Stats 545: Homework 5

Due before class on Nov 10.
All plots should have labelled axes and titles.

Important:`Rcode`**, tables and figures should be part of a single .pdf or .html files from** `R Markdown` **and** `knitr`**. See the class reading lists for a short tutorial. Any derivations can also be in Markdown, in Latex or neatly written on paper which you can give to me.**

1 Gradient descent for blind source separation [100 pts]

Let $\mathbf{Y}$ be a $3 \times T$ matrix consisting of three independent audio recordings. Each column $\mathbf{y}_t = (y_{1t}, y_{2t}, y_{3t})^\top$ consists of the intensities of each source at time t. Even though each audio signal is a time-series, we will model the y_{ij}'s as i.i.d. draws from the hyperbolic secant distribution $p(y) = \frac{1}{\pi \cosh(y)}$.

1. Plot $p(y)$. Superpose a Gaussian with the same mean and variance. [3 pts]

Instead of observing $\mathbf{Y}$, we observe $\mathbf{X} = \mathbf{A} \cdot \mathbf{Y}$, where $\mathbf{A}$ is a 3×3 'mixing'-matrix. Given $\mathbf{X}$, we want to recover *both* $\mathbf{A}$ and $\mathbf{Y}$. Note this problem is *ill-posed* i.e. there are an infinite collection of valid $(\mathbf{A}, \mathbf{Y})$ pairs. However, our model of $\mathbf{Y}$ (in particular, the source-independence and non-Gaussianity assumptions) will allow us to recover a sensible solution.

Assume, $\mathbf{A}$ is invertible, and define $\mathbf{W} = \mathbf{A}^{-1}$. Write $\mathbf{w}^s$ for row s of $\mathbf{W}$.

2. Show that $\log p(\mathbf{X}|\mathbf{W}) = T \log |\mathbf{W}| + \sum_{t=1}^{T} \sum_{s=1}^{3} \log p(\mathbf{w}^s \mathbf{x}_t)$. (Hint: see e.g. the last paragraph here: http://sccn.ucsd.edu/wiki/Random_Variables_and_Probability_Density_Functions). [5 pts]

We will find the maximum likelihood (ML) estimate of $\mathbf{W}$ by gradient ascent.

3. Show that $\frac{\partial \log p(\mathbf{X}|\mathbf{W})}{\partial w_{ij}} = T a_{ji} + \sum_{t=1}^{T} \frac{\partial \log p(y_{it})}{\partial y_{it}} x_{jt}$ (Hint: look at Wikipedia for the derivative of a determinant). [5 pts]

   It follows that
$$\frac{1}{T} \frac{\partial \log p(\mathbf{X}|\mathbf{W})}{\partial \mathbf{W}} = \mathbf{A}^\top + \frac{1}{T} \sum_{t=1}^{T} \frac{\partial \log p(\mathbf{y}_t)}{\partial \mathbf{y}_t} \mathbf{x}_t^\top$$
.

   For our choice of $p(y)$, we have $\frac{\partial \log p(\mathbf{y}_t)}{\partial \mathbf{y}_t} = -\tanh(\mathbf{y}_t) := [-\tanh(y_{1t}), -\tanh(y_{2t}), -\tanh(y_{3t})]^\top$.
   The summation over T on the RHS can be written as $-\tanh(\mathbf{Y}) \cdot \mathbf{X}^\top$: this is just one line of `R` code.

4. Gradient ascent iteratively tries to reach a local maximum of the likelihood according to the rule
$$\mathbf{W}_{new} = \mathbf{W}_{old} + \eta \left(\frac{1}{T} \frac{\partial \log p(\mathbf{X}|\mathbf{W})}{\partial \mathbf{W}} \bigg|_{\mathbf{W}_{old}} \right)$$

   The last term is the gradient evaluated at $\mathbf{W}_{old}$. Plug in the expression for the gradient to write the update rule. [5 pts]

5. Write a function that takes a matrix $\mathbf{X}$ as input, and performs gradient ascent to get an ML-estimate of $\mathbf{W}$. Your function should have an initialization rule and a stopping criteria (which you will describe later). You will also need to set the learning rate η. You can get better results by using a smaller η for later iterations. [15 pts]

On the course webpage are three wav files, `mike1.wav, mike2.wav` and `mike3.wav`. Download them, and load them into `R` using `load.wave` from the `audio` package. Hopefully, you'll get three 490000-length vectors, which you can store into a matrix `X`. Directly running your function on this probably won't work (but you can try).

6. Write the 3×3 covariance matrix of `X`. Also, for each pair of recordings, plot a scatterplot of values (i.e. for rows (1,2), (2,3) and (1,3)). [10 pts]

7. If `Cov` is the covariance matrix, calculate its square root using the `sqrtm` function from the `expm` package. Call this `sqCov`. Define a new 'whitened' matrix $\mathbf{X}_{white} = \mathbf{sqCov}^{-1} \cdot \mathbf{X}$. This should have covariance equal to the identity matrix. Verify this. [10 pts]

8. Run your gradient ascent algorithm on the whitened data. How did you initialize $\mathbf{W}$? Describe how you set the learning rate η as well as your stopping criterion (I set η to about 1). How many iterations did you need? Plot the evolution of the log-likelihood. What is the returned estimate $\hat{\mathbf{W}}$? Your actual demixing matrix for the original data $\mathbf{X}$ is this value multiplied by $\mathbf{sqCov}^{-1}$. Print this out. If you had to restart with different initializations, describe this too. [30pts]

9. Plot histograms of the raw data, the whitened data and the recovered data (a total of 9 histograms) [9pts]

10. Your estimate of the latent sources is given by $\hat{\mathbf{Y}} = \hat{\mathbf{W}} \cdot \mathbf{sqCov}^{-1} \cdot \mathbf{X}$. Plot a histogram of the marginals for each source. What is their covariance? Plot the three pairwise scatterplots. [5pts]

You can verify your answer by listening to the recovered sources. You can save these using the `save.wave` command. You should normalize these first (I divided by twice the maximum value for each source).

11. Is the maximum likelihood estimate of $\mathbf{W}$ you obtained unique? If so explain why, if not, explain some sources of degeneracy. [2 pts]