RICA

ICA Summary

Independent Component Analysis (ICA) allows us to generate sparse representations of whitened data by the following formula:

$$\begin{aligned} \text{minimize} \quad & \|Wx\|_1 \\ \text{s.t.} \quad & WW^T = I \end{aligned}$$

where W is our weight matrix and x is our input. In ICA, we minimize the L1 penalty (sparsity) of our hidden representation, Wx, while maintaining an orthonormal constraint for our weight matrix. The orthonormal constraint exists to ensure that our uncorrelated data remains uncorrelated within our feature representation. In other words, an orthonormal transformation of whitened data remains white.

The orthonormal constraint in ICA presents some drawbacks to the algorithm. Namely, difficulties arise when the number of features (rows of W matrix), exceed the dimensionality of input, x. Optimization also becomes more difficult with hard constraints, and thus requires longer training. How could we speed this up? What if the dimensionality of our data is too large to be whitened? Keep in mind, if $x \in R^n$ it requires an $n \times n$ whitening matrix.

RICA

One algorithm called Reconstruction ICA (RICA), was designed to overcome the drawbacks of ICA by replacing ICA's orthonormality constraint with a soft reconstruction penalty.

$$\min_W \quad \lambda\|Wx\|_1 + \frac{1}{2}\left\|W^TWx - x\right\|_2^2$$

To help understand the motivation behind this, we see that we can get a perfect reconstruction when the features are not over-complete. To achieve this, we constrain $W^TW = I$. It is also possible to recover ICA from RICA when features are not over-complete, data is whitened, and λ goes to infinity; at this point, perfect reconstruction becomes a hard constraint. Now that we have a reconstructive penalty in our objective and no hard constraints, we are able to scale up to over-complete features. However, will the result still be reasonable when we are using an over-complete basis? To answer this, we move to another common model, the sparse autoencoder.

To better interpret what happens when we move to an over-complete case, let's revisit sparse autoencoders. The objective is listed below:

$$\min_W \quad \lambda\|\sigma\left(Wx\right)\|_1 + \frac{1}{2}\left\|\sigma\left(W^T\sigma\left(Wx\right)\right) - x\right\|_2^2$$

There are different variations of autoencoders, but for the sake of consistency, this formula uses an L1 sparsity penalty and has a tied reconstruction matrix W. The only difference between this sparse autoencoder and RICA is the sigmoid non-linearity. Now, looking at the reconstructive penalty from the auto-encoder perspective, we can see that the reconstructive penalty acts as a degeneracy control; that is, the reconstructive penalty allows for the sparsest possible representation by ensuring that the filter matrix does not learn copies or redundant features. Thus we can see that RICA in the over-complete case is the same as a sparse autoencoder with an L1 sparsity constraint and without non-linearity. This allows RICA to scale to over-complete basis and be optimized with backprop like sparse auto-encoders. RICA has also been shown to be more robust to non-whitened data, which is again more similar to auto-encoder behavior.

Self-Taught Learning

Self-Taught Learning (http://ufldl.stanford.edu/tutorial/selftaughtlearning/SelfTaughtLearning)

Exercise: Self-Taught Learning (http://ufldl.stanford.edu/tutorial/selftaughtlearning/ExerciseSelfTaughtLearning)