Passage à l'échelle de la méthode de « Multidimensional Scaling » pour l'étude de la biodiversité

Emmanuel Agullo (HiePACS), Olivier Coulaud (HiePACS), Alain Franc (PLEIADE)

November 9, 2018

Contents

1 Encadrement

- Encadrants: Emmanuel Agullo (Inria HiePACS), Olivier Coulaud (Inria Hiepacs), Alain Franc (Inria Pleiade)
- Tél: EA 05 24 57 41 50 - OC 05 24 57 40 80 - AF 05 35 38 53 53
- Courriels: emmanuel.agullo@inria.fr; olivier.coulaud@inria.fr; alain.franc@inra.fr

2 Contexte

Les composantes méthodologiques et logicielles du high performance data analytics (HPDA) servent d'éléments de base pour des développements algorithmiques vers les applications de simulation ou d'analyse de données

massives. Ainsi, une activité cruciale sur la voie de la convergence entre calcul haute performance (HPC) et HPDA est l'extension des méthodes de calcul haute performance aux traitements de données issues de la biologie. Celle-ci est une discipline en pleine évolution via les besoins de traitements de données très massives en entrée, issues des séquenceurs de nouvelles générations (révolution dite des *omics*). Un aspect est l'étude de la biodiversité par méthodes d'apprentissage sur ces données (plusieurs millions de séquences à traiter). Les outils numériques sous-jacents sont essentiellement des méthodes de réduction de dimension faisant appel à des méthodes d'algèbre linéaire ou multi-linéaire [7] dont le passage à l'échelle reste un challenge scientifique ouvert, aussi bien du point de vue numérique (complexité arithmétique/mémoire et précision) pour la taille des problèmes envisagés que du point de vue informatique pour la taille des machines visées.

L'étude de la biodiversité, classiquement fondée sur une approche naturaliste et prudente de reconnaissances d'espèces et production d'inventaires, est entrée brutalement dans le monde de la valorisation de données massives, produites par des séquenceurs actuellement de troisième génération (domaine du metabarcoding). On connaît des marqueurs qui permettent d'identifier les espèces, et le séquençage massif permet actuellement la production de centaines d'inventaires (qui sont à la base des études en écologie) en quelque semaines uniquement (alors qu'il aurait fallu plusieurs décennies au siècle dernier).

La question est posée de traiter ces données, avec classiquement des méthodes d'analyses multivariées, donc du calcul matriciel (voir [5] pour les grandes dimensions). Les besoins concernent in fine des décompositions spectrales, soit en valeurs singulières, de très grandes matrices pleines (de l'ordre de 10^5 à 10^6 comme dimension). Les algorithmes *standards* sont cubiques selon la dimension n, et ne passent pas facilement à cette échelle. Dans ce cadre, une collaboration entre les équipes Pleiade et Hiepacs (dans le cadre d'une thèse, voir [1]) a permis de connecter le besoin de ces outils pour les études de biodiversité avec un champs récent pour le passage à l'échelle en calcul matriciel dense: les méthodes de projection aléatoires [3,4]. Cela s'est traduit par le développement d'une librairie (`fmr`, voir [1]), qui met en œuvre ces méthodes, et plus largement des sélections de colonnes, avec un parallélisme basé sur OpenMP. Elle est écrite en C++ et fait appel aux bibliothèques standards de calcul matriciel intensif (BLAS, Lapack, Arpack). Dans le cadre de l'approche Multidimensional Scaling (MDS) [2], elle permet de traiter des matrices de dimension 10^5. Les opérations *élémentaires* sont des produits matrice-matrice, matrice-vecteur, et une décomposition QR. Ces opérations peuvent se réaliser par blocs, ce qui permet une distribution

des calculs sur plusieurs nœuds de calcul.

3 Objectif

L'objectif du stage est d'explorer les opportunités offertes par les techniques numériques émergentes en algèbre linéaire et multi-linéaire « data sparse » [1, 3] et leur capacité de mises en oeuvre efficace sous forme d'algorithmes à base de tâches (en utilisant les support d'exécution StarPU/NewMadeleine développés à Inria Bordeaux) pour paralléliser l'algorithme MDS. Le candidat considérera le cas applicatif du metabarcoding. Les cas applicatifs challenges sont aujourd'hui hors de portée. Il faudra ainsi proposer de nouveaux outils numériques parallèles performants à cette communauté scientifique peu utilisatrice des outils HPC et HPDA, donc avec un accent mis également sur l'interface avec les utilisateurs, aussi accessible que possible. Nous nous appuierons sur le solveur d'algèbre linéaire dense à base de tâches chameleon [6], développé au sein de l'équipe HiePACS, afin d'offrir la possibilité de traiter des matrices pleines de dimension 10^6. Le traitement des opérations élémentaires produits matrice-matrice, matrice-vecteur, et une décomposition QR seront ainsi parallélisées en se basant sur chameleon. Les développements seront intégrés au sein de `fmr`. Dans la deuxième partie du stage, on étudiera l'influence de la décomposition sur différents jeux de données. Nous disposons pour cela de dix jeux de données, d'environ 10^5 séquences chacun, concernant des cortèges de diatomées du lac Léman. L'objectif de ce stage est de permettre le traitement de l'ensemble des ces échantillons conjointement, afin de réaliser une intercalibration entre échantillons (unités taxonomiques partagées, ou spécifiques). La taille du jeu de données à traiter pour cela est de 10^6 séquences.

La validation de l'approche se fera faite sur les machines de PlaFRIM et du GENCI pour les plus gros cas tests.

Ce sujet contribue à:

- un rapprochement entre le domaine de la biodiversité et du calcul intensif (biodiversité computationnelle);
- une connexion entre des méthodes issues du calcul HPC (algèbre linéaire, paradigmes de parallélisation MPI et à base de tâche) et des besoins actuels du calcul intensif dans le cadre du traitement de données massives (HTC, paradigme map-reduce).

Il s'inscrit donc pleinement dans le cadre d'une convergence HPC - big data.

4 Mot-clés

Décomposition en valeurs singulières; données massives; algèbre linéaire dense; données massives; parallélisation MPI/OpenMP; programmation à base de tâches; projections aléatoires.

5 Commentaires

Le stage sera effectué sur Bordeaux, au sein de l'équipe HiePACS en partenariat avec l'équipe Pleiade. Il sera possible de le poursuivre en **doctorat** dans le cadre du projet région `hpc-scalable-ecosystem` (*financement de thèse acquis*).

6 Références

[1] Blanchard, P. Fast hierarchical methods for the low-rank approximation of matrices. PhD thesis, Université de Bordeaux, 2016.

[2] T.F. Cox and M. A. A. Cox. Multidimensional Scaling - Second edition, volume 88 of Monographs on Statistics and Applied Probability. Chapman & al., 2001.

[3] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review, 53(2):217-288, 2011.

[4] S. Vempala. The Random Projection Method, volume 65 of DIMACS Series in Discrete Mathematics and Theoretical Computer Sciences. American Mathematical Society, 2004.

[5] J. Wang. Geometric structure in high-dimensional data and dimensionality reduction. Springer & Higher Education Press, 2012.

[6] E. Agullo, O. Aumage, M. Faverge, N. Furmento, F. Pruvost, M. Sergent, S. Thibault, Achieving High Performance on Supercomputers with a Sequential Task-based Programming Model, IEEE Transactions on Parallel and Distributed Systems; 2017

[7] A. J. Izenman. Modern Multivariate Statistical Techniques. Springer, NY, 2008.