi de Fisher centrée sur le pôle nord (vecteur $\overrightarrow{e_z}$) en fonction du paramètre θ.

Remarques :

- Le vecteur moyen de cette distribution est $\overrightarrow{e_z}$.
- Cette densité de probabilité ne dépend pas de φ, et donc possède une symétrie de révolution autour de l'axe Oz.

Cette densité de probabilité donc est centrée sur le pôle nord de la sphère. Cependant, les familles que nous cherchons à regrouper ne sont pas nécessairement centrées sur ce pôle. Il convient alors d'effectuer une rotation de cette densité, pour la centrer sur un vecteur moyen donné, qui sera un paramètre de la "nouvelle" loi de Fisher sur la sphère.

Nous remarquons également que la densité de Fisher telle qu'elle est définie est non nulle sur toute (là où elle est définie) la sphère. Au vu des familles de points qu'il faut regrouper, la zone située à l'antipode du vecteur

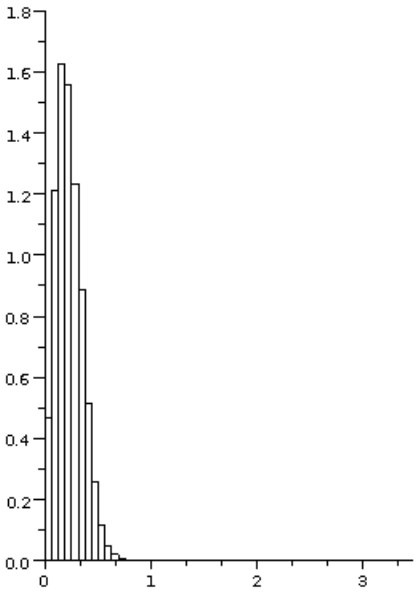

Figure 4: Histogramme de la densité de Fisher en fonction de θ pour $K = 30$

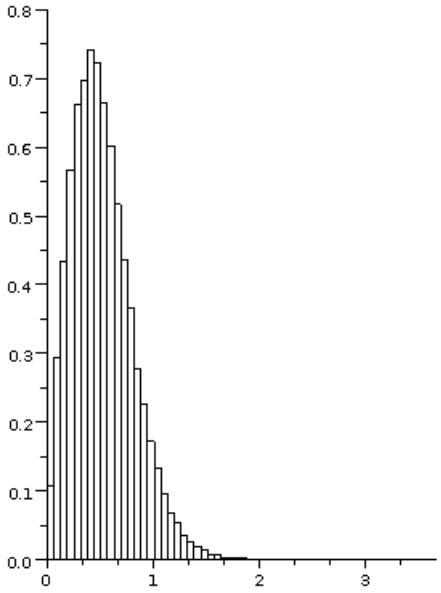

Figure 5: Histogramme de la densité de Fisher en fonction de θ pour $K = 6$

moyen d'une famille ne contient pas de point de cette famille. Pour insister sur ce constat, nous avons choisi de modifier un petit peu la loi de Fisher,

en la prenant nulle sur la demi-sphère antipodale au vecteur moyen. Elle s'obtient par rotation de la densité de loi suivante :

$$\begin{array}{ccc} [0,\pi[\times[0,2\pi[& \longrightarrow & \mathbb{R} \\ (\theta,\varphi) & \longmapsto & \begin{cases} C'\sin(\theta)e^{-K\cos(\theta)} \ \ si \ \ \theta \leq \frac{\pi}{2} \\ 0 \ \ sinon \end{cases} \end{array} \tag{4}$$

C' est la constante de normalisation et vaut : $C' = \frac{K}{2\pi(1-e^{-K})}$

Nous pourrons remarquer par la suite que cette forme de loi est avantageuse pour les algorithmes utilisés.

Les figures 6 et 7 représentent deux familles de points tirés selon des lois de Fisher avec les paramètres $K = 30$ et $K = 60$.

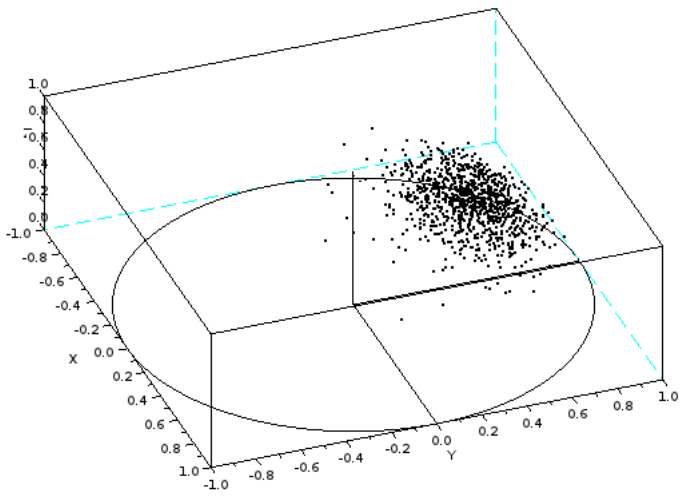

Figure 6: Représentation d'un échantillon tiré selon la loi de Fisher, $K = 30$

3.3 Projection stéréographique de lois multi-normales dans le plan tangent

Nous avons alors souhaité aller plus loin du point de vue des formes pouvant être reconnues par les algorithmes. Tandis que la loi de Fisher est symétrique par rotation autour de son vecteur moyen, nous envisageons à présent une loi sur la sphère unité permettant des formes plus variées et ne possèdant pas de symétrie de révolution.

C'est le cas par exemple des lois obtenues par projection stéréographique de lois multi-normales dans le plan tangent en un point de la sphère unité.

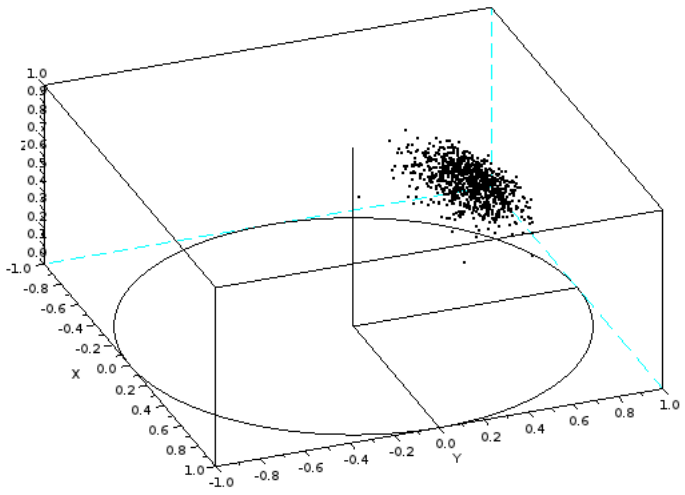

Figure 7: Représentation d'un échantillon tiré selon la loi de Fisher, $K = 60$

Plus précisément, on obtient un échantillon de ce type de loi en choisissant un vecteur moyen sur la sphère unité ; puis, en simulant dans le plan tangent (muni d'un repère dont l'origine est le vecteur moyen) à la sphère en ce point un échantillon issu d'une loi multi-normale plane de moyenne nulle ; enfin nous projetons cet échantillon sur la sphère stéréographiquement par rapport au point diamétralement opposé au vecteur moyen.

Les paramètres de cette loi sont donc :

- Un vecteur moyen.
- Une matrice de covariances.

Rappels mathématiques sur la projection stéréographique

Définition : Soit $\mathbb{S}$ la sphère unité de $\mathbb{R}^3$. Soit N et S deux points de $\mathbb{S}$ diamétralement opposés. Soit P le plan tangent à $\mathbb{S}$ en S, $P = \{Q \in \mathbb{R}^3 \ / \ \overrightarrow{OS} \cdot \overrightarrow{SQ} = 0\}$. On définit la projection stéréographique de $\mathbb{S}\backslash\{S\}$ sur P par l'application qui associe à un point de $Q \in \mathbb{S}\backslash\{S\}$ l'unique point d'intersection entre le plan P et la droite (NQ).

Pour simplifier, supposons que $S = (0, 0, -1)$ et $N = (0, 0, 1)$. On a alors l'écriture suivante de la projection stéréographique :

$$\begin{array}{rccc} \pi_N : & \mathbb{S}\backslash\{S\} & \longrightarrow & P \\ & (x, y, z) & \longmapsto & (\frac{-2x}{z-1}, \frac{-2y}{z-1}, -1) \end{array} \tag{5}$$

Il est également important de préciser quel repère nous choisissons pour le plan tangent. Soit $Q \in \mathbb{S}$ et P le plan tangent en Q. En cherchant la simplicité, nous avons choisi de prendre comme coordonnées dans P celles obtenues en considérant la base $(\overrightarrow{e_\theta}, \overrightarrow{e_\varphi})$ issues de la base des coordonnées sphériques en Q déjà présentées ($P = Vect_{\mathbb{R}}(\overrightarrow{e_\theta}, \overrightarrow{e_\varphi})$).

Déterminons la formule (que nous avons ensuite utilisée pour les simulations) pour passer d'un point de coordonnées $M(a, b) \in \mathbb{R}^2$ du plan tangent en $V_m(x_m, y_m, z_m) \in \mathbb{S}$ à sa projection stéréographique $M_S(x, y, z) \in \mathbb{S}$.

Par définition des coordonnées de M dans le plan tangent on a :

$$\overrightarrow{OM} = \overrightarrow{OV_m} + a\overrightarrow{e_\theta} + b\overrightarrow{e_\varphi} \tag{6}$$

Il existe alors $\lambda \in \mathbb{R}$ tel que

$$\overrightarrow{OM_S} = -V_m + \lambda\overrightarrow{Q_m M} \tag{7}$$

où Q_m est le point diamétralement opposé à V_m.

Cherchons alors l'expression de λ. On a $M_S \in \mathbb{S} \Leftrightarrow \|\overrightarrow{OM_S}\| = 1$. Et comme $(V_m, \overrightarrow{e_\theta}, \overrightarrow{e_\varphi})$ est une base orthonormée de $\mathbb{R}^3$, il est aisé de calculer $\|\overrightarrow{OM_S}\|$

$$\begin{array}{rcl} \|\overrightarrow{OM_S}\| = 1 & \Leftrightarrow & (2\lambda - 1)^2 + (a\lambda)^2 + (b\lambda)^2 = 1 \\ & \Leftrightarrow & \lambda = \frac{1}{1 + \frac{a^2+b^2}{4}} \end{array} \tag{8}$$

On repère V_m par ses coordonnées sphériques (θ, φ), on a alors l'expression des vecteurs $\overrightarrow{e_\theta}$ et $\overrightarrow{e_\varphi}$

$$\overrightarrow{e_\theta} = \begin{pmatrix} \cos(\theta)\cos(\varphi) \\ \cos(\theta)\sin(\varphi) \\ -\sin(\theta) \end{pmatrix} \tag{9}$$

$$\begin{array}{rl} Si\ \theta < \frac{\pi}{2} & \overrightarrow{e_\varphi} = \begin{pmatrix} -\cos(\varphi) \\ \sin(\varphi) \\ 0 \end{pmatrix} \\ Si\ \theta \geq \frac{\pi}{2} & \overrightarrow{e_\varphi} = \begin{pmatrix} \cos(\varphi) \\ -\sin(\varphi) \\ 0 \end{pmatrix} \end{array} \tag{10}$$

Nous trouvons, après calcul les expressions (11), qui permettront de simuler un échantillon de points sur la sphère à partir de la simulation directe (grâce au logiciel Scilab) d'une loi multi-normale dans un plan.

$$Si\ \theta < \frac{\pi}{2} \quad \begin{cases} x = (2\lambda - 1)x_m + \lambda(a\cos(\theta)\cos(\varphi) - b\sin(\varphi)) \\ y = (2\lambda - 1)y_m + \lambda(a\cos(\theta)\sin(\varphi) + b\cos(\varphi)) \\ z = (2\lambda - 1)z_m - a\lambda\sin(\theta) \end{cases} \tag{11}$$

$$Si\ \theta \geq \frac{\pi}{2} \quad \begin{cases} x = (2\lambda - 1)x_m + \lambda(a\cos(\theta)\cos(\varphi) + b\sin(\varphi)) \\ y = (2\lambda - 1)y_m + \lambda(a\cos(\theta)\sin(\varphi) - b\cos(\varphi)) \\ z = (2\lambda - 1)z_m - a\lambda\sin(\theta) \end{cases} \tag{12}$$

$$\text{où } \lambda = \frac{1}{1+\frac{a^2+b^2}{4}}$$

4 Premier algorithme de regroupement en familles

Nous possèdons déjà un algorithme de regroupement en familles non probabiliste. Nous le présenterons dans cette section afin de donner une idée de ce qui est faisable de façon relativement élémentaire.

Description de l'algorithme :

La distance utilisée sur la sphère est celle du plus court chemin, ou géodésique. Elle est donnée par la formule :

Pour $P, Q \in \mathbb{S}$ (de centre O) :

$$d(P,Q) = \arccos(\overrightarrow{OP} \cdot \overrightarrow{OQ}) \tag{13}$$

Nous ferons l'amalgame entre les points de la sphère et les vecteurs unitaires, la distance sera alors notée indifféremment $d(\overrightarrow{OP}, \overrightarrow{OQ})$.

A partir d'un échantillon de points de la sphère, que nous souhaitons regrouper en familles, nous appliquons :

- Nous formons autant de familles qu'il y a de points, un point dans chaque famille.
- Nous calculons les vecteurs moyens correspondant à chaque famille.
- Nous cherchons les deux vecteurs les plus proches.
- Nous fusionnons alors les deux familles associées.

Ce procédé est itéré jusqu'à ce qu'il n'y ait plus qu'une famille.

A chaque étape, nous calculons pour chaque famille un indicateur $\sigma_{i,j}$, où i est l'indice correspondant au numéro de la famille et k est celui correspondant au numéro de l'itération, selon la formule (14)

$$\sigma_{i,k} = \frac{1}{N_i}\sqrt{\sum_n \sum_m d(\overrightarrow{OP_n} \cdot \overrightarrow{OP_m})^2} \tag{14}$$

où N_i est le nombre de points de la famille i et $\overrightarrow{OP_n}$ est un point parcourant cette famille.

Cet indicateur nous permet alors d'harmoniser l'écart type de chaque famille, et de décider a posteriori le nombre de familles que nous souhaitons garder.

5 L'algorithme EM (Expectation Maximization)

5.1 Présentation de l'algorithme

L'algorithme EM (Expectation Maximisation), mis en place par Dempster, Laird et Rubin en 1977, est un algorithme d'estimation des paramètres d'un mélange de lois de probabilités de la forme $p_1\mu_1 + \cdots + p_n\mu_n$, où les mesures de probabilités μ_i sont dans un même modèle paramétrique et p_i sont les proportions du mélange. EM cherche alors à donner à partir d'un échantillon, les paramètres estimés de chaque mesure μ_i ainsi que sa proportion p_i dans le mélange.

Un exemple simple : le mélange de lois multi-normales dans un plan

Soit $n \in \mathbb{N}^*$ qui représentera le nombre de composantes du mélange. On considère le modèle paramétrique suivant

$$\mathcal{P} = \left\{ \sum_{i=1}^{n} p_i \mathcal{N}_i \ / \ p_i \in \mathbb{R}_+, \ \sum_i p_i = 1 \ et \ \mathcal{N}_i \sim \mathcal{N}_2(\mu, \Sigma) \ , \ \mu_i \in \mathbb{R}^2 \ et \ \Sigma_i \in \mathcal{S}_n^+(\mathbb{R}) \right\}$$

Soit $Y = (y_1, \ldots, y_m) \in (\mathbb{R}^2)^m$ un échantillon de points du plan pour lequel nous souhaitons reconnaître un mélange de notre moèle paramétrique $\mathcal{P}$.

L'algorithme EM donnera alors une estimation des proportions du mélange p_i ainsi que pour chaque loi $\mathcal{N}_i$ multi-normale une estimation de sa moyenne μ_i et de sa matrice de covariances Σ_i.

Description de l'algorithme

A partir d'un échantillon $Y = (y_1, \ldots, y_m)$, on cherche à estimer les paramètres d'un mélange issu du modèle statistique $\{\sum_{k=1}^{n} p_k \mathcal{L}_k(x_{k,1}, \ldots, x_{k,r})\}$ où les $\mathcal{L}_k$ sont des lois de probabilités dépendantes des paramètres $x_{k,1}, \ldots, x_{k,r}$.

Il y a néanmoins quelques conditions à l'application de l'algorithme EM :

- Connaître des estimateurs spécifiques (qui utilisent les probabilités a posteriori que nous verrons juste après) pour les $x_{k,j}$ que nous noterons $\hat{x}_{k,j}$.
- Chaque $\mathcal{L}_k$ possède une densité connue en fonction des paramètres $x_{k,1}, \ldots, x_{k,r}$ que nous noterons $f_k(x_{k,1}, \ldots, x_{k,r})$.

Nous avons besoin d'une famille de vecteurs dit de probabilité a posteriori qui à chaque point de l'échantillon y_j associe $(z_{j,1}, \ldots, z_{j,n}) \in \mathbb{R}^n$ où $z_{j,k}$ représente la probabilité estimée par l'algorithme que y_j soit attribué à la

loi $\mathcal{L}_k$. Ces probabilités a posteriori sont modifiées à chaque itération de EM.

L'algorithme consiste alors en une boucle d'un nombre d'itérations déterminé à l'avance contenant les étapes suivantes :

Début de la boucle

† Estimation des proportions du mélange : Pour chaque famille que nous voulons estimer $k \in [|\ 1, n\ |]$

$$p_k = \frac{1}{N_k} \sum_{j=1}^{m} z_{j,k}$$

où $N_k = \sum_{j=1}^{m} z_{j,k}$

† Estimation des paramètres de la loi :

$$\hat{x}_{k,l} = \ldots \quad l \in [|\ 1, r\ |]\ \ k \in [\|1, n\|]$$

Cette étape dépend de la loi utilisée et des probabilités a posteriori choisies.

† Estimation des probabilités a posteriori :

$$z_{j,k} = \frac{p_k f_k(\hat{x}_{k,1}, \ldots, \hat{x}_{k,r})(y_j)}{\sum_{q=1}^{n} p_q f_q(\hat{x}_{q,1}, \ldots, \hat{x}_{q,r})(y_j)} \quad j \in [|\ 1, m\ |]\ \ k \in [|\ 1, n\ |]$$

Fin de la boucle.

Remarques :

- On peut suivre l'évolution de la log-vraisemblance pour savoir si l'algorithme s'est stabilisé car cette dernière croît au fur et à mesure des itérations et finit par se stabiliser. La formule de la log-vraisemblance est donnée par :
$$Log\_v = \sum_{j=1}^{m} \ln(\sum_{k=1}^{n} p_k f_k(\hat{x}_{k,1}, \ldots, \hat{x}_{k,r})(y_j))$$
- Les probabilités a posteriori peuvent être utilisées pour établir un critère pour savoir si un point a bien été reconnu comme appartenant à une famille ou si l'algorithme n'arrive pas à trouver une famille pour ce point. Par exemple si pour le point y_j on a $z_{j,k} > 0.95$, alors on peut dire que l'algorithme a attribué la famille k (la loi $\mathcal{L}_k$ pour la simulation de ce point.

5.2 Essai de l'algorithme pour différents mélanges

Nous allons à présent expliciter la forme des estimateurs utilisés dans le cas de la loi de Fisher et dans celui des mélanges de lois multi-normales dans le plan tangent.

5.3 Mélange de lois multi-normales dans le plan

Achevons l'exemple simple des mélanges de lois multi-normales dans le plan qui a servi dans le paragraphe précédent. Gardons par conséquent les mêmes notations.

Dans ce cas les deux paramètres de la loi sont la moyenne et la matrice de covariances. Les estimateurs utilisés sont :

Pour la moyenne :

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{j=1}^{m} z_{j,k} y_j$$

Pour la matrice de covariances :

$$\hat{\Sigma}_k = \frac{1}{N_k} \sum_{j=1}^{m} z_{j,k} (y_j - \hat{\mu}_k)^t (y_j - \hat{\mu}_k)$$

Nous avons programmé en Scilab le cas des mélanges de lois multi-normales dans le plan pour tester l'algorithme EM et mettre en place une première structure de programme pouvant être réutilisée par la suite.

Pour tester le fonctionnement de l'algorithme EM de manière simple et lisible, nous avons choisi faire tourner EM sur un échantillon construit en simulant des points tirés selon un mélange de lois multi-normales dans le plan. Dans le cas d'une application à la géologie, il n'y a aucune raison évidente pour que l'échantillon de points que l'on possède soit issu d'un mélange de lois de probabilités. Cependant c'est une approche possible de classification des discontinuités.

Protocole

Nous choisissons dans un premier temps les "vrais" paramètres du mélange de lois. Nous simulons un échantillon à l'aide de Scilab. Après avoir exécuté EM sur cette échantillon, nous utilisons les paramètres estimés pour simuler un échantillon beaucoup plus conséquent (représenté par des petits points). Nous finissons par représenter la superposition des deux échantillons ainsi simulés sur un même graphique.

Représentations graphiques des résultats

Les figures 8,9 et 10 donnent des exemples de résultats que nous avons obtenus.

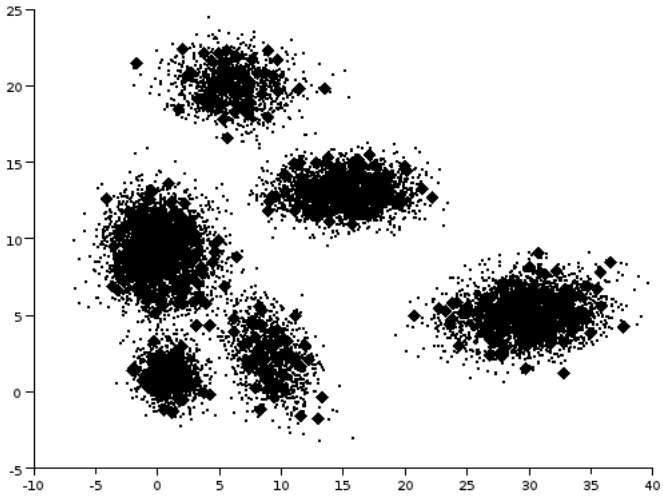

Figure 8: Résultats EM mélange de lois multi-normales avec 6 composantes nettement séparées

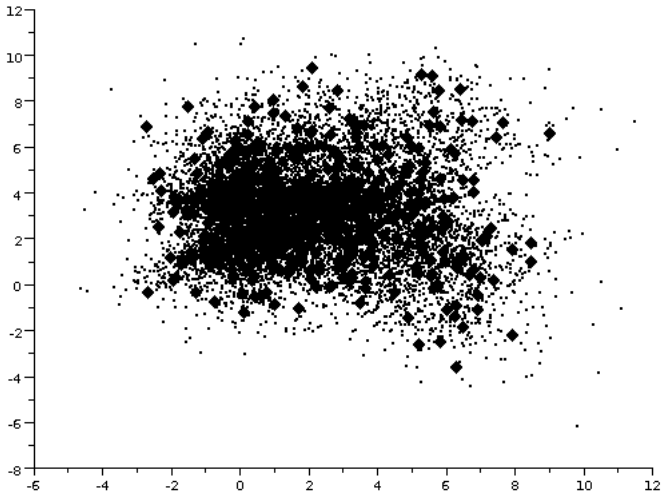

Figure 9: Résultats EM mélange de lois multi-normales avec 6 composantes

5.3.1 mélanges de lois de Fisher

Commençons par expliciter les formules des estimateurs pour les paramètres de la loi de Fisher. Pour cela, introduisons d'abord quelques notations.

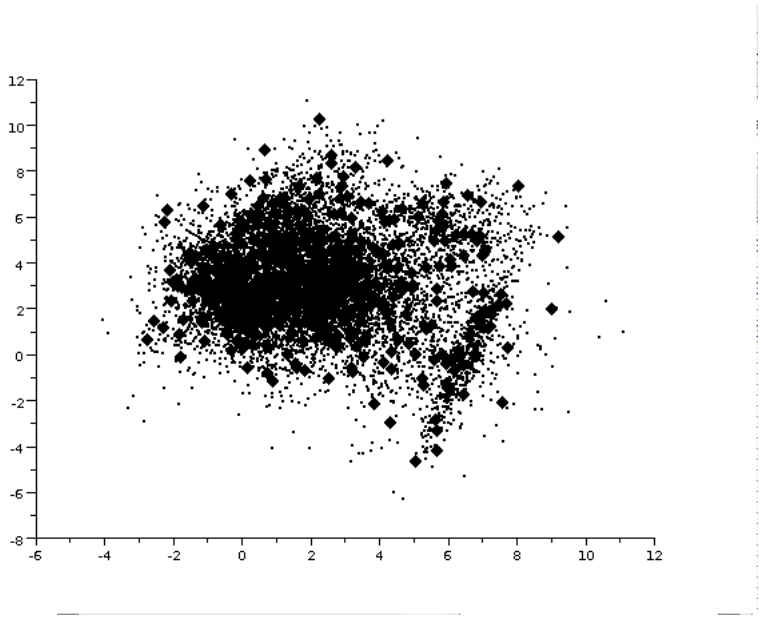

Figure 10: Résultats EM mélange de lois multi-normales avec 6 composantes

Nous notons notre échantillon $Y = (y_1, \ldots, y_m) \in \mathbb{S}^m$. Les composantes du mélange sont numérotées de 1 à n, pour la composante k, le vecteur moyen est noté $v_{m,k} \in \mathbb{R}^3$ et sa concentration K_k.

Le vecteur moyen est une moyenne pondérée des points de la sphère

$$v_{m,k} = \frac{1}{N_k} \sum_{j=1}^{m} z_{j,k} y_j$$

Pour la concentration K

$$K_k = \frac{1}{1 - \|v_{m,k}\|}$$

Remarques :

- Pour avoir une idée de ce qu'est la concentration K, nous pouvons remarquer que si les points attribués à une même composante sont très proches les uns des autres sur la sphère, alors la norme du vecteur moyen est proche de 1 (mais toujours inférieure) et la concentration est alors très grande.

- La densité de la loi de Fisher fait intervenir la fonction exponentielle dont l'argument a le même ordre de grandeur que K. Cela peut être

à l'origine de problèmes lors de l'exécution de l'algorithme. En effet, lors des premières itérations, il se peut que des familles aient très peu de points, ce qui entraîne des valeurs de K qui peuvent atteindre 100. Or les logiciels de calculs scientifiques ne maîtrisent pas avec précision de telles valeurs.

Visualisons à présent quelques résultats selon le même protocole que précédemment.

Représentations graphiques des résultats

Commençons par un exemple pour lequel les familles sont aisément reconnaissables à l'oeil nu. Dans la figure 11, nous avons représenté les vecteurs normaux des plans de fracture de notre échantillon. Nous pouvons remarquer que les points sont déjà dédoublés. La figure est donc symétrique par rapport au point O comme nous l'avions déjà mentionné. Le diagramme de Wulff est donné à la figure 12. On y distingue deux familles car le dédoublement n'apparaît pas sur le diagramme de Wulff à cause de la convention normale montante.

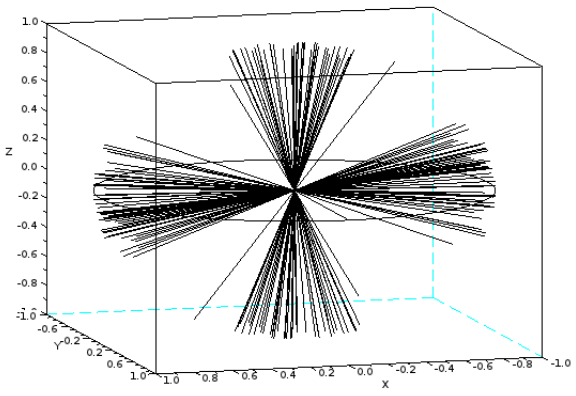

Figure 11: essai EM (Loi de Fisher) : Vecteurs normaux des plans de fracture de l'échantillon

Sur la figure 13, nous avons représenté les résultats obtenus après l'exécution de l'algorithme. Nous avons tracé dans différentes couleurs les familles que l'algorithme a formées.

En fait, l'algorithme ne départage pas aussi simplement les points. En effet, nous avons vu que EM ne donne qu'un vecteur de probabilités a posteriori pour chaque point de l'échantillon. Ainsi pour chaque point nous avons la probabilité qu'il soit issu d'une certaine composante. Pour départager les

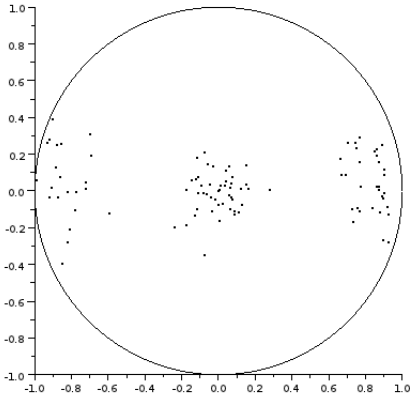

Figure 12: essai EM (Loi de Fisher) : Diagramme de Wulff de l'échantillon

points nous avons mis en place un critère simple qui consiste à dire qu'un point appartient à une famille (composante) si sa probabilité a posteriori d'appartenir à cette famille est supérieure à 0.95.

Cela nous donne également un certain nombre de points indécis pour lesquels nous n'avons pas pu déterminer à quelles familles ils pourraient appartenir. Ce taux de points indécis fournit une information sur l'imbrication des composantes estimées.

Résultats numériques :

Présentons à présent les résultats numériques que nous avons obtenus dans ce cas ci.

Premièrement, 100% des points ont été reconnus comme appartenant à une famille avec une probabilité a posteriori à plus de 0.95. Ce résultat n'est pas étonnant puisque les familles sont très séparées.

	paramètres de simulation de l'échantillon	paramètres évalués par EM
vecteurs moyens	$v^0_{m,1} = (0, 0, 1)$	$v_{m,1} = (-0.026, 0.062, 0.939)$
	$v^0_{m,2} = (1, 0, 0)$	$v_{m,2} = (0.966, 0.012, 0.019)$
proportions	$p^0_1 = 0.5$	$p_1 = 0.57$
	$p^0_2 = 0.5$	$p_2 = 0.43$
concentrations	$K^0_1 = 30$	$K_1 = 29.6$
	$K^0_2 = 20$	$K_2 = 17.2$

Passons maintenant à un mélange beaucoup plus difficile. Comme nous pouvons le voir sur le figure 14, les familles ne sont pas reconnaissables à l'oeil nu. Pour aider à leur visualisation nous les avons coloriées au moment de les

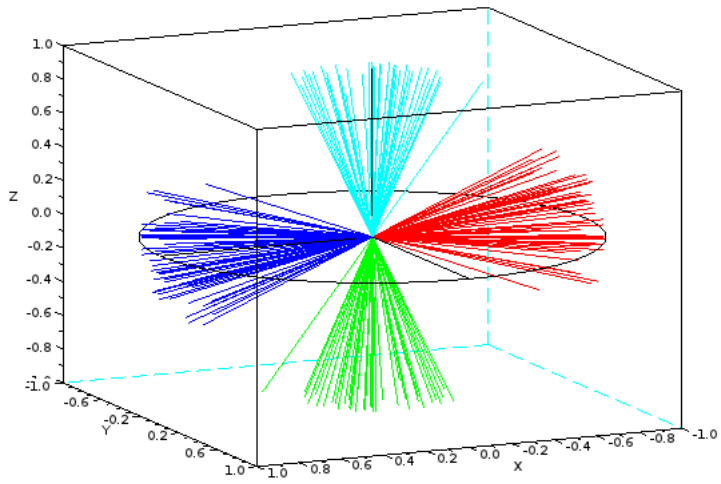

Figure 13: essai EM (Loi de Fisher) : Visualisation du partage en familles des vecteurs normaux des plans de fracture

simuler. Il y a dans ce mélange trois composantes. La figure 15 montre le diagramme de Wulff où l'on peut distinguer les trois familles plus aisément, mais cela reste difficile. Remarquons également que le dédoublement des points induit en réalité six familles à reconnaître pour l'algorithme.

Ce cas présente les limites d'utilisation de notre algorithme programmé en Scilab. Nous allons présenter des résultats obtenus à partir d'un échantillon très grand, 800 points au total. Il faut noter que ce nombre excède les tailles d'échantillonnage habituellement rencontrées dans nos application, qui sont aux alentours de 300. Avec un nombre de points plus élevé, nous donnons en quelque sorte plus d'informations à l'algorithme, qui par conséquent, va être capable d'estimer avec plus de précision les paramètres du mélange.

Avant de présenter les résultats numériques et graphiques obtenus avec un échantillon de 800 points présentons les difficultés rencontrées avec un échantillon de taille habituelle de 300 points, pour des mélanges difficiles tels que celui-ci.

Difficultés rencontrées :

- L'algorithme peut faire disparaître une composante du mélange. C'est-à-dire qu'une des proportions devient nulle ou trop petite, ou d'un autre point de vue, une des composantes prend le dessus sur les autres.

- Résultats non symétriques. Il arrive, pour des raisons que nous ne comprenons pas très bien, que les résultats trouvés n'aillent pas par

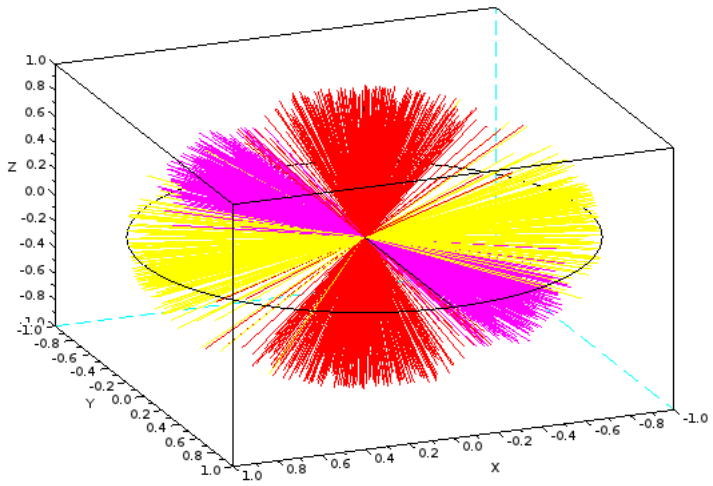

Figure 14: essai EM bis (Loi de Fisher) : Visualisation des vecteurs normaux des plans de fracture de l'échantillon

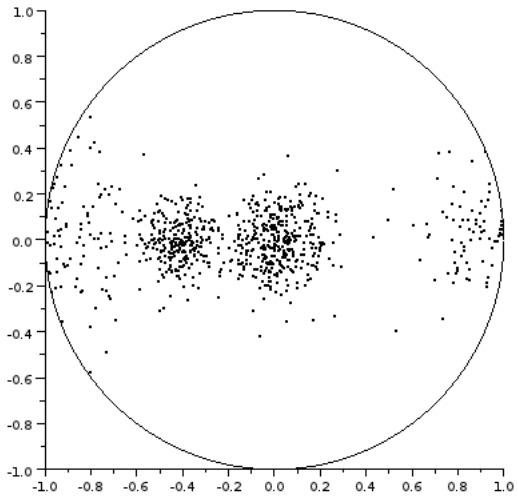

Figure 15: essai EM bis (Loi de Fisher) : Visualisation du diagramme de Wulff de l'échantillon

deux. L'échantillon, après avoir été dédoublé, est symétrique par rapport au point O, l'algorithme devrait alors travailler de la même façon

sur les composantes symétriques. Or, un résultat que nous ne pouvons pas regrouper deux par deux est inutilisable. En effet, nous nous attendons à ce que l'algorithme estime les mêmes concentrations et des vecteurs moyens opposés pour deux familles symétriques. Si le regroupement est impossible à la fin, nous ne pouvons pas identifier des familles de bipoints antipodaux, c'est-à-dire nos droites normales aux plans de fracture.

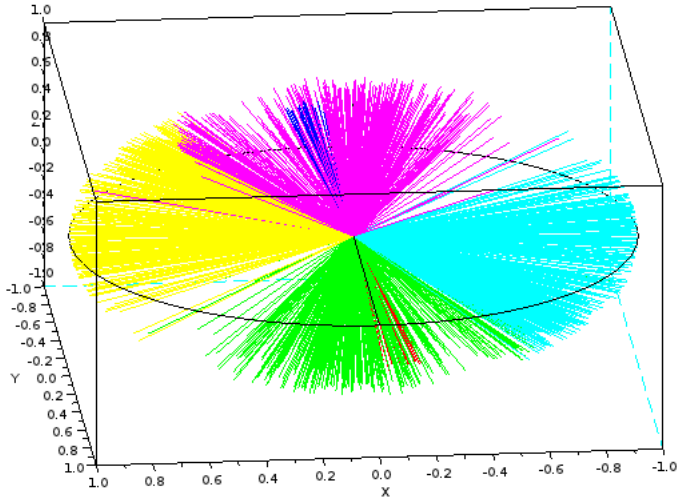

Figure 16: essai EM bis (Loi de Fisher) : Visualisation du partage en familles des vecteurs normaux des plans de fracture

Résultats numériques :

	paramètres de simulation de l'échantillon	paramètres évalués par EM
vecteurs moyens	$v^0_{m,1} = (0, 0, 1)$	$v_{m,1} = (0.0026; 0.0152; 0.9655)$
	$v^0_{m,2} = (0, 7071, 0, 0, 7071)$	$v_{m,2} = (0.1543; 0.0111; 0.8777)$
	$v^0_{m,3} = (1, 0, 0)$	$v_{m,3} = (0.8191; 0.0107; 0.3503)$
proportions	$p^0_1 = 0.5$	$p_1 = 0.10$
	$p^0_2 = 0.3$	$p_2 = 0.44$
	$p^0_3 = 0.2$	$p_3 = 0.46$
concentrations	$K^0_1 = 20$	$K_1 = 29.04$
	$K^0_2 = 50$	$K_2 = 9.16$
	$K^0_3 = 20$	$K_3 = 9.19$

Nous remarquons que les familles reconnues ne sont pas celles attendues. Une famille avec peu de points proches (grande concentration) s'est formée alors que les autres familles sont restées très étalées (faibles concentrations) regroupant la majorité des points.

L'algorithme ne donne pas ce regroupement avec une bonne précision, seulement 50% des points sont reconnus avec une probabilité a posteriori d'au moins 0.8.

5.3.2 Mélanges de lois multi-normales dans le plan tangent

Pour des raisons déjà évoquées d'une plus grande diversité de reconnaissance de formes, nous avons souhaité programmer l'algorithme EM avec la loi mélange de lois multi-normales projetées du plan tangent à la sphère en un point donné sur la sphère.

Rappelons brièvement les paramètres dont nous avons besoin pour décrire ces lois de probabilités.

- Un vecteur moyen : $v_{m,k}$ estimé avec :

$$v_{m,k} = \frac{1}{N_k} \sum_{j=1}^{m} z_{j,k} y_j$$

- Une matrice de covariances : Σ_k estimée avec :

$$\Sigma_k = \frac{1}{N_k} \sum_{j=1}^{m} z_{j,k}\, \tilde{y}_j^t \tilde{y}_j$$

  où $\tilde{y}_j$ est le vecteur projeté de y_j sur le plan tangent à la sphère au point indiqué par le vecteur moyen $v_{m,j}$.

Notons que nous utilisons le même estimateur de la moyenne que pour la loi de Fisher et le même estimateur de la matrice de covariances que pour une loi multi-normale dans le plan, à la différence près que les calculs se font dans le plan tangent au point de la sphère donné par le vecteur moyen.

Présentons à présent quelques résultats.

Un cas simple : composantes très séparées

La figure 17 montre l'échantillon de départ, les points ont déjà été dédoublés.

Sur la figure 18, qui présente le diagramme de Wulff de cet échantillon, on distingue nettemement les 4 familles.

Sans surprise, les familles trouvées et présentées en couleur sur la figure 19 sont les bonnes et nous verrons que les paramètres du mélange sont très

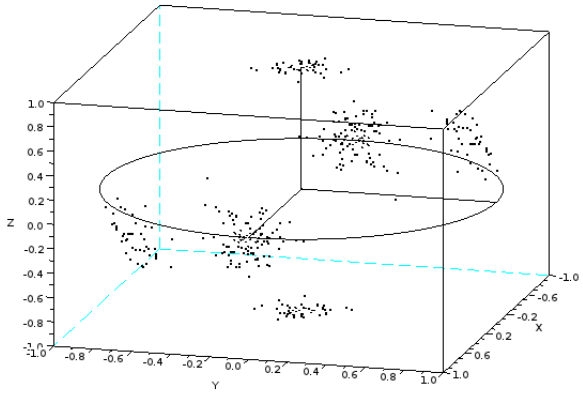

Figure 17: essai EM 1 : Projection Gaussienne : visualisation de l'échantillon de départ

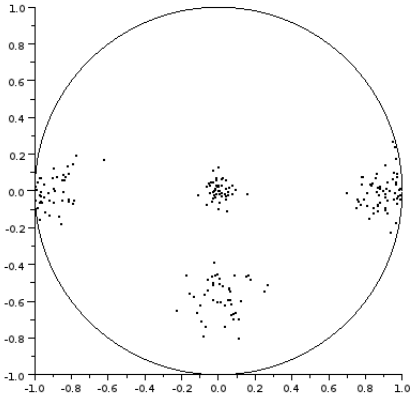

Figure 18: essai EM 1 : Projection Gaussienne : Diagramme de Wulff de l'échantillon de départ

bien estimés. Nous avons 100% des points qui sont attribués à une famille avec une probabilité a posteriori d'au moins 0.95.

Avec les figures 19 et 20, nous comparons l'échantillon de départ avec un échantillon simulé à partir des estimations, de taille plus importante. Des gros points sont utilisés pour représenter les points appartenant à l'échantillon de départ, les petits points correspondent à ceux de l'échantillon simulé.

Nous avons également tracé l'évolution de la log-vraisemblance au fur et

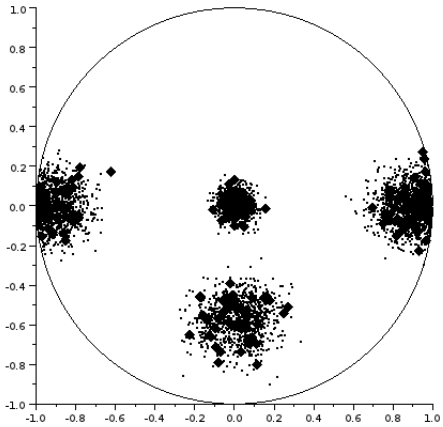

Figure 19: essai EM 1 : Projection Gaussienne : Visualisation des familles trouvées par l'algorithme

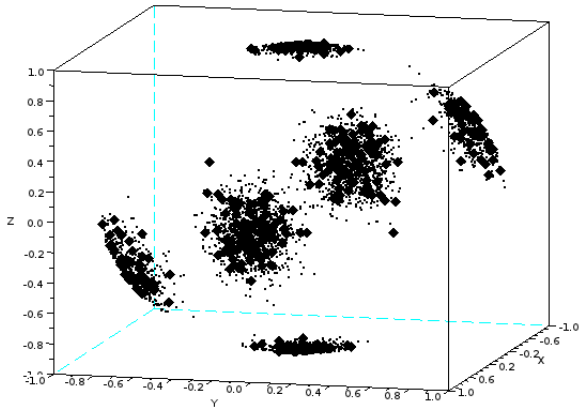

Figure 20: essai EM 1 : Projection Gaussienne : Comparaison de l'échantillon de départ avec un échantillon simulé

à mesure des itérations de l'algorithme EM. Nous pouvons alors remarquer que la log-vraisemblance croît, comme nous l'avions remarqué plus haut, et qu'elle atteint une valeur limite après une vingtaine d'itérations dans ce cas ci. Remarquons également que le nombre d'itérations qu'il faut pour atteindre une valeur limite pour la log-vraisemblance dépend fortement de la difficulté du mélange, mais nous sommes assurés que la fonction obtenue est croissante.

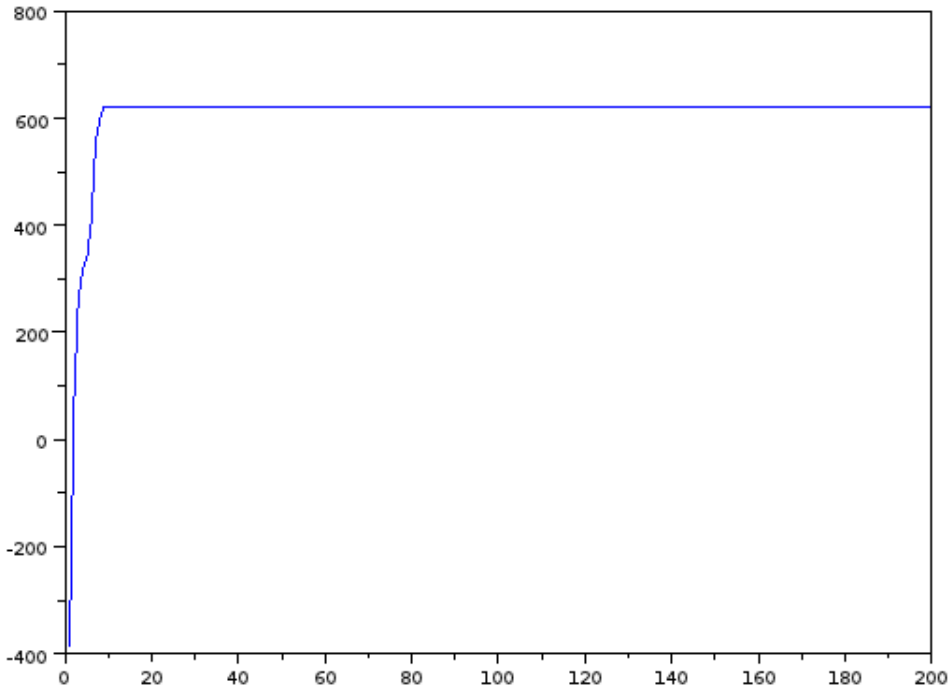

Figure 21: essai EM 1 : Projection Gaussienne : tracé de la log-vraisemblance

Résultats numériques :

Ici les résultats sont très bons, cependant le mélange est très simple et n'est pas très intéressant du point de vue des applications.

paramètres de simulation de l'échantillon	paramètres évalués par EM
$v^0_{m,1} = (0, 0, 1)$	$v_{m,1} = (-0.0132, -0.0069, 0.9999)$
$v^0_{m,2} = (0, 0.866, 0.5)$	$v_{m,2} = (0.0250, 0.8659, 0.4996)$
$v^0_{m,3} = (1, 0, 1)$	$v_{m,3} = (0.9998, 0.0033, 0.0186)$
$p^0_1 = 0.5$	$p_1 = 0.525$
$p^0_2 = 0.2$	$p_2 = 0.225$
$p^0_3 = 0.3$	$p_3 = 0.25$
$\Sigma^0_1 = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}$	$\Sigma_1 = \begin{pmatrix} 0.0081 & 0.0003 \\ 0.0003 & 0.0091 \end{pmatrix}$
$\Sigma^0_2 = \begin{pmatrix} 0.02 & 0 \\ 0 & 0.01 \end{pmatrix}$	$\Sigma_2 = \begin{pmatrix} 0.0198 & 0.0000 \\ 0.0000 & 0.0103 \end{pmatrix}$
$\Sigma^0_3 = \begin{pmatrix} 0.03 & 0 \\ 0 & 0.01 \end{pmatrix}$	$\Sigma_3 = \begin{pmatrix} 0.0209 & 0.0016 \\ 0.0016 & 0.0251 \end{pmatrix}$

Un cas plus difficile de mélange

Nous simulons un échantillon à quatre composantes. Nous rappelons qu'après dédoublement des points l'algorithme devra reconnaître un échantillon à huit composantes.

Les figures 22 et 23 donnent un aperçu de l'échantillon simulé. Nous pouvons remarquer que les familles ne sont pas facilement distinguables à l'oeil.

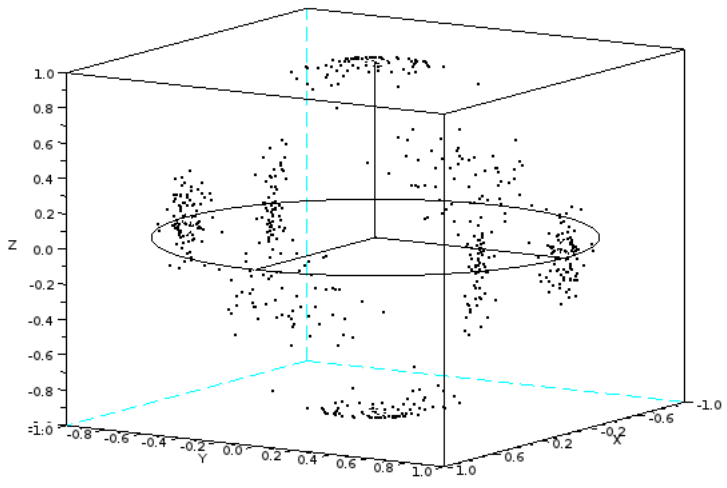

Figure 22: essai EM 2 : Visualisation des points de l'échantillon

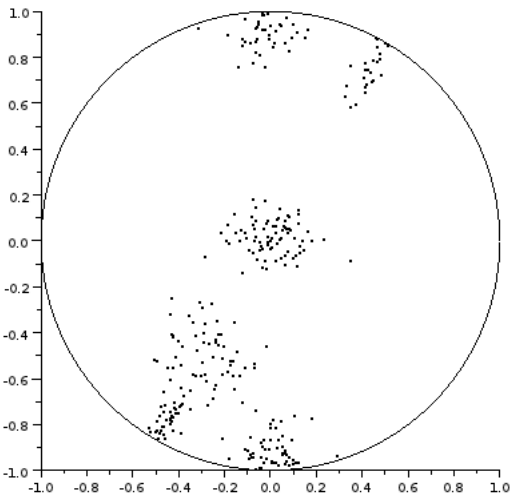

Figure 23: essai EM 2 : Diagramme de Wulff de l'échantillon

Pour aider le lecteur à visualiser où se trouvent les familles, nous avons tracé grossièrement à la main dans des couleurs différentes les familles qui apparaissent sur le diagramme de Wulff.

L'algorithme regroupe les familles comme nous l'attendions, c'est une réussite. Attendons maintenant les résultats numériques.

Comme pour l'exemple précédent, nous simulons un autre échantillon à partir des estimations des paramètres. Puis nous superposons cet échantillon avec celui de départ, dans un diagramme de Wulff (figure 26) et sur la sphère (figure 27). Les résultats sont convaincants à l'oeil.

Résultats numériques :

Avant tout, l'algorithme annonce avoir reconnu 94.6% des points avec une probabilité a posteriori d'au moins 0.95. Ce qui est très bon, compte tenu des exemples étudiés jusqu'à présent.

Résumons les résultats dans le tableau :

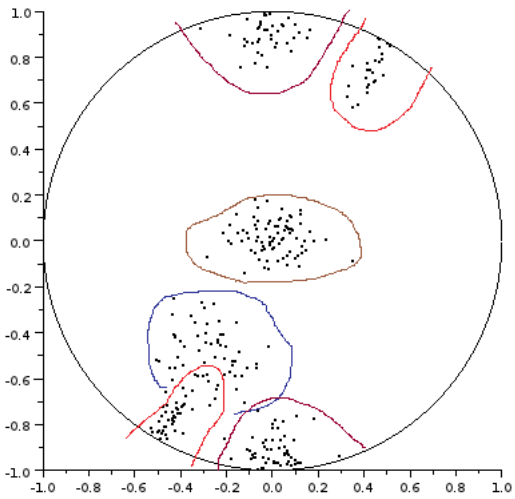

Figure 24: essai EM 2 : Visualisation manuelle des familles présentes sur le diagramme de Wulff

paramètres de simulation de l'échantillon	paramètres évalués par EM
$v^0_{m,1} = (0.5, 0.866, 0.5)$	$v_{m,1} = (0.4982, 0.8669, 0.0187)$
$v^0_{m,2} = (0, 1, 0)$	$v_{m,2} = (0.0189, 0.9998, 0.0080)$
$v^0_{m,3} = (0.433, 0.75, 0.5)$	$v_{m,3} = (0.4242, 0.7503, 0.5071)$
$v^0_{m,4} = (0, 0, 1)$	$v_{m,4} = (0.0170, 0.0286, 0.9994)$
$p^0_1 = 0.2$	$p_1 = 0.215$
$p^0_2 = 0.3$	$p_2 = 0.315$
$p^0_3 = 0.2$	$p_3 = 0.200$
$p^0_4 = 0.3$	$p_4 = 0.2700$
$\Sigma^0_1 = \begin{pmatrix} 0.02 & 0 \\ 0 & 0.01 \end{pmatrix}$	$\Sigma_1 = \begin{pmatrix} 0.0297 & 0.0009 \\ 0.0009 & 0.0012 \end{pmatrix}$
$\Sigma^0_2 = \begin{pmatrix} 0.02 & 0 \\ 0 & 0.01 \end{pmatrix}$	$\Sigma_2 = \begin{pmatrix} 0.0145 & 0.0009 \\ 0.0009 & 0.0093 \end{pmatrix}$
$\Sigma^0_3 = \begin{pmatrix} 0.03 & 0 \\ 0 & 0.03 \end{pmatrix}$	$\Sigma_3 = \begin{pmatrix} 0.0261 & 0.0044 \\ 0.0044 & 0.0357 \end{pmatrix}$
$\Sigma^0_4 = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.02 \end{pmatrix}$	$\Sigma_4 = \begin{pmatrix} 0.0249 & 0.0097 \\ 0.0097 & 0.0411 \end{pmatrix}$

Pour clore les exemples avec les projections de lois multi-normales dans le plan tangent, poussons EM à sa limite avec ce dernier cas qui pose problème.

On considère un mélange à trois composantes seulement, mais très enchevêtrées.

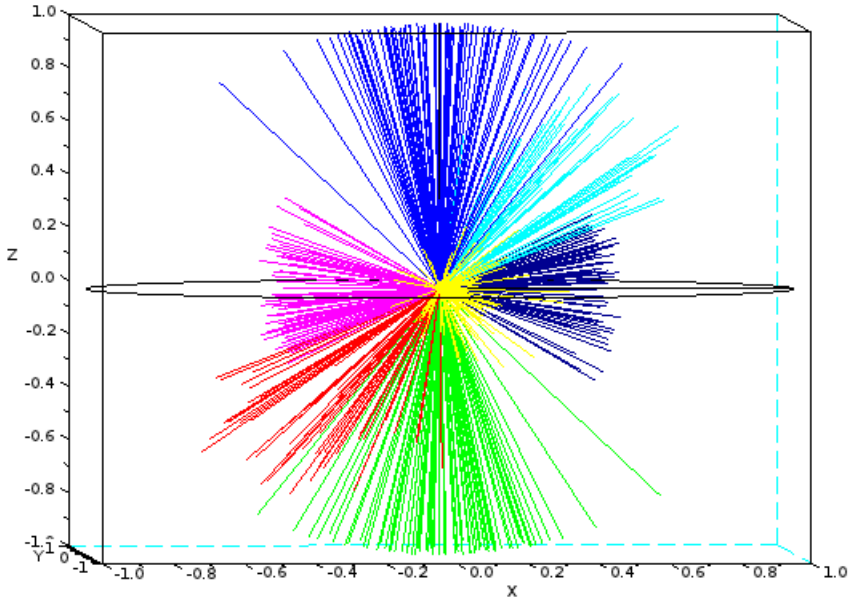

Figure 25: essai EM 2 : Visualisation des familles trouvées par l'algorithme

La figure 28 montre bien qu'il est très difficile de discerner trois composantes de mélange.

Nous remarquons que la répartition en familles proposée par l'algorithme n'est pas symétrique (figure 30). Nous nous retrouvons dans un cas où les données sont inexploitables, comme nous l'avions fait remarquer à la section précédente.

6 L'algorithme SEM (Stochastic Expectation Maximization)

6.1 Présentation de l'algorithme

L'algorithme SEM est une version stochastique de l'algorithme EM. Les pondérations intervenant dans les estimations des paramètres que constituaient les probabilités a posteriori sont remplacées par de vrais choix, 1 ou 0. Cependant ces choix ne sont pas définitifs, ils sont changés à chaque itérations selon une certaine probabilité correspondant aux probabilités a posteriori. Un point pour lequel l'algorithme a trouvé avec une grande probabiltié à quelle composante il appartient aura alors une très faible probabilité

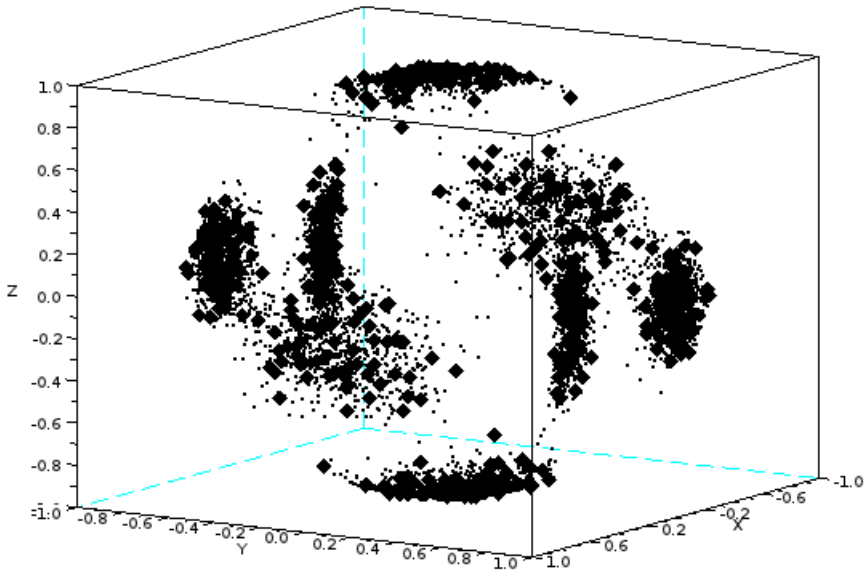

Figure 26: essai EM 2 : Simulation à partir des paramètres trouvés

de voir son choix changer à l'itération suivante. Au contraire, pour un point incertain, la probabilité que le choix change à l'itération suivante est plus importante.

Nous reprenons la description de l'algorithme EM en modifiant en rouge les parties spécifiques à l'algorithme SEM.

Notons $Z_{i,j} = 0$ *ou* 1 les choix faits à chaque itération.

Début de la boucle

† Estimation des proportions du mélange : Pour chaque famille que nous voulons estimer $k \in [\![1, n]\!]$

$$p_k = \frac{1}{N_k} \sum_{j=1}^{m} Z_{j,k}$$

où $N_k = \sum_{j=1}^{m} Z_{j,k}$

† Estimation des paramètres de la loi :

$$\hat{x}_{k,l} = \ldots \quad l \in [\![1, r]\!] \;\; k \in [\|1, n\|]$$

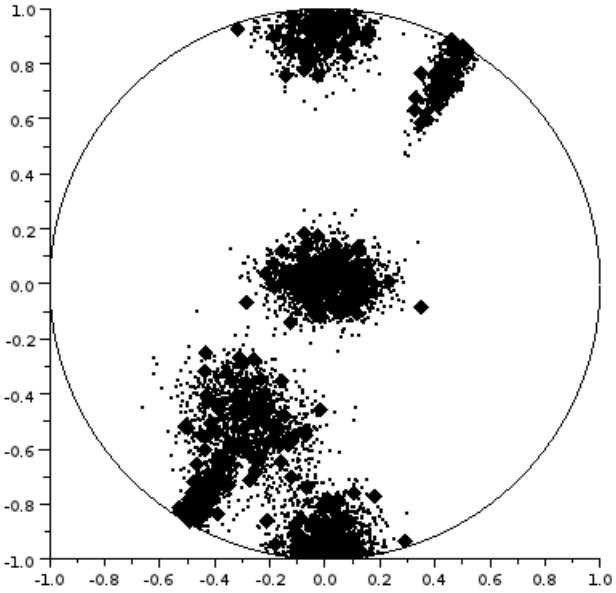

Figure 27: essai EM 2 : Simulation à partir des paramètres trouvés : diagramme de Wulff

Cette étape dépend de la loi utilisée et des probabilités a posteriori choisies.

† Estimation des probabilités a posteriori :

$$z_{j,k} = \frac{p_k f_k(\hat{x}_{k,1}, \ldots, \hat{x}_{k,r})(y_j)}{\sum_{q=1}^{n} p_q f_q(\hat{x}_{q,1}, \ldots, \hat{x}_{q,r})(y_j)} \quad j \in [\![\ 1, m\ ]\!]\ \ k \in [\![\ 1, n\ ]\!]$$

- Etape stochastique : détermination des choix : $Z_{j,k}$

  Pour chaque $k \in [\![\ 1, n\ ]\!]$, on simule un nombre entier entre 1 et n correspondant au choix de la famille selon le vecteur de probabilité $(z_{k,1}, \ldots, z_{k,n})$

  Alors si on note i le nombre obtenu de cette manière, $Z_{k,j} = \delta_{j,i}$

Fin de la boucle.

Remarques :

- Les formules que nous avons vues pour les estimateurs dans les différents cas de lois de probabilités restent les mêmes pour l'algorithme SEM à

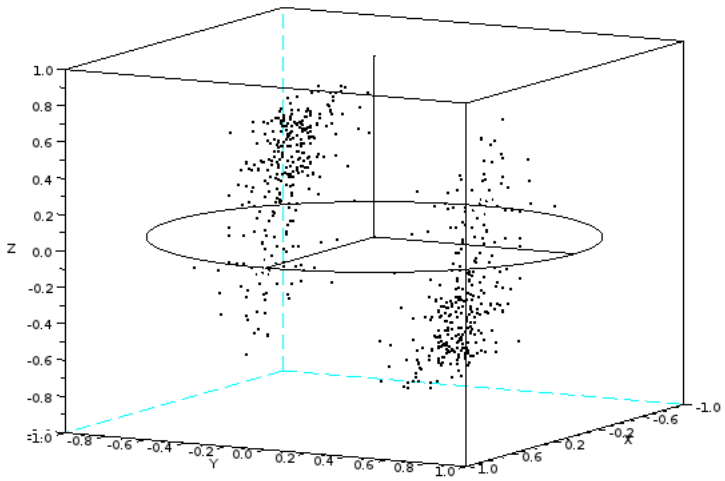

Figure 28: essai EM 3 : Visualisation de l'échantillon de départ

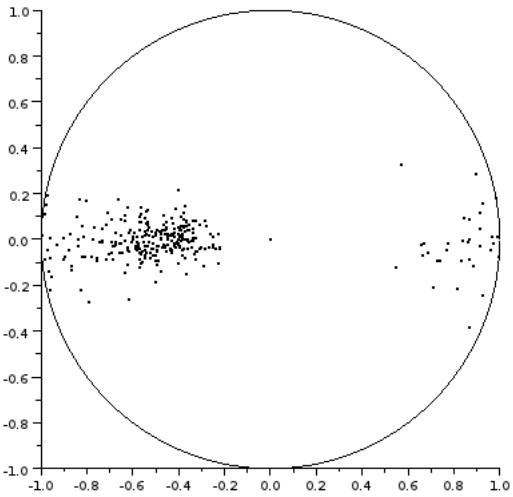

Figure 29: essai EM 3 : Visualisation de l'échantillon de départ : diagramme de Wulff

cela près qu'on remplace les $z_{j,k}$ par $Z_{j,k}$. Cela correspond donc à un choix binaire, le point est compté ($Z_{j,k} = 1$) ou pas ($Z_{j,k} = 0$).

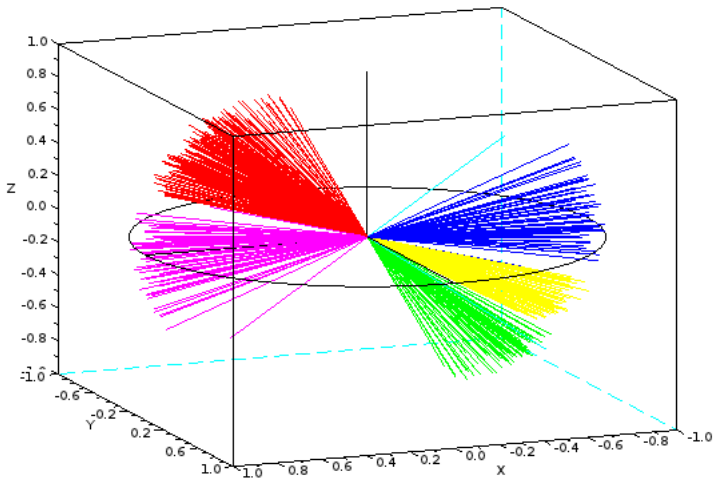

Figure 30: essai EM 3 : Visualisation des familles données par l'algorithme : pas de symétrie

- La log-vraisemblance n'est plus assurée de croître à cause de la partie stochastique de l'algorithme. Cependant, nous observons, à des oscillations près, le même comportement de la log-vraisemblance. Cette dernière a une tendance globale à l'accroissement et se stabilise autour d'une valeur limite à partir d'un certain nombre d'itérations, qui dépend de la difficulté du mélange.

Volontairement et par souci de concision, nous ne mettrons pas d'exemples pour l'utilisation de l'algorithme SEM, car les résultats trouvés sont très semblables à ceux de l'algorithme EM. Nous résumons néanmoins les différences que nous avons pu noter dans la section suivante.

6.2 Comparaison avec l'algorithme EM

Au cours des essais, nous avons pu noter quelques différences entre les utilisations et les résultats obtenus par EM et SEM.

Nous résumons ici les deux différences principales que nous avons pu remarquer.

- Du fait que la log-vraisemblance ne croît pas nécessairement en fonction des itérations dans le cas de l'algorithme SEM, la stabilisation des estimateurs est moins rapide.

- Nous avons pu remarquer que dans certains cas, qu'il nous est difficile d'expliciter, une initialisation différente de EM permettait d'avoir des résultats meilleurs qu'avec l'initialisation que nous prenions d'habitude. Nous n'avons pas remarqué de tels phénomènes avec l'algorithme SEM, car ce dernier s'affranchit de la situation initiale par son étape stochastique.

7 Test de validation des paramètres à l'issue de EM (ou de SEM)

L'objectif de cette section est de présenter un critère de validation des paramètres du mélange obtenus par les algorithmes EM et SEM.

Nous sommes partis du constat suivant ; avec certains échantillons simulés dont les composantes du mélange sont trop confondues, nos algorithmes pouvaient faire disparaître une ou plusieurs composantes. Nous nous sommes alors demandé si pour un échantillon de ce genre nous obtenions tout de même des simulations acceptables à partir des paramètres estimés. L'algorithme aurait alors trouvé une autre façon d'approcher notre échantillon, qui était trop compliqué pour être reconstitué dans tous ses détails.

Notre critère s'appuie sur une technique utilisée en reconnaissance de formes et basée sur la distance de Hausdorff.

C'est pourquoi nous commençons cette section par un petit rappel sur la distance de Hausdorff et par l'introduction de la distance de Hausdorff modifiée.

7.1 Présentation de la distance de Hausdorff et distance de Hausdorff modifiée

La distance de Hausdorff est une distance au sens mathématique sur l'ensemble des parties compactes d'un espace métrique. Cette distance dépend alors de la métrique choisie sur cet espace.

Définition : distance de Hausdorff Soit (E, d) un espace métrique, $A, B \subset E$ des parties compactes de E. La distance de Hausdorff entre A et B est définie par :

$$dH(A,B) = \sup(\inf_{a\in A} d(a,B)\ , \inf_{b\in B} d(b,A)\ )$$

où

$$d(a,B) = \inf_{b\in B} d(a,b) \quad , a \in A$$

Cette distance mesure l'éloignement entre A et B mais ne correspond néanmoins pas à la notion de proximité pour deux ensembles que nous voulons. La distance de Hausdorff accorde trop d'importance à un unique

point (sur 10000 par exemple) qui serait trop éloigné de la "forme de référence" comme le montre la figure 31.

Les deux ensembles A et B (bleu et rouge) de la figure 31, ont la même distance de Hausdorff. Pourtant, notre oeil nous indique que dans le premier cas, les deux ensembles se ressemblent plus. Ceci est dû à un effet de moyenne, et est encore plus prononcé pour un plus grand nombre de points.

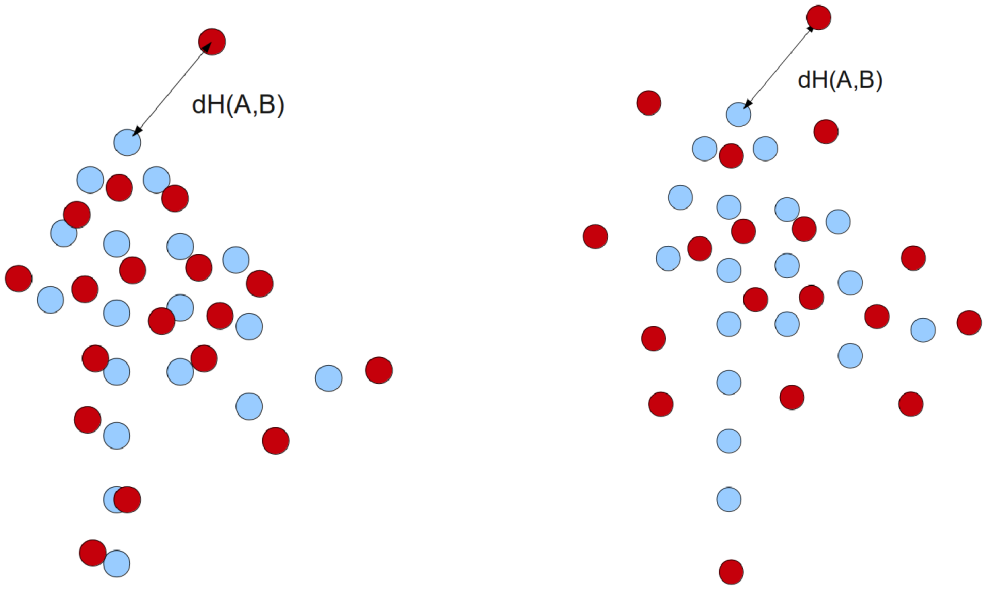

Figure 31: Deux ensembles ayant la même distance de Hausdorff, mais pas la même distance de Hausdorff modifiée

Pour prendre en compte l'effet de moyenne, on peut modifier la distance de Hausdorff dans le cas particulier des parties finies d'une espace métrique.

Soit (E, d) un espace métrique, $A, B \subset E$ des parties finies de E. La distance de Hausdorff modifiée entre A et B est donnée par :

$$d\bar{H}(A, B) = \sup(d\bar{H}(A \mid B), d\bar{H}(B \mid A))$$

où

$$d\bar{H}(A \mid B) = \frac{1}{Card(A)} \sum_{a \in A} d(a, B)$$

$d\bar{H}(A \mid B)$ est la distance de Hausdorff modifiée relative de A par rapport à B et décrit comment A s'insère dans B en moyenne. Par exemple si $A \subset B$ alors $d\bar{H}(A \mid B) = 0$, mais on n'a pas nécessairement $d\bar{H}(B \mid A) = 0$.

Une remarque importante est que la distance de Hausdorff modifiée n'est pas une distance au sens mathématique sur l'ensemble des parties finies de E car elle ne vérifie pas l'inégalité triangulaire.

Néanmoins, nous avons :

i)
$$d\bar{H}(A, B) = 0 \Leftrightarrow A = B$$

ii)
$$d\bar{H}(A, B) = d\bar{H}(B, A)$$

iii)
$$d\bar{H}(A \mid B) = 0 \Leftrightarrow A \subset B$$

7.2 Élaboration du test

La démarche est la suivante. Nous voulons à partir des paramètres trouvés par l'algorithme EM (ou SEM), simuler plusieurs mélanges et tester s'ils sont à une distance de Hausdorff "raisonnable" l'un de l'autre. Si c'est le cas, cela permettrait de valider les paramètres trouvés et dans le cas contraire, au moins d'avertir l'utilisateur et éventuellement relancer l'algorithme.

L'idée est alors de comparer des distances de Hausdorff modifiées d'une part entre l'échantillon de départ et des échantillons simulés selon les paramètres estimés et d'autre part entre deux échantillons simulés.

Cependant, les variables aléatoires obtenues ont des lois différentes a priori, et donc ne peuvent pas être soumises à un test d'adéquation classique. De plus elles dépendent de l'échantillon de départ, ce qui interdit d'avoir un critère fixe les concernant. Nous sommes donc obligés d'aller plus loin, en observant comment varie la différence entre ces lois.

Plus précisément, nous allons observer la différence entre des quantiles donnés pour les deux lois, i.e. la quantité $q_{1,\alpha_1} - q_{2,\alpha_2}$, où q_{i,α_i} est le quantile d'ordre α_i de la première loi ($i = 1$) ou de la deuxième ($i = 2$).

De façon heuristique, nous avons choisi de rejeter les paramètres du mélange lorsque :

$$q_{1,0.97} - q_{2,0.88} < 0$$

.

Les deux valeurs des quantiles à utiliser ont été choisies après une étude statistique qui sera présentée dans la section suivante, de manière à maîtriser les risques de première et de seconde espèce du mieux possible.

7.3 Statistiques du test de validation des paramètres à l'issue de EM et SEM

Etude de l'influence de la matrice de covariances sur la distance de Hausdorff modifiée

Nous avons d'abord pensé à un critère de validation des paramètres trouvés par l'algorithme de classification utilisant le protocole suivant : Nous simulons un échantillon à partir des paramètres estimés et nous calculons la distance de Hausdorff modifiée entre cet échantillon et celui qui nous avons au départ. Le problème est alors d'être en mesure de comparer la valeur obtenue à une valeur de référence. L'objectif de cette section est de mettre en évidence que nous ne pouvons pas obtenir directement une telle valeur de référence. En effet, comme nous allons le voir, la valeur de la distance de Hausdorff modifiée entre les deux échantillons dépend entre autres de l'ordre de grandeur des coefficients des matrices de covariances.

Afin de mettre en évidence cet effet, nous avons tracé les histogrammes des distances de Hausdorff modifiées entre une famille simulée de 100 points servant de référence et une série de familles de 100 points également simulées. Les simulations effectuées sont des projections de lois multi-normales dans le plan tangent à la sphère unité. Toutes les familles sont centrées sur le pôle nord. Nous effectuons trois séries de mesures, pour chacune d'entre elles la matrice de covariances est fixe.

La figure 32 est effectuée avec une matrice de covariances diagonale de coefficient 0.01, la figure 33 est effectuée également avec une matrice de covariances diagonale de coefficient 0.03 et nous avons choisi 0.08 comme coefficient pour la dernière matrice de covariances correspondant à la dernière série de mesures, figure 34.

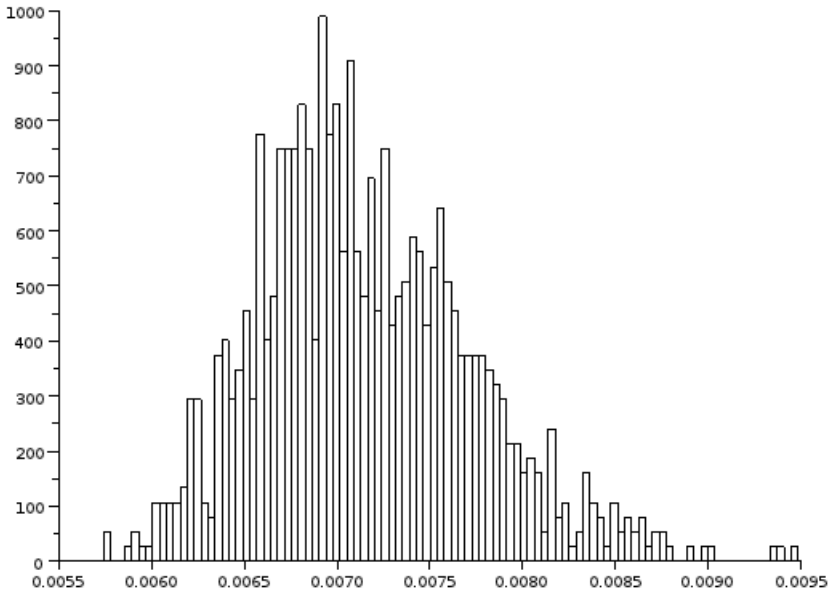

Figure 32: Histogramme des distances de Hausdorff modifiées : matrice de covariances diag(0.01,0.01)

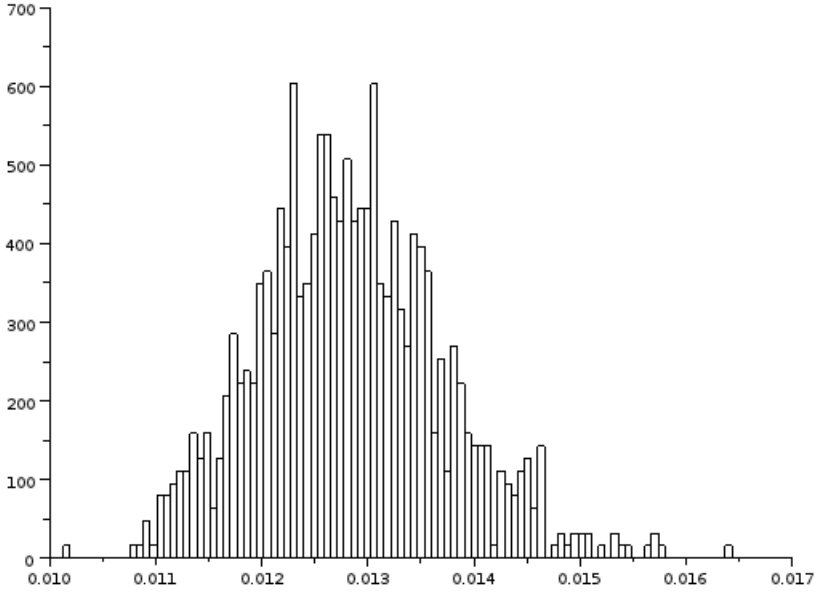

Figure 33: Histogramme des distances de Hausdorff modifiées : matrice de covariances diag(0.03,0.03)

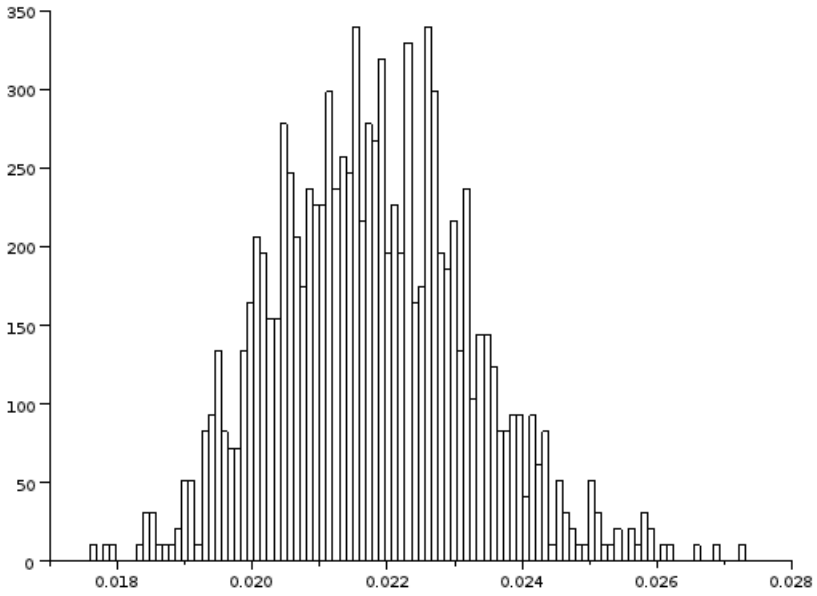

Figure 34: Histogramme des distances de Hausdorff modifiées : matrice de covariances diag(0.08,0.08)

Nous pouvons donc observer sur les trois figures 32, 33 et 34, que les valeurs pouvant être prises par la distance de Hausdorff modifiée varient en fonction de l'ordre de grandeur des termes de la matrice de covariances. Ceci nous pousse donc à chercher plus loin pour établir notre critère.

Influence de la position du vecteur moyen

Alors que nous venons de voir que la matrice de covariances a un effet sur les valeurs que peut prendre la distance de Hausdorff entre deux échantillons, nous allons à présent mettre en évidence que ce n'est pas le cas pour le vecteur moyen.

En suivant le même protocole qu'à la section précédente, mais en faisant varier le vecteur moyen cette fois-ci, nous traçons trois histogrammes correspondant à des séries de mesures de 1000 simulations.

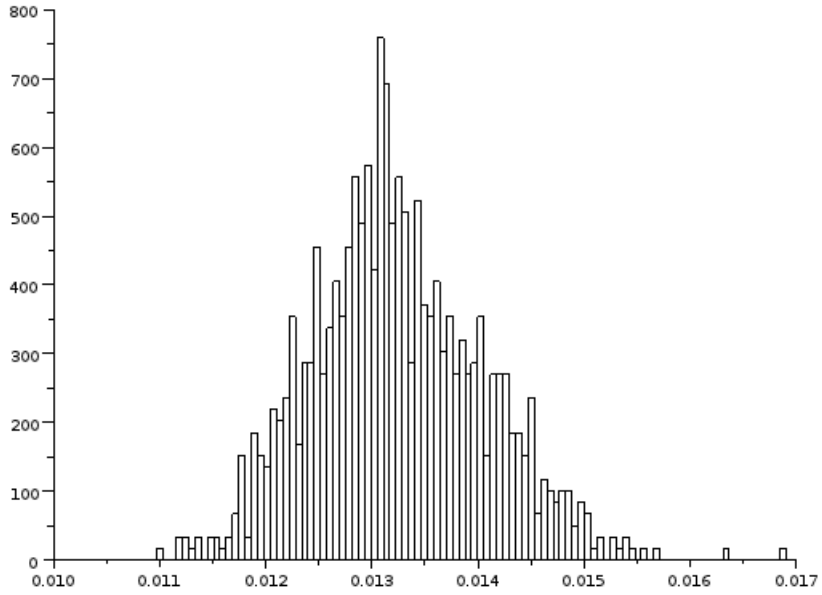

Figure 35: Histogramme des distances de Hausdorff modifiées : influence vecteur moyen : $(0, 0)$

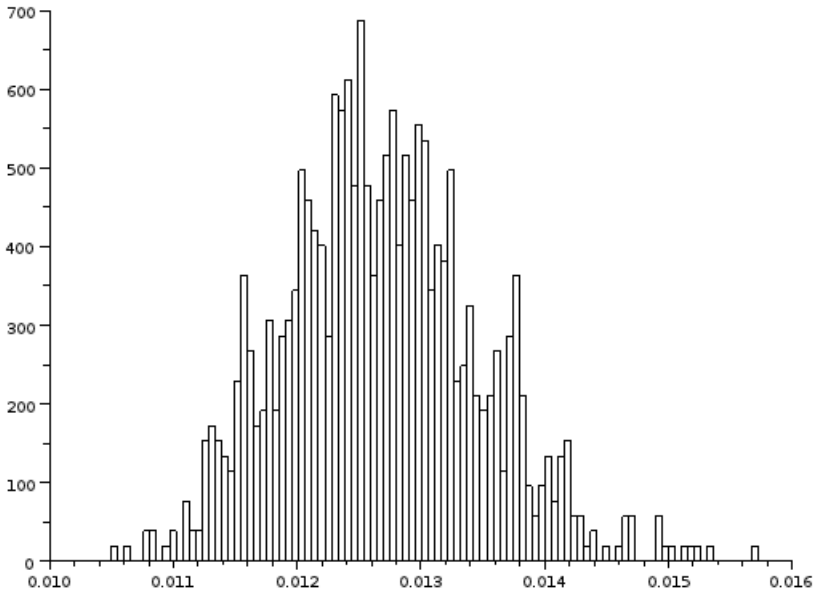

Figure 36: Histogramme des distances de Hausdorff modifiées : influence vecteur moyen : $(\frac{\pi}{2}, 0)$

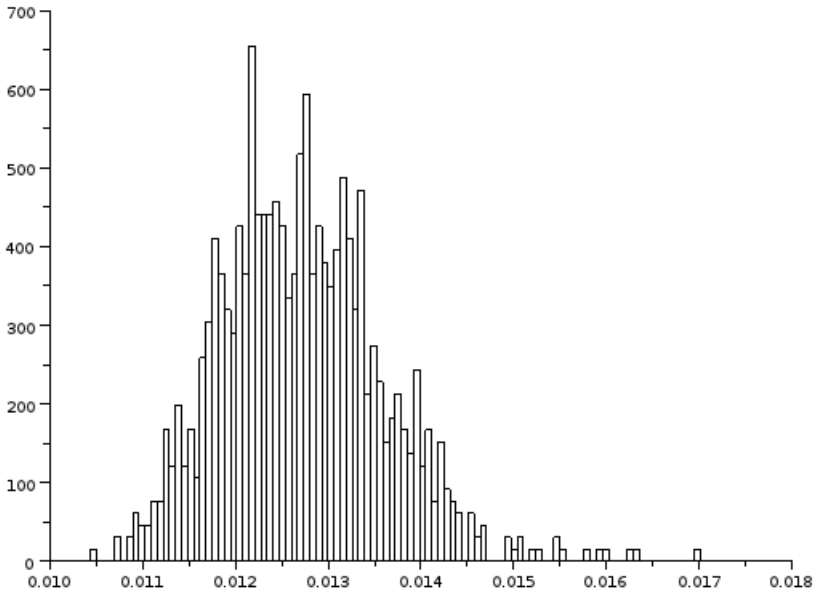

Figure 37: Histogramme des distances de Hausdorff modifiées : influence vecteur moyen : $(\frac{\pi}{2}, \frac{\pi}{3})$

Nous pouvons donc ignorer le vecteur moyen dans l'étude statistique de notre critère. L'ignorer dans le sens où il est transparent aux critères axés sur la distance de Hausdorff modifiée. Cependant, l'étude statistique qui va suivre montre que le critère que nous avons établi est sensible aux erreurs survenues sur le vecteur moyen.

Risque de première espèce : rejet à tort

Nous voulons dans cette section établir le risque de première espèce pour notre test statistique : la probabilité de rejeter à tort des paramètres qui seraient correctement estimés. Cependant, nous savons d'après les sections précédentes que si nous pouvons ne pas prendre en compte l'influence du vecteur moyen pour établir le risque de première espèce, il en est tout autrement de la matrice de covariances. Cependant, puisqu'il ne s'agit que de l'ordre de grandeur des coefficients de la matrice de covariances, nous allons établir le risque de première espèce pour l'ordre de grandeur le plus fréquemment rencontré dans la pratique. Cette étude n'est donc pas rigoureusement un calcul de risque de première espèce pour ce test, car la valeur obtenue dépend en quelque sorte du type de paramètres. Il convient de bien noter que la matrice de covariances prise pour établir le risque de première espèce est la matrice diagonale de coefficients 0.03, et que si dans la pratique on s'en écarte, le risque de première espèce un peu être différent.

Protocole :
Nous appliquons le test à deux échantillons simulés avec les mêmes paramètres un grand nombre de fois. C'est-à-dire :

- Simulation d'un échantillon de 30 points simulés selon une projection de loi multi-normale dans le plan tangent au pôle nord de la sphère, de matrice de covariances $diag(0.03, 0.03)$. C'est l'échantillon de référence.

- Nous appliquons 3000 fois les deux étape qui suivent :

  – Nous simulons trois échantillons de 30 points avec toujours les mêmes paramètres : l'échantillon 1, 2 et 3

  – Nous enregistrons alors deux distances de Hausdorff modifiées : entre l'échantillon de référence et l'échantillon 1 et entre les échantillons 2 et 3.

- Les deux distances calculées à chacune des 3000 itérations nous fournissent deux histogrammes, dans l'ordre d'apparition histogramme 1 et 2. Et nous calculons alors leur quantile respectifs : $q_{1,0.97}$ et $q_{2,0.88}$. Ce qui nous permet d'appliquer notre critère en effectuant la différence : $q_{1,0.97} - q_{2,0.88}$.

Ces étapes sont elles-même répétées 600 fois, afin d'optenir une proportion approchant la probabilité de rejet à tort. En effet tous les rejets trouvés lors de cette opération le sont à tort, car toutes les simulations ont été effectuées avec les mêmes paramètres.

Remarquons que pour le calcul de la distance de Hausdorff, nous sommes obligés de parcourir tout l'échantillon à chaque fois, ce qui est extrêmement coûteux pour l'ordinateur. Effectuer ce calcul avec Scilab n'est pas pensable car il y a beaucoup trop d'itérations. Cependant Scilab est très pratique pour l'aspect graphique, les fonctions d'affichage y sont déjà programmées. Nous avons dû avoir recours à la programmation en C++. A l'issue du calcul, le programme écrit en C++ enregistre dans un fichier les données collectées, ce qui permet à Scilab d'en faire un traitement graphique par la suite. Même programmé en C++, ce calcul prend environ une heure.

La figure 38 présente l'histogramme des écarts entre les quantiles : $q_{1,0.97} - q_{2,0.88}$. Après calcul, nous trouvons un risque de première espèce de 0.054, ce que nous visions.

Sensibilité du test

Maintenant que nous savons que le risque de première espèce n'est pas trop grand, examinons si notre test est précis. Nous cherchons à établir un certain intervalle contenant des erreurs acceptables tout en maîtrisant le risque de deuxième espèce.

Plus formellement, notons $A^0 = (a_1^0, \ldots, a_q^0) \in \mathbb{R}^q$ la vraie valeur des paramètres dans l'espace des paramètres. Notons ensuite $T : \mathbb{R}^q \longrightarrow \{0, 1\}$ notre test qui va de l'espace des paramètres possibles et qui donne une réponse positive 1 ou négative 0. Nous cherchons un ensemble de la forme $\mathcal{I} = \prod_{i=1}^{q}]a_i - \varepsilon_i, a_i + \varepsilon_i[$, $\varepsilon_i > 0$ tel que

$$\mathbb{P}(T(A) = 0 \mid A \notin \mathcal{I}) > 1 - \alpha$$

A est la valeur des paramètres trouvés, α est en quelque sorte la sensibilité du test. Par exemple on peut prendre $\alpha = 0.05$.

Dans ce cas, la probabilité d'avoir une grande erreur sur les paramètres avec un test positif est de moins de 5%.

Protocole : On procède de la même façon que dans la section précédente, seulement, au lieu de simuler les échantillons toujours avec les mêmes paramètres, à chaque nouvelle simulation, on administre une erreur volontaire et maîtrisée sur les paramètres. On étudie ensuite la probabilité que le test émette un résultat négatif. Nous souhaitons optenir une fréquence de refus de 0.90.

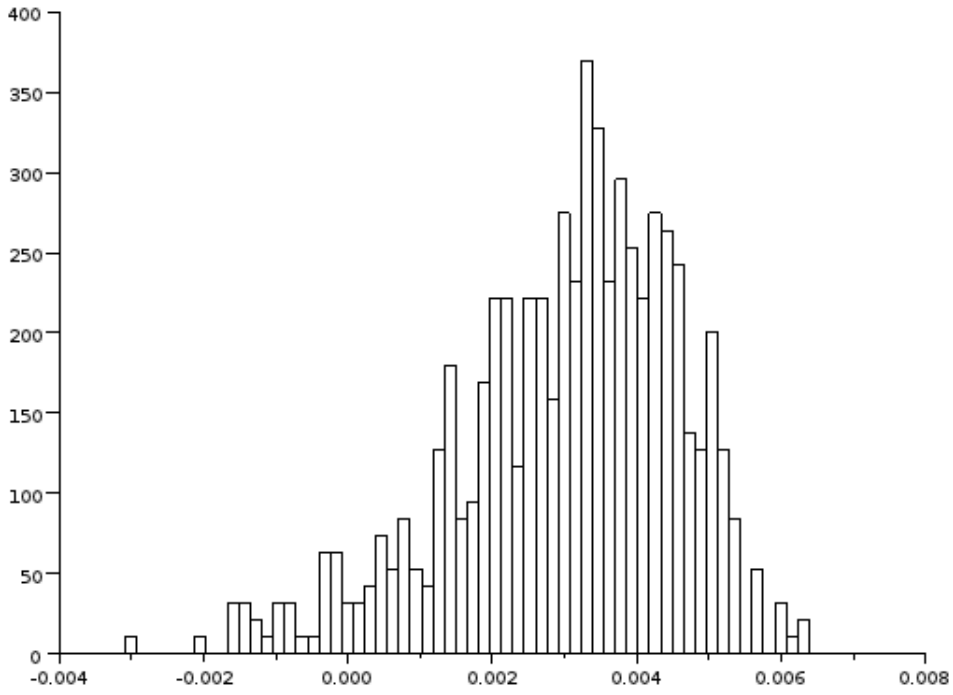

Figure 38: Histogramme des valeurs des écarts des quantiles : $q_{1,0.97} - q_{2,0.88}$, évaluation du risque de première espèce

Erreur sur la matrice de covariances :

Nous appliquons une erreur fixée à l'avance sur les coefficients non nuls de la matrice de covariances. On prendra toujours une matrice de covariances diagonale, afin d'avoir directement accès aux vecteurs propres qui régissent la forme de la loi multi-normale dans le plan tangent.

- Erreur de : 0.001 : le taux de rejet par le critère est de

6.83%

- Erreur de : 0.009 : le taux de rejet par le critère est de

36.8%

- Erreur de : 0.01 : le taux de rejet par le critère est de

44.2%

- Erreur de : 0.015 : le taux de rejet par le critère est de

64.3%

- Erreur de : 0.02 : le taux de rejet par le critère est de

83.0%

- Erreur de : 0.03 : le taux de rejet par le critère est de

97.2%

- Erreur de : 0.05 : le taux de rejet par le critère est de

100%

Remarques :

Nous sommes forcé d'admettre que les erreurs sur la matrice de covariances sont très difficiles à détecter avec ce critère. En effet, les erreurs donnant lieu à un taux de rejet raisonnable, 83.0% ou encore 97.2% correspondent à des erreurs relatives de près de 100%.

Ceci est dû au fait que le critère repère surtout les écarts de forme en moyenne, et que les erreurs sur la matrice de covariances telles que nous les avons appliquées ne font que faire grossir la forme générale par homothétie. De ce fait, les écarts de forme se situent à la périphérie de la tache et ne concernent qu'un nombre limité de points.

Pour les mêmes raisons, des erreurs appliquées de façon non symétrique sur les deux termes diagonaux auraient un effet plus important sur la forme moyenne de la tache, et seraient donc mieux repérées par le critère de la distance de Hausdorff modifiée. Nous nous sommes placés volontairement dans le cas le plus difficile à détecter, afin de majorer l'erreur faite.

Erreur sur le vecteur moyen :

De même que pour l'application des erreurs sur la matrice de covariances, dans le cas du vecteur moyen, nous appliquons la même erreur sur les deux composantes des coordonnées sphériques pour effectuer nos tests.

- Erreur de : 3° : le taux de rejet par le critère est de

11.3%

- Erreur de : 5° : le taux de rejet par le critère est de

29.3%

- Erreur de : 8° : le taux de rejet par le critère est de

76.6%

- Erreur de : 10° : le taux de rejet par le critère est de

95.2%

- Erreur de : 20° : le taux de rejet par le critère est de

100%

Nous observons donc que notre critère est sensible à des variation d'à peine une dizaine de degrés sur les coordonnées sphériques des vecteurs moyens des familles. C'est une assez bonne précision compte tenu des fluctuations d'échantillonnage importantes dues à des tailles d'échantillon assez faibles (30 points). Nous ne sommes néanmoins pas en mesure d'effectuer ces essais sur des familles beaucoup plus grandes car la complexité des algorithmes nous en empêche.

Remarque : Pour être tout à fait rigoureux, nous devrions mener une étude en conjuguant les types d'erreurs, sur la matrice de covariances et sur le vecteur moyen. En effet, il pourrait y avoir un phénomène de compensation des erreurs, qui les rendrait moins visibles par le critère de la distance de Hausdorff modifiée.

8 Étude de l'influence des erreurs de mesures

Les deux dernières sections vont être consacrées à une application de notre étude à la géologie. Nos travaux se prêtent en effet bien à la classification des orientations de discontinuités dans les roches. Les échantillons étant relevés à la main sur des carottes prélevées directement sur le terrain, les données comportent une incertitude. L'objet de cette section est alors d'étudier, en prenant un modèle simple pour la simulation des erreurs de mesures, leur influence sur la classification par la méthode EM.

8.1 Protocole

Tout d'abord, nous devons choisir un modèle pour les erreurs de mesures. Des études empiriques suggèrent qu'il est possible d'envisager d'appliquer les erreurs de mesures sur les coordonnées sphériques des points de l'échantillon de la façon suivante :

Etant donné un échantillon parfait (simulé) $Y = (y_1, \ldots, y_m)$, où $y_i = (\theta_i, \varphi_i)$ en coordonnées sphériques (décrites dans la première section de l'article), les erreurs sont calculées comme suit :

Pour θ, en notant $\bar{\theta}$ la nouvelle coordonnée prenant en compte les erreurs de mesures :

$$\bar{\theta}_i = \theta_i + \varepsilon_\theta$$

où ε_θ suit une loi normale de moyenne nulle et de variance $\frac{2\pi}{180}$:
$\varepsilon_\theta \sim \mathcal{N}(0, \frac{2\pi}{180})$

Pour φ, en notant de même $\bar{\varphi}$ la nouvelle coordonnée prenant en compte les erreurs de mesures :

$$\bar{\varphi}_i = \varphi_i + \varepsilon_\varphi$$

où ε_φ suit une loi normale de moyenne nulle et de variance $\frac{8\pi}{180}$:
$\varepsilon_\varphi \sim \mathcal{N}(0, \frac{8\pi}{180})$

Remarques :

- Remarquons que les erreurs sont indépendantes pour chacune des coordonnées sphériques, et également pour chacun des points de l'échantillon.
- Nous pouvons déjà remarquer également que les erreurs sur φ sont beaucoup plus importantes que les erreurs faites sur θ.

Visualisation de l'influence des erreurs de mesures :

Comme l'indique la remarque précédente, les erreurs sur φ ont un amplitude supérieure par rapport à celles faites sur θ. Ce phénomène est d'autant plus sensible que la moyenne de la famille considérée est proche de l'équateur.

Sur les figures de cette section nous avons fait apparaître la superposition des vecteurs normaux aux plans de fracture d'une famille simulée (en noir) pour une loi de projection gaussienne du plan tangent, et ceux de cette même famille après application du modèle pour les erreurs de mesures (en rouge). Nous pouvons alors observer la façon dont la famille est déformée par les erreurs de mesures ainsi que l'influence de la position du vecteur moyen dans cette déformation.

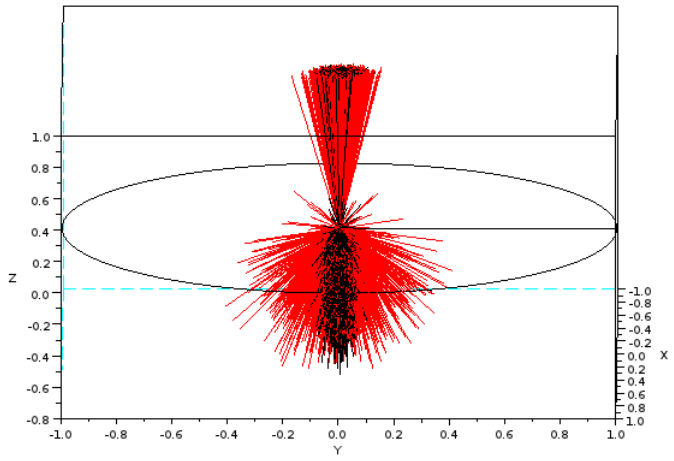

Figure 39: Visualisation de la déformation d'une famille peu étendue suivant φ 1

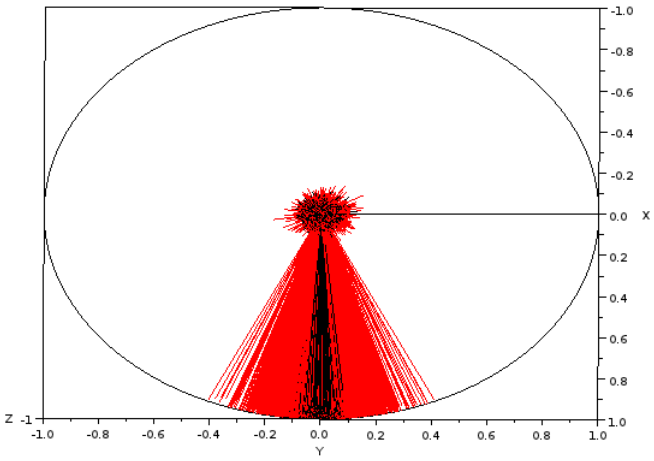

Figure 40: Visualisation de la déformation d'une famille peu étendue suivant φ 2

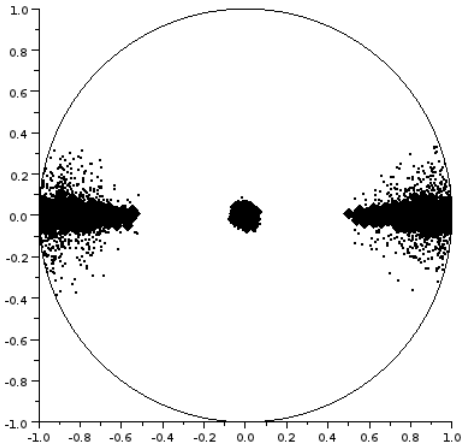

Figure 41: Visualisation de la déformation d'une famille très concentrée autour d'un pôle : diagramme de Wulff

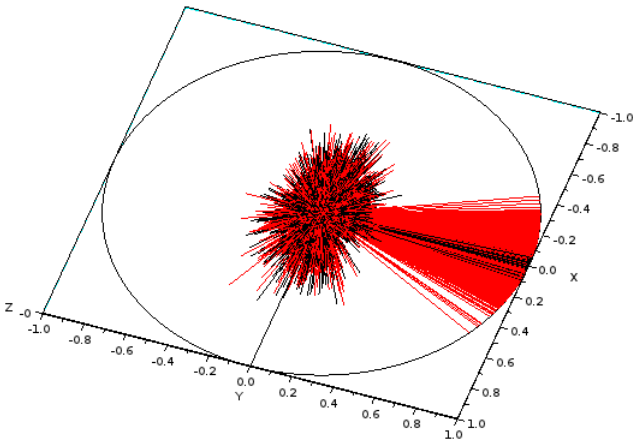

Figure 42: Visualisation de la déformation d'une famille centrée sur un pôle

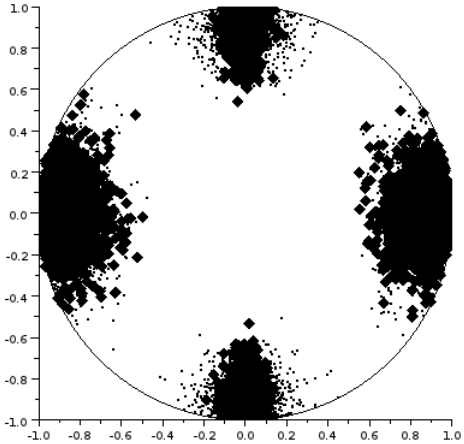

Figure 43: Visualisation de la déformation de familles à l'équateur : diagramme de Wulff

8.2 Présentation des résultats

Nous n'avons malheureusement que peu de résultats quantitatifs concernant un éventuel moyen de prendre en compte les erreurs de mesures en agissant directement sur l'échantillon. Néanmoins, nous avons quelques résultats d'observation du comportement des algorithmes EM et SEM suite aux erreurs ainsi introduites.

Observations :

- Les erreurs introduites perturbent comme nous l'attendions l'estimation des paramètres, qui devient beaucoup moins précise. Pour des familles proches l'une de l'autre, les erreurs peuvent aller jusqu'à les faire confondre en une seule famille par les algorithmes EM et SEM.
- Le degré d'appartenance à une famille estimé grâce aux probabilités a posteriori est également modifié, mais pas de façon uniforme pour tous les points. Ainsi certains choix faits par les algorithmes EM et SEM sont plus robustes que d'autre face aux erreurs de mesures.
- Nous n'avons pas remarqué de différence entre EM et SEM pour faire face aux erreurs de mesures.

9 Application à des données réelles, étude d'un forage

9.1 Présentation des objectifs

Nous souhaitons finir notre étude par une application à un cas réel. Nous disposons du recensement d'un réseau de fractures issues d'un forage. Les données sont enregistrées dans un fichier de type tableur, qui pouront être directement utilisées par les programmes. Les données y sont également classées par type de discontinuités. Dans le forage étudié, nous avons deux types de discontinuités, des fractures et des veines.

Un changement de coordonnées est nécessaire car les conventions utilisées en géologie, pour des questions pratiques de mesures sur le terrain, ne sont pas les coordonnées sphériques que nous avons utilisées. Ceci se fait sans mal, nous ne le détaillons pas ici.

Objectifs :

- Notre principal objectif est la simulation de ce réseau de discontinuités, dans son ensemble.
- Nous souhaitons également classer les familles au mieux à l'aide de l'algorithme EM.

Problématiques :

- A-t-on intérêt, pour la simulation, à séparer les données au préalable par rapport à l'origine géologique des discontinuités ?
- L'algorithme de classification automatique EM nous permet-il de classer les fractures selon leur origine géologique ?

9.2 Regroupement par l'algorithme EM des discontinuités de type fractures

Commençons par l'exécution de l'algorithme EM sur l'échantillon correspondant aux discontinuités de type fractures. Cet échantillon contient 99 points, ce qui est assez peu en comparaison aux essais que nous avons faits dans les sections précédentes.

La figure 44 présente le diagramme de Wulff de l'échantillon. La figure 45 est le diagramme de Wulff de la superposition d'un échantillon simulé à partir des paramètres estimés par EM (petits points) et de l'échantillon de départ (gros points). Pour les essais des sections précédentes avec les algorithmes EM et SEM, nous savions à l'avance quel nombre de familles chercher. Sur des données réelles, cela n'est plus le cas. Nous devons de ce fait essayer plusieurs nombres de familles possibles et décider a posteriori en fonction des estimations obtenues. Cette décision n'est pas automatisée, cependant elle pourrait être soumise à l'utilisation du critère de la distance de Hausdorff modifiée. En effet, le test nous donne une indication sur l'adéquation du mélange estimé à l'échantillon. Ici, nous avons choisi 6 familles.

Si les simulations obtenues semblent être satisfaisantes, le regroupement en familles n'est pas tout à fait symétrique. Nous avons déjà signalé que ce genre de problèmes peuvent être rencontrés avec des mélanges trop difficiles ou comptant un nombre de points trop faible. Pour ces raisons nous ne présenterons pas les données numériques, qui ne ne sont pas très significatives. Pour donner tout de même une idée au lecteur, le figure 46 représente les familles obtenues, avec des probabilités a posteriori d'au moins 0.70.

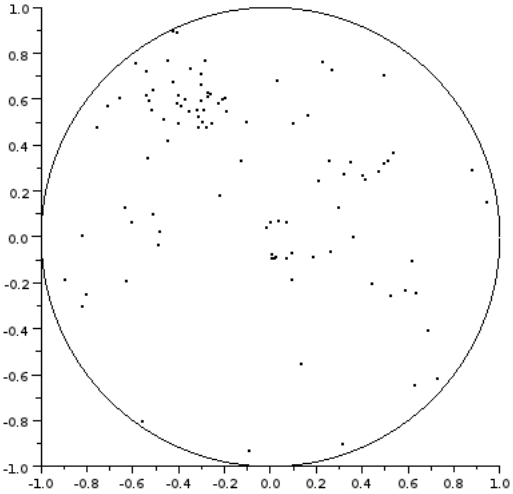

Figure 44: Diagramme de Wulf des données réelles pour les discontinuités de type fractures

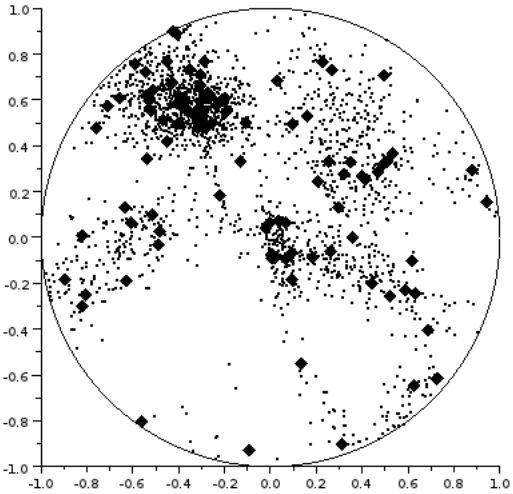

Figure 45: Diagramme de Wulf d'une simulation à partir des paramètres évalués par EM

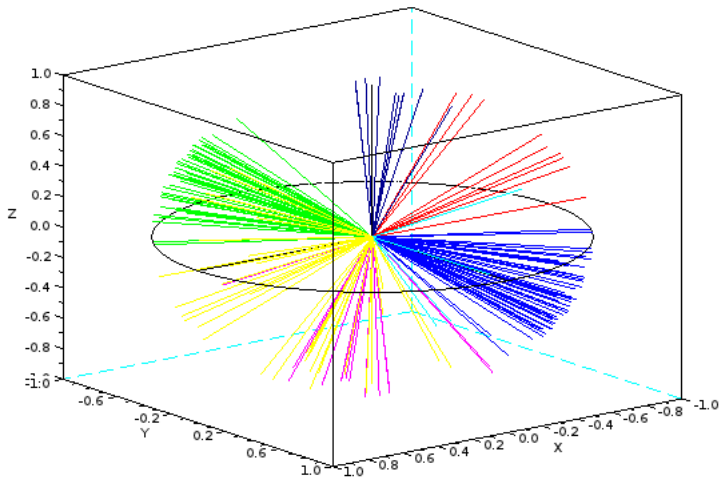

Figure 46: Représentation du regroupement en famille par l'algorithme EM : cas de fractures

9.3 Regroupement par l'algorithme EM des discontinuités de type veines

Cette section présente la même étude que précédemment sur les discontinuités de type veines. Nous avons rencontré les même problèmes. Les figures 47, 48 et 49 sont les analogues des figures 44, 45 et 46 pour ce type de discontinuités. Cet échantillon comporte 166 points. Nous avons opté ici pour 4 familles.

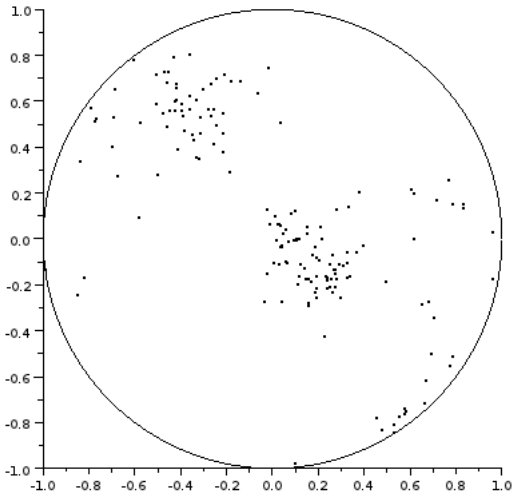

Figure 47: Diagramme de Wulf des données réelles pour les discontinuités de type veines

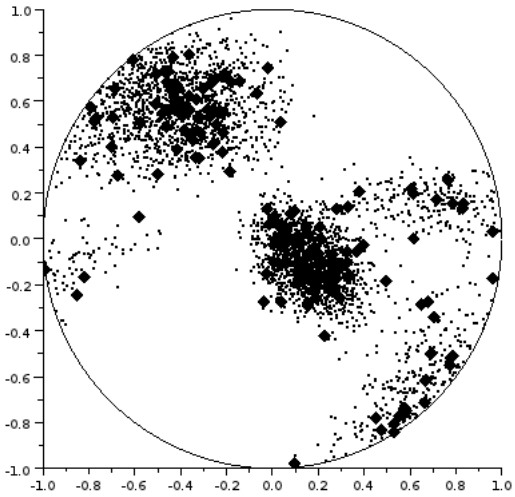

Figure 48: Diagramme de Wulf d'une simulation à partir des paramètres évalués par EM

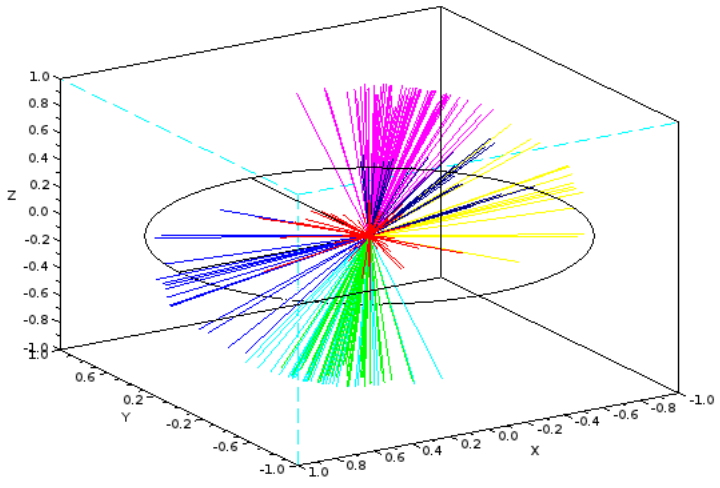

Figure 49: Représentation du regroupement en famille par l'algorithme EM : cas des veines

9.4 Regroupement des fractures tout type de discontinuités confondus

Dans cette section, nous exécutons l'algorithme EM non plus sur les données séparées, mais sur toutes les données en même temps. Cela fait beaucoup plus de points à traiter pour l'algorithme, ce qui devrait lui faciliter la tâche. Cependant, nous nous attendons à ce que le nombre de familles à trouver soit plus élevé.

La figure 50 présente le diagramme du Wulff de l'échantillon, qui n'est en fait que la superposition des diagrammes de Wulff des figures 44 et 47. La figure 51 représente sur un diagramme de Wulff la superposition d'un échantillon simulé à partir des paramètres évalués par EM (petits points) et de l'échantillon initial (gros points).

Volontairement nous ne présentons pas les données numériques. En effet, les données numériques que nous obtenons ne sont encore une fois pas symétriques. Nous présentons néanmoins une visualisation des familles de points trouvés, en différentes couleurs sur la figure 52. Pour cette échantillon, nous avons choisi 8 familles. Nous pouvons remarquer dès à présent que le nombre de familles choisi n'est pas la somme des nombres de familles pour chacune des données séparées. Nous en déduisons donc que des discontinuités d'origine géologique différente peuvent être comptabilisées dans une même famille.

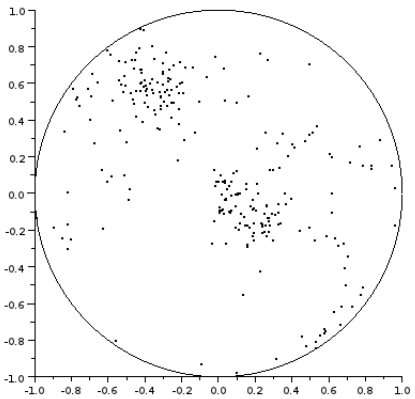

Figure 50: Diagramme de Wulff des données réelles

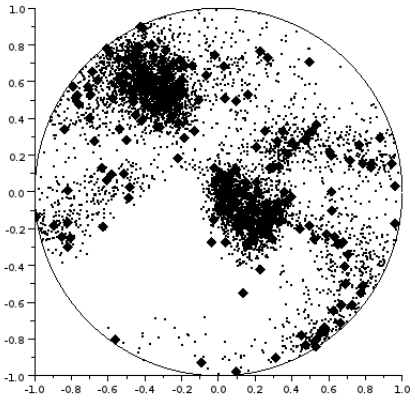

Figure 51: Diagramme de Wulff d'une simulation à partir des paramètres évalués par EM

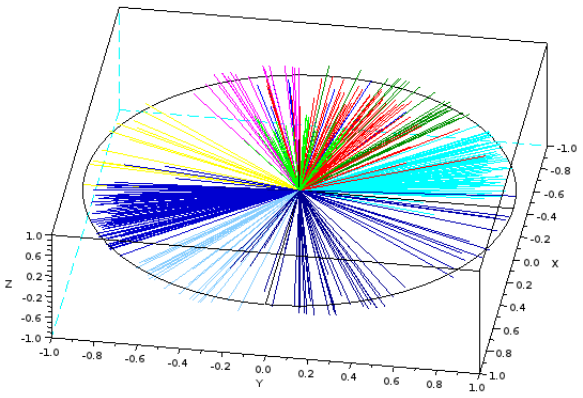

Figure 52: Représentation du regroupement en famille par l'algorithme EM : tout type de discontinuités confondu

9.5 Conclusion

L'étude de ce forage met en évidence quelques points qui restent à approfondir. Parmi ceux-ci nous pouvons noter qu'il est assez délicat de choisir un bon nombre de familles à chercher et que nos algorithmes tels qu'ils sont à l'heure actuelle ne permettent pas d'effectuer ce choix a priori. L'estimation des paramètres est elle aussi assez délicate. En effet, en s'affranchissant de la convention normale montante (ou descendante), nous nous sommes confrontés à la quasi-nécessité que l'estimation des paramtères soit faite de façon symétrique. Or nous avons pu observer que dans des mélanges trop difficiles, et de fait dans le cas réel, les estimations peuvent ne pas l'être. Cependant, la simulation à partir de paramètres non symétriques peut quand même aboutir, cela demande une étude plus approfondie de notre critère de la distance de Hausdorff modifiée.

Nous pouvons remarquer également à l'issue de cette essai que du point de vue de la simulation, il semble préférable d'avoir le maximum de points à proposer à l'algorithme EM, et donc de ne pas séparer les données par type de discontinuités. Cependant, il ne semble pas possible de retrouver les types de discontinuités dans des familles distinctes formées avec EM.

Conclusion

En conclusion, ce rapport contient l'étude d'une approche de classification automatique des orientations des réseaux de discontinuités dans la roche en vue d'une modélisation 3D de tels réseaux. L'approche choisie est celle de classification probabiliste avec les algorithmes d'estimation de paramètres de mélanges de lois de probabilités EM et SEM. Au cours de ce stage scientifique, j'ai écris un panel de programmes développés en Scilab ou en C++ pouvant être utiles dans la poursuite de l'étude de la classification des orientations des plans de discontinuités. Certains programmes sont disponibles à la fois en version Scilab et C++, car j'ai du faire face à des problèmes de temps d'exécution dus à la trop grande complexité de certains algorithmes. J'ai également commencé la mise en place d'un test servant de critère de validation des paramètres estimés à l'issue des algorithmes EM et SEM, et commencé l'étude statistique de ce test. Ce test est novateur et s'appuie sur des outils utilisés dans d'autres domaines tels que la reconnaissance de formes. Si le risque de première espèce est bien maîtrisé pour ce test, les travaux devront être poursuivis afin de mieux maîtriser le risque de deuxième espèce et de préciser l'étude qui à été faite sur la sensibilité du test. Le stage s'achève par l'étude de l'influence des erreurs de mesures, par modélisation probabiliste, sur les estimations fournies par EM et SEM et sur l'étude d'un cas réel de forage qui met en exergue encore des difficultés encore non surmontées. Nous n'avons pas eu le temps d'aboutir à une conclusion satisfaisante quand à cette influence, qui pourrait servir à poursuivre la comparaison des algorithmes EM et SEM. Cette étude peut donc être déjà servir de base pour la modélisation de RD, dont les applications peuvent aller de l'étude de l'infiltration de l'eau dans des massif rocheux jusqu'à l'étude mécanique du sol à l'échelle d'un ouvrage de génie civil.

Références

- Gasc-Barbier, M., O. Fouché, and C. Gaillard. "Etude comparée de la fracturation observable sur carottes de forage et obtenue par diagraphie. Application au marbre de Saint-Béat (Haute-Garonne)." Revue française de Géotechnique. 133 (2010): 37-49.

- Pouya, A., and O. Fouché. "Permeability of 3D discontinuity networks: new tensors from boundary-conditioned homogenisation." Advances in Water Resources. 32.3 (2009): 303-314.

- Fouché, O. "Caractérisation géologique et géométrique, et modélisation 3D, des réseaux de discontinuités d'un massif granitique reconnu par forages carottés (Charroux - Civray, Vienne)." Vol. 1 et 2. Ph. D. Thesis: Ecole des Ponts-ParisTech, 2000. 296 p. + annexes.

- Marcottea, D. and E. Henryb E. "Automatic joint set clustering using a mixture of bivariate normal distributions" International Journal of Rock Mechanics & Mining Sciences 39 (2002) 323-334.

- Sait Ismail Ozkaya "Fracture Length Estimation From Borehole Image Logs"

- Chaoshui Xu and Peter Dowd "A new computer code for discrete fracture network modelling" Computers & Geosciences.

Bilan personnel

Ce stage m'a avant tout fait découvrir le monde de la recherche, comment on y travaille, avec quelles méthodes, quels outils. J'y ai ainsi découvert la place primordiale que prend aujourd'hui l'outil informatique dans des domaines tels que les mathématiques appliquées, qui peuvent sembler d'un premier abord très théoriques. Au cours de ce stage, j'ai donc eu à apprendre à travailler les mathématiques non plus seulement sur le papier, mais aussi à travers l'expérimentation, l'exploration d'hypothèses par la modélisation rendue possible grâce à l'ordinateur et aux logiciels de calculs scientifiques. J'ai également touché du doigt les limites que peuvent imposer de tels outils, que seul un emploi réfléchi permet d'outrepasser. En effet, j'ai eu à choisir entre plusieurs méthodes informatiques, entre le logiciel de calculs scientifiques ou la programmation brute, et finir par associer les deux pour un résultat optimal. J'ai également beaucoup appris sur l'interdisciplinarité qu'offrent les mathématiques et le dialogue avec les autres chercheurs que cela suppose. Il m'a en effet fallu pendant ce stage comprendre la finalité de la problématique, qui ne concerne pas les mathématiques mais la géologie, afin de mieux répondre aux attentes du domaine. C'était la première fois que je pensais les mathématiques à travers une autre science, qui plus est m'est étrangère. Ce stage me conforte bien sûr dans le choix du département IMI (Ingénierie Mathématique et Informatique) mais également dans mon projet professionnel qui consisterait à faire de la recherche et du développement en tant qu'ingénieur mathématicien.

Appendices

A Liste des fonctions Scilab écrites

Le lecteur pourra trouver dans cette annexe une brève description des principales fonctions écrites en Scilab. Par soucis de concision, le code source des fonctions ne figure pas dans le rapport. Pour toutes informations complémentaires, me contacter aux l'adresses antoine.bensalah@eleves.enpc.fr ou antoine.bensalah@gmail.com.

A.1 Mélange de lois multi-normales dans le plan

Fonctions :

- `distance_H_relative.sci` Permet de calculer la distance de Hausdorff modifiée relative d'un ensemble fini par rapport à un autre.
- `fonct_EM_melange_gaussien.sci` Simule un mélange de deux lois normales, pour EM ou pour SEM.
- `fonct_EM_melange_multi_gaussien_plan.sci` Simule un mélange d'un nombre quelconque de lois multi-normales dans le plan, pour EM ou pour SEM.

Procédures :

- `EM_melange_gaussien.sce` Algorithme EM, dans le cadre d'un mélange à deux composantes de loi normale.
- `EM_melange_multi_gaussien_plan.sce` Algorithme EM, cas d'un mélange à plus de deux composantes de loi normale.
- `SEM_melange_gaussien.sce` Algorithme SEM, cas d'un mélange à deux composantes de loi normale.
- `SEM_melange_gaussien_plan.sce` Algorithme SEM, cas d'un mélange à plus de deux composantes de loi multi-normale.
- `SEM_melange_multi_gaussien_plan.sce` Algorithme SEM, cas d'un mélange à plus de deux composantes de loi multi-normale.
- `loi_distance.sce` Permet l'étude de la loi des distance de Hausdorff modifiée entre deux échantillons simulés selon une loi multi-normale dans le plan.

A.2 Loi de Fisher

Fonctions :

- `distance_H.sci` Calcule la distance de Hausdorff modifiée entre des ensembles de points de $\mathbb{R}$.
- `fisher_simulation.sci` Simule une loi de Fisher.
- `rotation.sci` Permet d'effectuer la rotation utile à la simulation de la loi de Fisher.
- `wulff.sci` Trace le diagramme de Wulff d'un ensemple de points en convention normale montante.

Procédures :

- `EM_melange_multi_fisher.sce` Algorithme EM, cas de plusieurs composantes suivant la loi de Fisher.
- `graqhe_distance_H.sce` Permet l'étude de la loi de la distance de Hausdorff modifiée entre échantillons suivant la loi de Fisher.
- `SEM_melange_multi_fisher.sce` Algorithme SEM, cas de plusieurs composantes suivant la loi de Fisher.

A.3 Mélange de projections normales du plan tangent sur la sphère unité

Fonctions :

- `distance_H.sci` Calcule la distance de Hausdorff modifiée pour des ensembles de points de $\mathbb{R}^3$.
- `simule_famille.sci` Simule un mélange de lois de projections multi-normale du plan tangent.
- `simule.sci` Simule un unique échantillon suivant une loi de projection multi-normale du plan tangent.

Procédures :

- `EM_melange_projection_gaussiennes.sce` Algorithme EM, pour des mélanges à plusieurs composantes.
- `fluctuations_echantillonage.sce` Permet l'étude des fluctutations d'échantillonnage qui n'apparaît pas dans ce rapport.

- `SEM_melange_projection_gaussiennes.sce` Algorithme SEM, pour des mélanges à plusieurs composantes.
- `simulation_erreurs.sce` Simule les erreurs de mesures sur un échantillon.
- `test_statistique.sce` Permet l'application du critère établi, mais très lent.

B Fonctions écrites en C++

Dand la même optique que la section précédente, cette annexe présente un bref aperçu des fonctionnalités des procédures écrites en C++. Le code source n'est pas fourni pour des raisons de concision, n'hésiter pas à me contacter pour de plus amples informations : antoine.bensalah@eleves.enpc.fr, antoine.bensalah@gmail.com

Ont été programmés en C++, le calcul du critère basé sur la distance de Hausdorff modifiée, l'algorithme EM intégrant ce critère. L'étude des fluctuations d'échantillonnages. L'étude statistique statistique du critère : risque de première espèce, sensibilité.