Compressive Gaussian Mixture Estimation by Orthogonal Matching Pursuit with Replacement

Nicolas Keriven, Rémi Gribonval

Compressive Gaussian Mixture Estimation by Orthogonal Matching Pursuit with Replacement

Nicolas Keriven, Rémi Gribonval
INRIA Rennes-Bretagne Atlantique
Campus de Beaulieu, 35042 Rennes, France
Email: firstname.lastname@inria.fr

***Abstract*—This work deals with the problem of fitting a Gaussian Mixture Model (GMM) to a large collection of data. Usual approaches such as the classical Expectation Maximization (EM) algorithm are known to perform well but require extensive access to the data. The proposed method compresses the entire database into a single low-dimensional sketch that can be computed in one pass then directly used for GMM estimation. This sketch can be seen as resulting from the application of a linear operator to the underlying probability distribution, thus establishing a connection between our method and generalized compressive sensing. In particular, the new algorithms introduced to estimate GMMs are similar to usual greedy algorithms in compressive sensing.**

I. Background

Traditional methods for diminishing the complexity of a learning task compress *each* individual data point to reduce dimension [1]. On the contrary, the database literature often summarizes an entire collection of data with objects referred to as "sketches", whose size does not depend on the number of items in the collection [2].

In [3], Bourrier *et al.* introduced a method to estimate isotropic Gaussian Mixture Models (GMM) with fixed variance from a sketch, using an algorithm similar to Iterative Hard Thresholding (IHT) [4]. Inspired by Random Fourier sampling, the sketch is formed by a sampling of the empirical characteristic function: given a collection of items $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ drawn *i.i.d.* from a density $p \in L^1(\mathbb{R}^n)$ and frequencies $\boldsymbol{\omega}_j \in \mathbb{R}^n$, $1 \leq j \leq m$, the sketch is defined as

$$\hat{\mathbf{z}} = \left[\hat{\mathbb{E}}\left(e^{-i\boldsymbol{\omega}_j^T \mathbf{x}}\right)\right]_{j=1,\dots,m} \tag{1}$$

where $\hat{\mathbb{E}}(f(\mathbf{x})) = \frac{1}{N}\sum_{i=1}^N f(\mathbf{x}_i)$. The inequality $m \ll nN$ is verified so that the database is heavily compressed. The sketch is formed by a collection of empirical moments, and thus approximates a vector of real moments that can be considered as resulting from a linear operator $\mathcal{A}$ applied to the distribution p. Our goal is to estimate p from the sketch $\hat{\mathbf{z}} \approx \mathcal{A}p$.

In compressive sensing, such underdetermined inverse problems are usually dealt with by assuming that the encoded signal belongs to a low-dimensional model such as the set of sparse vectors, i.e. vectors that are a combination of only a few elements from the canonical basis of $\mathbb{R}^n$. In our setting, the density p is said to be "sparse" if it is a GMM, i.e. a combination of a few elements from the set $\mathcal{G} = \{p_{\boldsymbol{\theta}}\}$ of Gaussian distributions: $p = \sum_{k=1}^K \alpha_k p_{\boldsymbol{\theta}_k}$.

For density estimation as an inverse problem, unlike previous approaches [5] [6] that assume a *finite* set of basic distributions with limited *coherence* between its elements, here the set $\mathcal{G}$ is uncountable, and two distributions in $\mathcal{G}$ can be infinitely close to each other, hence the need to come up with a novel approach.

II. Contribution

Our main goal is to extend the method in [3] to *non-isotropic* Gaussians with diagonal covariance. Many modifications were necessary to deal with the numerous challenges raised by relaxed variances.

A. Proposed algorithms

The proposed algorithms derive from the Orthogonal Matching Pursuit (OMP) algorithm, which progressively adds *atoms* $\mathcal{A}p_{\boldsymbol{\theta}_k}$ most correlated with the residual to the mixture, until the desired sparsity K is attained. As the dictionary of available atoms is uncountable, this requires an optimization step that only yields an atom *highly* correlated with the residual instead of the true maximum. We derive two algorithms from OMP, by successively adding two modifications.

- **OMP for Compressive GMM (OMPC)**: we add a parametric optimization of the whole mixture at each iteration, providing the adjustment required when adding a Gaussian to the mixture (Fig. 1).

- **OMP with Replacement for Compressive GMM (OMPRC)**: in addition to the previous modification, similar to [7], we run the algorithm for *more* than K iterations (typically $2K$ iterations), and at each iteration enforce K-sparsity by Hard Thresholding if necessary.

B. Construction of the sketch

Similar to classical Random Fourier sampling, the frequencies $\boldsymbol{\omega}_j$ are randomly chosen. A novel distribution $p_{\boldsymbol{\omega}}$ is introduced, based on a heuristic that maximizes the capacity of the sampled characteristic function to discriminate between two distinct GMM parameters. In particular, while previous choices [3] exhibit many problems in high dimension, this distribution leads to reconstruction results that are substantially more robust to dimensionality (Fig. 2).

III. Some Results

Through extensive testing of the method on synthetic data, we evaluate the precision of the reconstruction with the Kullback-Leibler (KL) divergence, by taking a median value over 30 experiments. A clear phase transition with respect to the number of measurements is observed (Fig. 5). On large databases, our MATLAB implementation of the compressive algorithms substantially outperforms a state-of-the-art C++ implementation of EM [8] in terms of time and memory complexity (Fig. 3). In terms of reconstruction precision, the OMPRC algorithm essentially matches EM in every setting, and even outperforms it in some cases (Fig. 4).

We also compare our method to EM on a speaker verification task [9] using the NIST 2005 database [10], which requires learning a GMM over a database with millions of items. Despite the heavy compression, the proposed compressive method yields results close to those of EM, in particular when increasing the number of measurements m (Fig. 6).

IV. Conclusion

We presented a practical compressive approach to GMM estimation which outperforms usual methods by orders of magnitude in terms of memory usage while preserving the precision of the estimation.

Moreover, the sketch can be easily updated with new data and preserves data privacy, which can be crucial for many applications. For instance, in the presented speech processing task, it allows not to store the spoken fragments, possibly of sensitive nature, while permitting the update of the database with new recordings.

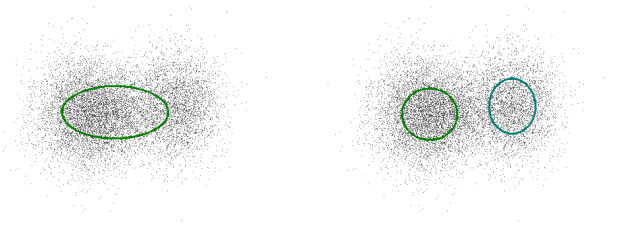

Fig. 1. Approximation of a distribution with a 1-GMM (left) and a 2-GMM (right). The 2-GMM cannot be derived from the 1-GMM by simply adding a Gaussian (as would do usual OMP), hence the need for a global optimization step in OMPC and OMPRC.

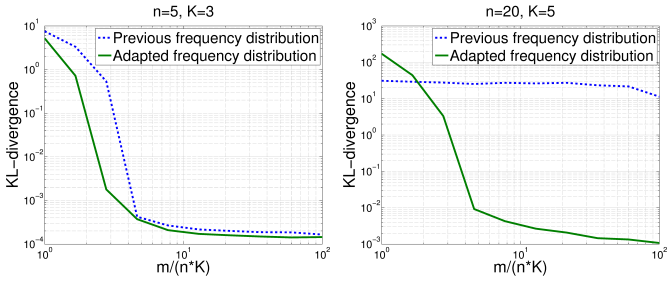

Fig. 2. KL-divergence results for OMPRC with respect to the relative number of frequencies $\frac{m}{Kn}$, drawn either from the previous frequency distribution [3] or our adapted distribution $p_{\boldsymbol{\omega}}$. The former choice is seen to be very sensitive to the dimension n, and fails to yield satisfying results in high dimension (right).

ACKNOWLEDGMENT

This work was supported in part by the European Research Council, PLEASE project (ERC-StG- 2011-277906).

REFERENCES

[1] R. Calderbank, S. Jafarpour, and R. Schapire, "Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain," *Preprint*, 2009.

[2] A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. J. Strauss, "How to summarize the universe: Dynamic maintenance of quantiles," *International Conference on Very Large Data Bases*, pp. 454–465, 2002.

[3] A. Bourrier, R. Gribonval, and P. Pérez, "Compressive gaussian mixture estimation," *International Conference on Acoustics, Speech and Signal Processing*, pp. 6024–6028, 2013.

[4] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 44, no. 0, pp. 1–11, 2009.

[5] F. Bunea, A. B. Tsybakov, M. H. Wegkamp, and A. Barbu, "SPADES and mixture models," *The Annals of Statistics*, vol. 38, no. 4, pp. 2525–2558, Aug. 2010.

[6] K. Bertin, E. Le Pennec, and V. Rivoirard, "Adaptive Dantzig density estimation," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 47, no. 1, pp. 43–74, Feb. 2011.

[7] P. Jain, A. Tewari, and I. S. Dhillon, "Orthogonal matching pursuit with replacement," *Advances in Neural Information Processing Systems 24*, pp. 1–9, 2011.

[8] A. Vedaldi and B. Fulkerson, "VLFeat - an open and portable library of computer vision algorithms," in *ACM International Conference on Multimedia*, 2010.

[9] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, Jan. 2000.

[10] "Nist speaker recognition evaluation plan," website: www.nist.gov/speech/test.htm.

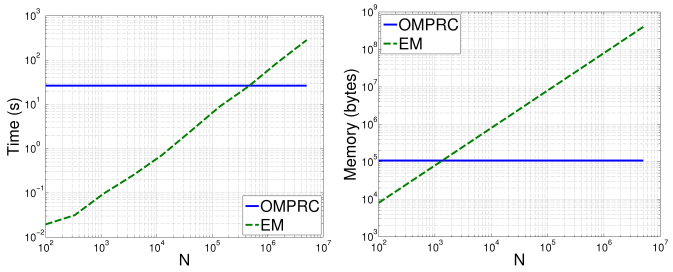

Fig. 3. Time (left) and memory (right) usage of OMPRC and EM with respect to the size of the database N, for $n = 10$ and $m = 1000$.

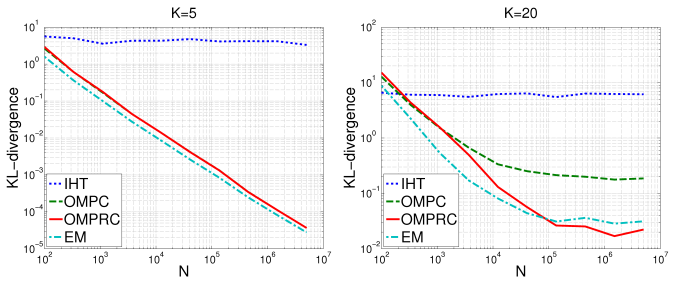

Fig. 4. KL-divergence results for the compressive algorithms and EM with respect to the size of the database N, for $n = 10$ and $K = 5$ (left) or $K = 20$ (right). The previous IHT [3] fails in our setting of non-isotropic Gaussians, while OMPRC matches the performance of EM.

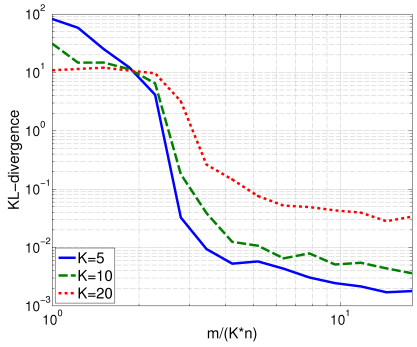

Fig. 5. KL-divergence results for OMPRC with respect to the relative number of measurements $\frac{m}{Kn}$, for $n = 15$ and $K = 5, 10, 20$.

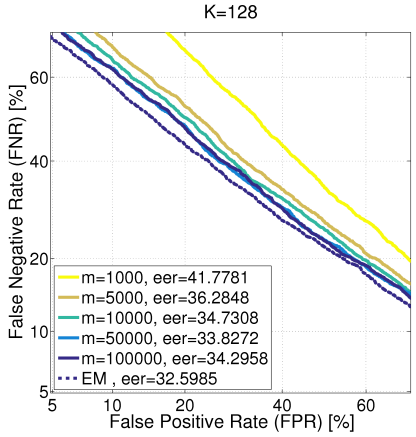

Fig. 6. Speaker verification results presented with DET-curves (on both axis, lower is better), for EM and OMPRC for various number of measurements m. At high m the performance of OMPRC approaches that of EM despite the heavy compression (here $nN \approx 10^8$).