Problemset (1),
Foundations of Machine learning,
Winter 2024

Maximilian Kasy

In this problem, you are asked to implement some simulations and estimators in R, or in Python. Your code should run from start to end in one execution, producing all the output. Output and discussion of findings should be integrated in a report generated in R-Markdown, or from a Jupyter Notebook. Figures and tables should be clearly labeled and interpretable. The findings should be discussed in the context of the theoretical results that we derived in class.

You will calculate the risk functions (MSE) of various estimators in the normal means setting, using simulations. To do so,

1. Pick a random vector $\boldsymbol{\theta}_1$ of length 1 (you can pick random normally distributed components and then normalize),
2. take $\boldsymbol{\theta} = r \cdot \boldsymbol{\theta}_1$ for $r \in [0, 6]$,
3. repeatedly (say, 10,000 times) draw $\boldsymbol{X} \sim N(\boldsymbol{\theta}, I)$,
4. calculate estimates $\widehat{\boldsymbol{\theta}}$,
5. evaluate loss $\frac{1}{k}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2$,
6. average loss over simulation draws,
7. and plot average loss as function of r.

Do this separately for $k = \dim(\boldsymbol{\theta}) = 2, 3, 10$ and for the following estimators:

1. The MLE,
2. the estimator $\widehat{\boldsymbol{\theta}} = (1 - 1/\overline{X^2}) \cdot \boldsymbol{X}$,
3. the James-Stein estimator,
4. the positive part James-Stein estimator,
5. the estimator shrinking to the grand mean using the optimal shrinkage factor $1 - \frac{(k-3)/k}{s_X^2}$.

For a given dimension, plot the risk functions of all these estimators in one figure. Discuss your results.