

Sur la taille du support de l'optimum OCS génomique

Adel Kaleche

Chercheur indépendant, France.

Version française ; note de référence en anglais.

11 octobre 2026

Résumé

Note compagne du solveur support-first. Pour la sélection à contribution optimale (OCS) génomique, on encadre la taille du support optimal : une borne existentielle du support total indépendante de n tient sans plafonds effectifs dans la limite sans ridge (Théorème 1), aucune borne universelle ne survit au ridge (Théorème 2), et le régime pratique entre les deux est gouverné par la géométrie conjointe du spectre, des valeurs génétiques et du cap, sans prédicteur scalaire fiable parmi ceux testés. Une identité KKT conditionnelle (Proposition 1) décrit les contributions libres. Les expériences historiques sont conservées avec leur contexte ; scripts et prérequis sont indiqués ci-dessous.

1 La question

L'OCS génomique résout, pour n candidats de valeurs génétiques $\mathbf{b} \in \mathbb{R}^n$ et de matrice de parenté $\mathbf{G} = \mathbf{Z}\mathbf{Z}^\top / s + \varepsilon \mathbf{I}$ ($\mathbf{Z}$ la matrice $n \times m$ des génotypes centrés, s une échelle, ε un petit ridge assurant la définie-positivité),

$$\max \mathbf{b}^\top \mathbf{c} \quad \text{s.c.} \quad \mathbf{A}\mathbf{c} = \mathbf{d}, \quad \mathbf{0} \leq \mathbf{c} \leq \mathbf{u}, \quad \mathbf{c}^\top \mathbf{G}\mathbf{c} \leq k, \quad (1)$$

où $\mathbf{A}$ est la matrice de budget $q \times n$ ($q = 1$ pour le simplexe $\mathbf{1}^\top \mathbf{c} = 1$; $q = 2$ pour la forme sexuée $\sum_{\text{mâles}} = \sum_{\text{femelles}} = \frac{1}{2}$). Empiriquement l'optimum $\mathbf{c}^*$ active très peu de candidats : son support $S = \{i : c_i^* > 0\}$ était $\approx 15\text{--}30$ dans les essais historiques à cap lâche. Aux points opérationnels du proxy de contributions du manuscrit, il va de quelques dizaines à quelques centaines (souris : 47/89/182 à $\Delta F = 2/1/0,5\%$) ; le balayage structuré en n donne environ 80–103 à proxy fixé. Ces observations ne prouvent pas de borne universelle. Qu'est-ce qui borne $|S|$?

La réponse est un *encadrement*. Sans plafonds effectifs et sans ridge, une borne existentielle indépendante de n existe (§2) ; dès que le ridge est présent, il n'y a prouvablement aucune borne universelle, le support pouvant atteindre toute la population (§3) ; et le régime pratique entre les deux est fixé par la géométrie conjointe de (spectre, $\mathbf{b}, k$), dont la carte empirique ne fournit pas de prédicteur scalaire fiable parmi ceux testés (§5). Une identité KKT conditionnelle (§4) décrit les contributions libres.

2 La borne $\varepsilon = 0$

Théorème 1. *Supposons $\varepsilon = 0$, $s > 0$, un ensemble réalisable non vide et les budgets simplex ou sexués, de sorte que $\{\mathbf{A}\mathbf{c} = \mathbf{d}, \mathbf{c} \geq 0\}$ est borné. Sans plafonds supérieurs effectifs, il existe un optimum vérifiant*

$$|S| \leq \text{rang} \begin{bmatrix} \mathbf{A} \\ \mathbf{Z}^\top \\ \mathbf{b}^\top \end{bmatrix} \leq q + r + 1, \quad r = \text{rang}(\mathbf{Z}) \leq m.$$

Démonstration. Partons d'un optimum atteint $\mathbf{c}^*$ et figeons son image et son objectif dans

$$P = \{\mathbf{c} \geq 0 : \mathbf{A}\mathbf{c} = \mathbf{d}, \mathbf{Z}^\top \mathbf{c} = \mathbf{Z}^\top \mathbf{c}^*, \mathbf{b}^\top \mathbf{c} = \mathbf{b}^\top \mathbf{c}^*\}.$$

Tout point a la même parenté $\|\mathbf{Z}^\top \mathbf{c}^*\|^2/s$ et le même objectif : il est optimal. P est non vide et borné. La matrice $M = [\mathbf{A}; \mathbf{Z}^\top; \mathbf{b}^\top]$ est de rang au plus $q+r+1$. Un sommet avec plus de rang(M) coordonnées positives admettrait une direction non nulle du noyau portée par ces coordonnées. Deux petites perturbations opposées conserveraient égalités et positivité, contredisant son caractère extrême. Un sommet fournit la borne. Aucune dualité ni qualification de contraintes n'est requise. $\square$

Extension aux plafonds finis. Intersectons P avec $\mathbf{c} \leq \mathbf{u}$. Un sommet a au plus rang(M) coordonnées *libres* $F = \{i : 0 < c_i < u_i\}$, par le même argument. Les coordonnées $U_+ = \{i : c_i = u_i > 0\}$ comptent également, donc $|S| = |F| + |U_+| \leq \text{rang}(M) + |U_+|$. Des plafonds finis arbitraires interdisent une borne du support total indépendante de $n : q = 1, r = 0, \mathbf{G} = 0, k = 0, u_i = 2/n$ imposent $|S| \geq \lceil n/2 \rceil$, alors que $q + r + 1 = 2$. Un $\mathbf{b}$ strictement décroissant atteint ce minimum en un sommet LP optimal plafonné.

Le théorème porte sur l'existence d'un sommet d'une *tranche optimale*. Rien ne prouve que l'ensemble actif implémenté renvoie ce sommet, ni qu'un solveur à point intérieur ou ADMM ne puisse rendre une solution parcimonieuse. Imposer cette garantie à une sortie arbitraire nécessiterait une extraction séparée de sommet.

3 Aucune borne universelle pour $\varepsilon > 0$

Théorème 2. *Pour $\varepsilon > 0$, aucune borne du support total $f(q, r)$ indépendante de n n'existe, même sans plafonds effectifs : le support peut valoir n .*

Démonstration (contre-exemple). Prenons l'instance OCS abstraite $q = 1, \mathbf{G} = \varepsilon \mathbf{I}, r = 0, u_i = 1$. Ce contre-exemple algébrique n'est pas un panel de marqueurs utilisable dans une interface publique. Sur le simplexe, $\|\mathbf{c}\|^2 \geq 1/n$ par Cauchy-Schwarz. Si une coordonnée est nulle, $\|\mathbf{c}\|^2 \geq 1/(n-1)$. Ainsi, pour

$$\varepsilon/n < k < \varepsilon/(n-1),$$

le plan uniforme est strictement réalisable, mais tout plan réalisable, donc tout optimum pour n'importe quel $\mathbf{b}$, a un support plein. Cela vaut pour tout $n \geq 2$. $\square$

L'ancien balayage (`bound_validation.py`, bloc 7, $n = 300$) donnait

$$|S| = 300, 295, 256, 151, 52, 12,$$

quand k se desserrait de $1,05\times$ à $50\times$ le minimum. Ce sont des observations pour un $\mathbf{b}$; $1,05\times$ est hors de l'intervalle analytique de support plein et n'est pas un témoin universel. Le bloc 8 vérifie cet intervalle et l'obstruction des plafonds finis.

Ici $\sum c_i^2$ mesure la concentration des contributions. Avec $N_e = 1/\sum c_i^2$, $\Delta F = 1/(2N_e)$ est un proxy fondé sur les contributions, pas un taux mesuré de consanguinité génomique réalisée. Le petit support observé dépend de la géométrie conjointe de l'instance.

4 Une identité KKT conditionnelle

Proposition 1. *Supposons $\varepsilon > 0$ et un optimum disposant de multiplicateurs KKT avec $\lambda > 0$ pour la parenté. Sur $F = \{i : 0 < c_i^* < u_i\}$,*

$$c_i^* = \alpha b_i + \mathbf{w}^\top \mathbf{z}_i + v^\top \mathbf{a}_i, \quad i \in F,$$

a_i étant la colonne i de $\mathbf{A}$. Pour les budgets d'incidence de sexe, $v^\top a_i = \beta_{\text{sexe}(i)}$.

Démonstration. Avec les multiplicateurs inférieur et supérieur $\ell, h \geq 0$, stationnarité et complémentarité sont

$$\mathbf{b} = \mathbf{A}^\top \mu + 2\lambda \mathbf{G} \mathbf{c}^* - \ell + h, \quad \ell_i c_i^* = 0, \quad h_i(u_i - c_i^*) = 0, \quad \lambda(\mathbf{c}^{*\top} \mathbf{G} \mathbf{c}^* - k) = 0.$$

Sur F , les deux multiplicateurs de borne s'annulent. Avec $\mathbf{y} = \mathbf{Z}^\top \mathbf{c}^*$ et $\mathbf{G} \mathbf{c}^* = \mathbf{Z} \mathbf{y} / s + \varepsilon \mathbf{c}^*$, la division par $2\lambda\varepsilon$ donne $\alpha = 1/(2\lambda\varepsilon)$, $\mathbf{w} = -\mathbf{y}/(s\varepsilon)$, $v = -\mu/(2\lambda\varepsilon)$. $\square$

L'identité comporte $m + q + 1$ coefficients ($m + 2$ en simplex, $m + 3$ pour deux budgets sexuels). Une parenté active ne suffit pas à imposer $\lambda > 0$. L'existence des multiplicateurs exige une qualification appropriée ou un certificat fourni ; elle est supposée ici. L'identité ne s'applique pas aux coordonnées au plafond et ne borne aucune cardinalité. Le passage $\varepsilon \rightarrow 0$ ne retrouve pas le Théorème 1 : supports et multiplicateurs peuvent changer ou diverger, et la stationnarité ne vaut que sur les coordonnées actuellement libres. La preuve par tranche optimale est indépendante.

5 Le régime empirique entre les théorèmes

Entre les deux énoncés propres se trouve le régime que les sélectionneurs utilisent, cartographié sur des spectres synthétiques, des alignements de $\mathbf{b}$ et des caps, ainsi que sur les panels réels :

- **L'alignement de $\mathbf{b}$ avec le spectre gouverne $|S|$, inversement.** À $\mathbf{G}$ et k fixés, placer $\mathbf{b}$ sur la direction propre dominante (chère en coancestrie) force un grand support (moyenne $|S| \approx 146$ sur $n = 800$ pour $\mathbf{b}$ le vecteur propre de tête) ; étaler $\mathbf{b}$ sur de nombreuses directions bon marché l'effondre à ≈ 4 . Le gain cherché dans une direction chère doit être dilué sur beaucoup de candidats pour respecter le cap (`bound_balign.py`).
- **Aucun scalaire unique ne prédit $|S|$ à travers les régimes.** Ni $\text{rang}(\mathbf{G}_0)$, ni le rang effectif / ratio de participation du spectre, ni le coût directionnel $\mathbf{b}^\top \mathbf{G} \mathbf{b} / \mathbf{b}^\top \mathbf{b}$, ni le support de la direction non contrainte $(\mathbf{G}^{-1} \mathbf{b})_+$ ne suit $|S|$ à travers les instances synthétiques et réelles ; la meilleure combinaison sans dimension, $|S| \sim (\mathbf{b}^\top \mathbf{G} \mathbf{b} / \mathbf{b}^\top \mathbf{b} \cdot 1/k)^{0,8}$, n'atteint que $R^2 \approx 0,58$ et rate le support *relatif* des panels réels (`bound_predictor.py`, `bound_lawfit.py`).
- **La courbe $|S|(k)$ est une loi de puissance par instance, d'exposant non universel.** $|S| \sim k^{-\alpha}$ ajuste bien chaque instance ($R^2 \approx 0,9\text{--}0,98$ là où le support est bien résolu), avec $\alpha \approx 1$ sur les deux panels réels mais variant de 0,5 à 1,7 selon les spectres et alignements synthétiques (`bound_curve.py`).
- **Panels réels.** Blé ($n = 599$) et souris ($n = 1814$) : $|S| = 25/26$ et $17/26$ (simplexe / optiSel sexé) au cap historique lâche. Ces deux panels fixes ne prouvent pas une stabilité en n ; les cibles opérationnelles du manuscrit ont des supports plus grands (`bound_real.py`).

Le Théorème 2 exclut une borne universelle en q, r seuls. Les balayages empiriques montrent séparément que les prédicteurs scalaires testés ne prédisent pas le support de façon fiable entre régimes.

6 Une borne conditionnelle : l'obstruction et la conjecture

Une borne conditionnelle reste ouverte. Si la contrainte de parenté est *redondante sur tout le polytope budgétaire*, un LP non plafonné admet un sommet optimal de support $|S| \leq \text{rang}(\mathbf{A}) \leq q$;

les plafonds finis donnent seulement $|F| \leq \text{rang}(\mathbf{A})$. Une parenté non saturée en un optimum ne suffit pas : pour $\mathbf{b}$ constant, $\mathbf{G} = \mathbf{I}$ et $1/n < k < 1/(n-1)$, l'optimum uniforme est strictement réalisable mais tout optimum réalisable a un support plein. Au rang exact r et à $\varepsilon = 0$, le Théorème 1 donne la borne existentielle non plafonnée ; son extension borne les coordonnées libres.

Une borne en *rang effectif* — l'objectif réaliste, puisque les vrais spectres décroissent — ne découle pas du même argument, et l'obstruction est précise. Le Théorème 1 linéarise la tranche optimale en épinglant $\mathbf{Z}^\top \mathbf{c}$, soit *toute* l'image de $\mathbf{G}_0$; pour un $\mathbf{G}$ de plein rang il faudrait épingler les n projections $u_i^\top \mathbf{c}$, ce qui ne donne que le trivial $|S| \leq q + n + 1$. On ne peut pas n'épingler que les d directions de tête et négliger la queue, car celle-ci contribue $\mathbf{c}^\top \mathbf{E} \mathbf{c} = \sum_{i>d} \lambda_i (u_i^\top \mathbf{c})^2$ — une *vraie quadratique*, pas un terme de bas rang — donc la fixer n'est pas une contrainte linéaire et la tranche n'est plus un polytope. La courbure de la queue est exactement ce qui laisse le support s'étaler ; le rang effectif ne peut pas entrer dans un comptage de sommets. Une borne conditionnelle exige donc un autre argument (perturbation spectrale).

Deux conditions qualitatives motivent un éventuel théorème conditionnel ; elles ne sont ni quantitativement spécifiées ni prouvées suffisantes :

- **(A1) Pas de grand plateau dégénéré bon marché.** Un bloc de nombreuses petites valeurs propres quasi-égales est un sous-espace bon marché où l'optimum s'étale ; $\mathbf{G} = \varepsilon \mathbf{I}$ est le cas extrême et donne $|S| = n$ (Théorème 2). Un spectre régulièrement décroissant n'a pas un tel plateau.
- **(A2) $\mathbf{b}$ évite les directions dominantes.** Même avec un gap, $\mathbf{b}$ sur le vecteur propre de tête (cher) force la dilution et un grand support (moyenne $|S| \approx 146$ sur 800 dans le balayage d'alignement).

Le contre-exemple du ridge pur et le balayage d'alignement motivent (A1) et (A2), sans prouver la nécessité de ces conditions informelles. Sous des versions quantitatives à point de fonctionnement spécifié, on conjecture un petit support ; les 15–30 historiques concernent les caps lâches, pas tout panel réel ; la moitié croissante ($|S|$ augmentant quand k approche le minimum réalisable) suit la filiation proxy de contributions $\Delta F = 1/(2N_e)$, distinct de la consanguinité génomique réalisée (Wray & Thompson 1990 ; Woolliams & Bijma 2000). La prouver — coupler le gap spectral à la projection de $\mathbf{b}$ sur le sous-espace dominant, en résistant à la courbure de la queue — est le problème ouvert, point de rencontre de l'optimisation (la face du cône perturbé) et de la génétique quantitative (lignées effectives).

Une prise exacte sur le support effectif. Écrivons le nombre *effectif* de contributeurs comme le ratio de participation $\text{PR}(\mathbf{c}^*) = 1/\|\mathbf{c}^*\|^2$ (égal à $|S|$ pour des contributions uniformes, moindre si inégales). Au cap actif positif $k > 0$ avec $\mathbf{1}^\top \mathbf{c}^* = 1$, une identité tient : $\text{PR}(\mathbf{c}^*) = R(\mathbf{c}^*)/k$, où $R(\mathbf{c}^*) = \mathbf{c}^{*\top} \mathbf{G} \mathbf{c}^* / \mathbf{c}^{*\top} \mathbf{c}^*$ est le quotient de Rayleigh propre de l'optimum (car $R\|\mathbf{c}^*\|^2 = \mathbf{c}^{*\top} \mathbf{G} \mathbf{c}^* = k$). Donc $\text{PR}(\mathbf{c}^*) \leq \lambda_1/k$ sous ces conditions de cap actif positif, et une borne du support *effectif* se réduit à borner $R(\mathbf{c}^*)$ — le coût directionnel de coancestrie que l'optimum est forcé d'encaisser. Dans les cas mesurés où $\mathbf{b}$ évite les directions dominantes, $R(\mathbf{c}^*)$ est petit (observé : $R \approx 0,02$ – $0,4$ sur les cas structurés contre $\lambda_1 \approx 45$ – 143 , donnant $\text{PR} \approx 2$ – 12), tandis que $\mathbf{G} = \varepsilon \mathbf{I}$ force $R = \varepsilon = \lambda_1$. Au cap actif, $\text{PR} = \varepsilon/k$ appartient à $(n-1, n)$ dans tout l'intervalle de support plein $\varepsilon/n < k < \varepsilon/(n-1)$, et tend vers n lorsque k décroît vers ε/n . Le support total vaut déjà n dans tout cet intervalle ; $\text{PR} = n$ exige des contributions uniformes. Par exemple, $n = 3$ et $k = 5\varepsilon/12$ donnent $|S| = 3$ mais $\text{PR} = 12/5$. La pièce ouverte résiduelle est désormais nette et scalaire — borner $R(\mathbf{c}^*)$ a priori sous (A1)+(A2). (La cardinalité brute valait historiquement $\approx 1,5$ – $2,5 \times \text{PR}$ dans ces cas. En général $\text{PR} \leq |S|$: borner PR supérieurement ne borne pas $|S|$ sans hypothèse supplémentaire sur les contributions.) Vérifié dans `bound_kernel.py`.

Attaquer ce résidu. Pour $\mathbf{G}$ définie positive, la relaxation simplex sans bornes inférieures ni supérieures a, dans le cas à parenté active avec $\lambda > 0$, l'expression du quotient de Rayleigh

$R_{\text{relax}} = (\beta^\top \mathbf{G}^{-1} \beta) / (\beta^\top \mathbf{G}^{-2} \beta)$ avec $\beta = \mathbf{b} - \mu \mathbf{1}$ — une moyenne des λ pondérée par β_i^2 / λ_i^2 , donc petite sous (A2). Mais elle ne prédit $R(\mathbf{c}^*)$ que lorsque $\mathbf{c} \geq 0$ est quasi-inactif : sur les panels réels l’optimum sans contrainte porte de grandes composantes négatives (min ≈ -1 à $-2,6$) et R_{relax} tombe $60\text{--}300\times$ sous $R(\mathbf{c}^*)$. Donc **la non-négativité est le cœur irréductible** — de fait, $|S|$ est la cardinalité de l’optimum d’un QP contraint en positivité, dur en général — et une preuve doit être *genuinely* $\mathbf{c} \geq 0$ -aware, pas un court-circuit par relaxation. La prise a priori (la formule de relaxation) et l’obstacle irréductible ($\mathbf{c} \geq 0$) sont désormais tous deux identifiés.

Évidence numérique (reproductible)

`bound_validation.py` (observations de bas rang avec ridge, pas un test du théorème de sommet à $\varepsilon = 0$, le balayage empirique en n , les balayages ridge et décroissance spectrale, et le balayage historique du Théorème 2 et témoins analytiques du bloc 8); `bound_real.py` (blé et souris, après `research/repro/*_export.R`); `bound_balign.py` (le balayage d’alignement de $\mathbf{b}$); `bound_predictor.py` et `bound_lawfit.py` (la recherche de prédicteur / de loi); `bound_curve.py` (l’ajustement de $|S|(k)$ par instance). Journal complet : `research/support_bound_sketch.md`.

Références

1. **Pataki, G. (1998).** On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Math. Oper. Res.* 23(2) :339–358. DOI 10.1287/moor.23.2.339.
2. **Markowitz, H. (1956).** The optimization of a quadratic function subject to linear constraints. *Naval Res. Logist. Q.* 3(1–2) :111–133. DOI 10.1002/nav.3800030110.
3. **Wray, N.R. & Thompson, R. (1990).** Prediction of rates of inbreeding in selected populations. *Genet. Res.* 55(1) :41–54. DOI 10.1017/S0016672300025180.
4. **Woolliams, J.A. & Bijma, P. (2000).** Predicting rates of inbreeding in populations undergoing selection. *Genetics* 154(4) :1851–1864. DOI 10.1093/genetics/154.4.1851.
5. **Yamashita, M., Mullin, T.J. & Safarina, S. (2018).** An efficient second-order cone programming approach for optimal selection in tree breeding. *Optim. Lett.* 12(7) :1683–1697. DOI 10.1007/s11590-018-1229-y.
6. **Waldmann, P. (2025).** Genomic optimum contribution selection and mate allocation using JuMP. *Bioinform. Adv.* 5(1) :vbaf259. DOI 10.1093/bioadv/vbaf259.