Suppression de ligne de base et débruitage de chromatogrammes par pénalisation asymétrique de positivité et dérivées parcimonieuses

Laurent DUVAL[1], Aurélie PIRAYRE[1], Xiaoran NING[2], Ivan W. SELESNICK[2]

[1]IFP Energies nouvelles
1 et 4 avenue de Bois-Préau, 92852 Rueil-Malmaison Cedex, France

[2]Polytechnic School of Engineering, New York University
6 Metrotech Center, Brooklyn, NY 11201, USA

laurent.duval@ifpen.fr, aurelie.pirayre@ifpen.fr
xn211@nyu.edu, selesi@nyu.edu

Résumé – Bon nombre de mesures expérimentales sont altérées par des fluctuations stochastiques et/ou systémiques, souvent inhérentes aux protocoles expérimentaux et aux méthodes d'acquisition. Dans ce contexte, cet article présente une méthode permettant de retrouver le signal d'intérêt débruité et corrigé de sa tendance ou ligne de base. En s'appuyant sur différents *a priori* sur les signaux (ainsi que sur leurs dérivées) comme la linéarité, la positivité, ou encore la parcimonie, ce problème est formulé comme la minimisation d'une fonction comportant un terme quadratique de fidélité aux données, d'une pénalité régularisée de type ℓ_1 asymétrique promotrice de positivité, ainsi qu'une contrainte de parcimonie. Cette modélisation étant assez générique, la méthode proposée (BEADS) est susceptible d'être appliquée à d'autres types de signaux. Nous nous intéressons ici au cas des analyses physico-chimiques, et particulièrement à la chromatographie, où les chromatogrammes sont bruités et présentent de fortes dérives de la ligne de base, biaisant ainsi les informations qui peuvent en être extraites. Les performances sont évaluées sur des données simulées et réelles.

Abstract – Many experimental measurements are altered by stochastic or systemic fluctuations, often inherent to experimental protocols and acquisition methods. In this context, the paper introduces a method allowing us to recover a noise-reduced and trend- or baseline-corrected signal. Based on several *a priori* on signals and their derivatives, i.e. linearity, positivity or sparsity, this problem is formulated as the minimization of a quadratic function for the data fidelity term, an asymmetric ℓ_1-like penalization enforcing positivity and a derivative-based sparsity constraint. This model being quite generic, the proposed method (BEADS) may be applied to different kinds of signals. In this work, we deal with physico-chemical analysis, particularly with gas chromatography where chromatograms are noisy and corrupted by a baseline. These artifacts lead to biases in the extracted information. Results obtained on simulated and real data are also presented.

1 Introduction

Un grand nombre de mesures expérimentales comporte, outre les signaux d'intérêt et des perturbations plus aléatoires, ou bruits, des variations plus lentes, à plus long terme. Ces variations s'appellent, suivant leur origine ou le domaine d'application, dérive, biais, tendance, composante saisonnière, arrière-plan ou fond en image. Elles se retrouvent également sous forme de fond spectral en analyse de Fourier. La combinaison de ces variations avec le bruit est parfois nommée « bruit de fond ». La suppression de ces artefacts ou leur correction par calibration est une étape essentielle de prétraitement pour l'interprétation, l'analyse et la quantification précise des paramètres calculés sur les mesures ou leurs transformées.

Dans le cas des signaux issus d'analyse physico-chimique [1], ces variations sont le plus souvent nommées « ligne de base ». Leur origine dépend de la modalité étudiée : spectroscopie infrarouge ou spectrométrie de masse, Raman, RMN, chromatographie en phase liquide ou gazeuse. Ces deux dernières visent à séparer les molécules de natures diverses d'un mélange potentiellement très complexe. Par exemple, dans le cas de la chromatographie en phase gazeuse, les molécules transitent au travers de colonnes capillaires et migrent différentiellement en fonction de leur masse ou de leur affinité chimique. Elles sont analysées en sortie de colonne afin de mesurer leur concentration initiale dans le mélange. La quantité associée à une molécule ainsi séparée s'exprime souvent dans le signal acquis par un motif en forme de pic (gaussien, lorentzien, de Voigt, etc. [2]), idéalement distinct des pics correspondant à d'autres molécules. Dans ce cas, sa concentration se calcule par intégration, ou calcul de l'aire comprise entre le bruit de fond instrumental et la courbe du pic, sur un intervalle choisi au-dessus du niveau de bruit, en l'absence de dérive.

En chromatographie gazeuse, une dérive est souvent présente (*cf.* figure 1), due à des changements de température du détecteur ou au relargage de colonne. Un phénomène analogue se retrouve sur la plupart des mesures physico-chimiques [3] et dans certaines modalités d'imagerie hyperspectrale [4], et est susceptible de susciter un intérêt élargi dans la communauté du

traitement de signal [5].

En effet, la majorité des signaux physico-chimiques présentent des *a priori* de linéarité, de positivité, de concentration unitaire ou de parcimonie. De nombreux travaux s'attachent à la séparation de pics partiellement ou entièrement superposés [2], et donc à la déconvolution ou au démélange (spectral) de pics [6]. Cependant, la présence d'une ligne de base décale de manière lente mais non stationnaire les pics. Cette ligne de base ne disposant pas toujours de modèle paramétrique simple à estimer et soustraire, elle dégrade les résultats d'analyses quantitatives ultérieures. Dans cet article, nous nous intéressons à l'aspect de prétraitement de signaux appliqué à des données de chromatographie bidimensionnelle [1]. Son objectif est d'atténuer à la fois le bruit et de soustraire une ligne de base estimée, en exploitant la linéarité, les propriétés de parcimonie des pics et de leurs dérivées et la positivité par pénalisation asymétrique. Il s'inscrit dans la lignée des techniques de décomposition morphologique [10, 11] en structure et texture, remplaçant continuité par morceaux et oscillations par dérive et pics. De surcroît, l'augmentation du volume des données générées par l'expérimentation haut-débit suscite l'usage de méthodes flexibles, dont les paramètres peuvent être relativement facilement inférés des données. La section 2 rappelle les travaux antérieurs, expose le modèle choisi et formule le problème de suppression de ligne de base et de débruitage conjoint. Les résultats de cette approche sont présentés dans la section 3, suivie d'une conclusion et de perspectives.

2 Antériorité, formulation du problème

La suppression de la ligne de base est en apparence un problème simple. Il est cependant étudié depuis longtemps [12]. Les approches standard emploient des filtrages linéaires ou non [13], des méthodes basées sur les ondelettes [14]. Les propriétés des signaux d'analyse chimique ont conduit au développement de modèles de régression plus poussés : les lentes variations de la ligne de base permettent de la modéliser par des polynômes de faible degré ou des splines (cubiques) [8, 15], potentiellement couplés à des approches bayésiennes ou de seuillage itéré. L'usage de dérivées du signal a également été proposé [16]. Plutôt que d'employer des modèles paramétriques peu vraisemblables, il est possible de s'appuyer sur les caractéristiques morphologiques attendues des signaux et d'en inférer des pénalités plus adaptées que les classiques moindres carrés [9]. Ce travail s'inscrit dans cette continuité : la séparation conjointe des trois composantes attendues [11]. Les hypothèses posées sont les suivantes. Le chromatogramme observé de longueur N est noté $\mathbf{c} = [c_0, c_1, \ldots, c_{N-1}]^\mathsf{T}$. Le signal composé uniquement des pics est noté $\mathbf{p}$, $\mathbf{l}$ et $\mathbf{b}$ représentent respectivement la ligne de base et le bruit. Ainsi l'hypothèse de linéarité se résume à $\mathbf{c} = \mathbf{p} + \mathbf{l} + \mathbf{b}$. Les pics $\mathbf{p}$ possèdent en théorie une forme gaussienne positive et symétrique, du fait de la séparation moléculaire, assimilée au mécanisme de la planche de Galton. Par ailleurs, leurs dérivée première et dérivée seconde présentent des caractéristiques de parcimonie. Cette observation est illustrée dans la partie gauche de la figure 1, sur un chromatogramme possédant une ligne de base presque constante. Le bruit en chromatographie est souvent considéré gaussien. La ligne de base pourrait être obtenue approximativement par filtrage passe-bas d'un chromatogramme sans pics. La partie droite de la figure 1 représente une ligne de base et un bruit simulés permettant d'obtenir le chromatogramme composite $\mathbf{c}$.

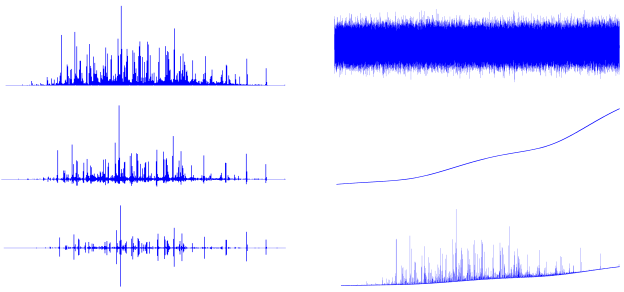

FIGURE 1 – Composantes morphologiques d'un signal chromatographique, de haut en bas. Colonne de gauche : somme de pics $\mathbf{p}$, ses dérivées première et seconde. Colonne de droite : bruit $\mathbf{b}$, ligne de base $\mathbf{l}$ et chromatogramme composite $\mathbf{c}$.

Ces différentes hypothèses sont regroupées dans une fonction de coût $F(\mathbf{p})$, dont le minimum fournit l'estimation de $\hat{\mathbf{p}}$:

$$\frac{1}{2}\|\mathbf{H}(\mathbf{c}-\mathbf{p})\|_2^2 + \sum_{n=0}^{N-1}\Big\{\lambda_0\theta_{\epsilon,r}(p_n) + \sum_{i=1}^{2}\lambda_i\phi\left([\mathbf{D}_i\mathbf{p}]_n\right)\Big\}.$$

Dans le premier terme, $\mathbf{H}$ est l'expression matricielle d'un filtre récursif, non-causal à phase nulle [17]. La ligne de base sera estimée par $\hat{\mathbf{l}} = (\mathbf{D}_0 - \mathbf{H})(\mathbf{c} - \hat{\mathbf{p}})$, où $\mathbf{D}_0$ désigne la matrice identité. Le deuxième terme correspond à une pénalité en norme ℓ_1. Le facteur r règle une pente qui, pour $r > 1$, pénalise asymétriquement les valeurs négatives. La fonction $\theta_{\epsilon,r}(c)$ est régularisée en 0 par un polynôme de degré 2 :

$$\theta_{\epsilon,r}(c) = \begin{cases} c, & c > \epsilon \\ \frac{1+r}{4\epsilon}c^2 + \frac{1-r}{2}c + \epsilon\frac{1+r}{4}, & |c| \le \epsilon \\ -rc, & c < -\epsilon \end{cases}$$

Le dernier terme joue le rôle d'une pénalisation basée sur la parcimonie des dérivées première et seconde, où $\mathbf{D}_1$ et $\mathbf{D}_2$ correspondent aux matrices de différences finies de premier et de second ordre, par exemple :

$$\mathbf{D}_2 = \begin{bmatrix} -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \end{bmatrix}.$$

Le choix de la fonction ϕ est $\phi(x) = |x| - \epsilon\log(|x| + \epsilon)$, de dérivée $\phi'(x) = x/(|x| + \epsilon)$. La minimisation de la fonction

1. Ce travail s'appuie sur [7], qui n'a pas été présenté en congrès, et présente des résultats comparatifs avec [8] et [9].

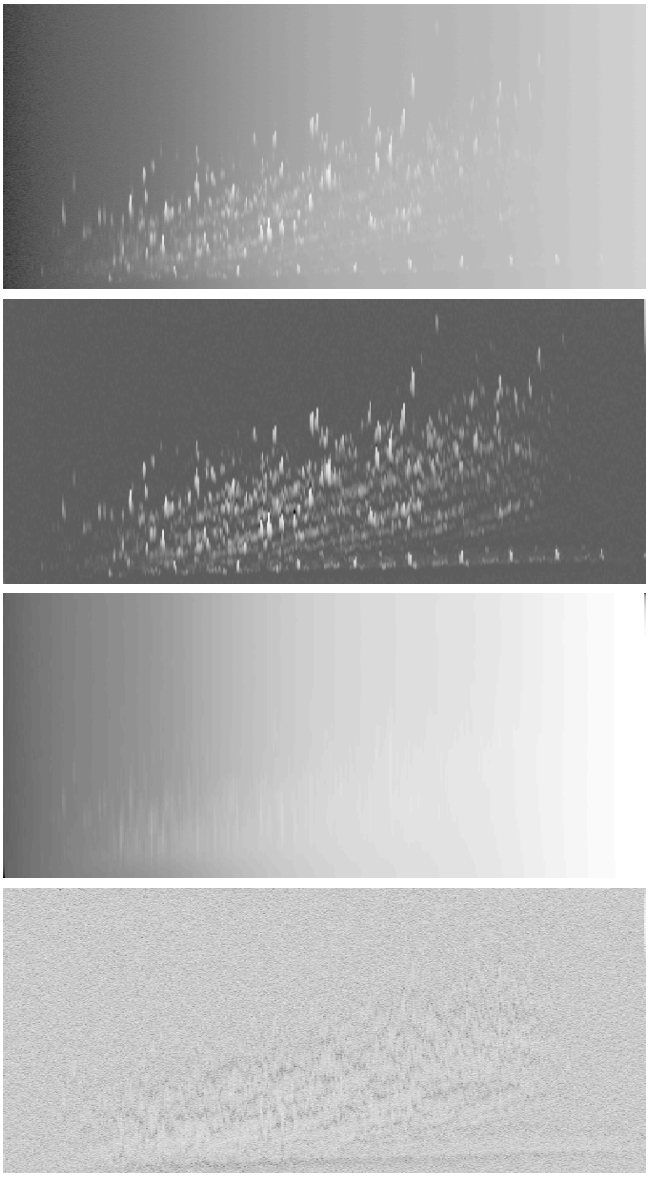

FIGURE 2 – Données, pics, ligne de base et bruit du chromatogramme bidimensionnel composite ($f_c = 0{,}01$, $\lambda = 0{,}4$).

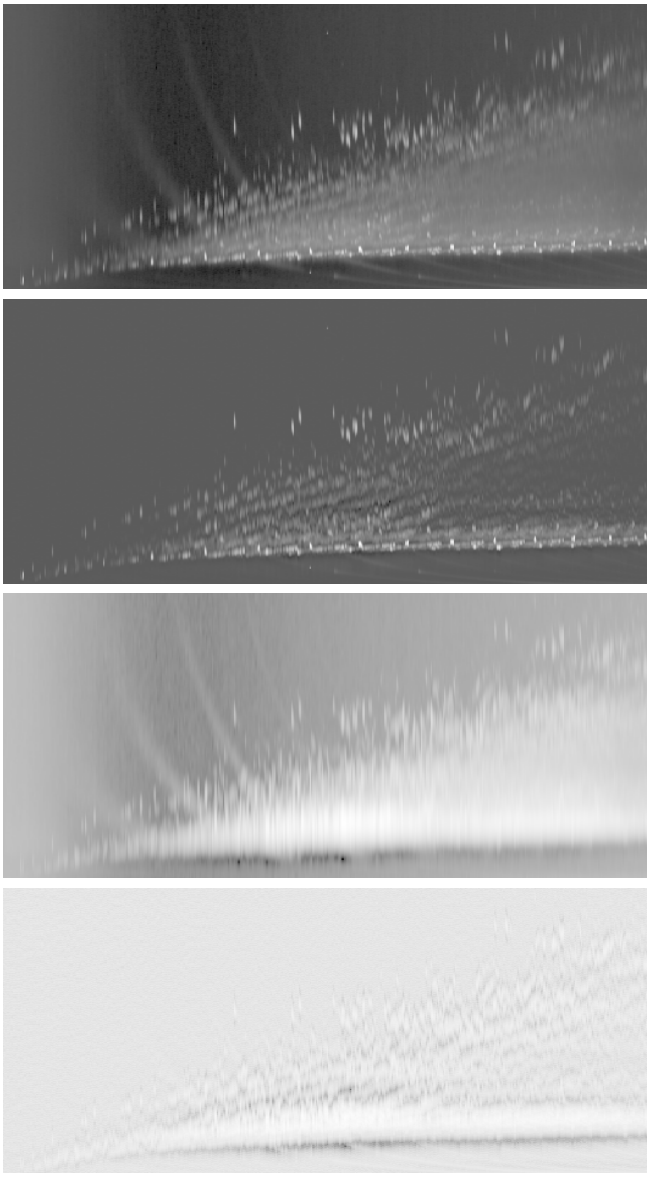

FIGURE 3 – Données, pics, ligne de base et bruit d'un chromatogramme bidimensionnel réel ($f_c = 0{,}08$, $\lambda = 0{,}8$).

$F(\mathbf{p})$ peut être obtenue par majoration-minimisation [18, 19], et nous renvoyons vers [7] pour les détails d'implémentation, ainsi que pour la comparaison avec deux algorithmes concurrents [8, 9] dans le cas de bruits gaussien et de Poisson. Nous évaluons maintenant les performances sur des données chromatographiques bidimensionnelles, issues d'un enroulement cyclindrique par injection sur deux colonnes "orthogonales".

3 Simulations et résultats

La validation est effectuée sur des données de chromatographie bidimensionnelle, où le mélange complexe est séparé par deux colonnes de propriétés différentes [20]. Elles sont segmentées en blocs contigus distincts en une dimension. Ces derniers sont juxtaposés pour former une image possédant une continuité cylindrique, reliant haut et bas. Les paramètres sont choisis de la manière suivante. Le filtre $\mathbf{H}$, combinaison de deux filtres récursifs de premier ordre, est uniquement paramétré par sa fréquence de coupure f_c. Le paramètre r est choisi égal à 6. Les paramètres λ_i, $i \in 0{,}1{,}2$, pour des pics chromatographiques modèles ou de motifs représentatifs, peuvent être reliés à l'inverse des normes ℓ_1 des dérivées. Il est donc possible de les ramener à un paramètre λ, tel que $\lambda_i = \lambda / \|\mathbf{D}_i \mathbf{p}\|_1$, essentiellement proportionnel au niveau de bruit.

Les premières données sont issues de la simulation du chromatogramme composite présenté en figure 1. Les pics $\mathbf{p}$ sont issus d'un chromatogramme bien séparé, auquel est ajouté une ligne de base simple croissante et une réalisation d'un bruit

gaussien. Le résultat est visible en haut de la figure 2, où l'on perçoit un fond s'éclaircissant de la gauche de l'image vers la droite. Le résultat $\hat{\mathbf{p}}$ situé en dessous montre un fond quasi-constant, faisant émerger des pics initialement noyés dans le fond, extrait dans la troisième sous-figure. Enfin, le bruit extrait, bien que concentré dans la zone centrale, apparaît relativement étalé et aléatoire. Le chromatogramme en haut de la figure 3 est un mélange plus complexe (brut pétrolier), composé de milliers de pics. On remarque également des traces en forme d'exponentielles décroissantes manifestement liées à des artefacts d'acquisition. La sous-figure suivante montre un fond plus homogène et des pics mieux séparés. L'image de ligne de base a absorbé les artefacts, et présente bien un fond relativement continu dans la zone centrale, correspondant effectivement à une tendance lente. Le terme résiduel en bas présente une structure plus aléatoire, nettement hétéroscédastique, mais suffisamment diffuse pour ne pas perturber l'interprétation des pics de la deuxième sous-figure.

4 Conclusions et perspectives

Nous avons présenté une méthode de suppression conjointe de ligne de base et de débruitage adaptée aux signaux de type physico-chimique présentant un mélange linéaire de pics positifs, avec des contraintes de parcimonie. Son usage est montré sur des données de chromatographie bidimensionnelle. Cette méthode peut être appliquée à d'autres types de mesures parcimonieuses, par exemple dans le domaine biomédical (ECG, EEG) ou plus généralement dans les sciences de la vie [21], ou pour la soustraction de fond spectral en analyse fréquentielle ou temps-fréquence. Son usage couplé à une déconvolution aveugle parcimonieuse par rapport de normes ℓ_1/ℓ_2 [22] est une perspective. La boîte à outils BEADS (*Baseline Estimation And Denoising with Sparsity*) est disponible à l'adresse http://lc.cx/beads.

References

[1] S. D. Brown, R. Tauler, and B. Walczak, editors. *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*. Elsevier, 2009.

[2] H. Rix and J.-P. Malenge. Séparation de deux pics visuellement confondus. In *Proc. GRETSI*, pages 1–6, Nice, France, 26-30 avr. 1977.

[3] N. Dobigeon, S. Moussaoui, and J.-Y. Tourneret. Séparation bayésienne de sources spectrales sous contraintes de positivité et d'additivité. In *Proc. GRETSI*, pages 1237–1240, Troyes, France, 11-14 sep., 2007.

[4] N. Dobigeon, S. Moussaoui, M. Coulon, J.-Y. Tourneret, and A. O. Hero. Extraction de composants purs et démélange linéaire bayésiens en imagerie hyperspectrale. In *Proc. GRETSI*, Dijon, France, 8-11 sep., 2009.

[5] L. Duval, L. T. Duarte, and C. Jutten. An overview of signal processing issues in chemical sensing. In *Proc. Int. Conf. Acoust. Speech Signal Process.*, pages 8742–8746, Vancouver, BC, Canada, May 26-31, 2013.

[6] V. Mazet, J. Idier, D. Brie, B. Humbert, and C. Carteret. Estimation de l'arrière-plan de spectres par différentes méthodes dérivées des moindres carrés. In *Chimiométrie*, Paris, France, 3-4 déc. 2003.

[7] X. Ning, I. W. Selesnick, and L. Duval. Chromatogram baseline estimation and denoising using sparsity (BEADS). *Chemometr. Intell. Lab. Syst.*, 139:156–167, Dec. 2014.

[8] V. Mazet, C. Carteret, D. Brie, J. Idier, and B. Humbert. Background removal from spectra by designing and minimising a non-quadratic cost function. *Chemometr. Intell. Lab. Syst.*, 76(2):121–133, 2005.

[9] Z.-M. Zhang, S. Chen, and Y.-Z. Liang. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, 135(5):1138–1146, 2010.

[10] J.-F. Aujol and A. Chambolle. Dual norms and image decomposition models. *Int. J. Comp. Vis.*, 63(1):85–104, 2005.

[11] L. M. Briceño-Arias, P. L. Combettes, J.-C. Pesquet, and N. Pustelnik. Proximal algorithms for multicomponent image processing. *J. Math. Imaging Vision*, 41(1):3–22, Sep. 2011.

[12] J. D. Wilson and C. A. J. McInnes. The elimination of errors due to baseline drift in the measurement of peak areas in gas chromatography. *J. Chrom. A*, 19:486–494, 1965.

[13] A. W. Moore and J. W. Jorgenson. Median filtering for removal of low-frequency background drift. *Anal. Chem.*, 65(2):188–191, 1993.

[14] S. Cappadona, F. Levander, M. Jansson, P. James, S. Cerutti, and L. Pattini. Wavelet-based method for noise characterization and rejection in high-performance liquid chromatography coupled to mass spectrometry. *Anal. Chem.*, 80(13):4960–4968, 2008.

[15] K. H. Liland, M. Høy, H. Martens, and S. Sæbø. Distribution based truncation for variable selection in subspace methods for multivariate regression. *Chemometr. Intell. Lab. Syst.*, 122:103–111, Mar. 2013.

[16] R. Danielsson, D. Bylund, and K. E. Markides. Matched filtering with background suppression for improved quality of base peak chromatograms and mass spectra in liquid chromatography-mass spectrometry. *Anal. Chim. Acta*, 454(2):167–184, 2002.

[17] I. W. Selesnick, H. L. Graber, D. S. Pfeil, and R. L. Barbour. Simultaneous low-pass filtering and total variation denoising. *IEEE Trans. Signal Process.*, 62(5):1109–1124, Mar. 2014.

[18] M. Nikolova and M. K. Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM J. Sci. Comput.*, 27(3):937–966, Jan. 2005.

[19] E. Chouzenoux, J. Idier, and S. Moussaoui. A majorize-minimize strategy for subspace optimization applied to image restoration. *IEEE Trans. Image Process.*, 20(6):1517–1528, Jun. 2011.

[20] C. Vendeuvre, R. Ruiz-Guerrero, F. Bertoncini, L. Duval, and D. Thiébaut. Comprehensive two-dimensional gas chromatography for detailed characterisation of petroleum products. *Oil Gas Sci. Tech.*, 62(1):43–55, 2007.

[21] M. Blanco-Velasco, B. Weng, and K. E. Barner. ECG signal denoising and baseline wander correction based on the empirical mode decomposition. *Comput. Biol. Med.*, 38(1):1–13, Jan. 2008.

[22] A. Repetti, M. Q. Pham, L. Duval, E. Chouzenoux, and J.-C. Pesquet. Euclid in a taxicab: Sparse blind deconvolution with smoothed ℓ_1/ℓ_2 regularization. *IEEE Signal Process. Lett.*, 22(5):539–543, May 2015.