De la réduction de dimension à t-SNE

Dans cet exposé, nous réaliserons un tour d'horizon de quelques méthodes statistiques de réduction de dimension exploitant les aspects géométriques des données. Nous observons une matrice X composée de n individus et de p variables descriptives et l'objectif est de résumer les variables descriptives des individus de sorte à visualiser ces derniers dans un espace de dimension très petite (typiquement 1, 2 voir 3). Nous commencerons par les méthodes géométriques les plus classiques (ex : l'ACP) et nous y ajouterons des outils statistiques (ex : l'ACP stochastique). Ces méthodes linéaires conservent les propriétés globales des données (dissimilarités entre les points) mais échouent lorsque la structure sous-jacente n'est pas linéaire. Nous verrons comment la modélisation dans des espaces de probabilité permet de révéler les structures locales des données (similarités entre les points) à travers la notion de "voisinage" et nous étudierons en particulier les méthodes SNE et t-SNE. Des structures telles que des clusters sont alors révélées.