Le 6 mai 2011

Rapport de Thèse

Apprentissage statistique non supervisé : grande dimension et courbes principales

Aurélie Fischer

La thèse de A. Fischer porte sur des méthodes de statistique non asymptotique pour l'apprentissage en grande dimension. Deux axes de recherche ont été développés. Le premier axe concerne le regroupement non supervisé (*clustering*) d'éléments d'un espace fonctionnel. Le second se rapporte à l'estimation de courbes principales. Le manuscrit est constitué de deux parties. La première partie est composée de trois chapitres. A Fisher y développe les résultats qu'elle a obtenus dans son premier axe de recherche. La seconde partie contient deux chapitres et concerne son second axe de recherche. Plusieurs annexes achèvent le manuscrit : les trois premières fournissent des compléments d'ordre technique (moyennes de Rademacher, sélection de modèles et courbes paramétrées), les trois suivantes sont des articles reprenant, sous une forme condensée, les principaux résultats obtenus dans la thèse. Un long chapitre introductif synthétise les principaux résultats et en donne un aperçu assez précis. Le manuscrit est particulièrement bien rédigé et d'une lecture très agréable.

Dans le chapitre 1 de la première partie, Aurélie Fischer développe un point de vue nouveau en quantification. Il s'agit ici de *résumer* une mesure de probabilité portée par un espace fonctionnel abstrait E en une probabilité supportée par un nombre fini (noté k), de points de E. Pour cela, elle propose d'utiliser la divergence de Bregman comme mesure de distorsion. Elle développe d'abord une partie introductive où elle explique, de façon très claire, la quantification et sa version empirique qu'est la classification non supervisée aux plus proches voisins. Elle rappelle ensuite la définition et les propriétés de la divergence de Bregman associée à une fonctionnelle convexe lisse. Le point de vue est d'abord fini-dimensionnel puis certaines propriétés, utiles pour la suite, sont établies dans le cadre plus général d'un espace de Banach réflexif. L'exemple générique est celui d'une dissemblance bâtie comme fonctionnelle

intégrale d'une divergence de Bregman sur $\mathbb{R}$. Les résultats principaux obtenus dans ce chapitre concernent la quantification optimale pour une distorsion de type divergence de Bregman. Elle montre d'abord un résultat d'existence puis la convergence de la valeur optimale de la fonction de coût empirique vers celle obtenue pour la quantification optimale. L'hypothèse principale dans le cas infini-dimensionnel est que la mesure de probabilité à quantifier est supportée par un ensemble borné. La vitesse de convergence de la valeur optimale de la fonction de coût empirique est ensuite étudiée. En utilisant une technique de symétrisation, elle établit une majoration de l'écart entre la valeur de l'optimum empirique et sa limite. Cet écart est de l'ordre du nombre de clusters sur la racine du nombre d'observations. Des simulations numériques fouillées complètent et illustrent ce premier chapitre.

Le second chapitre de la première partie est un article écrit en collaboration et accepté pour publication dans une revue de statistique appliquée. Il s'agit ici de la mise en œuvre, sur un code numérique industriel, d'une méthode de classification non supervisée de courbes d'un espace de Hilbert. L'idée naturelle utilisée ici est de projeter les fonctions à classifier sur un espace de dimension finie puis d'effectuer une classification en dimension finie. Aurélie Fischer replace ce problème dans le cadre de la quantification hilbertienne. Elle donne tout d'abord une borne sur l'écart entre le critère de quantification dans l'espace de Hilbert et son approximation empirique pour un échantillon de taille n obtenu par projection sur un espace de dimension d. La méthode de classification après projection est ensuite numériquement validée sur des cas tests puis dans le cadre de sorties fonctionnelles du code numérique CATHARE du CEA. La méthode de classification est testée sur plusieurs familles de sous-espaces de dimension finie.

Le dernier chapitre de la première partie est consacré au choix automatique du nombre d'atomes k dans la quantification en fonction de la taille n de l'échantillon. Le point de vue est original car il conjugue des résultats de statistique non paramétrique, non asymptotique (inégalités oracle pour le choix de modèles), et de la théorie de l'information (complexité). Le cadre est celui d'un échantillon à valeurs dans $\mathbb{R}^d$ (muni de la norme euclidienne), dont la loi est à support compact. Après une discussion sur des méthodes empiriques de choix de k, proposées dans la littérature, Aurélie Fischer met en place un critère pénalisé permettant l'estimation jointe des atomes et de leur nombre. Ici, la pénalité ajoutée à la distorsion empirique s'interprète comme un terme de variance *fictive*. Elle montre ensuite une inégalité de type oracle pour l'estimateur pénalisé proposé. Des simulations numériques illustrent la pertinence de cette approche.

La seconde partie porte sur les courbes principales. Elle débute par un premier chapitre dressant un état de l'art complet sur le sujet. Aurélie Fischer explore d'abord les différentes définitions possibles des courbes principales (invariance, modèles de courbes paramétrées, ...), puis donne un tour d'horizon très précis des méthodes statistiques permettant l'estimation de ces courbes.

Sa contribution sur les courbes principales est développée au second chapitre. Elle explique tout d'abord le modèle qu'elle va utiliser. Il s'agit d'un modèle gaussien homoscédastique dans $\mathbb{R}^d$ dont le vecteur des moyennes est inconnu. Les composantes de cette moyenne sont des points de la courbe principale. Elle propose ensuite, pour estimer la courbe principale, d'adopter une approche par critère pénalisé et choix de modèles. La partie d'ajustement est un critère de moindres carrés, la pénalité fait

intervenir la complexité géométrique de la courbe candidate. Sous certaines hypothèses, elle montre, dans un premier temps, l'existence d'un minimiseur du critère pénalisé et obtient une inégalité oracle pour les points d'observation de la courbe principale. Dans un deuxième temps, elle montre une inégalité oracle fonctionnelle. Ses résultats sont très bien illustrés sur des simulations numériques avec données synthétiques ou réelles. La grande technicité des mathématiques développées dans ce chapitre montre sa maturité scientifique.

Aurélie Fischer a développé des techniques de choix de modèles dans les contextes originaux de la quantification et de la construction de courbes principales. Elle a également montré son intérêt pour la mise en œuvre pratique des techniques statistiques élaborées qu'elle développe. Les travaux de recherche d'Aurélie Fisher constituent une excellente thèse en statistique je suis tout à fait favorable à sa soutenance.

Fabrice Gamboa Professeur