Détection de cibles spatialement structurées sous contrôle global d'erreur

Raphael BACHER[1,2], Florent CHATELAIN[1], Olivier MICHEL[1]

[1]Université Grenoble-Alpes, Gipsa-Lab, Grenoble, France

[2]CRAL, Observatoire de Lyon, Lyon, France

raphael.bacher@gipsa-lab.fr, florent.chatelain@gipsa-lab.fr
olivier.michel@gipsa-lab.fr

Résumé – Dans un grand nombre de problèmes de détection (en imagerie par exemple), la cible recherchée présente une certaine structure spatiale. Prendre en compte cet *a priori* doit permettre d'améliorer la puissance de détection. Par ailleurs la détection de cibles multi-pixels dans un contexte de tests multiples requiert l'utilisation de techniques spécifiques de contrôle global des erreurs. Dans cette étude, une approche de détection s'appuyant sur des contraintes de connexité tout en assurant un contrôle de type *False Discovery Rate* (FDR) est proposée. L'approche développée est générique et ouvre la possibilité à diverses implémentations. Une méthode originale est proposée et implémentée (COMET). COMET permet un gain important en puissance de détection tout en garantissant un contrôle du taux global d'erreur, ce qui est illustré sur des données simulées.

Abstract – In this paper, a target detection procedure with global error control is proposed. The novelty of this approach consists of taking into account spatial structures of the target while ensuring proper error control over pixelwise errors. A generic framework is discussed and a method based on this framework is implemented (COMET). Results on simulated data show conclusive gains in detection power for a nominal control level.

1 Introduction

Nous considérons dans cette étude un problème de détection de cible/source dans un jeu de données massives, impliquant un très grand nombre de tests statistiques. Dans ce contexte, le contrôle du risque d'erreur de type I s'avère inapproprié [1, 2] : le nombre d'hypothèses nulles rejetées à tort (fausses alarmes) peut rapidement devenir important (i.e. jusqu'à dépasser le nombre de bonnes détections) du fait du grand nombre de tests réalisés. Dans [3] Benjamini et Hochberg ont proposé de substituer au contrôle individuel de l'erreur de type I (ou contrôle de la probabilité de fausse alarme) une mesure du contrôle global des erreurs, nommée *False Discovery Rate* (FDR), ainsi qu'une procédure de contrôle de ce FDR. Cette procédure, devenue classique (ci-après notée BH), nécessite seulement la connaissance de la distribution de la statistique de test des observations de bruit seul. Toutefois, dans le cadre de données réelles, cette distribution est rarement parfaitement connue. Par exemple lorsque le bruit présente une structure de dépendance complexe, éventuellement hétéroscédastique, un estimateur robuste de la distribution à partir des observations peut être difficile voir impossible à obtenir. Dans notre contexte spécifique, les données sont des ensembles d'images hyperspectrales et les objets recherchés (halos galactiques dans un cadre astrophysique) possèdent une extension spatiale. Nous utiliserons par conséquent indifféremment les termes "pixel" et "échantillon". Ces objets étant rares dans les données, le problème de détection s'inscrit dans le cadre décrit ci-dessus nécessitant un contrôle au sens du FDR. Appliquer la procédure BH dans une approche pixelique ne prend cependant pas en compte les éventuels liens entre les pixels de la cible, ceux-ci pouvant être organisés en structures connexes. La prise en compte d'un tel *a priori* ne peut qu'améliorer la puissance de détection. Plusieurs approches par groupe d'échantillons ont été récemment développées [4] mais s'appuient sur une connaissance *a priori* de ces groupes. Dans notre contexte, la forme de la cible est inconnue et nous ne disposons donc pas d'un tel *a priori*. Par ailleurs, un inconvénient notoire des approches classiques de contrôle FDR est que la puissance de détection diminue par essence lorsque le nombre de tests s'accroît : la puissance de détection d'une unique cible, à un niveau de contrôle FDR donné, va donc dépendre de la taille de la région dans laquelle la source est recherchée.

Ainsi, afin de 1) assurer un contrôle robuste des erreurs, 2) prendre en compte la structure de la cible et 3) limiter l'influence du nombre de tests, nous développons une procédure originale nommée *COnnection accounting Method for Extracting Target* (COMET). COMET s'appuie notamment sur une nouvelle classe d'approches de contrôle du FDR récemment introduite dans [5, 6], que nous proposons d'étendre de sorte à prendre en compte la connexité des structures recherchées. La procédure de contrôle proposée s'appuie sur des statistiques construites à partir des données, de façon à satisfaire certaines

propriétés de symétrie : nous nous intéressons en particulier aux problèmes de détection dans le cadre desquels le bruit possède une distribution symétrique et la cible a une contribution nécessairement positive. Cela est le cas par exemple dans l'application envisagée de détection de halos galactiques dans des données hyperspectrales [7].

Le papier est organisé de la façon suivante : dans la section 2 nous présentons COMET, méthode pour la détection de cibles multi-pixels à forte connexité et dans la section 3 nous illustrons les performances de COMET sur données simulées.

Notations

Les vecteurs de caractéristiques associés à un pixel (un spectre dans un contexte hyperspectral) sont représentés par des caractères en gras ($\boldsymbol{y}$). La notation $a \vee b$ indique $\max(a, b)$.

2 Méthode

Nous nous intéressons ici à la détection d'un signal à valeurs strictement positives, à partir d'un vecteur bruité $\boldsymbol{y} \in \mathbb{R}^l$. Soient $\mathcal{H}_0$ et $\mathcal{H}_1$ les hypothèses dénotant respectivement l'absence ou la présence de contribution d'une source. Ce problème de détection s'exprime comme le test d'hypothèse unilatéral suivant :

$$\begin{cases} \mathcal{H}_0 : \boldsymbol{y} = \boldsymbol{\epsilon}, \\ \mathcal{H}_1 : \boldsymbol{y} = \alpha \boldsymbol{d} + \boldsymbol{\epsilon}, \quad \text{avec } \alpha > 0, \end{cases} \tag{1}$$

$\boldsymbol{\epsilon} \in \mathbb{R}^l$ est un vecteur de bruit de distribution inconnue mais supposée symétrique. $\boldsymbol{d}$ est un signal de référence connu. $\boldsymbol{y} \in \mathbb{R}^l$ peut être un spectre (imagerie hyperspectrale), une série temporelle (imagerie IRMf), ou une intensité ($l = 1$ et $\boldsymbol{d} = 1 \in \mathbb{R}$) en imagerie classique.

Le test décrit par (1) doit être conduit pour un très grand nombre n d'observations $\{\boldsymbol{y}_i\}_{1 \leq i \leq n}$. Dans ce contexte de tests multiples, il faut déterminer un seuil de décision adapté aux données qui permette de contrôler le FDR

$$\text{FDR} = \mathbb{E}\left[\frac{U}{R \vee 1}\right],$$

où R est le nombre total de tests pour lesquels l'hypothèse nulle est rejetée, tandis que U est le nombre de fausses découvertes parmi les R découvertes. Le FDR est donc la proportion moyenne de vraies hypothèses nulles rejetées à tort, appelées *fausses découvertes*, parmi tous les tests rejetés (*découvertes*).

Une méthode de contrôle du FDR a été récemment proposée par Barber et Candès dans [5] dans laquelle ils proposent de construire des "contrefaçons" (*knockoffs*). Ces contrefaçons sont des variables artificielles reproduisant la structure de corrélation des variables originales, dans un cadre de régression linéaire avec bruit blanc Gaussien. Elles doivent permettre de construire des statistiques de contrôle $\{w_i\}_{1 \leq i \leq n}$ pour chaque test, satisfaisant en particulier les propriétés suivantes :

- symétrie sous $\mathcal{H}_0$, i.e. $\mathbb{P}(w_i > 0 | i \in \mathcal{H}_0) = \mathbb{P}(w_i < 0 | i \in \mathcal{H}_0)$,
- la statistique doit être stochastiquement plus grande sous $\mathcal{H}_1$ que sous $\mathcal{H}_0$, i.e. $\mathbb{P}(w_i > 0 | i \in \mathcal{H}_1) > \mathbb{P}(w_i > 0 | i \in \mathcal{H}_0)$.

Barber et Candès établissent alors une procédure de contrôle du FDR que nous reformulons ici dans notre contexte :

Proposition 2.1. *Si les $\{w_j\}_{1 \leq j \leq n}$ sont distribuées de façon symétrique sous $\mathcal{H}_0$, et si leurs signes sont indépendants entre eux, alors pour un niveau nominal de contrôle donné q, un seuillage au niveau*

$$\widehat{t}_q = \inf\left\{ t \geq 0 : \frac{1 + \#\{w_j < -t\}}{1 \vee \#\{w_j > t\}} \leq q \right\} \tag{2}$$

assure un contrôle exact du FDR au niveau q pour l'ensemble de détections $\mathcal{D} = \{i : w_i > \widehat{t}_q\}$.

La démonstration est détaillée dans le Théorème 3 de [5], que l'on applique à des p-valeurs binaires

$$p_i = \begin{cases} 1/2, & \text{if } w_{(i)} > 0, \\ 1, & \text{if } w_{(i)} < 0, \end{cases}$$

où $w_{(i)}$ sont les statistiques de contrôle ordonnées en valeur absolue i.e. $|w_{(1)}| \geq |w_{(2)}| \geq \cdots \geq |w_{(n)}|$.

2.1 Construction des statistiques de contrôle

La construction de contrefaçons peut être problématique notamment en grande dimension. Afin d'éviter la construction de ces dernières, nous proposons ici une méthode plus directe de construction des statistiques, adaptée au problème de détection exposé en (1). La méthode proposée s'appuie sur la symétrie de la distribution du bruit et le fait qu'une cible a nécessairement une contribution positive ($\alpha > 0$). Pour une signature de cible $\boldsymbol{d}$ connue, la statistique w_i du $i$eme test, $1 \leq i \leq n$, peut être définie par :

$$w_i = \boldsymbol{d}^T \boldsymbol{y}_i \tag{3}$$

qui correspond au filtre adapté dans le cas d'un bruit blanc. Une forte valeur positive de w_i indique alors très certainement que l'échantillon testé satisfait l'hypothèse alternative $\mathcal{H}_1$; sous $\mathcal{H}_0$, les w_i sont distribués de façon symétrique (tout comme le bruit).

Notons que dans la construction de ces statistiques de contrôle, il n'est pour l'instant fait aucun usage de la possible structure spatiale de la source à détecter.

2.2 Algorithme générique

Considérons désormais un problème de détections multiples dans lequel les échantillons ciblés sont possiblement structurés (e.g. la cible est un objet large de plusieurs pixels dans une image). Nous proposons un algorithme générique prenant en compte cette connexité spatiale afin d'améliorer la puissance de détection du test (2), tout en assurant le même contrôle FDR.

La stratégie proposée est la suivante : la cible étant possiblement formée des zones connexes, seuls les voisins des pixels déjà détectés sont à tester : cela correspond à développer une approche de croissance de région. La réduction du nombre effectif de tests aux seuls voisins permet alors de limiter la perte de puissance pour un niveau de FDR donné. Cette approche est décrite dans la procédure "step-up" de l'algorithme 1. Le contrôle FDR s'appuie sur la propriété de symétrie de la distribution des statistiques de contrôle sous $\mathcal{H}_0$. La difficulté vient

de la possible structure de dépendance du bruit entre pixels, susceptible d'introduire un biais de sélection. A une étape de sélection k donnée, notons $\mathcal{A}_k \subset \{1,\dots,n\}$ l'ensemble des pixels d'"intérêt" sélectionnés [1]. La procédure de sélection de l'ensemble $\mathcal{A}_k$ doit préserver P1 :

P1 (Symétrie post-sélection)**.** *Pour tout $j \in \mathcal{A}_k$ correspondant à une vraie hypothèse nulle, w_j est distribué symétriquement.*

Notons S l'opérateur associé à la procédure de sélection. Cet opérateur promeut la connexité spatiale avec les pixels nouvellement sélectionnés. Le lemme suivant peut alors aisément être établi, du fait de la symétrie de la distribution du bruit :

Lemme 2.2. *Si l'opérateur de sélection S dépend des données uniquement via les valeurs absolues $|w_i|$ des statistiques de contrôle, pour $1 \leq i \leq n$, alors la propriété P1 est satisfaite.*

Une estimation de la proportion de fausses découvertes, ou FDP, (parmi les statistiques positives $w_i > 0$ sélectionnés dans $\mathcal{A}_k$) est défini par

$$\widehat{q}_k = \frac{1 + \#\{i \in \mathcal{A}_k, w_i < 0\}}{1 \vee \#\{i \in \mathcal{A}_k, w_i > 0\}}. \tag{4}$$

L'algorithme construit donc itérativement un ensemble de test $\mathcal{A}_{\widehat{k}}$ (défini à l'étape 8 de l'algorithme 1) à l'aide d'une procédure de sélection exploitant les valeurs absolues des statistiques de contrôle (lemme 2.2) tout en contrôlant à chaque itération la valeur estimée du FDP. L'ensemble $\mathcal{D}$ des pixels constituant la cible est ensuite obtenu par une procédure de test sur les statistiques de contrôle de $\mathcal{A}_{\widehat{k}}$.

Algorithm 1 Procédure COMET générique

1: *Entrée :* statistiques de contrôle $\boldsymbol{w} = \{w_j\}_{1\leq j\leq n}$, niveau de contrôle nominal q

2: $k \leftarrow 0, \mathcal{A}_0 \leftarrow \emptyset, \widehat{q}_0 \leftarrow 0$ ▷ Initialisation
3: **while** $\mathcal{A}_k \neq \{1,\dots,n\}$ **do**
4: $\quad \mathcal{A}_{k+1} \leftarrow S(\mathcal{A}_k, \boldsymbol{w})$ ▷ Étape de sélection vérifiant P1
5: $\quad$ Calcul de $\widehat{q}_{k+1}$ à l'aide de (4) ▷ Estimation FDP
6: $\quad k \leftarrow k+1$
7: **endWhile**
8: $\widehat{k} \leftarrow \max\{k : \widehat{q}_k \leq q\}$ ▷ Temps d'arrêt
9: *Sortie :* $\mathcal{D} \leftarrow \{i \in \mathcal{A}_{\widehat{k}} : w_i > 0\}$ ▷ Liste des détections

Cet algorithme peut s'appuyer sur de nombreuses implémentations de la procédure de sélection (sous réserve de préserver P1). Il s'agit principalement de trouver un bon (i.e. favorisant les pixels de la cible) ordonnancement des pixels à tester. Dans la suite nous nous focalisons sur une procédure de sélection simple qui donne de bons résultats expérimentaux.

2.3 Implémentation

L'approche gloutonne suivante est proposée : à chaque étape la plus grande statistique en valeur absolue est retenue parmi les pixels du voisinage. A l'étape k, notons $\mathcal{N}_k = G(\mathcal{A}_k)$ le voisinage externe de $\mathcal{A}_k$, où G est le gradient morphologique externe, i.e. une dilatation (ici pour une clique en 8-connexité) suivie d'une soustraction. La procédure de sélection est alors définie par

$$S(\mathcal{A}_k, \boldsymbol{w}) \equiv \mathcal{A}_k \cup \{j_0\}, \quad \text{où } j_0 = \arg\max_{j\in\mathcal{N}_k} |w_j|.$$

La propriété de symétrie P1 est assurée par le lemme 2.2. En pratique, pour des gains de temps de calcul, la boucle interne de l'algorithme 1 peut être arrêtée lorsqu'à la fois le nombre de pixels sélectionnés est grand et $\widehat{q}_k$ est significativement plus grand que q (e.g. $\widehat{q}_k \geq 1.2 \times q$). Cette approche gloutonne permet de s'adapter à n'importe quelle forme connexe mais également de surmonter des "trous" puisqu'on cherche le plus grand ensemble de pixels où $\widehat{q} \leq q$.

2.4 Contrôle en présence de bruit indépendant

Proposition 2.3 (Contrôle FDR de COMET)**.** *Supposons que les vecteurs de bruit $\boldsymbol{\epsilon}_1,\dots,\boldsymbol{\epsilon}_n$ sont distribués selon une loi symétrique et sont indépendants entre eux. Alors l' algorithme 1, où les statistiques de contrôle $\boldsymbol{w}$ sont construites à l'aide de* (3)*, permet un contrôle exact du FDR :* $\mathbb{E}\left[\frac{U}{R\vee 1}\right] \leq q$.

Démonstration. Selon la propriété P1, pour tout $i \in \mathcal{A}_{\widehat{k}}$ correspondant à un vrai $\mathcal{H}_0$, w_i est distribué de façon symétrique. De plus les signes des $\{w_i\}_{1\leq i\leq n}$ sont indépendants par indépendance des $\{\boldsymbol{\epsilon}_i\}_{1\leq i\leq n}$. Nous pouvons donc appliquer la proposition 2.1 avec un seuil $\widehat{t}_q = 0$. Cela conclut la preuve. □

2.5 Contrôle en présence de bruit corrélé

En cas de bruit corrélé, le contrôle exact n'est plus assuré. Néanmois un contrôle asymptotique peut être prouvé à l'aide des hypothèses suivantes.

A1 (Faible dépendance)**.** *Pour n'importe quel ensemble $\mathcal{S}$ de pixels sous $\mathcal{H}_0$,*

$$\frac{\#\{i \in \mathcal{S} : w_i > 0\}}{\#\mathcal{S}} \underset{\#\mathcal{S}\to\infty}{\overset{a.s.}{\to}} \mathbb{P}(w_i > 0)$$

Pour simplifier les notations, $\mathcal{S}_n$ indique désormais l'ensemble final de pixels sélectionnés $\mathcal{A}_{\hat{k}}$ par l' algorithme 1 pour un jeu de données de n pixels.

A2 (Croissance de région)**.** *Pour un niveau de contrôle $q > 0$ donné, $\mathcal{S}_n$ croît avec le nombre total d'échantillons n et $\#\mathcal{S}_n \underset{n\to+\infty}{\to} +\infty$.*

Proposition 2.4 (Contrôle asymptotique de COMET)**.** *Supposons A1, A2 et une distribution symétrique des statistiques de contrôle sous $\mathcal{H}_0$. Alors l'algorithme 1 permet un contrôle asymptotique du FDR pour le test* (1)*.*

Éléments de preuve. Pour un niveau de contrôle donné q et un jeu de n échantillons, notons $\widehat{\text{FDP}}_q = \frac{1+\#\{i\in\mathcal{S}_n, w_i<0\}}{\#\{i\in\mathcal{S}_n, w_i>0\}}$. A partir des hypothèses ci-dessus, il peut être montré que $\liminf_{n\to\infty}(\widehat{\text{FDP}}_q - \text{FDP}) \geq 0$.

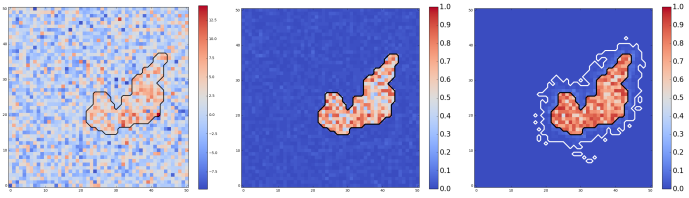
Puisque $\widehat{\text{FDP}}_q \leq q$ du fait du critère d'arrêt, le lemme de Fatou permet de conclure que $\limsup_{n\to\infty} \mathbb{E}[\text{FDP}] = \text{FDR} \leq q$ □

1. Par soucis de simplicité, chacun des n pixels est identifié par un indice $i \in \{1,\dots,n\}$ plutôt que par ses coordonnées spatiales. Il est à noter toutefois qu'ici la connexité s'entend en terme de coordonnés spatiales.

3 Performance

Dans cette section, la procédure COMET est comparée avec la méthode décrite en prop. 2.1, qui ne prend pas en compte d'information de connexité. Les données simulées sont construites sous forme de cubes hyperspectraux de dimension $31 \times 51 \times 51$ (une dimension spectrale et 2 spatiales). La cible est composée de 350 pixels connexes associés à un spectre. Chaque spectre est de forme gaussienne tronquée. Du bruit gaussien est ajouté, corrélé par convolution d'un noyau de taille 3×3. Le rapport signal-à-bruit pSNR est d'environ 7dB ($\text{pSNR} = 20 \log \frac{u_m}{\sigma}$ où u_m est la valeur maximale de la cible et σ est l'écart-type du bruit).

La figure 1 illustre la procédure de détection : la figure 1(a) représente une réalisation bruitée des données testées, intégrée le long de la dimension spectrale ; les figures 1(b) et 1(c) montrent les taux de détections sur 100 simulations Monte-Carlo pour respectivement l'approche non-connexe et COMET. Le code couleur illustre le gain de puissance obtenu avec COMET. La région explorée par COMET (contour blanc) reste à proximité de la cible, au contraire la méthode non-connexe (presque tous les pixels ont été détectés au moins une fois sur les 100 simulations dans la fig. 1(b)). La figure 2 détaille les performances de la méthode pour différents niveaux de FDR : la figure 2(a) souligne l'obtention d'un contrôle asymptotique en présence de bruit faiblement corrélé, pour les 2 approches ; la figure 2(b) illustre à nouveau le gain significatif de puissance lors de la prise en compte de la connexité. De plus, COMET garde la même puissance de détection lorsque la fenêtre de tests est agrandie alors que l'approche classique voit sa puissance diminuer.

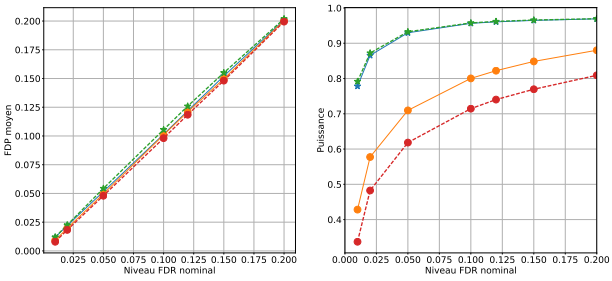


(a) Réalisation bruitée (b) Détection par méthode non-connexe (c) Détection par COMET

FIGURE 1 – (a) exemple d'une réalisation bruitée de cube simulé (intégré sur la dimension spectrale) ; (b) taux de détection avec un FDR de 10% pour la méthode de la prop. 2.1 (1 : toujours détecté, 0 : jamais détecté) ; (c) taux de détection par COMET pour un même niveau FDR de 10%. Résultats moyennés sur 100 simulations Monte-Carlo. En noir : position de la cible, en blanc : pixels détectés par COMET au moins une fois parmi les 100 simulations.

Conclusion

Dans cette étude, une méthode générique est proposée, permettant d'améliorer la puissance de détection des approches FDR sur des cibles multi-pixels connexes en présence de bruit



(a) Contrôle FDR (b) Puissance de détection

FIGURE 2 – (a) FDR empirique vs FDR nominal ; (b) : puissance vs FDR nominal. Méthode proposée : $*$, l'approche non-connexe : $\bullet$. En trait plein, la détection est faite sur une fenêtre 51×51, en pointillés sur une zone 71×71 (même cible de taille 360 pixels). Résultats moyennés sur 500 simulations Monte-Carlo.

corrélé. Cette méthode s'appuie sur une sélection des pixels à tester à l'aide de contraintes de connexité. Une implémentation de cette méthode via une procédure simple, nommée COMET, est décrite ; elle nécessite uniquement la symétrie de la distribution du bruit et la positivité des sources. Le contrôle exact du FDR par cette procédure est prouvé pour des échantillons indépendants, et un contrôle asymptotique est prouvé en cas de bruit faiblement dépendant. Comparée à une procédure FDR de l'état de l'art, COMET donne des résultats concluants sur données simulées, tant en gain de puissance qu'en robustesse à la taille de région à explorer. Cette méthode est non-paramétrique, donc robuste aux erreurs de modèle, et a un faible coût calculatoire (environ 1s pour traiter un cube $31 \times 51 \times 51$).

Les auteurs remercient l'ERC 339659-MUSICOS qui a permis le financement de ce travail.

Références

[1] C. Genovese et al. *Thresholding of statistical maps in functional neuroimaging using the false discovery rate*. Neuroimage, 2002.

[2] C. Meillier et al. *Error control for the detection of rare and weak signatures in massive data*. EUSIPCO, 2015.

[3] Y. Benjamini et al. *Controlling the false discovery rate : a practical and powerful approach to multiple testing*. Journal of the royal statistical society, 1995.

[4] R. Barber et al. *The p-filter : multi-layer FDR control for grouped hypotheses*. Journal of the royal statistical society, 2016.

[5] R. Barber et al. *Controlling the false discovery rate via knockoffs*. The Annals of Statistics, 2015.

[6] E. Candès et al. *Panning for Gold : Model-free Knockoffs for High-dimensional Controlled Variable Selection*. arXiv preprint arXiv :1610.02351, 2016.

[7] R. Bacher et al. *Robust control of varying weak hyperspectral target detection with sparse non-negative representation*. IEEE Trans. on Signal Processing, 2017.