Régression multivariée régularisée avec variance inconnue et intégration d'information a priori

Julien Chiquet[1,2], Tristan Mary-Huard[2,3] & Stéphane Robin [2]

[1] *UMR CNRS 8071/Université d'Évry Val d'Essonne – Évry, France*
[2] *MMIP – UMR INRA 518/AgroParisTech – Paris, France*
[3] *UMR de Génétique végétale du Moulon, Gif-sur-Yvette, France*

Résumé. Nous proposons un modèle de régression multivariée dont l'apprentissage s'appuie sur trois ingrédients: *i*) estimer la matrice de covariance résiduelle afin de tenir compte de la structure de dépendances entre les multiples variables de réponse; *ii*) sélectionner les liens directs entre réponses et prédicteurs, pour une meilleure interprétabilité; et *iii*) biaiser la sélection par un *a priori* structurel entre prédicteurs pour améliorer la prédiction. Ce modèle s'appuie sur une reformulation du modèle de régression multivariée en un modèle graphique gaussien conditionnel. Pour l'inférence, nous proposons un schéma de régularisation accompagné d'une stratégie d'optimisation efficace. Nous démontrons la bonne tenue de notre approche par rapport à ses concurrents en terme de prédiction à l'aide de simulations. Nous illustrons les capacités de cette approche en terme d'interprétabilité sur des exemples de spectroscopie et de génétique.

Mots-clés. analyse multivariée · sélection de variable · modèle graphique · méthode parcimonieuse · régularisation

Abstract. We propose a regularized method for multivariate linear regression when the number of predictors may exceed the sample size. This method is designed to strengthen the estimation and the selection of the relevant input features with three ingredients: *i*) it takes advantage of the dependency pattern between the responses by estimating the residual covariance; *ii*) it performs selection on direct links between predictors and responses; and *iii*) selection is driven by prior structural information. To this end, we build on a recent reformulation of the multivariate linear regression model to a conditional Gaussian graphical model and propose a new regularization scheme accompanied with an efficient optimization procedure. On top of showing very competitive performance on artificial and real data sets, our method demonstrates capabilities for fine interpretation of its parameters, as illustrated in applications to genetics, genomics and spectroscopy.

Keywords. multivariate analysis · variable selection · graphical model · sparse method · regularization

1 Modèle Statistique.

Le modèle de régression multivariée permet de prédire simultanément q réponses à partir du même ensemble de p prédicteurs. On note $(\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1,\dots,n})$ l'ensemble d'apprentissage où $\mathbf{y}_i \in \mathbf{R}^q$ et $\mathbf{x}_i \in \mathbf{R}^p$ représentent respectivement l'observation des q réponses et p prédicteurs pour l'individu i. Dans le cas gaussien, le modèle s'écrit

$$\mathbf{y}_i = \mathbf{B}^T \mathbf{x}_i + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \quad \forall i = 1, \dots, n, \tag{1}$$

où $\boldsymbol{\varepsilon}_i$ est un vecteur de bruit de matrice de covariance $\mathbf{R}$ inconnue de taille $q \times q$, et $\mathbf{B}$ est la matrice $p \times q$ des coefficients de régression. Ces coefficients décrivent les relations directes et indirectes existantes entre réponses et prédicteurs. Pour nous concentrer sur les liens directs uniquement, nous nous appuyons sur une paramétrisation alternative du modèle (1): en décomposant la matrice des coefficients de régression telle que $\mathbf{B} = \boldsymbol{\Omega}\mathbf{R}$, on peut montrer à l'aide d'arguments sur les vecteurs gaussiens que $\boldsymbol{\Omega}$ est lié aux corrélations partielles entre $\mathbf{x}$ et $\mathbf{y}$, concept statistique classiquement utilisé pour décrire des relations directes entre variables. Sous la paramétrisation alternative $(\boldsymbol{\Omega}, \mathbf{R})$, (1) s'écrit en terme de loi conditionnelle du vecteur de réponses

$$\mathbf{y}_i | \mathbf{x}_i \sim \mathcal{N}(\mathbf{R}\boldsymbol{\Omega}^T \mathbf{x}_i, \mathbf{R}). \tag{2}$$

La log-vraisemblance associée à (2) vérifie

$$-\frac{2}{n} \log L(\boldsymbol{\Omega}, \mathbf{R}) = \log |\mathbf{R}| + \operatorname{tr}\left(\mathbf{S}_{\mathbf{yy}} \mathbf{R}^{-1}\right) + 2\operatorname{tr}\left(\boldsymbol{\Omega}^\intercal \mathbf{S}_{\mathbf{xy}}\right) + \operatorname{tr}(\boldsymbol{\Omega}^\intercal \mathbf{S}_{\mathbf{xx}} \boldsymbol{\Omega}\mathbf{R}) + \text{cst.}, \tag{3}$$

où nous avons noté $\mathbf{S}_{\mathbf{yy}} = n^{-1} \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^T$, $\mathbf{S}_{\mathbf{xx}} = n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$, et $\mathbf{S}_{\mathbf{xy}} = n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{y}_i^T$ les matrices de covariance empirique. Dans [1], une formulation équivalente du modèle défini par (2) et (3) est présenté sous le nom de *modèle graphique gaussien conditionnel*.

2 Critère régularisé structuré et parcimonieux.

Notre schéma de régularisation est construit en deux temps:

1. Tout d'abord, nous cherchons à sélectionner les liens directs dans le modèle. À cet effet, nous utilisons une pénalité ℓ_1 à la manière du LASSO. Nous pénalisons la matrice $\boldsymbol{\Omega}$ liée aux corrélations partielles entre prédicteurs et réponses, et donc aux liens directs.

2. Ensuite, nous voulons introduire une forme de connaissance *a priori* sur les prédicteurs, pour traduire le fait que des prédicteurs similaires doivent avoir les mêmes liens forts avec les réponses. Supposons que la similarité entre prédicteurs peut être encodée à l'aide d'une matrice $p \times p$ notée $\mathbf{L}$. Dans le cadre de la régression bayésienne où

$q = 1$ (voir par exemple [2]), le prior conjugué serait $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{L}^{-1})$. En combinant avec la covariance $\mathbf{R}$ entre réponses, on a dans le cas général où $q \geq 1$,

$$\text{vec}(\mathbf{B}) \sim \mathcal{N}(\mathbf{0}, \mathbf{R} \otimes \mathbf{L}^{-1}),$$

où $\otimes$ est le produit de Kronecker. À l'aide des propriétés usuelles de l'opérateur vec, le prior correspondant pour $\boldsymbol{\Omega}$ s'écrit

$$\text{vec}(\boldsymbol{\Omega}) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}^{-1} \otimes \mathbf{L}^{-1}).$$

Le terme de pénalité correspondant – où régularisation – est lié au log du prior, i.e.

$$\log \mathbb{P}(\boldsymbol{\Omega}|\mathbf{L}, \mathbf{R}) = \frac{1}{2}\text{tr}\left(\boldsymbol{\Omega}^T \mathbf{L} \boldsymbol{\Omega} \mathbf{R}\right) + \text{cst}.$$

Ainsi, nous proposons un critère qui s'appuie sur la log-vraisemblance (3), pénalisée par une norme ℓ_1 pour la sélection et par une pénalité intégrant la connaissance *a priori* via $\mathbf{L}$. Notre fonction objectif est convexe jointement en $(\boldsymbol{\Omega}, \mathbf{R}^{-1})$ pour (λ_1, λ_2) fixés et s'écrit

$$J(\boldsymbol{\Omega}, \mathbf{R}) = -\frac{1}{n} \log L(\boldsymbol{\Omega}, \mathbf{R}) + \frac{\lambda_2}{2}\text{tr}\left({}^t\boldsymbol{\Omega} \mathbf{L} \boldsymbol{\Omega} \mathbf{R}\right) + \lambda_1 \|\boldsymbol{\Omega}\|_1.$$

Nous avons implémenté une procédure d'optimisation pour minimiser ce critère sur une grille de de pénalités (λ_1, λ_2) dans un package `R` prochainement disponible sur le `CRAN`.

2.1 Choix des paramètres de tuning

Dans les méthodes de régression pénalisée type LASSO, il est d'usage d'utiliser la validation croisée pour choisir le modèle final, c'est-à-dire les paramètres de tuning (λ_1, λ_2). Lorsque le nombre n d'échantillons le permet, une stratégie alternative est d'utiliser un critère type BIC, soit dans ce cas,

$$(\lambda_1^\star, \lambda 2^\star) = \arg\min_{\lambda_1, \lambda_2} \left\{ -2 \log L(\hat{\boldsymbol{\Omega}}^{\lambda_1, \lambda_2}, \hat{\mathbf{R}}^{\lambda_1, \lambda_2}) + \log(n) \cdot \text{df}_{\lambda_1, \lambda_2} \right\}.$$

Les degrés de libertés, notés df ci-dessus, sont compris au sens défini par Efron [3]. À l'aide des travaux de Tibshirani et Taylor [4] sur les degrés de liberté des problèmes de la famille du LASSO, nous pouvons montrer dans notre cas que

$$\text{df}_{\lambda_1, \lambda_2} = \text{card}(\mathcal{A}) - \lambda_2 \text{tr}\left((\hat{\mathbf{R}} \otimes \mathbf{L})_{\mathcal{A}\mathcal{A}} (\hat{\mathbf{R}} \otimes (\mathbf{S}_{\mathbf{xx}} + \lambda_2 \mathbf{L}))^{-1}_{\mathcal{A}\mathcal{A}} \right),$$

où $\mathcal{A} = \left\{ j : \text{vec}\left(\hat{\boldsymbol{\Omega}}^{\lambda_1, \lambda_2}\right) \neq 0 \right\}$ est l'ensemble des éléments actifs (non-nuls) dans $\hat{\boldsymbol{\Omega}}^{\lambda_1, \lambda_2}$.

2.2 Une application en sélection génomique

Contexte.

En sélection génomique, on cherche à prédire des traits phénotypiques en utilisant l'information génétique disponible sous forme de marqueurs. En génétique animale ou végétale, il est de première importance de disposer de prédictions précises sur des caractères complexes, afin de détecter en amont les individus au patrimoine génétique de grande valeur.

Nous illustrons notre méthode sur l'étude conduite dans [5] où $n = 103$ lignées de colza (*Brassica napus*) sont considérées, issues de 2 cultivars (ou "souche"), 'Stellar' ou 'Major'. Les prédicteurs correspondent à $p = 300$ marqueurs à valeur dans $\{0, 1\}$, chaque valeur correspondant à un cultivar. Les $q = 8$ caractères – ou traits – regroupent 5 pourcentages de survie des lignées au cours des hivers 1992, 1993, 1994, 1997 et 1999, et 3 durées en jours avant floraisons après vernalisation au bout de 0, 4 et 8 semaines.

Spécification de la structure *a priori*.

Pour des populations de lignée biparentale, la corrélation entre 2 marqueurs dépend de la distance génétique entre ces marqueurs, définie en terme de taux de recombinaison. Ainsi, on s'attend à ce que des marqueurs adjacents soient corrélés et induisent des relations directs similaires sur les traits.

En notant d_{12} la distance génétique entre deux marqueurs X^1 et X^2, et $\rho = .98$[1], on a

$$\mathrm{cor}(X^1, X^2) = \rho^{d_{12}} \ .$$

La covariance *a priori* $\mathbf{L}^{-1}$ peut donc être définie par $\mathbf{L}^{-1}_{ij} = \rho^{d_{ij}}$. De plus, en supposant que les événements de recombinaison sont indépendants entre X^1 et X^2 d'une part, et entre X^2 et X^3 d'autre part, on a $d_{13} = d_{12} + d_{23}$. Ainsi, La matrice $\mathbf{L}^{-1}_{ij}$ a un profile de type AR(1) no-homogène. En conséquence, $\mathbf{L}$ est tridiagonale de terme général

$$\begin{aligned} w_{i,i} &= \frac{1 - \rho^{2d_{i-1,i}+2d_{i,i+1}}}{(1 - \rho^{2d_{i-1,i}})(1 - \rho^{2d_{i,i+1}})}, \\ w_{i,i+1} &= \frac{-\rho^{d_{i,i+1}}}{1 - \rho^{2d_{i,i+1}}} \end{aligned}$$

et $w_{i,j} = 0$ si $|i - j| > 1$. Pour le premier (resp. le dernier) marqueur, la distance $d_{i-1,i}$ (resp. $d_{i,i+1}$) est infinie.

[1] Cette valeur vient directement de la définition de la distance génétique elle-même.

Résultats.

Sur le panneau de gauche de la figure 1, on représente simultanément les coefficients de régression $\hat{\mathbf{B}}$ estimés (en haut) et les liens directs $\hat{\mathbf{\Omega}}$. Les zones grises correspondent aux chromosomes 2, 8 et 10, respectivement. La localisation exacte des marqueurs au sein de ces chromosomes est représentée sur le panneau de droite, où la taille des points reflète le niveau d'intensité des coefficients de régression (en haut) et des liens directs (en bas). L'intérêt de considérer les effets directs plutôt que les coefficients de régression

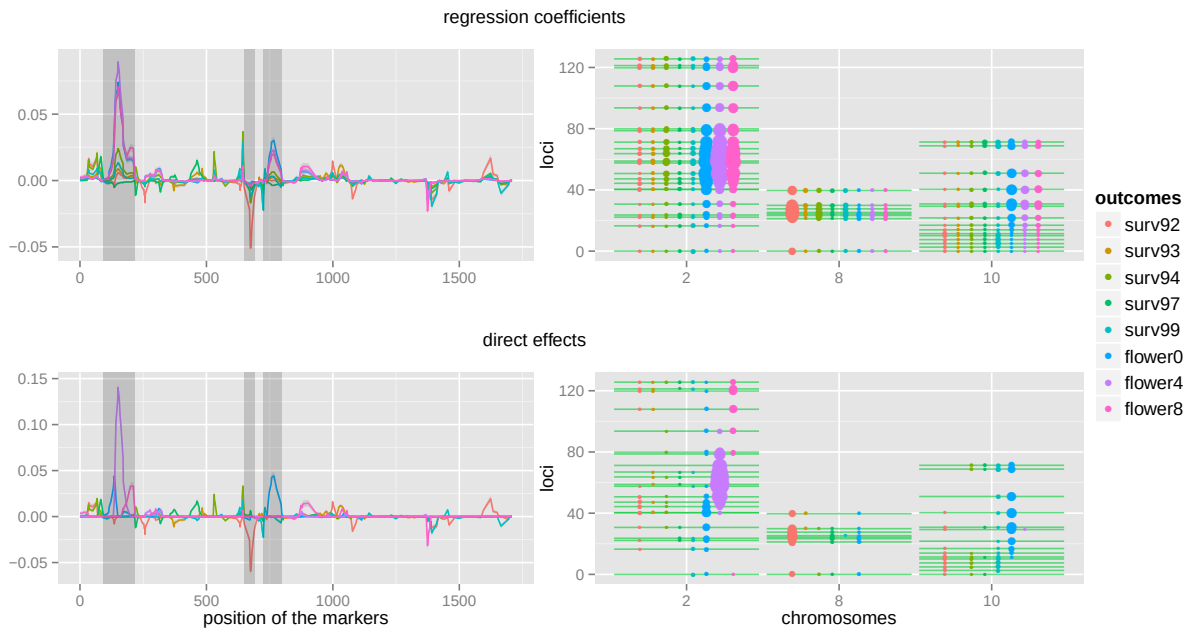

Figure 1: Estimation des effets génétiques directs $\mathbf{\Omega}$ et indirects $\mathbf{B}$ des marqueurs

apparaît évident à la figure 1, par exemple pour le chromosome 2: sur le graphe des coefficients de régression, trois larges régions chevauchantes sont visibles pour chaque traits. Une interprétation hâtive suggérerait que cette région contrôle le processus de floraison dans son ensemble. Le graphe des effets directs permet une interprétation plus fine, et montre que ces trois caractères sont en fait contrôlés par trois sous-régions séparées sur le chromosome. La confusion sur le graphe des coefficients n'est simplement dû qu'à de fortes corrélations observées entre les traits de floraison, que notre modèle a intégré via la matrice de covariance résiduelle $\mathbf{R}$.

Bibliographie

[1] Sohn, K., Kim, S.: Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. JMLR **W&CP**, 2012.

[2] Marin, J.-M., Robert, Ch. P. Bayesian Core: A Practical Approach to Computational Bayesian Statistics. Springer-Verlag: New-York, 2007.

[3] Efron, B. How biased is the apparent error rate of a prediction rule? J. Amer. Statist. Assoc. **81** 461–470, 1986

[4] Ryan J. Tibshirani and Jonathan Taylor: Degrees of freedom in lasso problems. Ann. Statist. **40**, 639–1284, 2012.

[5] Ferreira, M., Satagopan, J., Yandell, B., Williams, P., Osborn, T.: Mapping loci controlling vernalization requirement and flowering time in brassica napus. Theor. Appl. Genet. **90** (1995) 727–732