Linear Simplex Support Vector regression

Quentin Klopfenstein [1] & Samuel Vaiter [2]

[1] *Institut Mathématiques de Bourgogne, 9 avenue Alain Savary 21078 DIJON Cedex, quentin.klopfenstein@u-bourgogne.fr*

[2] *CNRS, Institut Mathématiques de Bourgogne, 9 avenue Alain Savary 21078 DIJON Cedex, samuel.vaiter@u-bourgogne.fr*

Résumé. La déconvolution est une méthode utilisée dans la recherche contre le cancer afin d'obtenir la composition cellulaire d'un échantillon tumoral. Cette modélisation mathématique permet à partir de l'expression des gènes de la tumeur d'obtenir les proportions de cellules présentes au sein de la tumeur. L'estimation de ces quantités repose sur des modèles linéaires et notamment sur un modèle de régression à vecteurs de support. De par la nature de ce qui est estimé, l'estimateur obtenu dans le modèle doit avoir des coefficients positifs dont la somme est égale à 1. La méthode la plus utilisée aujourd'hui ne gère ces contraintes que dans un second temps, dans une étape de post-normalisation. Nous proposons donc un nouvel estimateur appelé Linear Simplex Support Vector Regression (LSSVR) qui prend en compte les contraintes liées à l'estimation de proportions directement dans le modèle. Nous étudions l'impact de ce changement dans la qualité de l'estimation et la stabilité de l'estimateur.

Mots-clés. Support Vector Regression, Données de transcriptome, Modèle linéaire sous contraintes

Abstract. Deconvolution is a method used in the cancer research field to quantify cells populations present in a tumor sample. This mathematical modelization of the tumor allows the estimation of the proportions of cells inside the tumor from the genes expression. This quantification is done using linear models especially using a support vector regression (SVR) model. Because of what we are estimating, the estimator given by the model must have positive coefficients and their sum should be equal to 1. The gold standard method used today only manage these constraints in a second step, in a post-normalization step. We propose a new estimator called Linear Simplex Support Vector Regression (LSSVR) which takes into account these constraints directly. We study the impact of this change on the quality of the estimator and its stability.

Keywords. Support Vector Regression, Transcriptomic data, Constrained linear model

1 Introduction

En biostatistiques, la déconvolution est la modélisation linéaire de la relation entre les données d'expression de gènes de différentes cellules et l'expression des gènes d'un

échantillon tumoral. Cette modélisation permet l'estimation de proportions de cellules présentes au sein de la tumeur.

Soit $X \in \mathbb{R}^{l \times n}$ la matrice qui contient les signaux purifiés des populations cellulaires. Cette matrice est appelée matrice de signature. Le nombre de lignes l représente le nombre de gènes et n le nombre de populations cellulaires dont la quantitée est à estimer. Soit également $y \in \mathbb{R}^l$ un signal brouillé obtenu à partir d'un échantillon tumoral dont on veut connaitre la composition cellulaire. On peut alors modéliser la relation entre X et y par un modèle linéaire :

$$y = X\beta + \epsilon \tag{1}$$

où $\beta \in \mathbb{R}^n$ est le vecteur des quantités que l'on estimera à partir des données contenues dans X et y et $\epsilon \in \mathbb{R}^n$ est le terme d'erreur.

Cibersort [9] est le nom de la méthode de référence utilisée aujourd'hui pour faire de la déconvolution. Elle est basée sur l'algorithme de *Support Vector Machine Regression* (SVR) [5]. Cette méthode ne prend pas en compte les contraintes liées à l'estimation de proportions directement dans le modèle mais va en tenir compte seulement dans une étape de projections après l'estimation. Nous proposons un estimateur basé sur la SVR mais qui inclut les contraintes de positivités et de somme des coefficients égale à 1 dans le problème d'optimisation. L'estimateur obtenu peut donc directement être interprété comme des proportions de cellules sans avoir à le normaliser. Ce modèle de *Linear Simplex Support Vector Regression* (LSSVR) ajoute ces deux contraintes linéaires au problème quadratique classique de la SVR. Cela contraint l'estimateur à appartenir au simplexe de probabilités de l'espace auquel il appartient. Cette idée s'inscrit dans la même ligne que les travaux effectués dans [7] et [3] sur l'incorporation de connaissance préalable dans la SVR et la régression ordinale. Nous étudions l'impact de l'intégration de ces contraintes sur la qualité de l'estimation ainsi que la robustesse de l'estimateur quand les données sont bruitées.

2 Cibersort : l'outil de référence

Abbas et al. [1] sont les premiers à considérer le modèle linéaire (1) pour modéliser la relation entre les signaux purs de cellules et le transcriptome d'une tumeur. L'estimateur du vecteur β est calculé en utilisant les moindres carrés ordinaires (MCO). Cela revient à estimer β en résolvant le problème d'optimisation

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} ||y - X\beta||^2 \tag{2}$$

Afin de pouvoir interpréter les coefficients comme des proportions de cellules présentes dans la tumeur, ils appliquent deux étapes de projection :

1. $\hat{\beta}$ est projeté sur l'orthant positif de $\mathbb{R}^n$, cela revient à effectuer élément par élément

$$\hat{\beta}_+ = \max(0, \hat{\beta}) \tag{3}$$

2. Puis $\hat{\beta}_+$ est projeté sur la boule L_1 :

$$\hat{\beta}_S = \frac{\hat{\beta}_+}{||\hat{\beta}_+||_1} \tag{4}$$

Qiao et al. [11] décident d'intégrer les contraintes de positivité des coefficients, directement dans l'étape de l'estimation. Ce qui revient à résoudre le problème appelé Non Negative Least Squares. Ce problème a été très largement étudié et des algorithmes de résolutions efficaces existent [8]. L'estimateur est le résultat du problème suivant :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n_+}{\operatorname{argmin}} ||y - X\beta||^2 \tag{5}$$

L'estimateur final est obtenu en appliquant (4) au vecteur $\hat{\beta}$.

Gong et al. [6] ont proposé une version du problème d'optimisation combinant les deux contraintes liées à l'estimation de proportions. L'estimateur est alors obtenu comme solution du problème :

$$\hat{\beta} = \underset{\beta \in \Delta}{\operatorname{argmin}} ||y - X\beta||^2 \tag{6}$$

où $\Delta = \{x \in \mathbb{R}^n : x_i \geq 0, \forall i \in \{1, \ldots, n\} \text{ et } \sum_{i=1}^n x_i = 1\}$, on appelera cet ensemble Δ le simplexe de probabilités. Condat [4] propose un algorithme de résolution efficace pour ce problème d'optimisation.

Les modèles de classification de type machine à vecteur de support introduit par Vapnik [13] sont très utilisés pour la classification. Une variante du problème d'optimisation permet de proposer un estimateur pour des modèles de régression [5]. C'est un des ces algorithmes de régression que Newman et al. [9] ont utilisé pour leur méthode de déconvolution appelée Cibersort. La version du problème la plus utilisée est la ϵ-SVR mais ils ont choisi d'utiliser la version ν-SVR de Schölkopf [12]. Les différences entre ces deux algorithmes viennent notamment de leurs hyperparamètres. La version ϵ-SVR a deux hyperparamètres : ϵ qui contrôle le seuil à partir duquel une penalité sera appliquée à une observation et C qui contrôle l'erreur globale. La ν-SVR a deux hyperparamètres ν, qui permet de contrôler le nombres de vecteurs de support utilisés pour l'estimation et C qui contrôle l'erreur globale. Le paramètre ϵ est une variable dans le problème d'optimisation. Newman et al. [9] affirment que cet estimateur est moins sensible aux données aberrantes et est plus robuste au bruit. Il est obtenu en résolvant :

$$\begin{aligned} \min_{\beta,b,\xi_i,\xi_i^*,\epsilon} \quad & \frac{1}{2}||\beta||^2 + C(\nu\epsilon + \frac{1}{l}\sum_{i=1}^{p}(\xi_i + \xi_i^*)) \\ \text{sujet à} \quad & y_i - \beta^T x_i - b \leq \epsilon + \xi_i \\ & \beta^T x_i + b - y_i \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0, \epsilon \geq 0 \end{aligned} \tag{7}$$

Le choix des hyperparamètres C et ν est souvent réalisé par validation croisée en choisissant le couple de paramètres qui minimise l'erreur quadratique moyenne par exemple. L'algorithme utilisé par Newman et al. [9] pour obtenir une estimation des populations cellulaires commence par calculer l'estimateur de la SVR classique et applique les projections (3) et (4).

3 Linear Simplex Support Vector Regression (LSSVR)

Nous proposons un estimateur basé sur la SVR qui prend en compte les contraintes liées à l'estimation de proportions. Deux contraintes linéaires sont ajoutées au problème d'optimisation (7) qui vont contraindre l'estimateur à être dans le simplexe de probabilités. Cela permet donc de ne plus avoir à projeter l'estimateur de la SVR classique après l'estimation. L'estimateur que nous proposons est le résultat de :

$$\begin{aligned} \min_{\beta,\epsilon,\xi_i,\xi_i^*,b} \quad & \frac{1}{2}||\beta||^2 + C(\nu\epsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)) \\ \text{sujet à} \quad & (\beta^T x_i + b) - y_i \leq \epsilon + \xi_i \\ & y_i - (\beta^T x_i + b) \leq \epsilon + \xi_i^* \\ & \sum_{j=1}^{n} \beta_j = 1, \beta_j \geq 0, \xi_i, \xi_i^* \geq 0, \epsilon \geq 0 \end{aligned} \tag{8}$$

La solution du problème d'optimisation de la SVR classique est obtenue via des algorithmes de type "optimisation séquentielle minimale"(SMO) [10]. L'algorithme proposé est basé sur une selection de deux variables à optimiser dans le dual à chaque itération. Le point clé de cette méthode est que l'on peut trouver une écriture explicite de la solution du problème à 2 variables. L'implémentation de cet algorithme permet une résolution rapide du problème d'optimisation quadratique et est utilisé dans la librairie libSVM [2] notamment. Pour résoudre (8), il est possible d'utiliser des méthodes basées sur les algorithmes de points intérieurs ou les algorithmes du simplex mais ces méthodes sont lentes et ne permettent pas d'effectuer des tests avec des données de taille conséquente. De plus

ces deux contraintes changent la structure du problème et l'application directe de l'algorithme SMO ne permet pas d'obtenir la solution. Nous utilisons donc un algorithme qui alternera entre des étapes d'optimisation de type SMO de la SVR classique et des étapes d'optimisation des variables du dual correspondant aux contraintes linéaires ajoutées. Cette implémentation permet de résoudre ce problème de manière plus efficace que les "solvers" classiques.

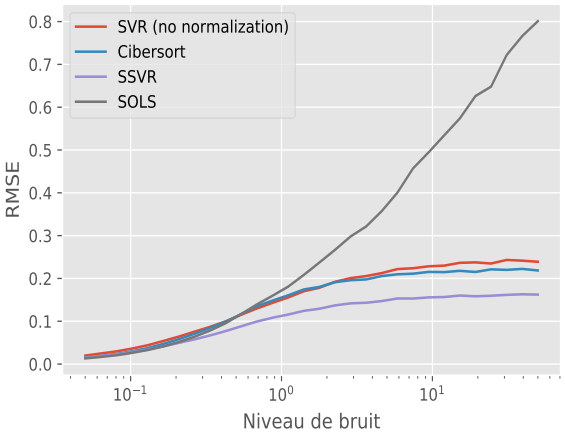

FIGURE 1 – Root Mean Squared Error pour différents estimateurs en fonction de l'écart type de la loi gaussienne utilisée pour bruiter les données.

Nous avons testé la robustesse de l'estimateur aux données bruitées sur des données simulées. Pour cela, nous avons simulé des données qui ont une relation linéaire entre elles et dont les coefficients appartiennent au simplexe de probabilités. Nous avons ajouté à ces données du bruit gaussien. Ce bruit gaussien est centré et des valeurs de 0 à 1 ont été choisies pour l'écart type afin d'augmenter la variation dans les données. Nous avons comparé la qualité de l'estimation pour chaque estimateur en se basant sur la Root Mean Squared Error (RMSE). Les résultats de la figure 1 montrent que l'estimateur LSSVR est celui qui garde la meilleure estimation alors que le bruit augmente. L'algorithme de Cibersort et la SVR sans normalisation sont très proches l'un de l'autre mais leur RMSE reste supérieure à LSSVR. L'estimateur Simplex Ordinary Least Squares (SOLS) (6) est l'estimateur qui perd le plus en qualité d'estimation alors que le bruit augmente. Ce n'est pas vraiment surprenant puisque les méthodes basées sur les MCO sont très sensibles aux données aberrantes et au bruit.

4 Conclusion

L'algorithme LSSVR permet d'éviter les étapes de projections de l'algorithme Cibersort pour estimer des proportions. On observe également une meilleure robustesse de l'estimateur lorsque le bruit dans les données devient plus conséquent. Il faudra confirmer ces observations sur des données réelles. Une des questions sous-jacentes à la modélisation est de savoir si le modèle linéaire est le modèle le plus adapté pour effectuer la quantification des populations cellulaires. Les méthodes de type SVM permettent facilement d'adapter la méthode à d'autres types de relation, grâce notamment au "kernel trick". Il pourra être intéressant de voir s'il est possible d'adapter la LSSVR à d'autres noyaux (gaussien, RBF, sigmoïde, ...)

Références

[1] A. R. Abbas, K. Wolslegel, D. Seshasayee, Z. Modrusan, and H. F. Clark. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLOS ONE*, 4(7) :1–16, 07 2009.

[2] C. Chang and C. Lin. Training v-support vector regression : Theory and algorithms. *Neural Comput.*, 14(8) :1959–1977, August 2002.

[3] W. Chu and S. S. Keerthi. Support vector ordinal regression. *Neural Comput.*, 19(3) :792–815, March 2007.

[4] L. Condat. Least-squares on the simplex for multispectral unmixing. 2017.

[5] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support vector regression machines. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 155–161. MIT Press, 1997.

[6] T. Gong, N. Hartmann, I. S. Kohane, V. Brinkmann, F. Staedtler, M. Letzkus, S. Bongiovanni, and J. D. Szustakowski. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLOS ONE*, 6(11) :1–11, 11 2011.

[7] F. Lauer and G. Bloch. Incorporating prior knowledge in support vector regression. *Machine Learning*, 70, 01 2008.

[8] C. Lawson and R. Hanson. *Solving Least Squares Problems*. Society for Industrial and Applied Mathematics, 1995.

[9] A. M. Newman, C.L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, and A. A. Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5) :453–457, May 2015.

[10] J. Platt. Sequential minimal optimization : A fast algorithm for training support vector machines. page 21, April 1998.

[11] W. Qiao, G. Quon, E. Csaszar, M. Yu, Q. Morris, and P. W. Zandstra. Pert : A method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLOS Computational Biology*, 8(12) :1–14, 12 2012.

[12] B. Schölkopf, P. Bartlett, A. Smola, and R. Williamson. Shrinking the tube : A new support vector regression algorithm. In *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*, pages 330–336, Cambridge, MA, USA, 1999. MIT Press.

[13] V Vapnik and A Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24, 1963.