Statistique en grande dimension pour la génomique
Projets 2013-2014
L. Jacob, F. Picard, N. Pustelnik, V. Viallon

Contents

1. Optimizing amino acid substitution matrices with a local alignment kernel

Description. Les noyaux définis positifs sont des mesures de similarités entre objets. On peut montrer qu'un noyau entre deux objets est équivalent à un produit scalaire pris sur une certaine description (vectorielle ou fonctionnelle) de ces objets. Cette propriété fait des noyaux un outil utile pour appliquer des outils statistiques ne dépendant des données que via leurs produits scalaires à des objets complexes comme des polymères ou des molécules se prêtant difficilement à des descriptions vectorielles explicites. Ce papier présente un noyau pour séquences biologiques, et son application à la prédiction d'homologies entre protéines.

Travail demandé. (1/2) Oral ∼10 minutes + 5 min de questions. Expliquer le principe des noyaux définis positifs. Expliquer le local alignment kernel: principe et algorithme. (2/2) Pour chaque famille de protéine dans la base COG (`http://www.ncbi.nlm.nih.gov/COG/`), construire une fonction prédisant si une nouvelle protéine appartient à la famille. Evaluer les performances de ces fonctions. Discuter le résultat: certaines familles sont-elles très faciles ou très difficiles à prédire? Code: `http://cbio.ensmp.fr/~jvert/software/LAkernel/LAkernel-0.3.2.tar.gz`, FASTA contenant les proteines: `ftp://ftp.ncbi.nih.gov/pub/COG/COG/myva`, Fichier contenant les classes: `ftp://ftp.ncbi.nih.gov/pub/COG/COG/whog`

Fichier. saigo.2006.pdf

Tuteur. L. Jacob

2. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis

Description. Ce papier présente une famille de méthodes d'analyse exploratoire, qui maximisent certains critères statistiques (variance, covariance) sur des données. Ces maximisations sont faites sous certaines contraintes (sparsité, positivité), censées conduire à des résultats plus interprétables biologiquement.

Travail demandé. (1/2) Oral ∼10 minutes + 5 min de questions. Expliquer le principe de la CCA, et de la sparse CCA (section 4 du papier biostatistics). Principes statistiques, motivation biologique, algorithme. (2/2) Implémentation: sparse CCA sur données synthétiques (code dispo sur la page de D. Witten). Generer des données X, Y dont certaines combinaisons linéaires (éventuellement sparse) de colonnes sont corrélées. Etudier l'effet du nombre d'échantillons, de la dimension de X et Y, des contraintes sur le résultat obtenu.

Fichier. witten.2009.pdf

Tuteur. L. Jacob

3. Efficient RNA Isoform Identification and Quantification from RNA-Seq Data with Network Flows

Description. Ce papier présente un algorithme permettant de calculer efficacement un estimateur de l'expression des isoformes d'un gène dans des données RNA-Seq. Cet es-

timateur (de type maximum a posteriori) implique la maximisation d'une vraisemblance sous contraintes l1 dans un espace de très grande dimension. Le calcul de l'estimateur est rendu possible par l'équivalence de ce problème de maximisation comme un problème d'optimisation de flot sur un graphe.

Travail demandé. (1/2) Oral ∼10 minutes + 5 minutes de questions. Expliquer le problème biologique et l'estimateur proposé (loss de Poisson et contrainte l1). Montrer l'équivalence entre le problème de maximisation de la vraisemblance a posteriori et le problème de flot. (2/2) Implémentation: consignes précises à venir. Utiliser le package flipflop pour estimer l'ensemble d'isoformes présents dans les données qui seront fournies.

Fichier. besnard.2013.pdf

Tuteur. L. Jacob

4. Spatial smoothing and hot spot detection for CGH data using the fused lasso

Description. Cet article propose une méthode d'analyse des données provenant de la technologie des microarrays CGH. Cette technologie permet de mesurer le nombre de copie des gènes le long du chromosome en une seule expérience. La méthode proposée consiste à considérer un modèle de régression particulier (avec X =Identité) dont les paramètres sont sujets à deux contraintes, une sur la valeur absolue des coefficients, et l'autre sur la valeur absolue des différences entre coefficients successifs.

Travail demandé. (1/2) Oral ∼10 minutes + 5 min de questions. Présentez la problématique biologique et la nature des données à analyser, et motivez la stratégie proposée dans l'article. (2/2) Reprendre le schéma de simulation de l'article et étudiez les propriétés empiriques de la méthode en terme de détection des points de rupture.

Fichier. tibshirani.2007.pdf

Tuteur. F. Picard

5. Penalized logistic regression for high-dimensional DNA methylation data with case-control studies

Description. Cet article présente une méthode pour analyser les données de méthylation dans le cadre d'études d'association. Le modèle utilisé est la régression logisique, et la méthode de sélection proposée consiste à utiliser une généralisation de la pénalité l'elastic-net à des variables dont les coefficients sont supposés structurés le long d'un graphe.

Travail demandé. Oral ∼10 minutes + 5 minutes de questions. Présentez la problématiques des données de méthylation et leur originalité par rapport aux données classiques d'expression ou de SNP. Après avoir présenté et discuté la méthode proposée dans l'article (notamment concernant le choix de la pénalisation), vous discuterez de l'influence du graphe sur les performances de la méthode. Que se passe-t-il ? Proposez des interprétations.

Fichier. sun.2012.pdf

Tuteur. F. Picard

6. On Estimating many means, selection bias and the bootstrap

Description. Dans les études d'association sur données génomiques (GWAS, etc..), pour chacun des nombreux marqueurs disponibles, on peut effectuer un test et estimer une *grandeur d'effet* (e.g., odds-ratio conditionnel ou non). Généralement, un intérêt tout particulier est porté aux grandeurs d'effet estimées pour les marqueurs les plus significatifs. On peut cependant montrer que les estimateurs standard de ces grandeurs d'effet sont biaisés, d'autant plus que le nombre de marqueurs initiaux est élevé. Dans cet article, les auteurs proposent une approche pour corriger ce biais et fournir ainsi de meilleurs estimateurs des grandeurs d'effet associées aux marqueurs retenus par l'étude.

Travail demandé. (1/2) Oral ∼10 minutes + 5 min de questions. Expliquer le principe du problème sous-jacent (montrer notamment en quoi ce problème est une illustration de la *régression vers la moyenne*; vous pourrez également illustrer le fait que la loi de $\{\max_i X_i, 1 \leq i \leq n\}$ n'est pas la même que celle des $(X_i)_{1\leq i\leq n}$, par exemple en prenant $X_i \sim \mathcal{U}_{[0,1]}$). Présenter ensuite le principe de la méthode (First order Bias). (2/2) Implémentation: Implémenter la méthode et l'illustrer sur quelques exemples simples (plutôt que le MSE, considérer les biais pour les variables sous (H_0), et sous (H_1) ; étudier ces biais en fonction du design, de n, de p, etc.).

Fichier. simonsimon-2013.pdf

Tuteur. V. Viallon

7. High-dimensional graphs and variable selection with the lasso

Description. Cet article s'intéresse au cadre de l'estimation de la structure dans les modèles graphiques gaussiens, qui permettent l'étude des relations d'indépendance conditionnelle entre les composantes d'un vecteur gaussien. Les auteurs proposent une méthode approchée pour estimer cette structure, qui repose sur l'utilisation de régressions linéaires pénalisées (Lasso). Des résultats théoriques sont obtenus, et la méthode est illustrée sur des données simulées.

Travail demandé. (1/2) Oral ∼10 minutes + 5 min de questions. Après avoir fait quelques rappels sur les vecteurs gaussiens (notamment le théorème de corrélation normale), présenter le principe général de la méthode (en particulier, expliquer pourquoi les indépendances conditionnelles correspondent aux zéros de la matrice de concentration et aux zéros dans les vecteurs de coefficients des régressions linéaires considérées.) Résumer les résultats théoriques du papier et les hypothèses sous lesquelles ils ont été obtenus (en particulier, les rapprocher des hypothèses vues en cours pour établir les propriétés du Lasso). (2/2) Implémentation: Implémenter la méthode et la comparer aux résultats du package glasso sur quelques exemples.

Fichier. meinshausenbuhlmann-2006.pdf

Tuteur. V. Viallon

8. Convex relaxation for permutation problems

Description. Cet article s'intéresse au problème de sériation, qui consiste à ordonner des variables le long d'une chaîne pour laquelle la similarité entre variables décroît avec la distance dans la chaîne. Cette opération s'effectue à partir d'une matrice de similarité (similarité entre deux variables) non organisée et pouvant être bruitée. Le problème associé est un problème d'optimisation non-convexe. Ce travail propose plusieurs relaxations convexes de ce problème. Une application au séquençage de gènes est proposée.

Travail demandé. (1/2) Oral $\sim$ 10 minutes + 5 min de questions. Expliquer le principe de sériation, sa formulation dans le cas de matrices CUT ainsi que l'intérêt de cette formulation. (2/2) Décrire la relaxation convexe basée sur une formulation quadratique (QP), le principe de l'algorithme "block-coordinate descent" et le détail des itérations pour résoudre le problème relaxé.

Fichier. fogel.2013.pdf

Tutrice. N. Pustelnik

---

9. Supervised Feature Selection in Graphs with Path Coding Penalties and Network Flows

Description. Description. Cet article s'intéresse au problème d'apprentissage supervisé en combinant l'hypothèse usuelle de parcimonie avec une hypothèse de structure de graphe. Plus précisément pour un graphe donné a priori sur les variables, les pénalités proposées conduisent à des estimateurs linéaires parcimonieux à l'échelle des groupes de variables, où les groupes correspondent à des chemins sur le graphe. Autrement dit, le support des estimateurs obtenus en minimisant le risque empirique pénalisé de cette manière correspondra typiquement à un petit nombre de chemins sur le graphe des variables. Une application est proposée en pronostic de cancer du sein, où l'on souhaite construire un estimateur n'impliquant qu'un petit nombre de composantes connexes sur le graphe décrivant les interactions connues entre les gènes.

Travail demandé. (1/2) Oral $\sim$ 10 minutes + 5 min de questions. Décrire l'intérêt de cette régularisation, sa formulation, sa relaxation convexe, l'opérateur proximal associé, et l'algorithme utilisé dans le cas convexe.(2/2) Générer des données synthétiques à partir d'un modèle linéaire, où le support de la fonction linéaire correspond à différents nombres de composantes connexes sur un graphe donné. Utiliser le code disponible sur `http://spams-devel.gforge.inria.fr/` pour évaluer sur ces données l'impact de ce type de régularisation en fonction du nombre de composantes connexes. Évaluer également l'impact du paramètre λ sur les propriétés de l'estimateur.

Fichier. mairal.2013.pdf

Tutrice. N. Pustelnik

10. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems

Description. Cet article propose un algorithme itératif permettant de minimiser une somme de deux fonctions convexes dont une est lisse. Ce type de critère est typiquement utilisé pour classifier des données d'expression des gènes (échantillons métastasé versus non-métastasé dans l'étude du cancer du sein). Cette approche est une version rapide de l'algorithme Iterative Soft Thresholding (ISTA).

Travail demandé. (1/2) Oral $\sim$ 10 minutes + 5 min de questions. Présenter les caractéristiques du critère pouvant être résolu par FISTA. Détailler les itérations de l'algorithme et préciser les garanties théoriques de la séquence générée par FISTA. Détailler les avantages/inconvénients de FISTA/ISTA/Forward-Backward. (2/2) En utilisant l'exemple étudié en TP (minimisation logistique + l1), comparer l'algorithme forward-backward et FISTA.

Fichier. beck.2009.pdf

Tutrice. N. Pustelnik