Fast algorithms for informed source separation

Augustin Lefèvre augustin.lefevre@uclouvain.be

July, 10th 2013
Source separation in 5 minutes

We consider the single-channel setting:

\[ x_t = s_t^{(1)} + s_t^{(2)}. \]

Ill-posed problem, need prior information.
Read mix waveform

\[
x \xrightarrow{\text{STFT}} C \xrightarrow{\cdot^2} Y
\]

\[
S_g \xleftarrow{i\text{STFT}} S_g \xrightarrow{\text{masking}} X_g
\]

\[g = 1 \ldots G\]
Short time Fourier transform

\[ C_{fn} = \sum_{t=1}^{F} x_{t+(n-1)H} w_t \exp \left( - \frac{2(f - 1)\pi(t - 1)}{F} \right) \]
Remove phase information

\[ x \xrightarrow{\text{STFT}} C \xrightarrow{| \cdot |^2} Y \]

\[ s_g \xleftarrow{i\text{STFT}} S_g \xrightarrow{\text{masking}} X_g \]

\[ g = 1 \ldots G \]
Output of source separation program

\[ x \xrightarrow{\text{STFT}} C \xrightarrow{| \cdot |^2} Y \]

\[ s_g \xleftarrow{i\text{STFT}} S_g \xleftarrow{\text{masking}} X_g \]

\[ g = 1 \dots G \]
Time-frequency masking

Estimates of each source’s complex STFT are obtained by:

\[ S_{g, fn} = \frac{X_{g, fn}}{\sum_l X_{l, fn}} C_{fn} \]
Estimate waveforms from STFT

\[ x \xrightarrow{\text{STFT}} C \xrightarrow{| \cdot |^2} Y \]

\[ S_g \leftarrow \text{iSTFT} \quad \text{masking} \quad X_g \]

\[ g = 1 \ldots G \]
Annotation informed source separation

[Lefèvre et al., 2012, Bryan and Mysore, 2013]: interaction between user and source separation software.

[Lefèvre et al., 2012]: detector trained on development database (random forest, SVM, nearest-neighbour, etc.).

Figure: Detections in the spectrogram
Annotation informed source separation.

Information is used as additional constraints: \( M_g \odot X_g = M_g \odot T_g \).

[Lefèvre et al., 2012]: nonnegative matrix factorization (nmf) with constraints:

\[
\begin{align*}
\min_{D,A} & \quad \left\| Y - \sum_g D_g A_g \right\|_F^2 \\
\text{s.t.} & \quad D \in \mathbb{R}^{F \times K}_+, A \in \mathbb{R}^{K \times N}_+ \\
& \quad M_g \odot (D_g A_g) = M_g \odot T_g \\
Y \in \mathbb{R}^{F \times N}_+ & \text{ is the input spectrogram.}
\end{align*}
\]

Need only \( D_1 A_1 \geq 0 \), but impose stronger constraint: \( D_1 \geq 0, A_1 \geq 0 \) (NMF).

nmf is hard ... see talk by Nicolas Gillis.
Informed source separation: \( X_1, \ldots, X_G \in \mathbb{R}^{F \times N} \).

\[
\begin{align*}
\min_X & \quad \frac{1}{2} ||Y - \sum_{g=1}^{G} X_g||_F^2 + \lambda \sum_{g=1}^{G} ||X_g||_* \\
\text{s.t.} & \quad M_g \odot X_g = M_g \odot T_g \\
& \quad X_g \geq 0
\end{align*}
\]

The rank of a matrix is revealed in its SVD: \( X = P\Sigma Q^T \).

\( \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_F \geq 0 \) singular values.

\( ||X_g||_* = \sum_{f=1}^{F} \sigma_f \).

Projecting on \( X_g \geq 0 \) is straightforward.

Instead of one \text{nmf}, we will make repeated calls to \text{svd} to compute \( ||X_g||_* \) and additional information.
Convex but nonsmooth problem.

Related approaches if no noise and no inequality constraints (Recht et al., 2010):

$$\min \|X\|_*$$

s.t. $$A(X) = b$$

where $$A : \mathbb{R}^{F \times NG} \rightarrow \mathbb{R}^p$$, $$b \in \mathbb{R}^p$$ ($$p \ll m \times n$$) is linear.

Link with SDP optimization:

$$\min t$$

s.t. $$A(X) = b$$

$$\begin{pmatrix} tl & X \\ X & tl \end{pmatrix} \succeq 0$$

Use interior-point solver, which has superlinear convergence rate.

BUT Hessian has size $$O(F^2N^2)$$, i.e. $$10^{10}$$ for a ten seconds audio track. This is too large!
Subgradient descent

Objective function $f$ is convex so it admits derivatives in all directions:

$$f'(X; D) = \lim_{t \downarrow 0} \frac{f(X + tD) - f(X)}{t}$$

Subgradients generalize the gradient:

$$Z \in \partial f(X) \iff f'(X; D) \geq \langle Z, D \rangle$$

$$\langle Z, D \rangle = \sum_g \text{Tr} \ Z_g^\top D_g$$

Projected subgradient descent: $X^{(t+1)} = \Pi(X^{(t)} - \mu_t Z^{(t)})$.

Warning: $f(X^{(t+1)}) \not\leq f(X^{(t)})$.

Guarantee: $\mu_t = \mu_0 (1 + t)^{-\frac{1}{2}} \Rightarrow \|X^{(t)} - X^*\| \downarrow 0$. 
Controlled experiments

**Figure:** (Left) Evolution of SDR as a function of CPU time (in seconds), for (green) our method and (red) NMF started from several initial points.

SDR is a measure of how well we have separated sources (the higher the better).
Shrinkage of singular values

**Figure**: Magnitude of singular values in decreasing order, for various values of $\lambda$. Dotted line is the true singular value profile.
Smoothing technique [Nesterov, 2003]

\[
\min_X \frac{1}{2} \| Y - \sum_{g=1}^{G} X_g \|_F^2 + \lambda \sum_{g=1}^{G} \| X_g \|_{*,\mu} \\
\text{s.t.} \quad M_g \odot X_g = M_g \odot T_g \\
X_g \geq 0
\]

\( \| \cdot \|_{*,\mu} \) is \( C^{1,1} \) with Lipschitz constant \( \frac{1}{\mu} \) and:

\[
\| X \|_{*,\mu} \leq \| X \|_* \leq \| X \|_{*,\mu} + \mu C \quad \forall X \in \mathbb{R}^{F \times N}
\]

\[
\| X \|_* = \max\{ \text{Tr} \ U^\top X, \sigma_1(U) \leq 1 \}
\]

\[
\| X \|_{*,\mu} = \max\{ \text{Tr} \ U^\top X - \| U \|_F^2, \sigma_1(U) \leq 1 \}
\]

Apply accelerated gradient descent to the smooth minimization problem.

\( \mu = 0 \) : slow convergence but accurate solutions.

Large \( \mu \) : fast but inaccurate solutions.
Comparison with subgradient

Figure: Decrease of the objective function as a function of the allowed CPU time, for various algorithms
Effect of $\mu$

Figure: Decrease of the objective function as a function of the allowed CPU time, for various values of $\mu$.

We display the original objective function:

$$\frac{1}{2} \| Y - \sum_{g=1}^{G} X_g \|_F^2 + \lambda \sum_{g=1}^{G} \| X_g \|_*.$$
Conclusion

Our formulation contributes to the field of informed source separation methods, where knowledge is directly relevant to the query audio track, and involves interaction with the user. These methods are the state of the art in single-channel source separation benchmarks.

Our convex formulation compares well with its NMF counterpart, even with a subgradient algorithm. The smoothing technique allows to retrieve more accurate solutions for a given CPU budget.

More complex constraints? E.g., source estimates should classify correctly: \( \langle W, X_g \rangle + b \leq 0 \).
Proximal operator:

$$\text{prox}(\bar{X}) = \arg \min_X \frac{1}{2} \| \bar{X} - X \|_F^2 + \lambda \| X \|_*, \quad \text{s.t.} \quad M_g \odot X_g = M_g \odot T_g,$$

Necessary and sufficient conditions:

$$0 \in X - \bar{X} + \lambda (PQ^T + W) + M \odot E$$
$$W^T X = 0$$
$$WX^T = 0$$
$$M \odot X = M \odot T$$
$$\| W \|_{op} \leq 1$$

where $E \in \mathbb{R}^{F \times N}$ are Lagrangian multiplicators associated with the constraint $M \odot X = 0$. Note that here, $X = P\Sigma Q^T$ is an economy-size SVD of $X$ and not $\bar{X}$, so $P$ and $Q$ depend on $X$. 

