SURE Estimation
in a
Heteroscedastic Hierarchical Model

Lawrence D Brown
Statistics Department, Wharton School
University of Pennsylvania

Joint work with S. Kou and X. Xie (Harvard)

SAHD conference, Duke Univ,
July 28, 2011

Normal Location Problem (Heteroscedastic)

Observe:

$$X_i \sim N\left(\theta_i, \sigma_i^2\right), \text{ independent, } i = 1,..,p.$$

σ_i^2 known (*for this talk*), and not necessarily equal.

Goal:

$$\text{Estimate } \theta = \left(\theta_1,..,\theta_p\right)' \text{ by } \hat{\theta} = \left(\hat{\theta}_1,..,\hat{\theta}_p\right)$$

Measure quality of estimation procedure by

$$R\left(\theta,\hat{\theta}\right) := E_\theta\left(p^{-1} \| \hat{\theta} - \theta \|^2\right) := E_\theta\left(L\left(\theta,\hat{\theta}\right)\right).$$

Variations and generalizations are of interest.

An Example: Predicting Batting Averages

$R_i =$ Batting average of Major Leaguer i for 1st half of 2005 season. (*Validate estimates using 2nd half batting records.*)

- Model: $R_i \sim \text{Bin}(n_i, p_i)$. Restrict sample to $n_i \geq 11$.
- Let $$X_i = \sin^{-1}\sqrt{\frac{R_i + 1/4}{n_i + 1/2}} \approx N\left(\theta_i, \frac{1}{4n_i}\right)$$
$$\ni \quad \theta_i = \sin^{-1}\sqrt{p_i}\,.$$
- Use $\{X_i\}$ to estimate $\{\theta_i\}$ under S.E. loss.

Hierarchical Bayes Estimators

- E-B hierarchy

$$X_i \mid \theta_i \sim N\left(\theta_i, \sigma_i^2\right)$$

$$\theta_i \mid \lambda \sim G_\lambda, \text{ independent}.$$

- Standard, Conjugate structure

$$G_\lambda = N\left(0, \lambda\right).$$

- Bayes Estimate given G_λ: $\hat{\theta}_i^{\lambda} = E_{G_\lambda}\left(\theta_i \mid X_i\right)$.
- E-B estimator: Estimate λ from the sample & plug in.

- Full Bayes hierarchy

$$\lambda \sim H \text{ } (H \text{ is assumed known}).$$

- Full Bayes estimator: $\hat{\theta}_i^{\text{Bayes}_H} = E_{\lambda \sim H}\left(\theta_i \mid X_i\right)$

SURE

- For an estimator of the form

$$\hat{\theta}_i(x) := x_i + \xi_i(x),$$

$$R\left(\theta,\hat{\theta}\right) = E_\theta\left(\text{SURE}\right) := E_\theta\left[p^{-1}\sum\left(\sigma_i^2 + 2\sigma_i^2\frac{\partial}{\partial X_i}\xi_i(X) + \xi_i^2(X)\right)\right].$$

- Conjugate prior Bayes estimator (with known λ):

$$\hat{\theta}_i^{\lambda} = \frac{\lambda}{\lambda+\sigma_i^2}X_i.$$

- Then,

$$\text{SURE}(\lambda) = p^{-1}\sum\left[\frac{\sigma_i^2}{\sigma_i^2+\lambda}X_i^2 + \sigma_i^2\frac{\lambda-\sigma_i^2}{\sigma_i^2+\lambda}\right].$$

(*Since Bayes estimator is linear this is also Mallow's C_P.*)

- The SURE estimator is $\hat{\theta}^{\text{SURE}} := \theta^{\hat{\lambda}_{\text{SURE}}}$ with

$$\hat{\lambda}_{\text{SURE}} := \operatorname{argmin}\left\{\text{SURE}(\lambda)\right\}.$$

so

$$\hat{\lambda}_{\text{SURE}} \text{ solves } \sum\left[\frac{\sigma_i^4}{\left(\sigma_i^2+\lambda\right)^3}X_i^2 - \frac{\sigma_i^4}{\left(\sigma_i^2+\lambda\right)^2}\right] = 0\ *.$$

- For comparison, the E-B MLE estimator is similar but

$$\hat{\lambda}_{\text{MLE}} \text{ solves } \sum\left[\frac{\sigma_i^2}{\left(\sigma_i^2+\lambda\right)^2}X_i^2 - \frac{1}{\sigma_i^2+\lambda}\right] = 0\ *.$$

- And, a conventional E-B Method of Moments

$$\hat{\lambda}_{\text{MM}} \text{ solves } \sum\left[X_i^2 - \sigma_i^2 - \lambda\right] = 0\ *.$$

The Homoscedastic Case

- Assume $\sigma_i^2 \equiv \sigma^2$ (σ^2 known, for now).
- Then

$$\hat{\theta}^{\text{SURE}} = \hat{\theta}^{\text{MLE}} = \hat{\theta}^{\text{MM}} = \left(1 - \frac{\mathbf{p}}{\|X\|^2}\sigma^2\right)_+ X.$$

- By contrast, the JAMES-STEIN$^+$ estimator is

$$\hat{\theta}^{\text{JS}} = \left(1 - \frac{\mathbf{p-2}}{\|X\|^2}\sigma^2\right)_+ X.$$

- The slight difference is a reflection the SURE logic. (*SURE can derive the choice* ***p-2*** *if used differently*.)

"SURESHRINK"
See Donoho and Johnstone (1995)

- Created for a related context.
- In our context uses

$$\hat{\theta}_i^{\mathrm{SS}} = \mathrm{sgn}\left(X_i\right)\left(\left|X_i\right| - \hat{\xi}\sigma\right)_+ .$$

And chooses $\hat{\xi}$ to minimize SURE (among all such soft threshold estimators with fixed ξ). [+ Truncated at $\sigma\sqrt{2\log p}$ to work better in a severely sparse setting.]

- Is regularized estimator for an L_1 penalty; and also
- Empirical-Bayes posterior mode under a Laplace prior with scale ξ.

Generalizations:

- Change standard E-B hierarchy to

$$X_i|\theta_i \sim N\left(\theta_i, \sigma_i^2\right)$$
$$\theta_i|\lambda \sim N\left(\boldsymbol{\mu}, \lambda\right).$$

- Two possible estimators:

1. Estimate μ by $\bar{X}$ and then apply SURE to get "best" estimator of the form $\hat{\theta}_i^{\lambda,\bar{X}} = \frac{\lambda}{\sigma_i^2+\lambda} X_i + \frac{\sigma_i^2}{\sigma_i^2+\lambda} \bar{X}$
2. Apply the argmin in SURE over **both** hyperparameters μ and λ to get "best" among $\hat{\theta}_i^{\lambda,\bar{X}} = \frac{\lambda}{\sigma_i^2+\lambda} X_i + \frac{\sigma_i^2}{\sigma_i^2+\lambda} \mu$.

- Estimators notated as $\hat{\theta}^{\text{SURE}\mathbf{1}}$ & $\hat{\theta}^{\text{SURE}\mathbf{2}}$, resp.

Generalization: A Monotone Estimator

- Motivation: A semi-parametric complete hierarchy:

$$X_i \mid \theta_i \sim N\left(\theta_i, \sigma_i^2\right) \quad \theta_i \mid \lambda \sim N\left(\mu, \lambda\right)$$

and $\lambda \sim H$ (H unknown).

- Then $\hat{\theta}_i = \left(1 - \beta_i\right) X_i + \beta_i \mu$ where $\beta_i = E_H\left(\dfrac{\lambda}{\sigma_i^2 + \lambda} \middle| X\right)$

- For a broad class of hyperpriors, H, & any fixed X

$$\beta \in \text{MON} := \left\{ b\left(\sigma_i^2\right) : b \text{ non-decreasing in } \sigma_i^2 \right\}$$

- Hence we propose $\hat{\theta}_i^{\text{MON}} = \left(1 - \hat{\beta}_i^{\text{MON}}\right) X_i + \hat{\beta}_i^{\text{MON}} \mu$ ∋

$$\hat{\beta}_i^{\mathrm{SM}}, \hat{\mu}^{\mathrm{SM}} = \mathrm{argmin}_{\beta_i \in \mathrm{MON}, \mu} \mathrm{SURE} =$$

$$\mathrm{argmin}_{\beta_i \in \mathrm{MON}, \mu} \left(p^{-1} \sum \left(\beta_j \left(X_j - \mu \right)^2 + \left(1 - 2\beta_j \right) \sigma_i^2 \right) \right) .$$

- Note: Many hierarchical Bayes estimators satisfy MON. Some others do as well:

1. If G has a log concave density, $\hat{\beta}_G \in \mathrm{MON}$.
2. D & J's SURESHRINK soft threshold estimator also satisfies MON.

- $\hat{\theta}^{\mathrm{MONM}}$ is a variant in which μ is estimated by $\bar{X}$.

- Asymptotics

- All our SURE estimators are asymptotically optimal within their respective classes, under weak conditions:

1. $\limsup_{p\to\infty} p^{-1}\sum_{i=1}^{p}\sigma_i^4 < \infty.$
2. $\limsup_{p\to\infty} p^{-1}\sum_{i=1}^{p}\theta_i^2 < \infty.$
3. $\limsup_{p\to\infty} p^{-1}\sum_{i=1}^{p}\sigma_i^2\theta_i^2 < \infty.$

- For any of our estimator classes, $\Im$, define the oracle

$$\tilde{\theta}^{\Im} = \operatorname{argmin}_{d(X,\theta)\in\Im} L\left(\theta, d\left(X,\theta\right)\right). \qquad \textbf{THEN},$$

$$L\left(\theta,\hat{\theta}^{SURE,\Im}\right) = L\left(\theta,\tilde{\theta}^{\Im}\right) + o_P\left(1\right) \text{ \& also}$$

$$\limsup_{p\to\infty}\left[R\left(\theta,\hat{\theta}^{\mathrm{SURE},\Im}\right) - E_\theta\left(L\left(\theta,\tilde{\theta}^{\Im}\right)\right)\right] = 0.$$

Baseball Data

- All E-B estimators are of the form

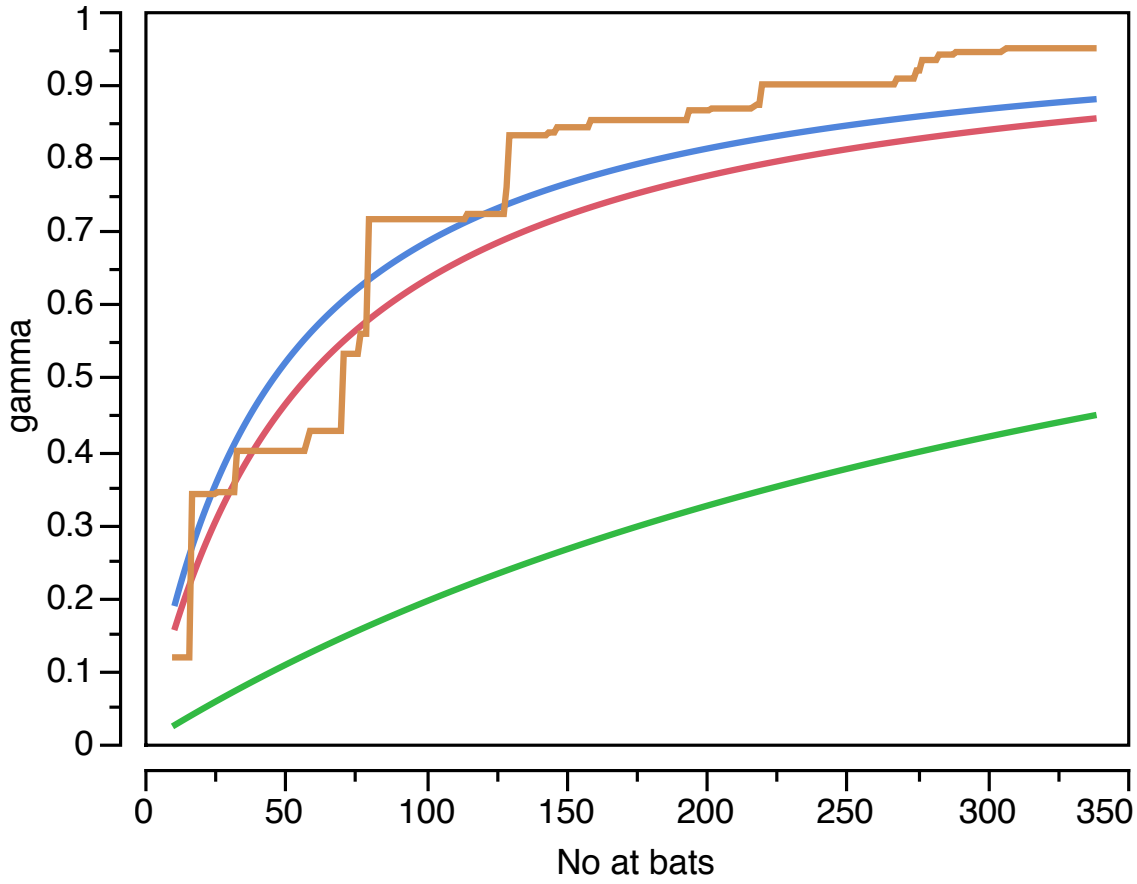
$$\hat{\theta}_i = \left(\hat{\gamma}\left(n_i\right)\right) X_i + \left(1 - \hat{\gamma}\left(n_i\right)\right) \hat{\mu}.$$

Table of Values of $\hat{\mu}$

Method	$\hat{\mu}$
$\bar{X}$	0.5095
MM	0.5276
MLE	0.5382
SURE1	0.5096
SURE2	0.4557
MON	0.5291
MONM	0.5096

Plot of values of $\hat{\gamma}$ for some estimators

$$\hat{\theta}_i = \left(\hat{\gamma}\left(n_i\right)\right) X_i + \left(1 - \hat{\gamma}\left(n_i\right)\right) \hat{\mu}$$

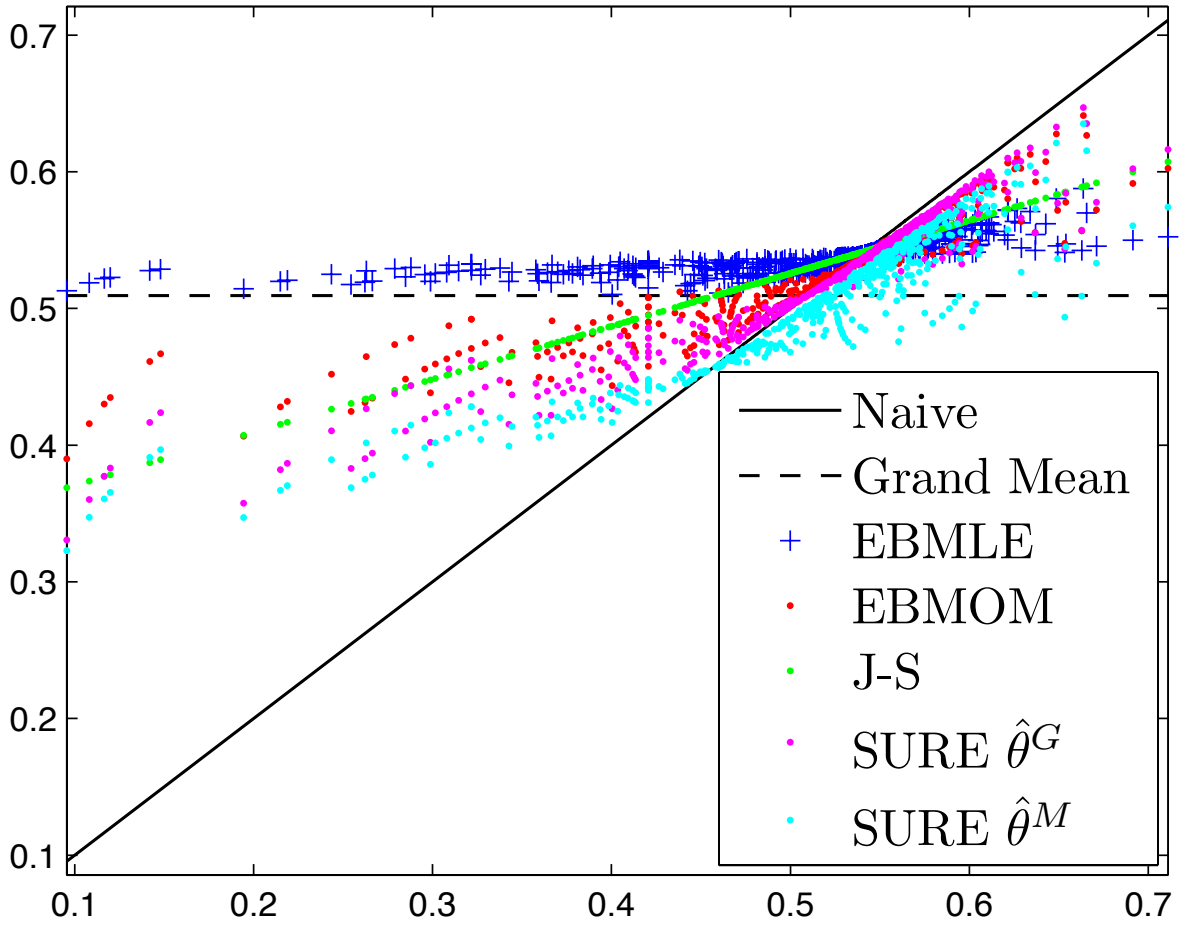


EB-MLE **EB-MM** **SURE2** **MON**

Relative Risk of Various Estimators
(as estimated from "holdout" sample = 2nd half of season)

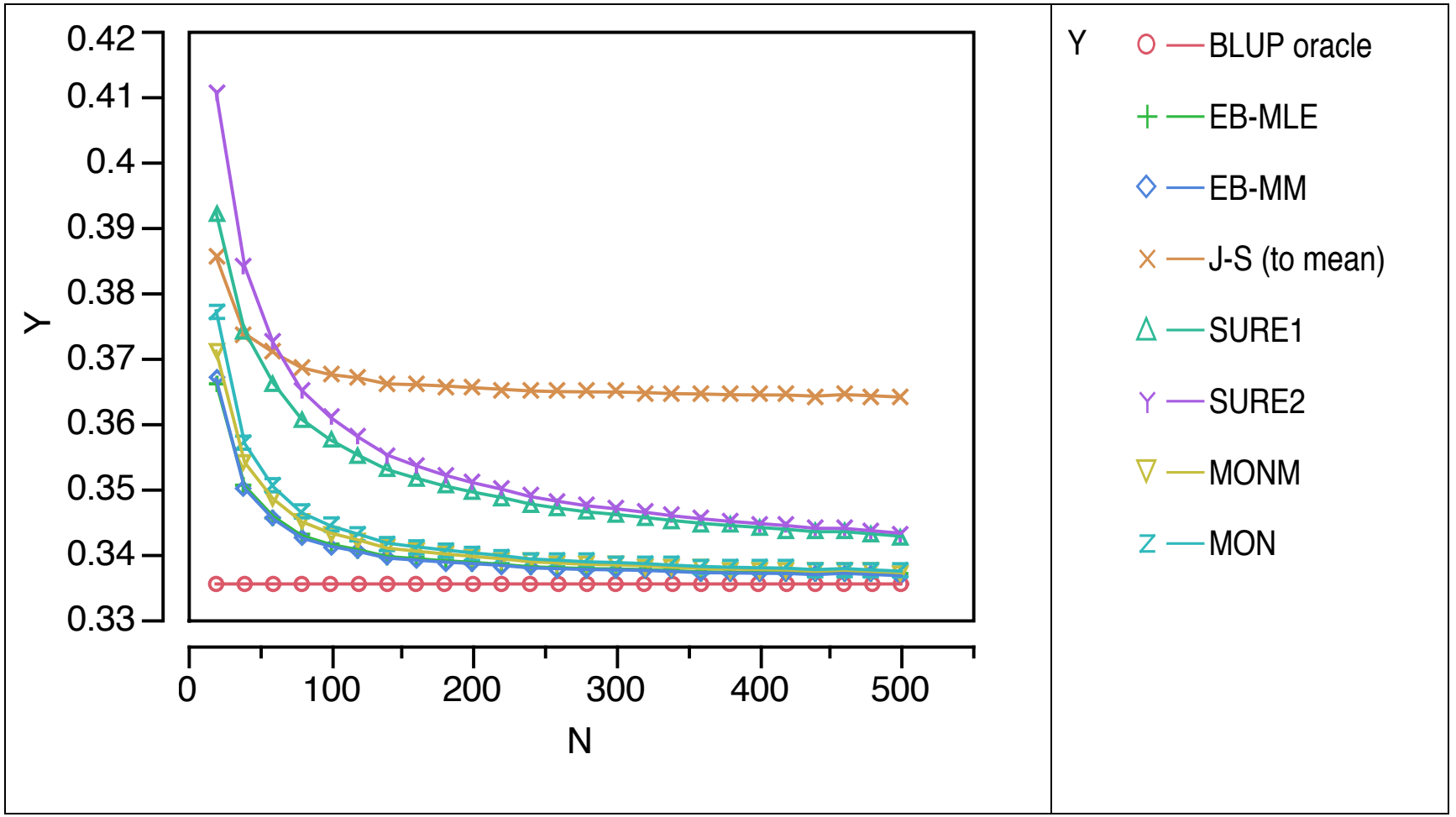
	All Batters	Non-pitchers	Only Pitchers
Naive	**1**	**1**	**1**
Grand Mean	0.853	0.378	0.127
EB-MM	0.585	0.357	0.129
EB-MLE	0.888	0.398	0.118
J-S (to mean)	0.535	0.348	0.165
SURE1	0.505	0.278	0.123
SURE2	0.422	0.282	0.123
MON	0.419	0.278	0.077
MONM	0.409	0.261	0.081

Plot of Some Estimators



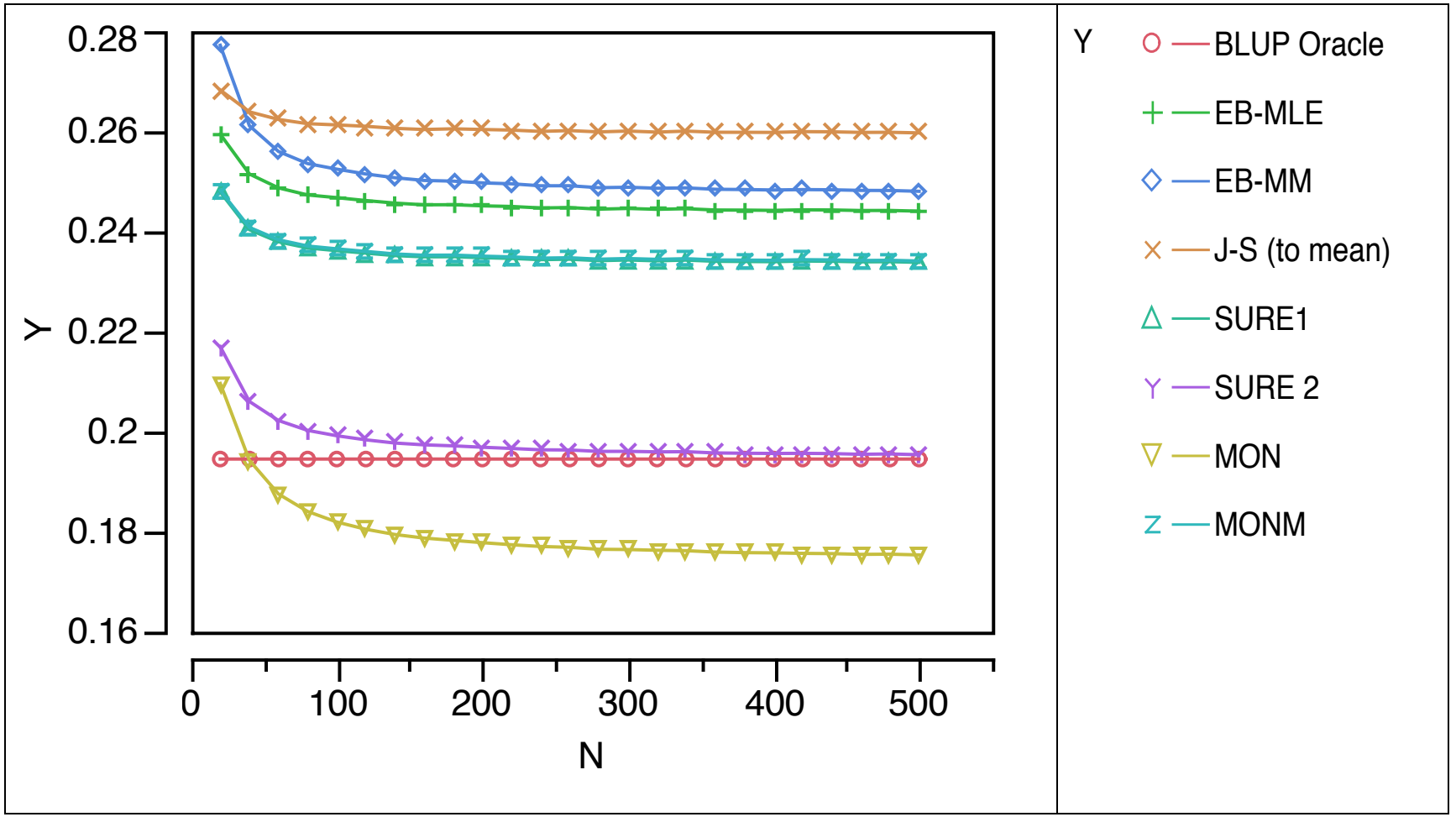
Simulation #1: An ideal situation:

$X_i \sim N\left(\theta_i, \sigma_i^2\right), \theta_i \sim N(0,1), \sigma_i^2 \sim \text{Unif}(0.1,1)$ [observe X_i, σ_i^2]



Simulation #5: A two-groups situation

$$X_i \sim N\left(\theta_i, \sigma_i^2\right),\ \theta_i \sim N\left(2, 0.1\right) \text{ or } N\left(0, 0.5\right),\ \sigma_i^2 \sim 0.1 \text{ or } 0.5$$



Summary

- Possibly heteroscedastic model
  [Normal, or approx. normal, variances known or estimable.]
- Consider a family of estimators, $\hat{\theta}^{\lambda}$.
  [Maybe generated from a hierarchical Bayes structure.]
- Write the SURE estimate of risk for fixed λ.
- Minimize this over λ at observed X.
- Estimator has desirable asymptotic properties [for all the families we consider], and
- Estimator performs well in practical examples and simulations