Weekly Paper Summary (25 points total)

Paper Title	Solving Large-Scale Sparse PCA to Certifiable (Near) Optimality
Authors	Dimitris Bertsimas Ryan Cory-Wright Jean Pauphilet
Student Name	Alex Ojemann
Student ID	109722375

1. **What do you think the paper is about in layman's terms? What did the research focus on, what did the authors find and what are the main conclusions (if any) [5 points]**

Principal component analysis, or PCA, is a popular technique to reduce the dimensionality of a data set. For example, if a data set has six features, you may want to use PCA to get five or fewer principal components that explain the maximum possible variation in the data. It's very useful for data visualization for example because we can only look at data plots in two dimensions, three in rare cases. Sparse PCA is a form of PCA that doesn't have nonzero coefficients for all features like PCA usually does. This is useful because it's more interpretable when the number of features is very large and it can help avoid overfitting by only selecting the most important features. However, generating sparse principal components is difficult because many of the existing algorithms for doing so either don't provide proof of near optimality or can't execute in polynomial time. The authors propose methods that relax the assumptions of the problem so that it's easier to solve while still providing numerical proof of near optimality. Ultimately their methods maintain comparable performance to the more exact methods that are much more computationally expensive to solve.

2. **How would you extend the research paper – what new area(s) would you focus the paper on? [5 points]**

I would extend this paper by exploring how well these methods perform when selecting multiple principal components. The research the authors presented in this paper was exclusively focused on finding principal component 1, which is the most important because it usually explains much more of the variation in the data than the successive principal components, but limiting because two principal components are needed to visualize data and even more are potentially used to further learn from the data. The authors indicated that the assumption relaxations and cutting plane method should scale for numbers of principal components greater than one so this is a feasible extension of their research.

3. **Discuss at least two real-world applications (not mentioned in the paper) that would benefit from the focus of / applications mentioned in the paper and why [15 points]**

   One potential real world application of this research is in bioinformatics. Data sets in this field may contain thousands of genes or proteins and people may study how those contribute to certain conditions. Sparse PCA would be useful here because it can find principal components without using all the features which makes the principal components much more interpretable. These methods for reducing the computational cost of finding sparse components would be especially useful because of the high number of features. This research could also be useful in finance because financial datasets often have a large number of variables, such as stock prices, interest rates, and economic indicators. Sparse PCA can be useful in predicting stock prices based on these other indicators because sparse principal components can parse out which of them are the most important and the methods described in the paper can help substantially reduce computational cost without losing accuracy. This is especially useful because even a small increase in stock market prediction can have a huge impact on return on investment.