Nonnegative CCA for Audiovisual Source Separation

Christian Sigg, Bernd Fischer, Volker Roth

ETH Zurich, Institute of Computational Science
Universität-Str. 6, CH-8092 Zurich
{chrsigg,bernd.fischer,vroth}@inf.ethz.ch

Abstract. We present a method for finding correlated components in audio and video signals. The concept of canonical correlation analysis is reformulated such that it allows us to incorporate non-negativity constraints on the coefficients. This additional requirement ensures that projection directions obey the non-negativity requirements of energy signals. By finding multiple orthogonal directions we finally obtain a component-based decomposition of both data modalities. Experiments for simultaneous source separation in both video and audio streams effectively demonstrate the benefits of this approach.

1 Introduction

In difficult auditory environments, such as a discussion in a cafeteria where talkers interfere and reverberation is high, humans often make use of visual cues to facilitate understanding and separate a speaker's voice from the background. In contrast to the auditory signal, the visual input is free of reflections, and regions of the visual field can be uniquely assigned to a source. So if there is considerable overlap in the spatial or frequency domain between acoustic sources, we expect a benefit from incorporating video into a source separation or signal enhancement algorithm, as compared to only relying on information available in the audio signal.

In recent years, there have been several proposals to exploit the statistical dependence of synchronous audio and video signals. Methods of this kind typically find projections of both data modalities that either maximize mutual information [1, 2] or correlation [3]. These methods, however, are limited in several aspects, e.g. the restriction to smooth L_2 penalties that ensure differentiability [1, 2], or the asymmetric treatment of audio and video [3]. A common drawback of all these methods is their sensitivity to outliers and the possible occurrence of negative projection coefficients which therefore cannot be interpreted as energy signals. The latter aspect is particularly important if one is interested not only in one pair of projections, but in *several* such pairs of highly correlating components of the audio and the video signal. Non-negativity of the coefficients, on the other hand, assures that individual projections define valid energy signals that successively decompose the total audio and video information. On the video side, this means that a pixel can at most be part of one source, whereas in the unconstrained case a pixel can be part of many projections (with mixed signs of coefficients) to explain the correlation structure.

We present a method that obeys such non-negativity constraints of energy signals. The key idea is to include these constraints in a generalized version of *canonical correlation analysis* (CCA). The method is highly flexible in that it allows the choice of

individual regularization strategies for the different data modalities such as sparseness constraints for video and smooth L_2 penalties for audio. Furthermore, it allows us to diminish the influence of outliers by substituting least-squares functionals with robust regression procedures.

Method Overview. We perform Canonical Correlation Analysis (CCA) to locate sources in video and separate their corresponding audio signals by filtering. Using a multi-dimensional representation of both audio (A) and video (V), we seek linear projection vectors α and β that maximize the correlation between the two projected signals: $\arg\max_{\alpha,\beta} corr(A\alpha, V\beta)$. To locate a source in the video signal, we identify those pixels whose coefficients contribute most to the projection. On the audio side, a properly defined projection can be interpreted as a frequency-domain filter, amplifying frequencies contained in the source and attenuating others.

We require the projection coefficients to be nonnegative: $\alpha(i), \beta(i) \geq 0 \ \forall i$. When working with pixel intensity information, this guarantees that a weighted video frame $\tilde{V}(i) = V(i)\beta(i)$ can again be interpreted as a proper image.

For typical video resolutions (e.g. 320x240 pixels) and frame-rates (e.g. 25Hz), the CCA problem will be severely under-determined and we need to include a regularization term to find nontrivial correlations. Concerning the video signal, it is desirable to have sparse projections $V\beta$ so that only those pixels have nonzero weights, that are associated with the source in question. L_1 regularization on the $\beta(i)$'s will do just that. On the audio side, sparsity is probably not desirable, because leaving out whole frequency bands can lead to audible artifacts. If we need to regularize we can add an L_2 term, or a smoothness penalty on coefficients $\alpha(i), \alpha(i+1)$ of adjacent frequency bands. It is an advantage of our approach that one can choose regularization type and tuning parameters individually for audio and video signals, each best suited for its data domain.

2 Nonnegative Canonical Correlation Analysis

The classical CCA method finds linear projections α, β of two multidimensional random variables such that their correlation is maximized

$$\arg\max_{\alpha,\beta} corr(A\alpha, V\beta). \tag{1}$$

A and V are matrices of size $n \times d_a$ and $n \times d_v$, where each row corresponds to one realization of the random variable. In practice, these different realizations are mimicked by using successive frames in the audio and video signal.

It can be shown that maximizing the correlation in (1) is equivalent to minimizing

$$\arg\min_{\alpha,\beta} E[A\alpha - V\beta]^2, \text{ s.t. } \|A\alpha\|_2^2 = 1 \wedge \|V\beta\|_2^2 = 1. \tag{2}$$

The solution is readily obtained using the eigenvalue decomposition of the (sample) covariance matrix (for centered A and V)

$$C = \begin{bmatrix} A^\top A & A^\top V \\ V^\top A & V^\top V \end{bmatrix} = \begin{bmatrix} C_{aa} & C_{av} \\ C_{va} & C_{vv} \end{bmatrix}. \tag{3}$$

A full derivation of the procedure can be found in [4]. The eigenvalue decomposition gives us not one, but all projection pairs $(A\alpha_k, V\beta_k)$ having maximum correlation, under the condition that subsequent projection pairs are orthogonal to each other. That is $\alpha_k^\top C_{aa}\alpha_l = \beta_k^\top C_{vv}\beta_l = \alpha_k^\top C_{av}\beta_j = 0$ for $k \neq l$. It is also possible to include an L_2 regularization penalty, to deal with the case that $n < d$.

Holding α fixed, the optimization criterion (2) is just the minimum mean-squared error criterion for regression coefficients $\beta(i)$. This formulation suggest an alternative solution approach to the CCA problem: we alternately hold one set of parameters (e.g. α) constant and perform a regression step to find the corresponding set of coefficients (β). This procedure is iterated until convergence. After each regression step, the coefficients have to be renormalized to satisfy (2).

An iterative regression solver is attractive for several reasons, and has therefore been proposed several times in the literature (e.g. [5]). We can perform ridge regression (L_2), the Lasso (L_1) or any other regression method, and choose the appropriate penalty for each data modality. As a second benefit, techniques for robust regression can be incorporated: the quadratic error in (2) can be replaced with more robust measures such as the Huber loss [6] in order to diminish the effect of outliers in the data. Finally, it is straightforward to include non-negativity constraints on the projection coefficients α and β.

Nonnegative regularized regression. When correlating audio to video, the number of pixels d_v typically exceeds the number of video frames n by far. As a consequence, the regression problem becomes ill-posed, i.e. there always exists a solution that provides a perfect regression fit with zero error. In order to find nontrivial correlations it is, thus, necessary to include a regularization penalty. An L_1 penalty seems suitable for the video signal, because it leads to a sparse solution where only pixels corresponding to the audio source have non-zero coefficients. On the audio side, a L_2 penalty is preferable because completely zeroing out bands leads to undesired and audible artifacts in the reconstruction.

We require the projection coefficients α and β to be nonnegative, since non-negativity ensures that all correlation vectors themselves are valid image- or audio energy signals and that successively found correlation directions decompose the two data modalities into additive energy components. Nonnegative regression can be solved directly by quadratic programming algorithms. Fast approximative techniques that in addition allow the inclusion of both L_1 and L_2 penalties have been proposed recently. One particularly interesting such method is the *monotone incremental forward stagewise regression* approach described in [7]. In its original formulation, it approximates the *monotone LASSO* that computes L_1-constrained regression fits in which the norms of the weights monotonically increase when relaxing the L_1-constraint. This algorithm inherently finds nonnegative weights, which for standard applications (where this feature is undesirable) is compensated for by replicating the input data with negative sign. For our purposes, we simply drop this data replication step which leaves us with a highly efficient iterative method for nonnegative L_1-penalized regression fits.

Finding all CCA projections. For the source separation task, we are naturally interested in more than one projection direction, expecting that distinct sources are retrieved in

different projections. We incorporate orthogonality constraints on subsequent projections by means of deflation. After every regression step, the projection vector α_{k+1} is adjusted so that the projection $A\alpha_{k+1}$ is uncorrelated (and therefore orthogonal) to all previously found projections $(A\alpha_l, V\beta_l),\ l < k+1$. For our special case of allowing only *non-negative* coefficients, orthogonal projections are only possible if the same column is not chosen more than once. Since we prefer sparse projections only on the video data and not in the audio domain, we only orthogonalized the β vectors by requiring that $\beta_{k+1}(i) > 0 \Rightarrow \beta_l(i) = 0\ \forall l < k+1$. This constraint corresponds to removing all previously selected pixels (and possibly all pixels in a close neighborhood thereof) before searching for the next correlations $(A\alpha_2, V\beta_2)$.

3 Experiments

We tested our method on a short video stream in which 2 persons speak simultaneously. The audio signal was represented as a vector of 50 frequency bands spaced in mel scale in the range 100 Hz - 8kHz, while for the video signal we simply worked on the pixel-intensity vectors. Nonnegative CCA was performed on sliding windows of size 50 frames. We used L_1 regularization for the video in order to identify single pixels, and L_2 regularization on the audio side. The first canonical correlation found clearly identified the left speaker, as can be seen in the middle panel of figure 1. We then searched for a second orthogonal projection vector β_2 by excluding all pixels within small windows around the identified correlating areas. The second correlation direction then clearly identified the second speaker, see the right panel in figure 1.

Fig. 1. Original scene (left), extracted image areas in the direction of highest correlation (middle), extracted image areas in the second correlating projection (right).

Future work. While we have shown that nonnegative CCA performs well in finding distinct areas in the image which e.g. correspond to different speakers, the reconstructed audio signals did not allow a good source separation which, however, could not be expected by solely using frequency bands to represent the audio signal. On the relevant time scale for correlating audio and video the frequency representation is no longer discriminative for separating concurrent speakers. We plan to address this problem by using spatial audio features derived from a microphone array with adaptive beam-forming.

4 Conclusion

We have presented the nonnegative CCA method for jointly analyzing audio and video streams. Compared to existing approaches of this kind, this technique allows us to find a series of orthogonal projections with nonnegative weights which successively decompose the signal into single components.

References

1. J. Hershey and J. Movellan. Audio vision: Using audiovisual synchrony to locate sounds. In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 813–819. MIT Press, 2000.
2. JW Fisher III and T. Darrell. Speaker association with signal-level audiovisual fusion. *Multimedia, IEEE Transactions on*, 6(3):406–413, 2004.
3. E. Kidron, Y.Y. Schechner, and M. Elad. Pixels that sound. *Proc. of CVPR*, pages 88–95, 2005.
4. D.R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical Correlation Analysis: An Overview with Application to Learning Methods. 2004.
5. T. Landelius M. Borga and H. Knutsson. A Unified Approach to PCA, PLS, MLR and CCA. Report LiTH-ISY-R-1992, ISY, SE-581 83 Linkoping, Sweden, November 1997.
6. P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.
7. T. Hastie, J. Taylor, R. Tibshirani, and G. Walther. Forward Stagewise Regression and the Monotone Lasso. Unpublished.