Supervised group nonnegative matrix factorisation with similarity constraints and applications to speaker identification

Romain Serizel\textsuperscript{1}, Victor Bisot\textsuperscript{2}, Slim Essid\textsuperscript{2}, Gaël Richard\textsuperscript{2}

\textsuperscript{1}LORIA, Université de Lorraine, Inria, CNRS (France)
\textsuperscript{2}LTCI, Télécom ParisTech, Université Paris-Saclay (France)

Monday 6\textsuperscript{th}, March 2017
1 Speaker identification: What? Why?

2 Task-driven group NMF

3 Conclusions
Speaker identification

Main goal
Identify a person from an audio recording
Speaker identification

Main goal
Identify a person from an audio recording
Session concept

Recording variability

Icons made by Freepik from www.flaticon.com
Session concept

Recording variability

- Aging,
Session concept

Recording variability

- Aging,
- Perturbations,
Session concept

Recording variability

- Aging,
- Perturbations,
- Microphones...
Session concept

Recording variability
- Aging,
- Perturbations,
- Microphones...
Applications

Audio indexing (broadcast show, conferences, . . .)
- Content retrieval
- Rich-text transcription

Robust speech transcription
- Speaker adaptive training
- Speaker related feature/model adaptation

Voice-based identification
- Soft biometrics
Standard classification chain (1)

- Train the dictionary

\[ V \approx W \times H \]

\[ \min_{W, H} D(V \mid WH) \]
Standard classification chain (2)

- Project data on the new space

\[
\min_{\mathbf{W}} D\left(\mathbf{V} \mid \mathbf{WH}\right)
\]
Standard classification chain (3)

- Train the classifier on projected data

\[
\begin{align*}
V & \approx W \times H \\
\min D(V \mid WH) & \min_{A} E_{y,H}[I_s(y, A, h)]
\end{align*}
\]
Standard classification chain (4)

- Use the dictionary and classifier

\[ V \approx W \times h \]

\[ \min_H D(V | WH) \]

Predicted class-probabilities

Problem

\[ W \] and \[ H \] are optimised according to a reconstruction criterion

Serizel et al.
Task-driven group NMF for speaker identification
3/6/2017
Standard classification chain (4)

- Use the dictionary and classifier

\[
\min_{H} D(V \mid WH)
\]

\[
V \approx W \times h
\]

Predicted class-probabilities

Problem

\( W \) and \( H \) are optimised according to a \textit{reconstruction} criterion
Task-driven NMF (1)

General idea
Learn the dictionaries together with classifier parameters:
- Nested optimisation problem

Dictionary divergence
- Euclidean norm: **closed form solution** for dictionaries
  - Task driven dictionary learning (Mairal et al., 2012)
  - Application to audio scene analysis (Bisot et al., 2016)
- General $\beta$-divergence:
  - Application to source separation (Sprechmann et al., 2014)
  - Application to event detection (Bisot et al., 2017)
Task-driven NMF : General idea

\[ V \approx \min \limits_{W, A} E_{y, v} [l_s (y, A, \hat{h}(v, W))] \]
Task-driven NMF : General idea

\[
\min_{W, A} E_{y, v}[l_s(y, A, \hat{h}(v, W))] + \frac{\nu}{2} \| A \|^2
\]
Task-driven NMF: algorithm (1)

- For each new sample ($v$)

\[
\min_h \frac{1}{2} \| v - Wh \|_2^2 + \lambda_1 \| h \|_1 + \frac{\lambda_2}{2} \| h \|_2^2
\]
Task-driven NMF: algorithm (2)

- For each new sample ($v$)

Update the classifier parameters $A$

$$
\min_{A, v} E_{y, v}[l_s(y, A, \hat{h}(v, W))] + \frac{\nu}{2} \| A \|_2^2
$$
Task-driven NMF: algorithm (3)

- For each new sample \( \mathbf{v} \)

Update the dictionary \( \mathbf{W} \)

\[
\frac{\min_{\mathbf{w}}}{E_{y,v}} \left[ l_s(y, A, \hat{h}(v, W)) \right] + \frac{\nu}{2} \| A \|_2^2
\]
Task-driven NMF in practice

Implementation details

- Can be applied to sample or mini-batch
- Supports nonnegativity constraints for $W$ and $H$
- Dictionary ($W$) initialisation:
  - Random
  - NMF
  - Concatenated group NMF dictionaries (Serizel et al., 2016)
Task-driven group NMF (1)

Include group NMF in a task-driven framework (Serizel et al., 2016)

- Subdictionaries (related to a single speaker/session)
Task-driven group NMF (1)

Include group NMF in a task-driven framework (Serizel et al., 2016)

- Subdictionnaires (related to a single speaker/session)
- Impose speaker/session similarity constraints
Task-driven group NMF (1)

Include group NMF in a task-driven framework (Serizel et al., 2016)

- Subdictionaries (related to a single speaker/session)
- Impose speaker/session similarity constraints
Task-driven group NMF: algorithm (1)

- For each new sample ($v$)

Source code is available at https://github.com/rserizel/TGNMF
Task-driven group NMF: algorithm (2)

- For each new sample ($v$)

Update the classifier parameters $A$

Source code is available at https://github.com/rserizel/TGNMF
Task-driven group NMF: algorithm (3)

- For each new sample ($v$)

Update the corresponding dictionary $W^{(cs)}$

$$
\min_{W^{(cs)}} E_{y,v} \left[ l_s (y,A,\hat{h}(v,W)) \right] + \frac{\nu}{2} \| A \|_2^2
$$

Source code is available at https://github.com/rserizel/TGNMF
Task-driven group NMF: algorithm (3)

- **Speaker** similarity constraint

Source code is available at https://github.com/rserizel/TGNMF
Task-driven group NMF: algorithm (3)

- **Session** similarity constraint

Update the corresponding dictionary $W^{(cs)}$

$$
\min_{W^{(cs)}} E_{y,v} \left[ l_s(y, A, \hat{h}(v, W)) \right] + \frac{\nu}{2} || A ||^2 + \mu_1 J_{SPK} + \mu_2 J_{SES}
$$

Source code is available at [https://github.com/rserizel/TGNMF](https://github.com/rserizel/TGNMF)
Experiments

Experiment setup

- Subset of the ESTER corpus (≈ 6 hours training data)
- 132 constant-Q transform coefficients
- Initial dictionary obtained with (group-)NMF : 100 iterations
- Projection on \( h \) with SPAMS toolbox\(^a\)
- Classifier : multinomial logistic regression
- After 5 epochs : fix \( W \), train \( A \) alone for 50 epochs

\(^a\) http://spams-devel.gforge.inria.fr/
### Results (1)

#### Weighted F1-scores

<table>
<thead>
<tr>
<th>Initialisations</th>
<th>I-vector</th>
<th>NMF</th>
<th>$\text{GNMF}_0$</th>
<th>$\text{GNMF}_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td>76.1%</td>
<td>75.6%</td>
<td>80.7%</td>
<td>81.7%</td>
</tr>
<tr>
<td>TNMF Tuning</td>
<td>–</td>
<td>79.9%</td>
<td>81.1%</td>
<td>81.9%</td>
</tr>
</tbody>
</table>

- $\text{GNMF}_0$ : group NMF **without** similarity constraints
- $\text{GNMF}_c$ : group NMF **with** similarity constraints (speaker and session)
## Results (2)

### Weighted F1-scores

<table>
<thead>
<tr>
<th>Initialisations</th>
<th>GNMFO</th>
<th>GNMFc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td>80.7%</td>
<td>81.7%</td>
</tr>
<tr>
<td>Tuning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TNMF</td>
<td>81.1%</td>
<td>81.9%</td>
</tr>
<tr>
<td>TGNMFO</td>
<td>81.7%</td>
<td>82.1%</td>
</tr>
<tr>
<td>TGNMFc</td>
<td>82.0%</td>
<td><strong>82.2%</strong></td>
</tr>
</tbody>
</table>

- **(T)GNMFO**: (task-driven) group NMF *without* similarity constraints
- **(T)GNMFc**: (task-driven) group NMF *with* similarity constraints (speaker and session)
Conclusions and future work

NMF for speaker identification
- Can be competitive with I-vectors

Task-driven NMF
- Large improvements for small dictionaries
- TGNMF\(_c\) best performance to date on the corpus

Future work
- Experiment with \(\beta\)-divergence
- Extend the framework to deep learning...
Further readings

V. Bisot, R. Serizel, S. Essid, and G. Richard. Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification. HAL-archives ouvertes : working paper or preprint (hal-01362864), September 2016. URL https://hal.archives-ouvertes.fr/hal-01362864.


