PCA SPARSE

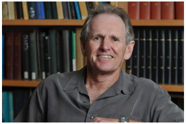
Non supervisé avancé: cours de C.Keribin

Malkiel Riveline

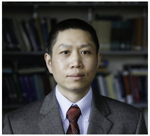
30 novembre 2023

Auteurs

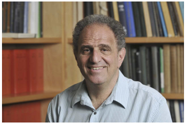
Sparse Principal Component Analysis *Journal of Computational and Graphical Statistics* (2006)



Trevor Hastie



Hui Zou

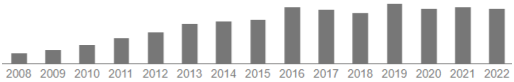


Robert Tibshirani

STANFORD UNIVERSITY

Bibliométrie

- Article plutôt théorique en sept parties dont la plus importante est la 3.
- Cité 3789 fois selon Google Scholar.
- Indice à combiner avec l'article introduisant elastic net (plus de 19000 citations).
- Utilisation récente (300 par an).



Plan

Pourquoi la réduction de dimension?

- Pour n observations et p variables, l'objectif est de transformer un nuage de points et des variables corrélés en des variables décorrélées.
- On considère donc une matrice de covariances des données $X^T X$, centrée.
- On veut réduire la dimension tout en gardant le plus d'informations possibles.

Représentations équivalentes

On a deux estimations, qui sont équivalentes pour la PCA:

1. Minimiser l'erreur de reconstruction pour une matrice de données centrées (problème des moindres carrés) $\rightarrow$ Décomposition en valeurs singulières (SVD).
2. Maximiser la variance expliquée sur la combinaison linéaire pour une matrice de covariance donnée (problème des valeurs propres).

Formulation mathématique du problème de minimisation

Soit x_i la i-ème ligne de X. Considérant les k premières composantes principales conjointement $V^k = [V_1|\ldots|V_k]$, V^k est une matrice orthogonale de dimensions $p \times k$. Une manière de définir la meilleure projection est de minimiser l'erreur totale d'approximation ℓ_2 :

$$\min_{V^k} \sum_{i=1}^{n} \|x_i - V^k(V^k)^T x_i\|_2^2 \tag{1}$$

Interprétation géométrique

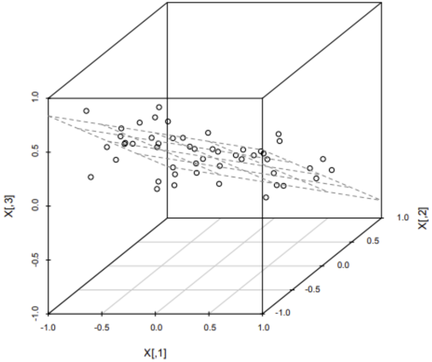


Figure: Intuition géométrique

Maximisation de la variance

En introduisant $\hat{\Sigma} = \frac{X^T X}{n}$, on définit la première composante principale:

$$\alpha_1 = \arg\max_{\alpha} \alpha^T \hat{\Sigma} \alpha \quad \text{sous la contrainte} \quad \|\alpha_1\| = 1 \tag{2}$$

Inconvénient de la méthode

- La PCA utilise une combinaison linéaire de toutes les variables d'origine: probléme d'interprétabilité ou de complexité quand p est grand.
- D'où l'idée naturelle d'en choisir une petite partie (c'est la sparsity).

Plan

LASSO- Least Absolute Shrinkage and Selection Operator

Le LASSO est une méthode de régularisation utilisée dans la régression linéaire.

1. Idée: une pénalité ℓ_1 sur les coefficients du modèle pour favoriser la sparsity **et** sélectionner des variables.
2. $$\min_{\beta_0,\beta}\left\{\frac{1}{2n}\sum_{i=1}^{n}(y_i-\beta_0-x_i^T\beta)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

Elastic net

1. Extension du LASSO qui combine à la fois les termes de régularisation ℓ_1 et ℓ_2.
2. $$\min_{\beta_0,\beta}\left\{\frac{1}{2n}\sum_{i=1}^{n}(y_i-\beta_0-x_i^T\beta)^2+\lambda_1\sum_{j=1}^{p}|\beta_j|+\lambda_2\sum_{j=1}^{p}\beta_j^2\right\}$$

LARS-Least angle regression (Efron en 2004)

1. **Initialisation** : Commence avec tous les coefficients à zéro ($\beta_1 = \beta_2 = \ldots = \beta_p = 0$) et le résidu $r = y$.
2. **Étape a** : Trouve le prédicteur x_j le plus corrélé avec le résidu r.
3. **Étape b** : Ajuste le coefficient β_j dans la direction de x_j jusqu'à ce qu'un autre prédicteur x_k atteigne son coefficient dans le modèle.
4. **Étape c** : après l'étape b), ajuste simultanément dans la direction jointe de x_k jusqu'à ce qu'un autre prédicteur atteigne son coefficient.
5. **Répétition**

Plan

Transformation du problème

Idée: reformuler la PCA avec une forme de relaxation.

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \|x_i - \alpha\beta^T x_i\|_2^2$$

sous contrainte que $\|\alpha\|_2^2 = 1$ et $\alpha = \beta$.

Apport théorique

Idée: éliminer la contrainte d'égalité.

Theorem

Pour tout $\lambda_0 > 0$*, soit:*

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \|\mathbf{x}_i - \alpha\beta^T\mathbf{x}_i\|_2^2 + \lambda_0\|\beta\|_2^2$$

sous contrainte $\|\alpha\|_2^2 = 1$*.*
Alors $\hat{\beta} \propto V_1$*.*
Avec α *fixé, le problème d'optimisation sur* β *est un problème de régression.*

Théorie 2

Supposons que nous considérons les k premières composantes principales. Soit $A_{p\times k} = [\alpha_1, \ldots, \alpha_k]$ et $B_{p\times k} = [\beta_1, \ldots, \beta_k]$. Pour tout $\lambda_0 > 0$, posons

$$(\hat{A}, \hat{B}) = \arg\min_{A,B} \sum_{i=1}^{n} \left\| x_i - AB^T x_i \right\|^2 + \lambda_0 \sum_{j=1}^{k} \|\beta_j\|^2$$

sous contrainte $A^\top A = I_{k\times k}$.
Alors, $\hat{\beta}_j \propto V_j$ pour $j = 1, 2, \ldots, k$.

Critère SPCA

Ces théorèmes permettent d'introduire la fonction objectif (ou le critère) pour les premières k composantes principales (SPCA) est définie comme

$$(\hat{A}, \hat{B}) = \arg\min_{A,B} \sum_{i=1}^{n} \left\| x_i - AB^T x_i \right\|^2 + \lambda_0 \sum_{j=1}^{k} \|\beta_j\|^2 + \sum_{j=1}^{k} \lambda_{1,j} \|\beta_{j,1}\|$$

sous contrainte $A^\top A = I_{k\times k}$, où différents $\lambda_{1,j}$ sont autorisés pour pénaliser les chargements des différentes composantes principales.

Algorithme SPCA

1. Soit A initialisée à $V[:, 1:k]$, les loadings des k premières composantes principales ordinaires.
2. Étant donné un A fixé $= [\alpha_1, \ldots, \alpha_k]$, résoudre le problème de régularisation suivant pour $j = 1, 2, \ldots, k$:

$$\beta_j = \arg\min_{\beta} \left((\alpha_j - \beta)^T X^T X (\alpha_j - \beta) + \lambda \|\beta\|_2^2 + \lambda_{1,j} \|\beta\|_1 \right)$$

3. Pour un B fixé $= [\beta_1, \ldots, \beta_k]$, calculer la SVD de $X^T X = UDV^T$, puis mettre à jour $A = UV^T$.
4. Répéter les étapes 2-3 jusqu'à convergence.
5. Normalisation : $\hat{V}_j = \frac{\beta_j}{\|\beta_j\|}$, $j = 1, \ldots, k$.

Résumé

On tente de minimiser la fonction suivante :

$$f(A,B) = \frac{1}{2}\|X - XBA^T\|_F^2 + \psi(B)$$

où B est la matrice de poids, et A est une matrice orthonormale. ψ désigne un régularisateur induisant la sparsity, en l'occurence elastic net (une combinaison des normes ℓ^1 et ℓ^2). L'idée "computationelle" est d'alterner suivant que l'on ait sous la main A (un problème de régression pour estimer B selon elastic net) ou B (un problème pour estimer A, qui se résoud par une SVD).

Plan

pitprops

- Les données consistent en des corrélations entre des propriétés physiques des rondins de bois.
- Indiqué pour l'apprentissage non supervisé.
- Cas $p < n$.

Figure: `pitprops`

Table: Pitprops : PCA

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.404	0.218	-0.207	0.091	-0.083	0.120
length	-0.406	0.186	-0.235	0.103	-0.113	0.163
moist	-0.124	0.541	0.141	-0.078	0.350	-0.276
testsg	-0.173	0.456	0.352	-0.055	0.356	-0.054
ovensg	-0.057	-0.170	0.481	-0.049	0.176	0.626
ringtop	-0.284	-0.014	0.475	0.063	-0.316	0.052
ringbut	-0.400	-0.190	0.253	0.065	-0.215	0.003
bowmax	-0.294	-0.189	-0.243	-0.286	0.185	-0.055
bowdist	-0.357	0.017	-0.208	-0.097	-0.106	0.034
whorls	-0.379	-0.248	-0.119	0.205	0.156	-0.173
clear	0.011	0.205	-0.070	-0.804	-0.343	0.175
knots	0.115	0.343	0.092	0.301	-0.600	-0.170
diaknot	0.113	0.309	-0.326	0.303	0.080	0.626
Variance (%)	32.4	18.3	14.4	8.5	7.0	6.3
Cumulative Variance (%)	32.4	50.7	65.1	73.6	80.6	86.9

PCA sparse en R

On utilise le package `elasticnet` du même auteur. On trouve le tableau:

	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.4773598	0.00000000	0.00000000	0	0	0
length	-0.4758876	0.00000000	0.00000000	0	0	0
moist	0.0000000	0.78471386	0.00000000	0	0	0
testsg	0.0000000	0.61935898	0.00000000	0	0	0
ovensg	0.1765675	0.00000000	0.64065264	0	0	0
ringtop	0.0000000	0.00000000	0.58900859	0	0	0
ringbut	-0.2504731	0.00000000	0.49233189	0	0	0
bowmax	-0.3440474	-0.02099748	0.00000000	0	0	0
bowdist	-0.4163614	0.00000000	0.00000000	0	0	0
whorls	-0.4000254	0.00000000	0.00000000	0	0	0
clear	0.0000000	0.00000000	0.00000000	-1	0	0
knots	0.0000000	0.01333114	0.00000000	0	-1	0
diaknot	0.0000000	0.00000000	-0.01556891	0	0	1

Tableau des composantes principales sparses (SPCA) Avec une variance cumulée de **75.8** pour cent.

Autres expériences de l'article

- Comparaison avec du soft-thresholding (angle mort numérique) et SCoTLASS.
- Étude d'un modèle synthétique pour retrouver un modèle caché à partir de données observées.
- Comparaison sur des données médicales de soft-thresholding et SPCA pour un problème d'identification de gènes où le soft-thresholding performe bien.

Plan

Idée naturelle de la sparsity

- Zou & al n'ont pas été les premiers à introduire de la sparsity dans les problèmes de PCA.
- Pourquoi ne pas considérer:

$$\max_{\|v\|_2=1} v^T X^T X v$$

sous contrainte que $\|v\|_0 \leq t$ où $\|v\|_0$ compte le nombre de valeurs non nulles dans le vecteur v.

SCotLASS (Joliffe)

Problème NP complet! D'où la relaxation naturelle de cet objectif, en remplaçant la norme ℓ_0 par la norme ℓ_1, conduisant à

$$\max_{\|v\|_2=1} v^T X^T X v \tag{3}$$

sous contrainte que $\|v\|_1 \leq t$. Le problème global reste **non convexe**.

Plan

Développements

De nombreux développements reprennent l'idée de reformulation de la PCA.

1. D'aspremont & al. se placent dans le cadre d'une optimisation pour des matrices semi-définies positive. Ils se démarquent de Zou.
2. Certains ont généralisé l'idée de la relaxation et l'algorithme alterné (Penalised Matrix decomposition par exemple).

Applications

1. Écologie (Gravuer & al. en 2008 pour donner les raisons de l'invasion d'une espèce en Nouvelle Zélande)
2. En neuroscience (Baden & al. en 2016 comme étape avant un modèle de mélange)
3. En imagerie médicale (Sjostrand & al. en 2007).

Problèmes ouverts

Plusieurs problèmes ouverts restent néanmoins à étudier:

1. Une procédure générique pour choisir les setting sparses (i.e, les contraintes les plus efficaces).
2. L'application au Deep Learning.
3. D'avantage d'analyses empiriques pour permettre une méta-analyse rigoureuse des différents types de PCA sparse pour mieux cerner quelle méthode utiliser.

Conclusion

- La méthode SPCA a quasiment vingt ans, mais continue d'être utilisée.
- La multiplication des problèmes en grande dimension conduit naturellement à considérer des méthodes sparses.
- Les algorithmes de relaxation convexes sont très efficaces mais pas les seuls méthodes.

Soft-Thresholding

Introduisons :
$S(Z, \gamma)$ est l'opérateur soft-thresholding sur un vecteur $Z = (z_1, ..., z_p)$ avec le paramètre de seuillage γ et défini par:

$$S(Z, \gamma)_j = (\|z_j\| - \gamma)_+ \text{sgn}(z_j)$$

Special SPCA

Soit $(\widehat{V}_j(\lambda_0))_j$ les k premiers loadings définis par le critère SPCA. Soit $(\widehat{A}, \widehat{B})$, la solution du problème d'optimisation

$$\arg\min_{A,B} -2\mathrm{Tr}\left(A^T X^T X B\right) + \sum_{j=1}^{k} \|\beta_j^k\|_2^2 + \sum_{j=1}^{k} \lambda_{1,jk} \|\beta_j^k\|_1$$

soumis à $A^T A = I_{k\times k}$. Lorsque $\lambda_0 \to \infty$, $\widehat{V}_j(\lambda_0) \to \hat{\beta}_j$(normé).

SPCA modifié

La résolution de ce nouveau problème peut également être effectuée avec l'algorithme SPCA légèrement modifié. Pour A fixé, nous avons que pour chaque j,

$$\hat{\beta}_j = \arg\min_{\beta_j} -2\alpha_j^T (X^T X)\beta_j + \|\beta_j\|_2^2 + \lambda_{1,j}\|\beta_j\|_1,$$

et la solution est donnée par

$$\hat{\beta}_j = S(X^T X \alpha_j, \frac{\lambda_{1,j}}{2}),$$

où $S(Z, \gamma)$ est l'opérateur de Soft-thresholding.

Détail de l'alternance

- On a un algorithme pour estimer B quand A est fixé.
- Étant donné B, la solution de A est à nouveau $\hat{A} = UV^T$ où U, V proviennent de la SVD de $(X^TX)B$: $(X^TX)B = UDV^T$.

Exemple synthétique

On génère des données avec trois facteurs cachés :

$$\begin{aligned}
&V1 \sim N(0, 290),\\
&V2 \sim N(0, 300),\\
&V3 = -0.3V1 + 0.925V2 + \epsilon, \quad \epsilon \sim N(0, 1).
\end{aligned}$$

$V1$, $V2$, et ϵ sont indépendants.

Suite

On crée 10 variables observées X_i basées sur ces facteurs. L'idée est de fortement corréler ces variables par groupe.
Résultats de la Simulation

- La SPCA identifie correctement les variables importantes, fournissant des représentations sparses idéales.
- La seuillage inclut à tort des variables moins importantes en raison d'une corrélation élevée.

LARS-Lasso: Description

Algorithme 1 LARS-Lasso

Entrées: X, Y, n, p
$i \longleftarrow 1$
$\lambda \longleftarrow +\infty$
$\beta \longleftarrow 0$
$\text{sign}(\beta) \longleftarrow 0$
$\Gamma \longleftarrow \emptyset$
Resultats $\longleftarrow \emptyset$
Tant que $\lambda > 0$ **faire**
« On commence par chercher la première variable insatisfaisante »
$\lambda_{\max} \longleftarrow 0$
$j_{\max} \longleftarrow 0$
Pour $j = 1, ..., p$ **faire**
Si $\text{sign}(\beta)_j = 0$ **alors**
$\lambda_{\text{test}} \longleftarrow \max \left\{ \frac{X_j^T [X_\Gamma (X_\Gamma^T X_\Gamma)^{-1} (X_\Gamma^T Y) - Y]}{[X_j^T X_\Gamma (X_\Gamma^T X_\Gamma)^{-1} (\text{sign}(\beta_\Gamma(\lambda)))] \pm 1} \right\}$
Sinon
$\lambda_{\text{test}} \longleftarrow \left(\frac{((X_\Gamma^T X_\Gamma)^{-1} (X_\Gamma^T Y))_{j'}}{((X_\Gamma^T X_\Gamma)^{-1} (\text{sign}(\beta_\Gamma(\lambda))))_{j'}} \right)$
Fin si
Si $\lambda_{\text{test}} < \lambda$ **et** $\lambda_{\text{test}} > \lambda_{\max}$ **alors**
$\lambda_{\max} \longleftarrow \lambda_{\text{test}}$
$j_{\max} \longleftarrow j$
Fin si
Fin pour
« On a dans $(\lambda_{\max}, j_{\max})$ le résultat d'insatisfaction cherché »
$\beta \longleftarrow \beta_\Gamma \longleftarrow (X_\Gamma^T X_\Gamma)^{-1} (X_\Gamma^T Y - \lambda_{\max} \text{sign}(\beta_\Gamma))$
$\lambda \longleftarrow \lambda_{\max}$
Si $j_{\max} > 0$ **alors**
Si $\text{sign}(\beta)_{j_{\max}} = 0$ **alors**
$\text{sign}(\beta)_{j_{\max}} \longleftarrow \text{sign}(X_{j_{\max}}^T (Y - X\beta))$
Sinon
$\text{sign}(\beta)_{j_{\max}} \longleftarrow 0$
$\beta_{j_{\max}} \longleftarrow 0$
Fin si
$\Gamma \longleftarrow$ non-nuls$(\text{sign}(\beta))$
Resultats $\longleftarrow$ Resultats $\cup \{(j_{\max}, \lambda_{\max}, \beta)\}$
Sinon
Resultats $\longleftarrow$ Resultats $\cup \{(\lambda_{\max}, \beta)\}$
Fin si
$i \longleftarrow i + 1$
Fin tant que
Retourner Resultats

Figure: Algorithme LARS-LASSO selon [Pierre GAILLARD, 2009]