Advanced Regression: 3a Principles of high-dimensional data analysis

Verena Zuber

Epidemiology and Biostatistics, Imperial College London

31st January 2019

Principles of high-dimensional data analysis

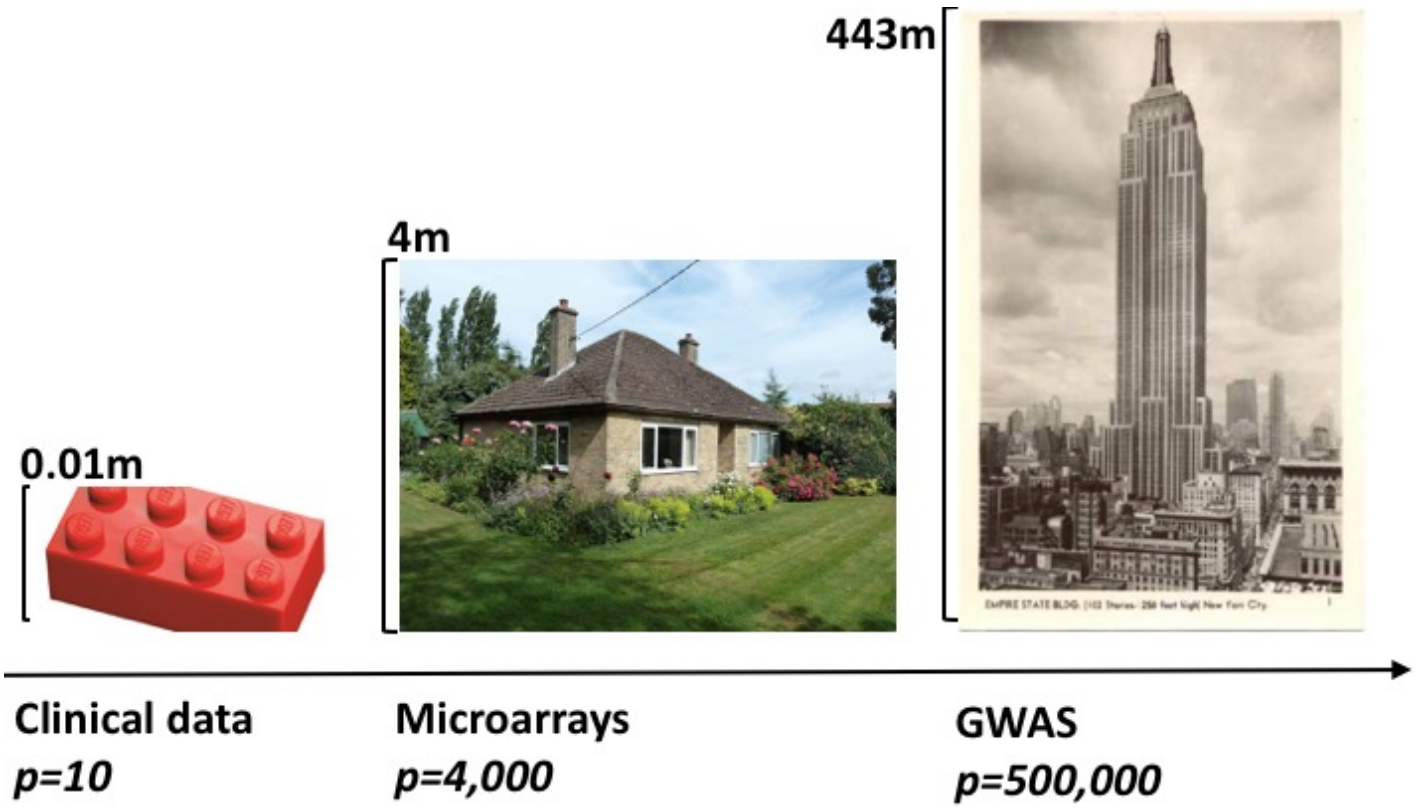
- Measuring the quality of fit
- Bias-variance trade-off in prediction
- Overfitting

Shrinkage estimation

- Bias-variance trade-off in estimation
- Stein's paradox
- A general form for the shrinkage estimate
- Comparing 2-groups: Shrinkage t-score
- Shrinkage t-score in practice
- Estimation of high-dimensional covariance matrices
- Shrinkage covariance estimation in practice

Principles of high-dimensional data analysis

- Classical statistical methods like linear models, glms or maximum likelihood estimates were designed for data-sets with more observations than variables $n > p$.
- They were not designed for handling many variables.



Principles of high-dimensional data analysis

Curse of dimensionality

Covered so far:

- Ill-conditioned $x^t x$ matrix:
  - Singular: Computation of OLS estimate not possible.
  - Multi-collinearity: Distorted results with high variability.
- Multiple testing: Number of tests acts as a penalty.

Concepts we cover today:

- Bias-variance trade-off in prediction
- Cross-validation (Lecture 3c)
- Shrinkage estimates
- Regularised regression (Lecture 3b)

Training and test data

Aim of prediction

Define a prediction rule that is accurate, but also generalises to new data.

When performing prediction we split the data into the following three subsets:

- Training data to fit the models.
- (Validation data to estimate extra parameters of the prediction rule.)
- Test data to assess the generalization properties.

Training and test data

Assume we have the following data:

- Training data
  - y_i, where $i \in 1, ..., n$.
  - x_i, vector of $j \in 1, ..., p$ predictors for observation i
- Test data
  - y_k^{Test}, where $k \in 1, ..., m$.
  - x_k^{Test}, vector of $j \in 1, ..., p$ predictors for observation k
- We assume the following general model to hold for both training and test data

$$y = f(x) + \epsilon$$

Measuring the quality of fit

1. Based on the training data we build a prediction rule $\hat{f}(x)$

$$\hat{y}_i = \hat{f}(x_i).$$

For example the ordinary least squares prediction rule is defined as

$$\hat{y} = hy = x(x^t x)^{-1} x^t y = x\beta.$$

2. We evaluate the prediction rule $\hat{f}(x)$ (derived from the training data) on the test data x_k^{Test} and obtain the prediction $\hat{y}_k^{Test}$

$$\hat{y}_k^{Test} = \hat{f}(x_k^{Test}).$$

Mean squared error (MSE)

- It is easy to derive the MSE on the training data, this is equivalent to the residual sum of squares.
- The residual sum of squares do not tell us how well the prediction rule generalises to new data, the test data.

MSE evaluated on the test data

$$MSE = \frac{1}{m}\sum_{k=1}^{m}\left(\underbrace{y_k^{Test}}_{\text{Observed}} - \underbrace{\hat{f}(x_k^{Test})}_{\text{Predicted}}\right)^2$$

Decomposition into bias and variance

$$
\begin{aligned}
E(MSE) &= E\left(\frac{1}{m}\sum_{k=1}^{m}\left(y_k^{Test} - \hat{f}(x_k^{Test})\right)^2\right) \\
&= \underbrace{\sigma^2}_{\text{Noise}} + \underbrace{E[\hat{f}(x_k^{Test} - E(\hat{f}(x_k^{Test}))]^2}_{\text{Variance}} \\
&\quad + \underbrace{[E(\hat{f}(x_k^{Test})) - \hat{f}(x_k^{Test})]^2}_{\text{Bias}^2}
\end{aligned}
$$

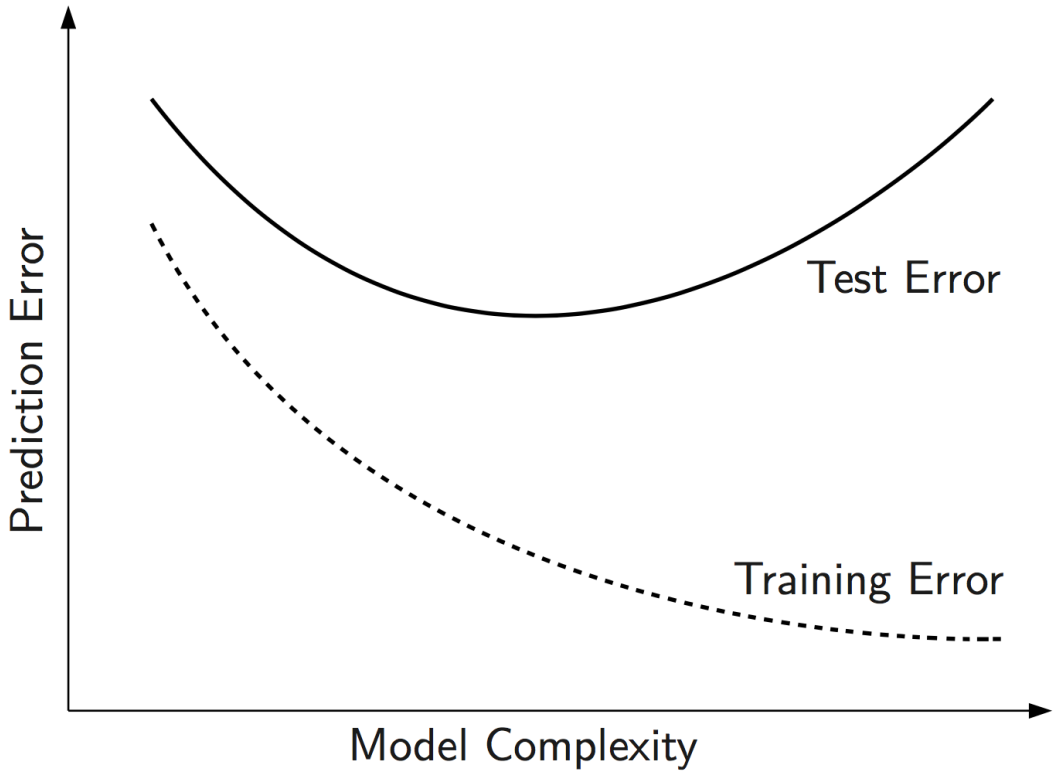
- Noise or irreducible error σ^2
- Variance $E[\hat{f}(x_k^{Test} - E(\hat{f}(x_k^{Test}))]^2$
- Bias $[E(\hat{f}(x_k^{Test})) - \hat{f}(x_k^{Test})]$

Bias-variance trade-off in prediction

.

- **Bias**: The error that is introduced by fitting the model.
  - → More variables reduce the residual sum of squares.
  - → We can reduce bias by adding more variables (higher complexity).
- **Variance**: The amount by which $\hat{f}(x)$ would change if we estimated it using a different training data set.
  - → The more variables we include, the more likely $\hat{f}(x)$ will differ for a new training data
  - → We can reduce variance by removing variables (lower complexity).

Overfitting



- Training MSE: We can always reduce the MSE by adding more variables (higher complexity).
- Test MSE: After the model is saturated, we will increase the MSE by adding more variables.

The problem of overfitting

Overfitting the data

- Complex models may be too precise and tailored only to the specific data used as training data.
- They follow the error or noise too closely.
- Complex models may provide perfect fit and very low MSE on the training data.
- But when used to build a prediction rule for new data they will have a high MSE on the test data.
- Thus, they do not generalise well to new data and do not provide a good prediction rule.

Measuring the quality of fit: Binary outcome

- Quantitative outcomes: Mean squared error (MSE)
- Binary outcome:
  - Misclassification error rate: Proportion of misclassified observations
  - Positive predictive value (PPV)

$$\begin{aligned} \text{PPV} &= \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false positives}} \\ &= \frac{\text{Number of true positives}}{\text{Number of positive calls}} \end{aligned}$$

Bias-variance trade-off in estimation

- Consider an estimate $\hat{\theta}$ for a parameter θ.
- Examples:
  - Sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ for the population mean.
  - Sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ for the population variance.

Bias of an estimate

$$Bias(\hat{\theta}) = E\left(\hat{\theta} - \theta\right)$$

Bias-variance trade-off in estimation

Mean squared error (MSE) of an estimate $\hat{\theta}$

MSE is the squared average difference between an estimate $\hat{\theta}$ and the true parameter θ.

$$
\begin{aligned}
MSE(\hat{\theta}) &= E\left(\hat{\theta} - \theta\right)^2 \\
&= (E(\hat{\theta}^2) - 2E(\hat{\theta})\theta + \theta^2) \\
&= (E(\hat{\theta}^2) - \underbrace{E(\hat{\theta})^2 + E(\hat{\theta})^2}_{0} - 2E(\hat{\theta})\theta + \theta^2) \\
&= (E(\hat{\theta}^2) - E(\hat{\theta})^2) + (E(\hat{\theta})^2 - 2E(\hat{\theta})\theta + \theta^2) \\
&= Var(\hat{\theta}) + \left(Bias(\hat{\theta})\right)^2
\end{aligned}
$$

Bias-variance trade-off in estimation

Also in estimation there is a trade-off between bias and variance

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \Big(Bias(\hat{\theta})\Big)^2$$

where

- $Var(\hat{\theta}) = E(\hat{\theta}^2) - E(\hat{\theta})^2$
- $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta = E(\hat{\theta} - \theta)$

- Many classical techniques are designed to be unbiased (BLUE) or consistent (Maximum Likelihood).
- Shrinkage estimates trade unbiasedness against a much reduced variance.

Maximum likelihood estimate

- Assume we have observed one observation $i = 1$ for p variables of which each follows a Normal distribution

$$\begin{aligned} x_1 &\sim N(\theta_1, 1) \\ x_j &\sim N(\theta_j, 1) \\ x_p &\sim N(\theta_p, 1) \end{aligned}$$

- The unknown parameters are θ_1 to θ_p.
- The Maximum Likelihood estimate is

$$\begin{aligned} \hat{\theta}_1 &= x_1 \\ \hat{\theta}_j &= x_j \\ \hat{\theta}_p &= x_p \end{aligned}$$

- Maximum Likelihood estimates are consistent, but not generally unbiased.
- In this particular case $\hat{\theta}$ is unbiased.

Stein's paradox (1955)

When $p \geq 3$

the Maximum Likelihood is suboptimal. There exists a different estimator with lower MSE than the Maximum Likelihood estimate.

- James-Stein estimate (1960) for the population mean of variable $j \in 1, ..., p$ with variance s^2

$$\hat{\theta}_j^{JS} = \underbrace{(1 - \frac{(p-2)}{\sum_{j=1}^{p} x_j^2})}_{c} x_j,$$

- The James-Stein estimate has lower MSE than the Maximum Likelihood estimator.

James-Stein estimate

$$\hat{\theta}_j^{JS} = \underbrace{(1 - \frac{(p-2)}{\sum_{j=1}^{p} x_j^2})}_{c} x_j$$

- When $\sum_{j=1}^{p} x_j^2$ is small, there is minimal heterogeneity ($c = 0$) $\rightarrow$ this supports the implicit assumption that all $\theta_1, ..., \theta_p$ might be similar (Target $t_k = 0$).
- When $\sum_{j=1}^{p} x_j^2$ is large, there is a huge heterogeneity among the p variables ($c = 1$)
  $\rightarrow$ there is no evidence for a common target, so it is better not to shrink.

$$\hat{\theta}_j^{JS} = \begin{cases} 0 & \text{if } c = 0 \\ x_j & \text{if } c = 1 \end{cases}$$

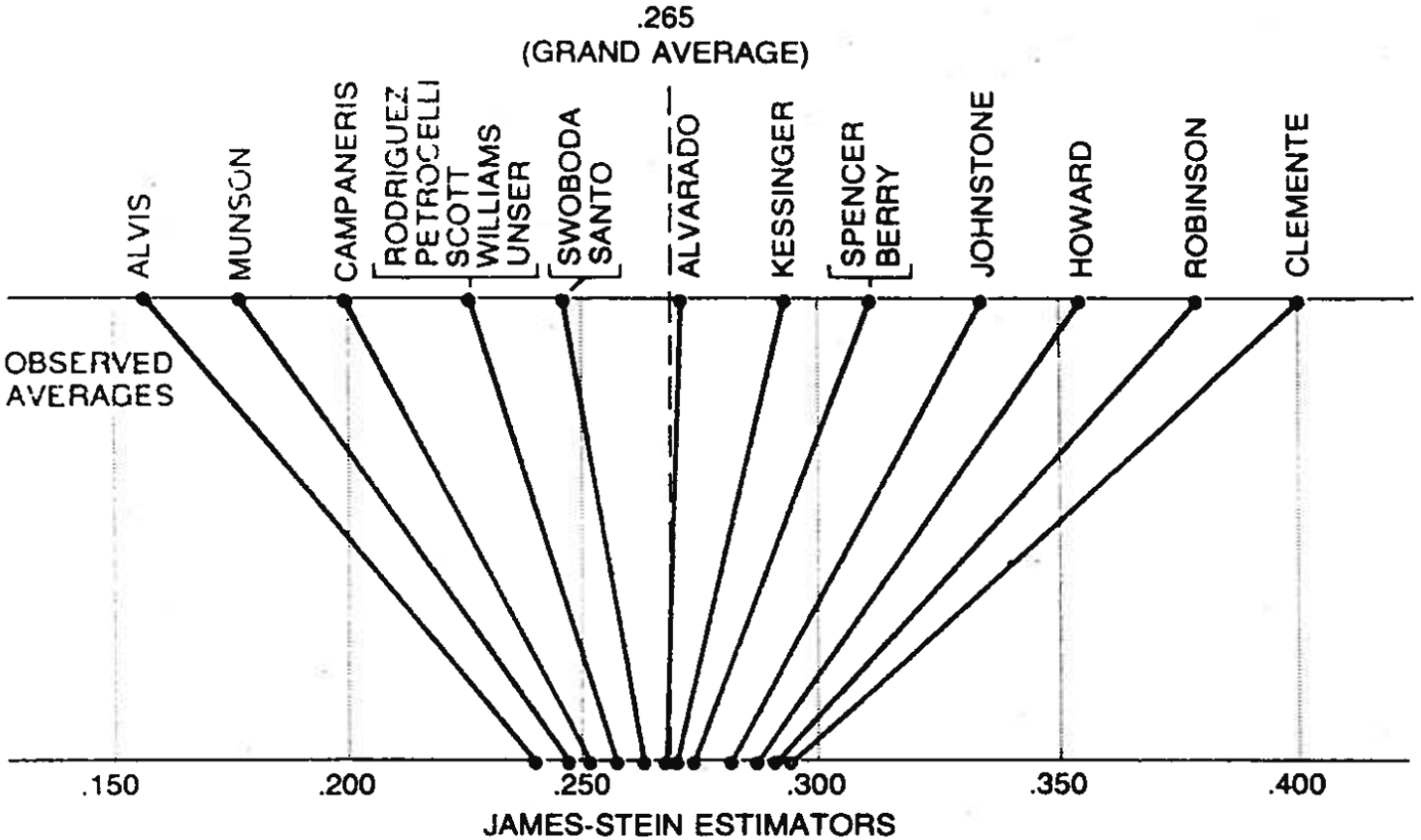
Efron-Morris estimate

$$\hat{\theta}_j^{EM} = \bar{x} \underbrace{(1 - \frac{(p-3)}{\sum_{j=1}^{p} x_j^2})}_{c}(x_j - \bar{x}),$$

- where $\bar{x} = \frac{1}{p}\sum_{j=1}^{p}$ grand mean over all variables
- $c \to 0$ when $\sum_{j=1}^{p} x_j^2$ is small (Shrinkage to 0)
- $c \to 1$ when $\sum_{j=1}^{p} x_j^2$ is large (No shrinkage)
- If $c = 0$ there is an implicit assumption that all $\theta_1, ..., \theta_p$ might be similar (Target $t_k = \bar{x}$).

$$\hat{\theta}_j^{EM} = \begin{cases} \bar{x} + 0 = \bar{x} & \text{if } c = 0 \\ \bar{x} + x_j - \bar{x} = x_j & \text{if } c = 1 \end{cases}$$

Efron-Morris estimate



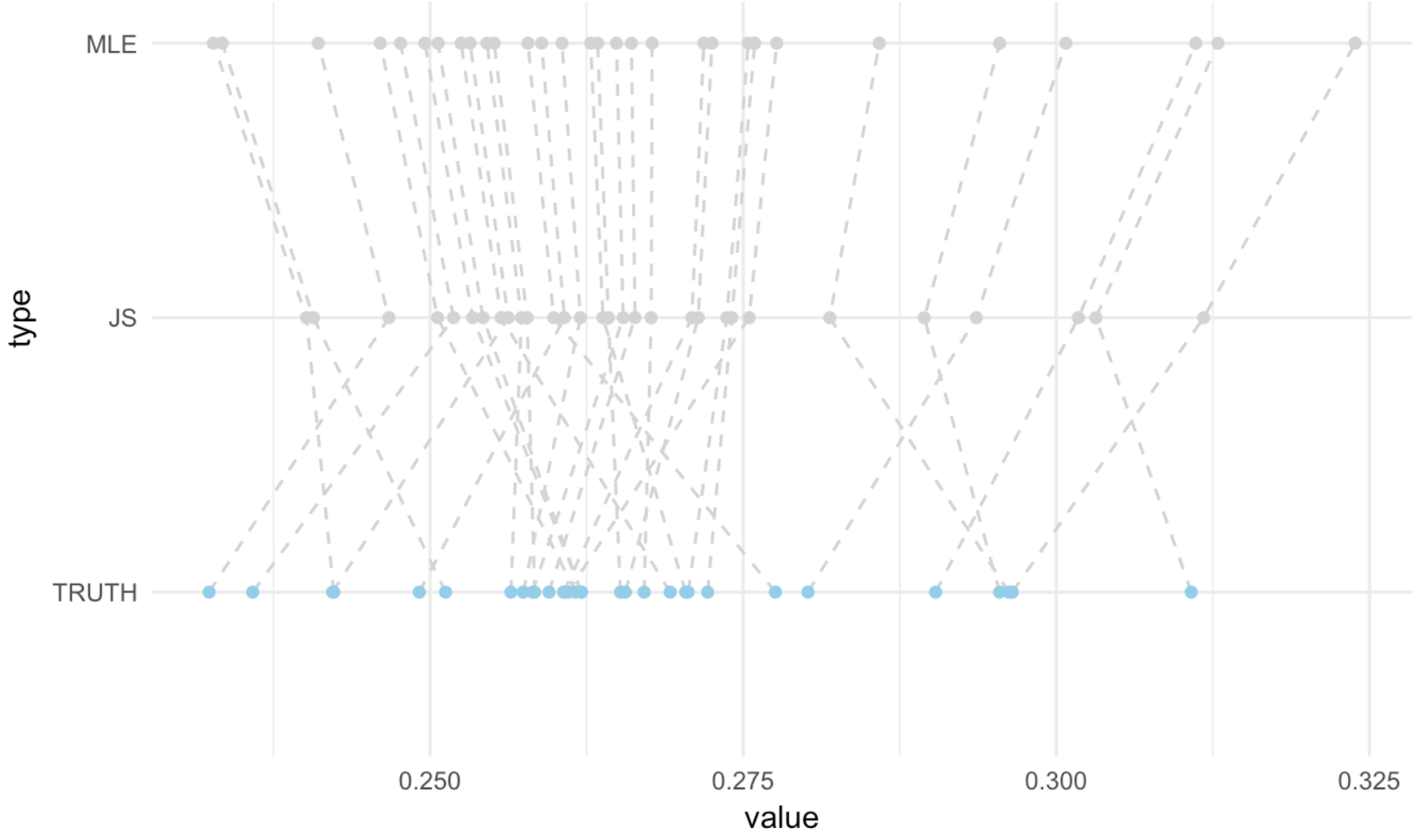
JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value.

Efron and Morris 1977

Efron and Morris 1977

The best guess about the future is usually obtained by computing the average of past events.

Efron-Morris estimate



https://bookdown.org/content/927/james-stein.html

Shrinkage estimate

Shrinkage estimate $\hat{\theta}^{\text{shrink}}$ for an unknown parameter vector θ of length p is based only on three components

$$\hat{\theta}^{\text{shrink}} = (1 - \lambda)\hat{\theta} + \lambda\theta^{\text{target}}$$

with

- $\hat{\theta}$ as an unregularized estimate for θ,
- θ^{target} as the target, and
- λ shrinkage parameter $[0, 1]$

Bias-variance trade-off

- The unregularised estimate $\hat{\theta}$ has no or little bias, but large variance.
- The target θ^{target} has large bias, but no or little variance.

Shrinkage parameter

- Choose the shrinkage parameter $\lambda^\star$ to minimise the MSE.
- The optimal shrinkage is avaialable in closed form

$$\lambda^\star = \frac{\sum_{j=1}^{p} var(\hat{\theta}_j) + cov(\hat{\theta}_j, \theta_j^{\text{target}})}{\sum_{j=1}^{p} E(\hat{\theta}_j - \theta_j^{\text{target}})^2}.$$

Interpretation: $\lambda^\star$ gets smaller

- the smaller the variance of the unregularised estimate $var(\hat{\theta}_j)$.
- the larger the mean squared difference between $\hat{\theta}_j$ and θ_j^{target}.

2-groups or binary outcome: t-test

Differential effects

For each variable x_j, where $j \in 1, \ldots p$ we fit a t-test t_j

$$t_j = \frac{\mu_j(1) - \mu_j(2)}{s_j^2}$$

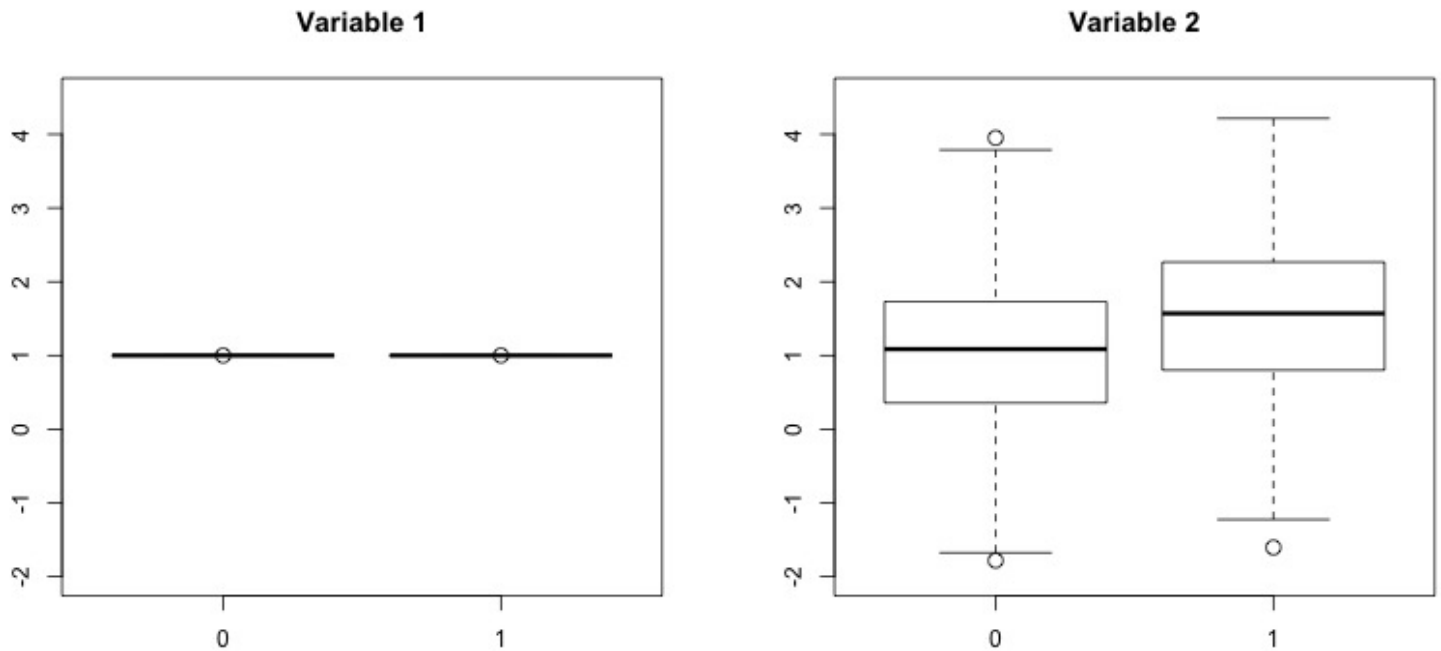
- $\mu_j(1)$: mean of x_j in group 1
- $\mu_j(2)$: mean of x_j in group 2
- $\mu_j(1) - \mu_j(2)$: fold change
- s_j^2 sample variance of the fold change, a weighted mixture of the sample variance for variable j in group 1 and group 2

$$s_j^2 = \frac{1}{n_1 + n_2 - 2}((n_1 - 1)var(x_1(j)) + (n_2 - 1)var(x_2(j)))$$

Shrinkage t-score

Variables with no heterogeneity can have an inflated t-score

- No heterogeneity means small denominator $s_j \to 0$, thus t_j will be inflated irrespective of the fold change.



- Variable 1 has a t-score of 15, variable 2 of 7.

Shrinkage t-score

Shrinkage variance estimate s^2_{shrink} of length p is based only on three components

$$s^2_{\text{shrink}} = (1 - \lambda)s^2 + \lambda s^2_{\text{median}}$$

with

- s^2 is the sample variance,
- s^2_{median} is the median over $(s^2_1, ..., s^2_p)$ as the target,
- λ shrinkage parameter $[0, 1]$

$$\lambda^\star = \frac{\sum_{j=1}^p var(s^2_j)}{\sum_{j=1}^p E(s^2_j - s^2_{\text{median}})^2}.$$

Example: Breast cancer data

- y: benign or aggressive tumour (binary)

Benign	Aggressive	Total
185	121	306

- x: gene expression of $p = 22,283$ genes
- $n = 306$: sample size
- Truly big data $n << p$
- Data taken from Hatzis et al 2011 `https://jamanetwork.com/journals/jama/fullarticle/899864`

Shrinkage *t*-score st()

The `st` package contains the following *t*-statistics:

- Fold-change: `diffmean.stat()`
- Standard *t*-score: `studentt.stat()`
- Shrinkage *t*-score: `shrinkt.stat()`

Arguments:

- Data matrix `X`
- Binary indicator as factor `L`
- Shrinkage parameter `lambda.var`, if not specified this is estimated from the data

Shrinkage *t*-score shrinkt.stat()

```
> shrinkt.out = shrinkt.stat(X=as.matrix(x), L=severity)
Number of variables: 22283
Number of observations: 306
Number of classes: 2

Estimating optimal shrinkage intensity lambda.freq (frequencies): 0.0717
Estimating variances (pooled across classes)
Estimating optimal shrinkage intensity lambda.var (variance vector): 0.0168

>
> output=cbind(colnames(x),shrinkt.out)
> colnames(output) = c("gene","shrink t")
> output=output[order(abs(shrinkt.out), decreasing=TRUE),]
> head(output,n=5)
     gene          shrink t
[1,] "200883_at"   "4.54773598774052"
[2,] "205780_at"   "4.53359329211625"
[3,] "201376_s_at" "4.53041589963766"
[4,] "219569_s_at" "-4.52783295373426"
[5,] "202731_at"   "4.48689261674946"
```

Shrinkage covariance matrix

Honey, I Shrunk the Sample Covariance Matrix

Olivier Ledoit
Equities Division
Credit Suisse First Boston
One Cabot Square
London E14 4QJ, UK
olivier@ledoit.net

Michael Wolf*
Department of Economics and Business
Universitat Pompeu Fabra
Ramon Trias Fargas, 25–27
08005 Barcelona, Spain
michael.wolf@upf.edu

November 2003

Shrinkage covariance matrix

Shrinkage covariance matrix Σ_{shrink} of dimension $p \times p$ is based only on three components

$$\Sigma_{\text{shrink}} = (1 - \lambda)\Sigma + \lambda diag(s^2)$$

with

- Σ is the sample covariance matrix,
- $diag(s_j^2)$ is the $p \times p$ diagonal matrix of $(s_1^2, ..., s_p^2)$ as the target,
- λ shrinkage parameter $[0, 1]$

The shrinkage covariance matrix is well-conditioned and invertible.

Shrinkage partial correlation matrix

- Gaussian graphical models define sparse networks.
- Interpretation: Association between variable i and j conditional on all other variables.
- Gaussian graphical models can be defined using partial correlations

$$\Omega = P^{-1} = \omega_{ij}$$

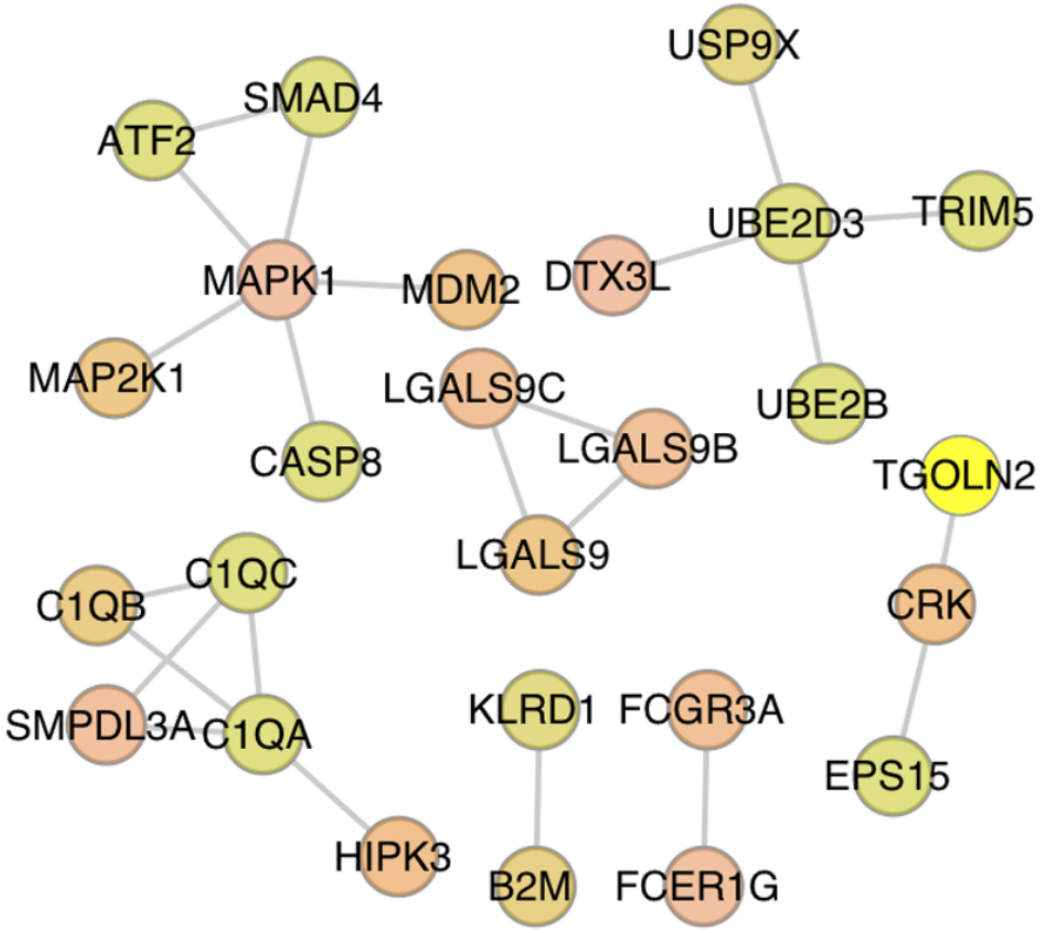
with

- P correlation matrix of dimension $p \times p$
- Ω inverse correlation matrix of dimension $p \times p$

Partial correlation τ_{ij}

$$\tau_{ij} = -\omega_{ij} / \sqrt{(\omega_{ii}\omega_{jj})}$$

Gaussian graphical models



Xiao et al 2014

Shrinkage covariance matrix in R

The `corpcor` package implements algorithms for the estimation of:

- Correlation matrix: `cor.shrink()`
- Partial correlation matrix: `pcor.shrink()`
- Covariance matrix: `cov.shrink()`

Argument:

- Data matrix `X`

```
> corr_mat = cor.shrink(as.matrix(x))
Estimating optimal shrinkage intensity lambda (correlation matrix): 0.1862

>
> dim(corr_mat)
[1] 22283 22283
```

Take away: High-dimensional data analysis

- When performing prediction there is an intrinsic trade-off between bias and variance:
  The higher the complexity of our model, the lower the bias in our training data, but the higher the variance in the test data.
- Overfitting refers to models that fit too precisely the training data (low training error) and do not generalise well to new observations (high test error).
- Shrinkage estimates trade bias for a reduced variance.

$$\hat{\theta}^{\text{shrink}} = (1 - \lambda)\hat{\theta} + \lambda\theta^{\text{target}}$$