Introduction to Empirical Bayes

Can Yang

Department of Mathematics
The Hong Kong University of Science and Technology

Spring, 2018

Charles M. Stein

- Charles M. Stein (March 22, 1920 - November 24, 2016), an American mathematical statistician, is emeritus professor of statistics at Stanford University. He received his Ph.D in 1947 at Columbia University with advisor Abraham Wald. He is known for Stein's paradox in decision theory which shows that ordinary least squares estimates can be uniformly improved when many parameters are estimated, and for Stein's method, a way of proving theorems such as the Central Limit Theorem.
- See `http://en.wikipedia.org/wiki/Charles_Stein_(statistician)`.

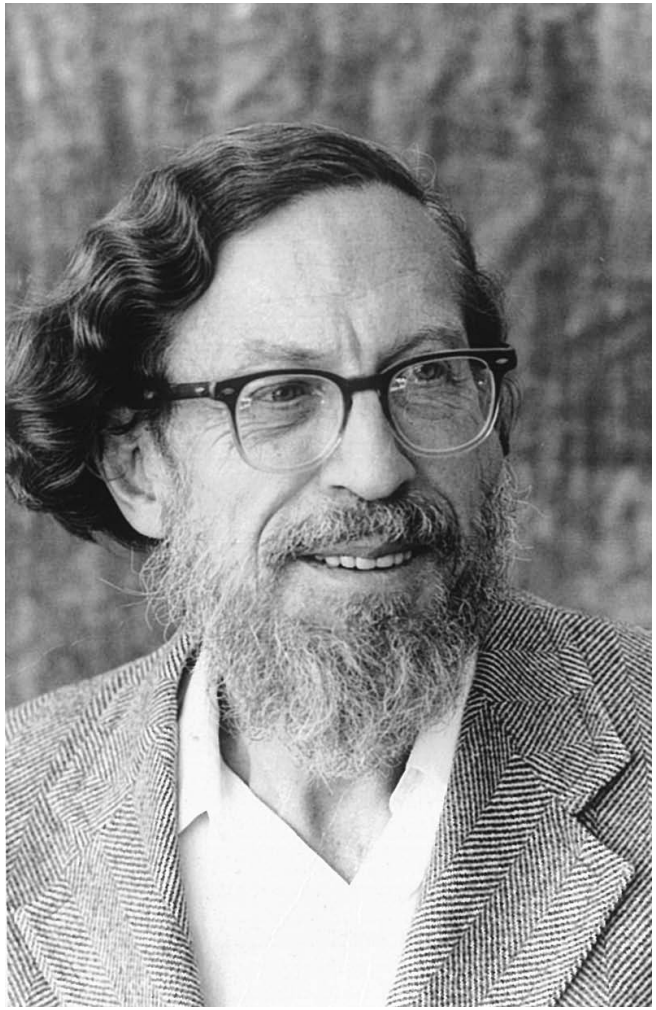


Figure 1: Charles M. Stein.

James-Stein Estimator

The problem

$z_i|\mu_i \sim \mathcal{N}(\mu_i, 1)$, $i = 1, \ldots, N.$, how to obtain a good estimate of $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_N]^T$ from $\mathbf{z} = [z_1, \ldots, z_N]^T$?

- The MLE is $\boldsymbol{\mu}_{MLE} = \mathbf{z}$.
- James-Stein Estimation is

$$\boldsymbol{\mu}_{JS} = \left(1 - \frac{N-2}{||\mathbf{z}||^2}\right)\mathbf{z}. \tag{1}$$

- James and Stein have shown that $\boldsymbol{\mu}_{JS}$ is always better than $\boldsymbol{\mu}_{LS}$ in terms of minimizing $E\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2\right)$ when $N \geq 3$. A proof can be found in Chapter 2 of (Efron, 2010).

B. Efron

Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction

Cambridge University Press, 2010.

MLE vs. JSE

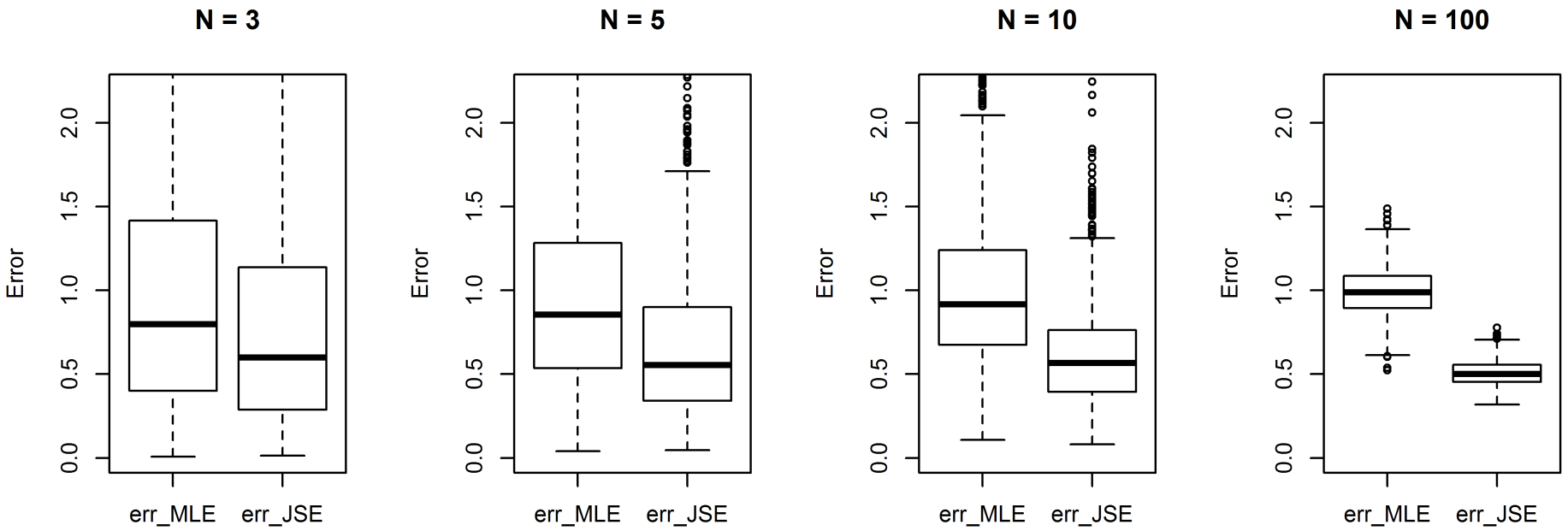


Figure 2: Comparison between MLE and JSE, $N = 3, 5, 10, 100$, where $\mu_i \sim \mathcal{N}(0, 1)$.

Derivation of James-Stein Estimator

- Known: $z_i|\mu_i \sim \mathcal{N}(\mu_i, 1)$, $i = 1, \ldots, N$, i.e., the conditional probability $f(z_i|\mu_i)$).
- Suppose we have prior distribution $g(\mu)$: $\mu_i \sim \mathcal{N}(0, \sigma^2), i = 1, \ldots, N$.
- How to get the posterior mean $\mathbb{E}(\mu_i|z_i)$?
- The idea is using Bayes rule as follows:
  - The marginal distribution of z_i: $z_i \sim \mathcal{N}(0, 1 + \sigma^2)$
  - The posterior distribution:

$$\mu_i|z_i \sim \mathcal{N}\left((1 - \frac{1}{1+\sigma^2})z_i, \frac{\sigma^2}{1+\sigma^2}\right). \tag{2}$$

.

- The posterior mean

$$\mathbb{E}(\mu_i|z_i) = \left(1 - \frac{1}{\sigma^2 + 1}\right) z_i. \tag{3}$$

The property of Gaussian distribution you should know

Marginal and Conditional Gaussians

Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.113}$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{2.114}$$

the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \tag{2.115}$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \tag{2.116}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}. \tag{2.117}$$

Figure 3: Bishop (2006). Pattern recognition and machine learning

Empirical Bayes estimates

We don't know the value of $1/(\sigma^2+1)$, we estimate it in the following way:

- We have known $z_i \sim \mathcal{N}(0, 1+\sigma^2)$.
- We have

$$Q = \Sigma_{i=1}^{N} \frac{z_i^2}{1+\sigma^2} \sim \chi_N^2, \tag{4}$$

  where χ_N^2 is the chi-square distribution with N degrees of freedom.
- Note that $Q \sim \chi_N^2$, $1/Q$ follows Inverse-χ^2 with $df = N$, and $\mathbb{E}(1/Q) = \frac{1}{N-2}$.
- Therefore, $\mathbb{E}(\frac{1}{\Sigma_{i=1}^{N} \frac{z_i^2}{1+\sigma^2}}) = \frac{1}{N-2}$, and then we use $\frac{N-2}{\sum_{i=1}^{N} z_i^2}$ as an estimate of $\frac{1}{1+\sigma^2}$.
- Now we obtain the James-Stein Estimator (1).

Exercise

Consider the gamma function

$$\Gamma(x) \equiv \int_0^\infty u^{x-1} e^{-u} du, \tag{5}$$

which is widely involved in probability density function, such as the Beta distribution, the student t distribution, and the χ^2 distribution.

- Using integration by parts, prove the relation

$$\Gamma(x+1) = x\Gamma(x). \tag{6}$$

  Also show that $\Gamma(1) = 1$ and hence that $\Gamma(x+1) = x!$ when x is an integer.
- Calculate the expectation of the random variable $1/x$, where x is a random variable with chi-square distribution χ^2_ν, where ν is the degree of freedom. Note that this property is used in derivation of the James-stein estimator.

Proof

- Using integration by parts we have

$$\begin{aligned}\Gamma(x+1) &= \int_0^\infty u^x e^{-u} du \\ &= [-e^{-u}u^x]_0^\infty + \int_0^\infty x u^{x-1} e^{-u} du = 0 + x\Gamma(x).\end{aligned} \tag{7}$$

For $x = 1$ we have

$$\Gamma(1) = \int_0^\infty e^{-u} du = [-e^{-u}]_0^\infty = 1. \tag{8}$$

If x is an integer we can apply proof by induction to relate the gamma function to the factorial function. Suppose that $\Gamma(x+1) = x!$ holds. Then from the result (7) we have $\Gamma(x+2) = (x+1)\Gamma(x+1) = (x+1)!$. Finally, $\Gamma(1) = 1 = 0!$, which completes the proof by induction.

Proof

- The probability density function (pdf) of the χ^2_ν distribution is

$$\frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}. \tag{9}$$

So we can write this problem as

$$\begin{aligned} & \int_0^\infty \frac{1}{x} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} dx \\ = & \int_0^\infty \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-2} e^{-x/2} dx \\ = & \int_0^\infty \left[\frac{1}{2^{(\nu-2)/2}\Gamma((\nu-2)/2)} x^{(\nu-2)/2-1} e^{-x/2} \right] (v-2) dx \\ = & \, v-2. \end{aligned} \tag{10}$$

For the second equality, we used $\Gamma(\frac{\nu}{2}) = \left(\frac{\nu}{2}-1\right)\Gamma(\frac{\nu}{2}-1)$.

Take-home-message

- Empirical Bayes methods are procedures for statistical inference in which the prior distribution is estimated from the data (quoted from wiki).
- Learning from the experience of others: EB allows incorporating indirect information into the prior distribution and adaptively updating its parameters.
- Statistics has a set of rules to share information and account for uncertainties in a statistically optimal way.

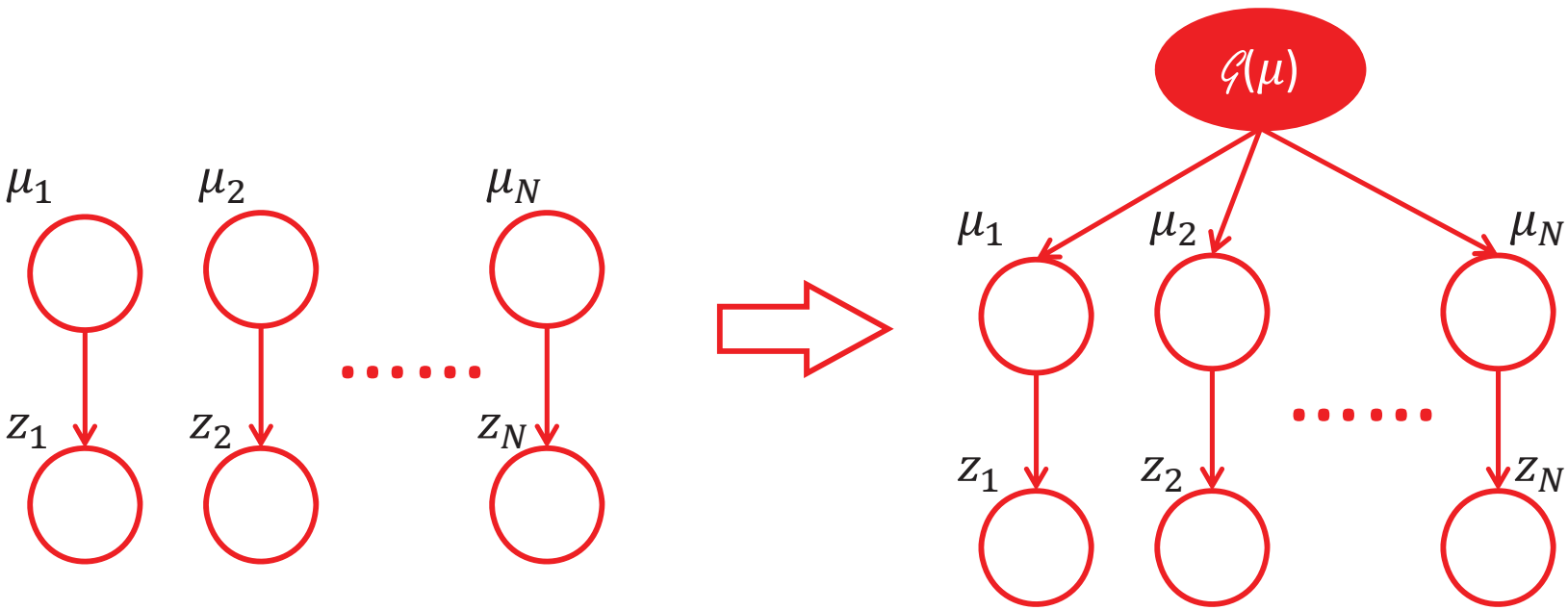


Figure 4: JSE uses the hierarchical structure to borrow information from each other.

How about μ_i from other prior distributions

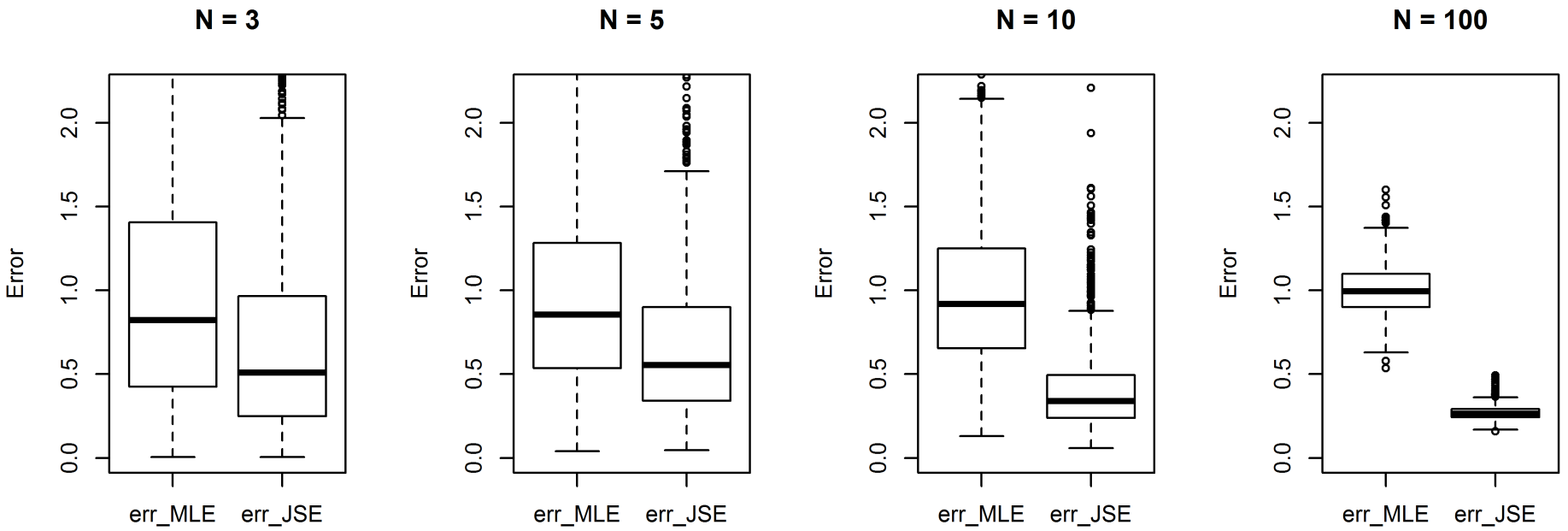


Figure 5: Comparison between MLE and JSE, $N = 3, 5, 10, 100$, where $\mu_i \sim$ Unif[0,1].

Tweedie's formula

- In James-Stein Estimator, we assumed $\mu_i \sim \mathcal{N}(0, \sigma^2)$.
- If without assuming a specific form of the prior distribution but only its existence, say $\mu \sim g(\cdot)$, can we still get $\mathbb{E}(\mu_i|z_i)$?
- The answer is "yes": Tweedie's formula.

$$\mathbb{E}(\mu_i|z_i) = z_i + \frac{d}{dz_i} \log f(z_i), \tag{11}$$

where $f(z)$ is the marginal density of z. Clearly, the first term z is the MLE, and $\frac{d}{dz} \log f(z)$ can be viewed as Bayes correction.

More details of Tweedie's formula

- For Tweedie's formula, we only need to assume $\mu \sim g(\cdot)$, where $g(\cdot)$ is an arbitrary probability density function (To keep notation unclutter, we drop the subscript i).

$$\mu \sim g(\cdot), \quad z|\mu \sim \mathcal{N}(\mu, 1). \tag{12}$$

- The marginal distribution of z is

$$f(z) = \int_{-\infty}^{\infty} \varphi(z-\mu) g(\mu) d\mu \tag{13}$$

where $\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ is the density for standard normal distribution.

- The posterior distribution of μ is

$$g(\mu|z) = \varphi(z-\mu) g(\mu) / f(z). \tag{14}$$

- Notice that we are only interested in the posterior mean $\mathbb{E}(\mu|z)$.

Exponential family

- The density function of Exponential family is

$$h(x) = \exp(\eta x - \psi(\eta))h_0(x). \tag{15}$$

where η is called the natural parameter,$\psi(\eta)$ is called the cumulant generating function.

- Example: *Normal distribution*

$$h(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right) = \exp(\mu x - \frac{\mu^2}{2})\varphi(x). \tag{16}$$

Comparing (15) with (16), we know $\eta = \mu$,$\psi(\eta) = \eta^2/2$.

- Example: *Poisson distribution*

$$h(x) = \frac{\exp(-\mu)\mu^x}{x!} = \frac{\exp\left(\log(\mu)x - \mu\right)}{x!} \tag{17}$$

Comparing (15) with (17), we have $\eta = \log\mu, \psi(\eta) = \exp(\eta)$.

Cumulant generating function

- Since $h(x)$ is a probability density function, it satisfies

$$\exp(-\psi(\eta)) \int \exp(\eta x) h_0(x) dx = 1. \tag{18}$$

- Take derivative w.r.t η on both sides of (18):

$$-\frac{d\psi(\eta)}{d\eta} \exp(-\psi(\eta)) \int \exp(\eta x) h_0(x) dx + \exp(-\psi(\eta)) \int \exp(\eta x) h_0(x) x dx = 0. \tag{19}$$

- Based on (18), from (19) we have

$$-\frac{d\psi(\eta)}{d\eta} + \exp(-\psi(\eta)) \int \exp(\eta x) h_0(x) x dx = 0. \tag{20}$$

and

$$\exp(-\psi(\eta)) \int \exp(\eta x) h_0(x) x dx = \int x h(x) dx = \mathbb{E}(x). \tag{21}$$

- Using (21), (20) can be rewritten as

$$\frac{d\psi(\eta)}{d\eta} = \mathbb{E}(x). \tag{22}$$

Derivation of Tweedies' Formula

- Now we write $g(\mu|z)$ into a density function of exponential family.

$$\begin{aligned} g(\mu|z) &= \varphi(z-\mu)g(\mu)/f(z) \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(z-\mu)^2}{2}\right)g(\mu)/f(z) \\ &= [\exp(z\mu)]\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)/f(z)\right]\left[\exp\left(-\frac{\mu^2}{2}\right)g(\mu)\right] \\ &= \left[\exp\left(z\mu - \log\frac{f(z)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)}\right)\right]\left[\exp\left(-\frac{\mu^2}{2}\right)g(\mu)\right]. \\ &= \exp(z\mu-\psi(z))h_0(\mu). \end{aligned} \tag{23}$$

  where z can be considered as the natural parameter, $\psi(z) = \log\frac{f(z)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)}$ is the cumulant generating function, and $h_0(\mu) = \exp\left(-\frac{\mu^2}{2}\right)g(\mu)$.

- Using (22),we have

$$\mathbb{E}(\mu|z) = z + \frac{d}{dz}\log f(z). \tag{24}$$

James-Stein Estimator vs. Tweedie's formula

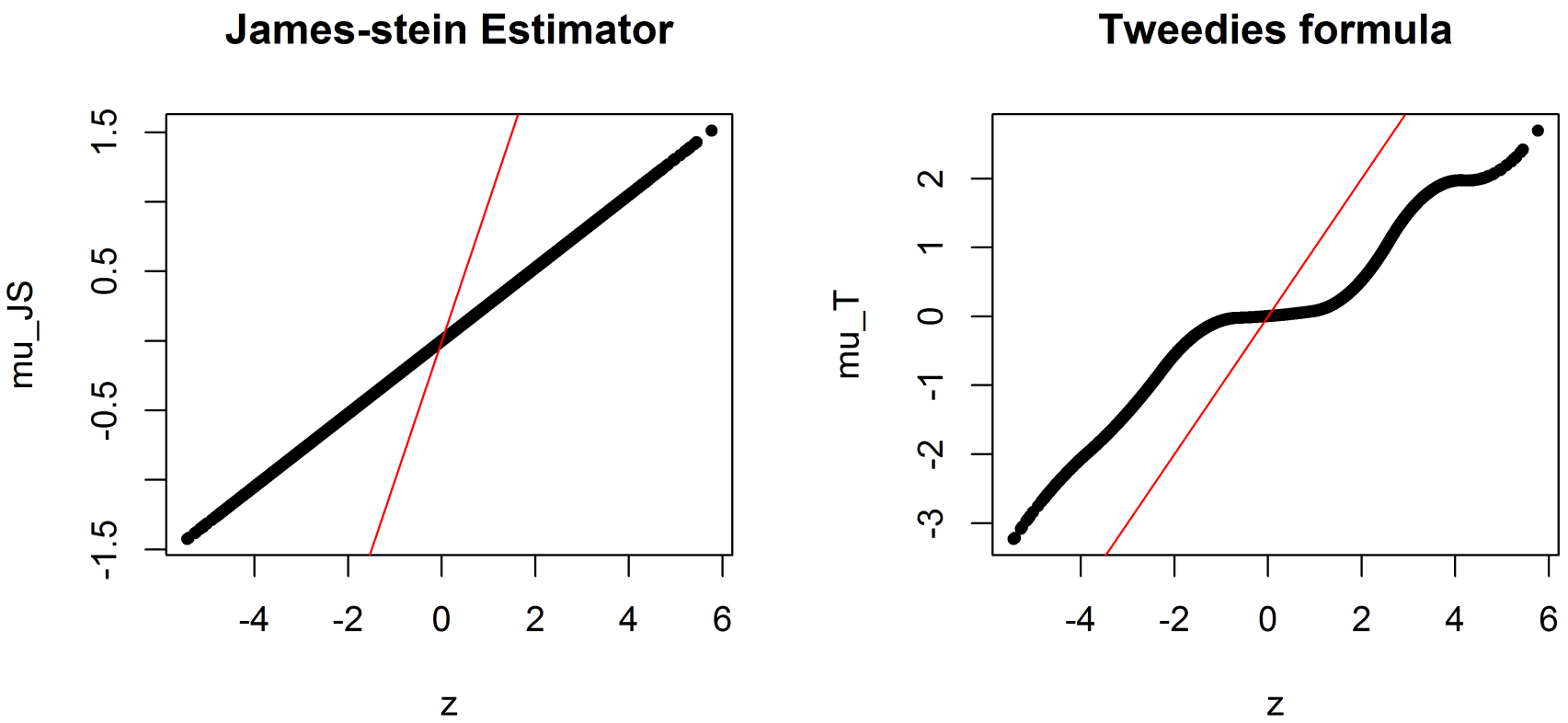


Figure 6: Different ways of shrinkage: James-Stein Estimator vs. Tweedie's formula. Here the true μ consist of {0,-2,2}, with their corresponding numbers $n_0 = 100000$, $n_{neg} = 5000$, and $n_{pos} = 5000$ (See R code next slide).

The R code for Tweedie's formula

- Try the following code for fun.

```
library(splines)
n0 = 100000
n_neg = 5000
n_pos = 5000
N = n0+n_neg+n_pos
z = c(rnorm(n0), rnorm(n_pos,2,1),rnorm(n_neg,-2,1))
bins = seq(min(z)-.1,max(z)+.1, len = 100)
h = hist(z, bins, plot = F)
x = h$m
g  = glm(h$c~ns(x,df=7),family = poisson)
ss = splinefun(x,log(dnorm(x)/g$fit),method = "natural")
mu_T = -ss(z,deriv=1)
mu_JS = (1-(N-2)/(t(z)%*%z))*z
#generate figure
#png(filename='z_mu.png',res=600, width = 20, height = 10, units="cm")
par(mfrow=c(1,2))
plot(z,mu_JS,pch=20)
abline(0,1,col='red')
title('James-stein Estimator')
plot(z,mu_T,pch=20)
title('Tweedies formula')
abline(0,1,col='red')
#dev.off()
```

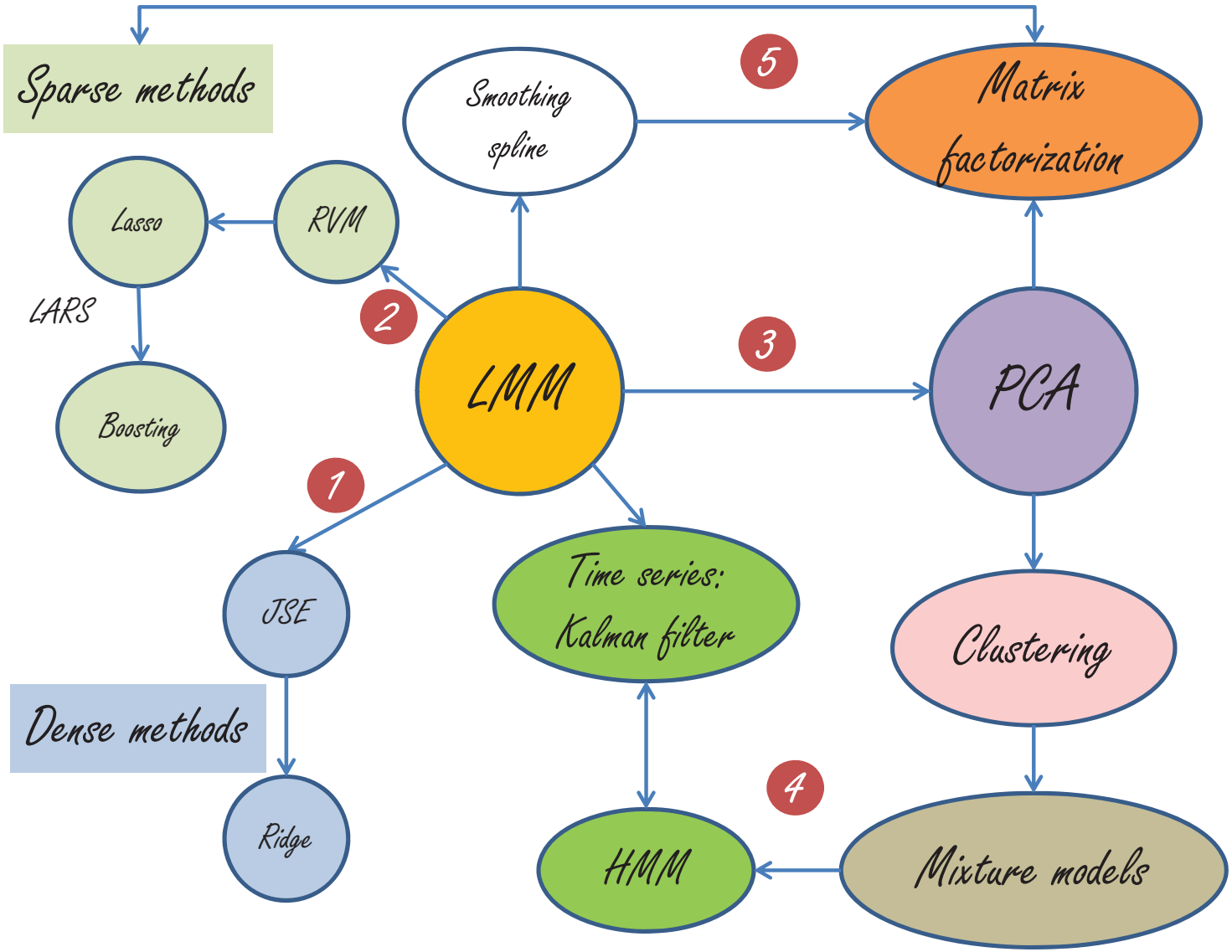
Connection

- James-Stein Estimator is a special case of LMM.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u} + \mathbf{e},$$
$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \mathbf{I}),$$
$$\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \mathbf{I}),$$

- A population machine learning model, Relevant Vector Machine (RVM), is an extension of LMM.
- The probabilistic interpretation of principal component analysis (PCA) is based on random effects.
- Kalman Filtering can also be written into a random-effects model.
- .....

A big picture

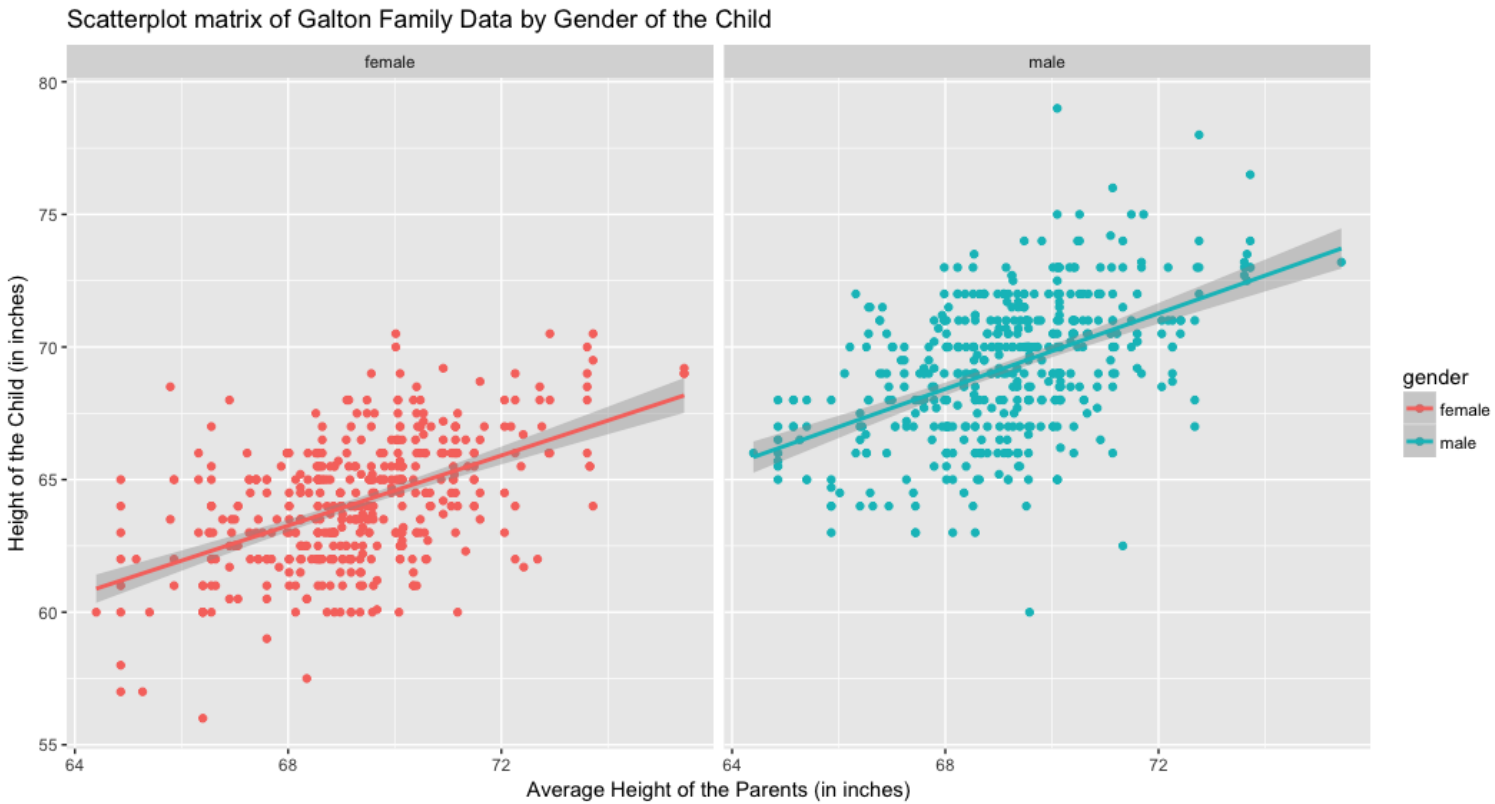


Galton Family Data

```
> library(HistData)
> data(GaltonFamilies)
> head(GaltonFamilies,10)
   family father mother midparentHeight children childNum gender childHeight
1     001   78.5   67.0           75.43        4        1   male        73.2
2     001   78.5   67.0           75.43        4        2 female        69.2
3     001   78.5   67.0           75.43        4        3 female        69.0
4     001   78.5   67.0           75.43        4        4 female        69.0
5     002   75.5   66.5           73.66        4        1   male        73.5
6     002   75.5   66.5           73.66        4        2   male        72.5
7     002   75.5   66.5           73.66        4        3 female        65.5
8     002   75.5   66.5           73.66        4        4 female        65.5
9     003   75.0   64.0           72.06        2        1   male        71.0
10    003   75.0   64.0           72.06        2        2 female        68.0
```

The origin of Regression

- "The farther backward you can look, the farther forward you can see." - Winston Churchill



References

D. Salsburg
The Lady Tasting Tea: How **Statistics** revolutionized science in the twentieth century.
W.H Freeman and Company, 2002.

C. Bishop
Pattern recoginition and Machine learning
Springer, 2006.

B. Efron
Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction
Cambridge University Press, 2010.

Yang C.
Stories of statistical learning,
COS, 2011, https://cosx.org/2011/12/stories-about-statistical-learning.

Yang C.
Chasing after EB,
COS, 2012, http://cos.name/2012/05/chase-after-eb/.

Yang C.
LMM and me: a romantic journey,
COS, 2014, http://cos.name/2014/04/lmmandme/.