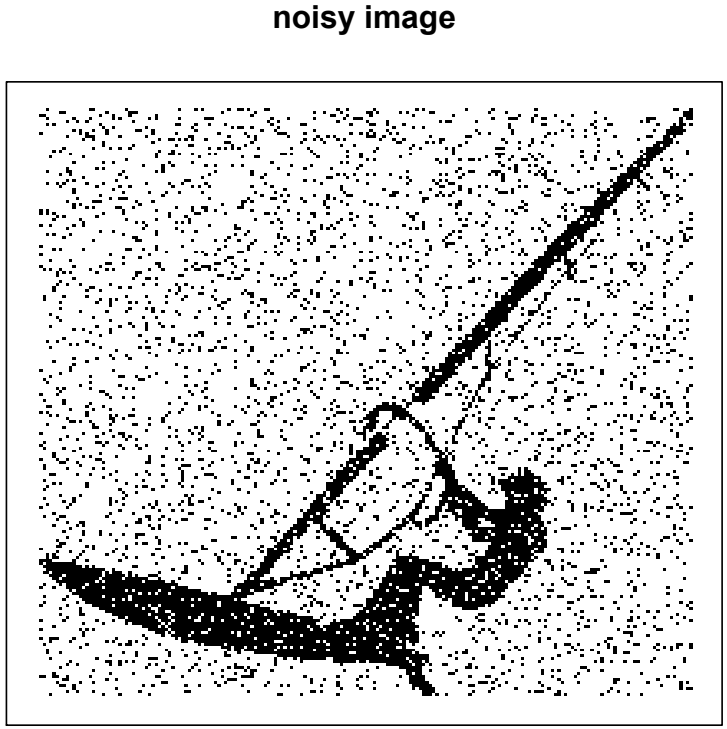
Asymptotic Statistics

Lecture 11

High-dimensional models

High-dimensional linear regression

Image with additive noise:



How to recover the image?

High-dimensional classification

Automatically classify 28×28 images.

Can use parametric logistic regression?

James-Stein theorem

Estimation in the normal means model

Observe $Y \sim N_n(\theta, I)$, $\theta \in \mathbb{R}^n$ unknown. Goal: estimate θ.

Immediately see that $\hat{\theta}_{MLE} = Y$.

Makes a lot of sense, right? So should we indeed use the MLE Y?

Estimation in the normal means model

Observe $Y \sim N_n(\theta, I)$, $\theta \in \mathbb{R}^n$ unknown. Goal: estimate θ.

Immediately see that $\hat{\theta}_{MLE} = Y$.

Makes a lot of sense, right? So should we indeed use the MLE Y?

$n = 1$: YES (see Exercise)

Estimation in the normal means model

Observe $Y \sim N_n(\theta, I)$, $\theta \in \mathbb{R}^n$ unknown. Goal: estimate θ.

Immediately see that $\hat{\theta}_{MLE} = Y$.

Makes a lot of sense, right? So should we indeed use the MLE Y?

$n = 1$: YES (see Exercise)
$n = 2$: YES (but more difficult to prove)

Estimation in the normal means model

Observe $Y \sim N_n(\theta, I)$, $\theta \in \mathbb{R}^n$ unknown. Goal: estimate θ.

Immediately see that $\hat{\theta}_{MLE} = Y$.

Makes a lot of sense, right? So should we indeed use the MLE Y?

$n = 1$: YES (see Exercise)
$n = 2$: YES (but more difficult to prove)

$n \geq 3$: NO!!!!!

You should not just use Y_i to estimate θ_i: Stein's phenomenon.

(Well, that is if you use mean squared error or something similar as measure of quality.)

James-Stein theorem - motivation - 1

Observe $Y \sim N_n(\theta, I)$, $\theta \in \mathbb{R}^n$ unknown. Goal: estimate θ.

For any estimator $\hat{\theta}$ with a finite covariance have bias-variance decomposition

$$E_\theta \|\hat{\theta} - \theta\|^2 = \|E_\theta \hat{\theta} - \theta\|^2 + \operatorname{tr} Cov_\theta \hat{\theta}.$$

(Here $\|\cdot\|$ is the Euclidean norm, $CovX = E(X - EX)(X - EX)^T$ is the covariance matrix of a random vector X and $\operatorname{tr} A$ is the trace of the matrix A.)

James-Stein theorem - motivation - 2

For $\hat{\theta}_c = cY$, have

$$E_\theta \|cY - \theta\|^2 = (c-1)^2 \|\theta\|^2 + c^2 n.$$

Mininimal for

$$c = c_\theta = \frac{\|\theta\|^2}{\|\theta\|^2 + n},$$

minimal value is

$$E_\theta \|\hat{\theta}_{c_\theta} - \theta\|^2 = \frac{n\|\theta\|^2}{\|\theta\|^2 + n} = \frac{\|\theta\|^2}{\|\theta\|^2 + n} E_\theta \|Y - \theta\|^2.$$

Factor is less than 1, but depends on θ.

James-Stein theorem

Theorem. (James-Stein)

Define

$$\hat{\theta}_{\mathrm{JS}} = \left(1 - \frac{n-2}{\|Y\|^2}\right) Y.$$

For $n \geq 3$, we have $E_\theta \|\hat{\theta}_{\mathrm{JS}} - \theta\|^2 < E_\theta \|\hat{\theta}_{\mathrm{MLE}} - \theta\|^2$ for all $\theta \in \mathbb{R}^n$.

Proof.

Compute bias and variance for every component. Use JS Lemma.

(See Exercises) □

James-Stein theorem - comments - 1

JS estimator has different bias and variance compared to the MLE:

- The JS estimator shrinks the MLE to 0.
- This reduces the variance, at the cost of increasing the bias.
- The net effect is that the MSE, or risk, is smaller.

James-Stein theorem - comments - 2

Improved bias-variance trade-off by reducing effect of outliers:

- Shrinking factor depends on all observations. So to estimate θ_i don't just use Y_i, but also borrow strength from other observations.
- The degree of shrinkage depends on how large $\|Y\|^2$ is compared to n. $\|Y\|^2 \gg n$ indicates there are outliers. Shrinking reduces terms in $\|\hat{\theta}_{\mathsf{JS}} - \theta\|^2$ corresponding to outliers, possibly at the cost of increasing the remaining terms. Net effect is that squared error improves on average.
- Essential that MSE simultaneously takes all coordinates of θ into account. Allows us to trade off gains in one coordinate with losses in others.

James-Stein theorem - comments - 3

Many generalizations possible:

- Away from $Y \sim N_n(\theta, I)$.
- Other norms
- Shrinking to a point other than 0 (Exercise)

General message:

In high-dimensional/nonparametric problems:

Advantageous to reduce variance by shrinking, or otherwise regularising.

James-Stein theorem - comments - 4

JS theorem implies the MLE in this model is inadmissible:

There exists another estimator $\hat{\theta}$ such that
$E_\theta\|\hat{\theta} - \theta\|^2 \leq E_\theta\|\hat{\theta}_{\text{MLE}} - \theta\|^2$ for all $\theta \in \mathbb{R}^n$, with strict inequality for at least one $\theta \in \mathbb{R}^n$.

James-Stein theorem - comments - 4

JS theorem implies the MLE in this model is inadmissible:

There exists another estimator $\hat{\theta}$ such that
$E_\theta\|\hat{\theta} - \theta\|^2 \leq E_\theta\|\hat{\theta}_{\text{MLE}} - \theta\|^2$ for all $\theta \in \mathbb{R}^n$, with strict inequality for at least one $\theta \in \mathbb{R}^n$.

The JS estimator is inadmissble as well:

$$\hat{\theta}_{\text{JS}+} = \left(1 - \frac{n-2}{\|Y\|^2}\right)_+ Y$$

is an estimator with strictly smaller risk for all $\theta \in \mathbb{R}^n$.

James-Stein theorem - comments - 4

JS theorem implies the MLE in this model is inadmissible:

There exists another estimator $\hat{\theta}$ such that
$E_\theta\|\hat{\theta} - \theta\|^2 \leq E_\theta\|\hat{\theta}_{\text{MLE}} - \theta\|^2$ for all $\theta \in \mathbb{R}^n$, with strict inequality for at least one $\theta \in \mathbb{R}^n$.

The JS estimator is inadmissble as well:

$$\hat{\theta}_{\text{JS}+} = \left(1 - \frac{n-2}{\|Y\|^2}\right)_+ Y$$

is an estimator with strictly smaller risk for all $\theta \in \mathbb{R}^n$.

$\hat{\theta}_{\text{JS}+}$ is inadmissble as well...

Message from the JS theorem

In high-dimensional/nonparametric problems:

Advantageous to reduce variance by shrinking, or otherwise regularising.