Optimal Shrinkage Estimation in Heteroscedastic Hierarchical Models: Beyond Gaussian

Samuel Kou

Joint with L. Brown, X. Xie & J. Yang

IMS Workshop: Meeting the Statistical Challenges in High Dimensional Data and Complex Networks

Shrinkage Estimate

- Ever since Stein's groundbreaking work, shrinkage estimation not only revolutionized statistics, but also has wide-ranging applications in science and engineering.
- James-Stein estimator

$$X_i|\theta_i \overset{\text{ind.}}{\sim} N(\theta_i, A), \quad i = 1, 2, \ldots, p$$

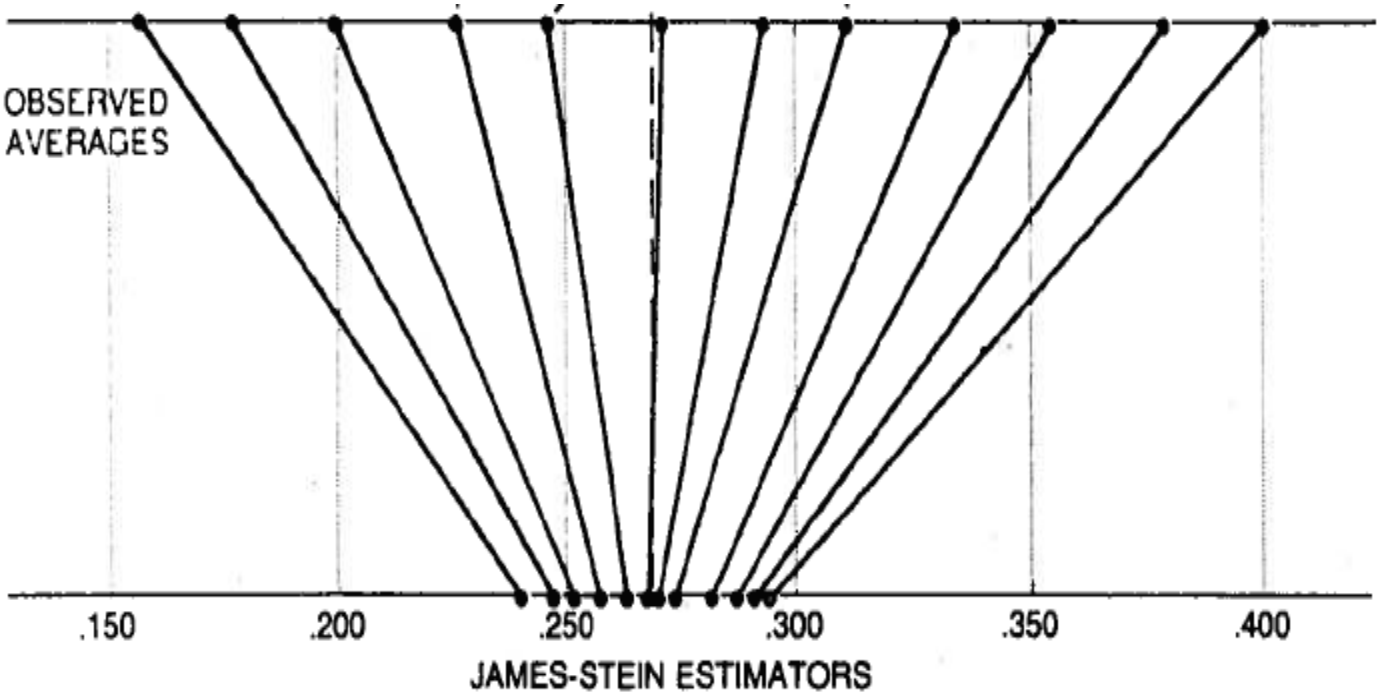
$$\hat{\theta}_i = \frac{\hat{\lambda}}{\hat{\lambda} + A} X_i, \qquad \hat{\lambda} = (\frac{1}{p}\sum_{i=1}^{p} X_i^2 - A)^+$$

Shrinkage est. Emp. Bayes est.

- Hierarchical model: a natural way to arrive at it

$$\begin{aligned} X_i|\theta_i &\overset{\text{ind.}}{\sim} N(\theta_i, A) \\ \theta_i &\overset{\text{ind.}}{\sim} N(0, \lambda) \end{aligned} \quad \Rightarrow \quad \begin{aligned} \theta_i|X_i &\sim N\left(\frac{\lambda}{\lambda + A} X_i, \frac{\lambda A}{\lambda + A}\right) \\ X_i &\sim N(0, \lambda + A), \end{aligned}$$

The picture



Efron and Morris (1977)

Heteroscedastic Hierarchical Model

- Distinct variances (e.g., different accuracy, sample sizes)

$$X_i|\theta_i \overset{\text{ind.}}{\sim} N(\theta_i, A_i)$$

$$\theta_i \overset{\text{ind.}}{\sim} N(0, \lambda)$$

- Shrinkage estimate $\theta_i|X_i \sim N\left(\frac{\lambda}{\lambda + A_i}X_i, \frac{\lambda A_i}{\lambda + A_i}\right)$

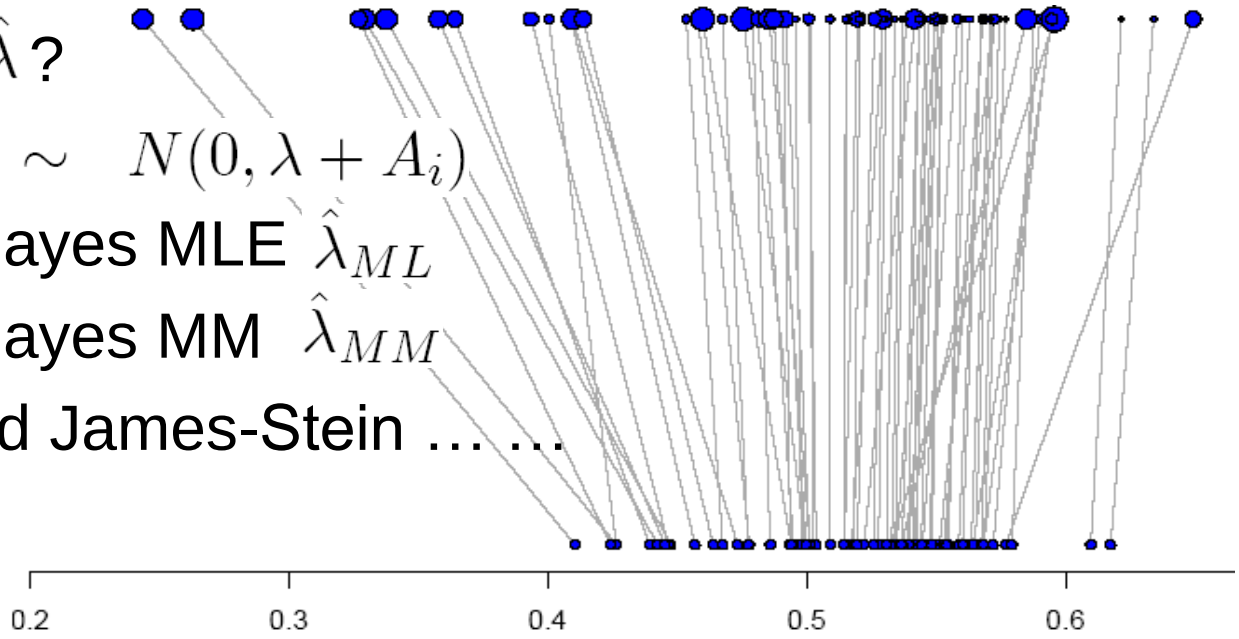
$$\hat{\theta}_i^{\lambda} = \frac{\lambda}{\lambda + A_i}X_i$$

- How to obtain $\hat{\lambda}$?

  Could use $X_i \sim N(0, \lambda + A_i)$

  - empirical Bayes MLE $\hat{\lambda}_{ML}$
  - empirical Bayes MM $\hat{\lambda}_{MM}$
  - Generalized James-Stein … …

- Optimal One??



SURE estimate

- With sum of squared error loss $l(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{1}{p}\sum_i (\hat{\theta}_i - \theta_i)^2$

  for shrinkage estimator $\hat{\theta}_i^{\lambda} = \frac{\lambda}{\lambda + A_i} X_i$,

  an unbiased estimate of its risk $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda}) = E[l(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda})]$

  is $$\mathrm{SURE}(\lambda) = \frac{1}{p}\sum_i \left[\left(\frac{A_i}{A_i + \lambda} \right)^2 X_i^2 + \frac{A_i(\lambda - A_i)}{A_i + \lambda} \right]$$

- URE estimate:

$$\hat{\lambda}_{\mathrm{SURE}} = \arg\min_{\lambda \geq 0} \mathrm{SURE}(\lambda)$$

$$\hat{\theta}_i^{\mathrm{SURE}} := \hat{\theta}_i^{\hat{\lambda}_{\mathrm{SURE}}} = \frac{\hat{\lambda}_{\mathrm{SURE}}}{\hat{\lambda}_{\mathrm{SURE}} + A_i} X_i$$

Risk properties of SURE estimate

- SURE(λ) is uniformly close to $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda})$:

$$\sup_{0 \le \lambda < \infty} \left| \mathrm{SURE}(\lambda) - R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda}) \right| \overset{L^2}{\to} 0 \quad as\ p \to \infty$$

- In fact, can strengthen to the loss function:

$$\sup_{0 \le \lambda < \infty} \left| \mathrm{SURE}(\lambda) - l(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda}) \right| \to 0 \ \ in \ \ L^2$$

Risk properties of SURE estimate (cont.)

- **Optimality**

For any $\hat{\lambda}_p$ and the corresponding $\hat{\boldsymbol{\theta}}^{\hat{\lambda}_p} = \frac{\hat{\lambda}_p}{\hat{\lambda}_p + \mathbf{A}} \mathbf{X}$ we **always** have

$$\lim_{p \to \infty} P\left(l(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{SURE}}) \geq l(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\lambda}_p}) + \varepsilon \right) = 0 \quad for\ any\ \varepsilon > 0$$

and

$$\limsup_{p \to \infty} \left[R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{SURE}}) - R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\lambda}_p}) \right] \leq 0.$$

Regularity conditions

$$(A): \limsup_{p\to\infty} \frac{1}{p} \sum_{i=1}^{p} A_i^2 < \infty,$$

$$(B): \limsup_{p\to\infty} \frac{1}{p} \sum_{i=1}^{p} A_i \theta_i^2 < \infty,$$

$$(\mathrm{C}): \limsup_{p\to\infty} \frac{1}{p} \sum_{i=1}^{p} \theta_i^2 < \infty$$

The oracle property

- Oracle parameter: parameter that minimizes the loss

$$\tilde{\lambda}^{OL} = \tilde{\lambda}^{OL}(\boldsymbol{\theta}; X_1, \ldots, X_p) = \arg\min_{\lambda \geq 0} l_p(\theta, \hat{\theta}^{\lambda})$$

$$\tilde{\boldsymbol{\theta}}^{OL} = \frac{\tilde{\lambda}^{OL}}{\tilde{\lambda}^{OL} + \boldsymbol{A}} \boldsymbol{X}$$

depends on each realization.

gives the theoretical limit for any shrinkage estimator

- This theoretical limit is achieved by the SURE estimate:

Assume conditions (A) and (B). Then

$$\lim_{p \to \infty} P\left(l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{SURE}}) \geq l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OL}) + \varepsilon \right) = 0 \quad \textit{for any } \varepsilon > 0.$$

$$\lim_{p \to \infty} \left[R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{SURE}}) - E\left(l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OL}) \right) \right] = 0.$$

General shrinkage estimate

- So far, shrink toward origin as we put

$$\theta_i \overset{\text{ind.}}{\sim} N(0, \lambda)$$

- Can shrink toward the grand mean $\bar{X}$
- In general, starting from $\theta_i \overset{\text{ind.}}{\sim} N(\mu, \lambda)$ Bayes estimate

$$\hat{\theta}_i^{\lambda,\mu} = \frac{\lambda}{A_i + \lambda} X_i + \frac{A_i}{A_i + \lambda} \mu$$

- Determine λ and μ from the data?

SURE estimate

- An unbiased estimate of the risk $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\mu})$ is

$$\mathrm{SURE}^M(\mu,\lambda) = \frac{1}{p}\sum_i \frac{A_i}{(A_i+\lambda)^2}\left(A_i(X_i-\mu)^2 + \lambda^2 - A_i^2\right)$$

- SURE estimate:

$$\hat{\theta}_i^M = \frac{\hat{\lambda}_M}{A_i+\hat{\lambda}_M}X_i + \frac{A_i}{A_i+\hat{\lambda}_M}\hat{\mu}_M$$

where

$$(\hat{\lambda}_M, \hat{\mu}_M) = \arg\min_{\lambda,\ \mu} \mathrm{SURE}^M(\lambda,\mu)$$

Risk properties

- SURE$^M(\lambda, \mu)$ is uniformly close to $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\mu})$:

$$\sup_{0\leq\lambda\leq\infty,\ |\mu|\leq\max_i|X_i|} \left|\mathrm{SURE}^M(\lambda,\mu) - R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\mu})\right| \to 0 \ in\ L^1$$

- Can be strengthened to loss function:

$$\sup_{0\leq\lambda\leq\infty,\ |\mu|\leq\max_i|X_i|} \left|\mathrm{SURE}^M(\lambda,\mu) - l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\mu})\right| \to 0 \ in\ L^1$$

Risk properties (cont.)

- **Optimality**

For any $(\hat{\lambda}_p, \hat{\mu}_p)$ and corresponding $\hat{\boldsymbol{\theta}}^{\hat{\lambda}_p,\hat{\mu}_p} = \frac{\hat{\lambda}_p}{\hat{\lambda}_p+\mathbf{A}}\mathbf{X} + \frac{\mathbf{A}}{\hat{\lambda}_p+\mathbf{A}}\hat{\mu}_p$ we **always** have

$$\lim_{p\to\infty} P\left(l(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{M}) \geq l(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\lambda}_p,\hat{\mu}_p}) + \varepsilon \right) = 0 \quad for\ any\ \varepsilon > 0$$

and

$$\limsup_{p\to\infty} \left[R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{M}) - R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\lambda}_p,\hat{\mu}_p}) \right] \leq 0$$

The oracle property

- Oracle parameters:

$$(\tilde{\lambda}^{\mathrm{OL}}, \tilde{\mu}^{\mathrm{OL}}) = \arg\min_{\lambda,\mu} l(\theta, \hat{\theta}^{\lambda,\mu})$$

$$\tilde{\boldsymbol{\theta}}^{MOL} = \frac{\tilde{\lambda}^{MOL}}{\tilde{\lambda}^{MOL} + \boldsymbol{A}} \boldsymbol{X} + \frac{\boldsymbol{A}}{\tilde{\lambda}^{MOL} + \boldsymbol{A}} \tilde{\mu}^{MOL}$$

depends on each realization.

gives the theoretical limit for any shrinkage estimator

- This theoretical limit is achieved by the SURE estimate:

Assume conditions (A)-(C). Then

$$\lim_{p\to\infty} P\left(l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^M) \geq l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{MOL}) + \varepsilon\right) = 0 \quad \textit{for any } \varepsilon > 0\,.$$

$$\lim_{p\to\infty} \left[R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^M) - E\left(l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{MOL})\right)\right] = 0.$$

Semi-parametric SURE estimate

- The properties of the SURE shrinkage estimate does not involve the prior distribution
- However, the general form of $\hat{\theta}_i = \frac{\lambda}{A_i+\lambda} X_i + \frac{A_i}{A_i+\lambda}\mu$ motivated from normal prior
- Question: can one do better?
- Yes, if we move on to a larger class:

$$\hat{\theta}_i^{b_i,\mu} = (1 - b_i) \cdot X_i + b_i \cdot \mu$$

$$\text{Requirement (MON)} : b_i \leq b_j \quad \text{for any } i \text{ and } j \text{ such that } \; A_i \leq A_j$$

- Semi-parametric

Semi-parametric SURE estimate

- An unbiased estimate of the risk is

$$\mathrm{SURE}^M(\mathbf{b},\mu) = \frac{1}{p}\sum_{i=1}^{p}\left[b_i^2(X_i-\mu)^2 + (1-2b_i)A_i\right]$$

- Semi-parametric SURE estimate

$$\hat{\theta}_i^{SM} = (1-\hat{b}_i^{SM})\cdot X_i + \hat{b}_i^{SM}\cdot\hat{\mu}^{SM},$$

where

$$(\hat{\mathbf{b}}^{SM},\hat{\mu}^{SM}) = \text{minimizer of } \mathrm{SURE}^M(\mathbf{b},\mu)$$

$$\text{subject to } b_i \in [0,1] \text{and Requirement (MON)}$$

Risk properties

- SURE^M($\mathbf{b}$, μ) is uniformly close to $l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathbf{b},\mu})$:

$$\sup \left| \mathrm{SURE}^M(\boldsymbol{b}, \mu) - l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathbf{b},\mu}) \right| \to 0 \; in \; L^1$$

- **Optimality**

For any shrinkage estimator $\hat{\boldsymbol{\theta}}^{\hat{\mathbf{b}}_p,\hat{\mu}_p} = (1-\hat{\mathbf{b}}_p)\cdot\mathbf{X}+\hat{\mathbf{b}}_p\cdot\hat{\mu}_p$ with $\hat{\mathbf{b}}_p \in [0,1]$ and MON, we **always** have

$$\lim_{p\to\infty} P\left(l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{SM}) \geq l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\mathbf{b}}_p,\hat{\mu}_p}) + \varepsilon\right) = 0 \quad for\ any\ \varepsilon > 0$$

and

$$\limsup_{p\to\infty} \left[R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{SM}) - R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\mathbf{b}}_p,\hat{\mu}_p})\right] \leq 0$$

Beyond Gaussian: Distributions with quadratic variance function (QVF)

- Distributions with QVF: Y_i satisfies

  $E(Y_i) = \theta_i \quad \mathrm{Var}(Y_i) = V(\theta_i)/\tau_i \quad V(\theta_i) = \nu_0 + \nu_1\theta_i + \nu_2\theta_i^2$

  τ_i : within-group sample size
- Include: (**exponential families**) normal, Poisson, binomial, neg-binomial, gamma distributions

  (**location-scale families**) *t,* logistic, uniform, Laplace, Pareto, extreme value distributions
- Want simultaneous inference of θ_i, $i = 1, 2, \ldots, p$
- Semi-parametric shrinkage estimator

$$\hat{\theta}_i^{\mathbf{b},\mu} = (1 - b_i) \cdot Y_i + b_i \cdot \mu$$

with $b_i \in [0, 1]$ satisfying

Requirement (MON) : $b_i \le b_j$ for any i and j such that $\tau_i \ge \tau_j$

Semi-parametric URE estimate

- An unbiased estimate of the risk is

$$\mathrm{URE}(\mathbf{b}, \mu) = \frac{1}{p} \sum_{i=1}^{p} \left[b_i^2 \cdot (Y_i - \mu)^2 + (1 - 2b_i) \cdot \frac{V(Y_i)}{\tau_i + \nu_2} \right]$$

- Semi-parametric URE estimator

$$\hat{\theta}_i^{\mathrm{SM}} = (1 - \hat{b}_i) \cdot Y_i + \hat{b}_i \cdot \hat{\mu}^{\mathrm{SM}}$$

where

$$(\hat{\mathbf{b}}^{SM}, \hat{\mu}^{SM}) = \text{minimizer of } \mathrm{URE}(\mathbf{b}, \mu)$$

$$\text{subject to } b_i \in [0, 1] \text{ and Requirement (MON)}$$

Optimality

THEOREM *Assume regularity conditions* (A)–(E). *Then for any shrinkage estimator* $\hat{\boldsymbol{\theta}}^{\hat{\mathbf{b}},\hat{\mu}} = (1-\hat{\mathbf{b}})\cdot\mathbf{Y}+\hat{\mathbf{b}}\cdot\hat{\mu}$, *where* $\hat{\mathbf{b}}\in[0,1]$ *satisfies Requirement* (*MON*), *and* $|\hat{\mu}| \le \max_i |Y_i|$, *we always have*

$$\lim_{p\to\infty} P\left(l_p\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\mathrm{SM}}\right) \ge l_p\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\hat{\mathbf{b}},\hat{\mu}}\right)+\varepsilon\right)=0 \qquad \textit{for any } \varepsilon>0$$

and

$$\limsup_{p\to\infty}\left[R\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\mathrm{SM}}\right)-R\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\hat{\mathbf{b}},\hat{\mu}}\right)\right]\le 0.$$

Parametric URE estimate

- For exponential family with conjugate priors

  (e.g., binomial-beta, Poisson-gamma, neg.binom-beta, gamma-inv.gamma, normal-normal)

  ⟹ **parametric** empirical Bayes shrinkage estimators

- Can also construct the URE estimator

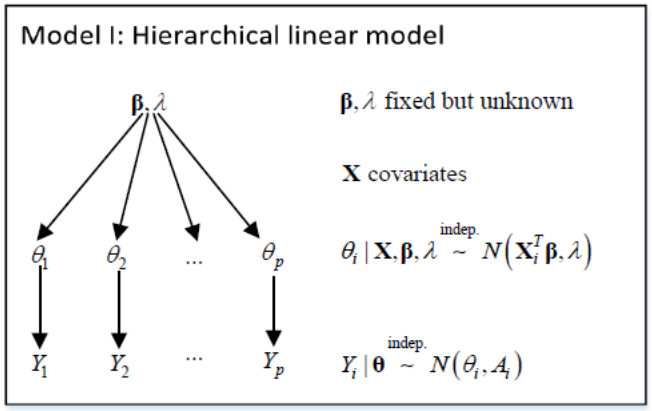
- The resulting **parametric** URE estimator is asymptotically risk **optimal** among the parametric estimators

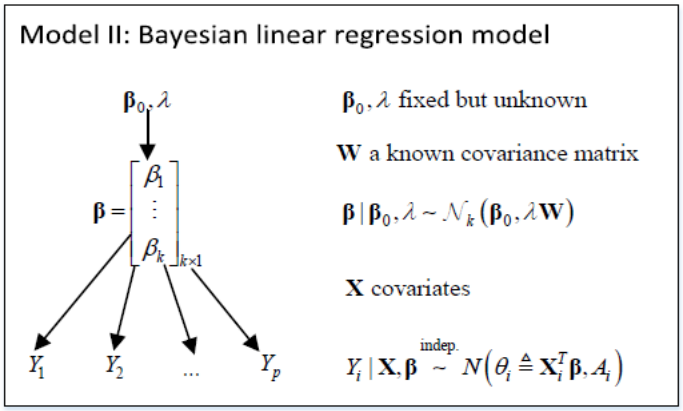
Linear models

- In many statistical applications, in addition to the heteroscedastic response variable, often has predictors
  ➡ extension to linear models
- $Y_i|\boldsymbol{\theta} \overset{\text{indep.}}{\sim} \mathcal{N}(\theta_i, A_i), \quad i = 1, \ldots, p$

  p independent statistical units $\{Y_i, \boldsymbol{X}_i\}_{i=1}^{p}$

  $$\boldsymbol{X} = [\boldsymbol{X}_1 | \cdots | \boldsymbol{X}_p], \quad \boldsymbol{X}_1, .., \boldsymbol{X}_p \in \mathbb{R}^k$$
- **Two** formulations:
  - Hierarchical linear model: $\theta_i \overset{\text{indep.}}{\sim} \mathcal{N}(\boldsymbol{X}_i^T \boldsymbol{\beta}, \lambda)$
  - Bayesian linear regression model: $\boldsymbol{\theta} = \boldsymbol{X}^T \boldsymbol{\beta} \qquad \boldsymbol{\beta} \sim \mathcal{N}_k(\boldsymbol{\beta}_0, \lambda \boldsymbol{W})$





Linear models

- Under Model **I** (hierarchical linear model) $\theta_i \overset{\text{indep.}}{\sim} \mathcal{N}\left(\boldsymbol{X}_i^T \boldsymbol{\beta}, \lambda\right)$

$$\hat{\theta}^{\lambda,\boldsymbol{\mu}} = \frac{\lambda}{\lambda + A_i} Y_i + \frac{A_i}{\lambda + A_i} \mu_i, \quad \text{with } \boldsymbol{\mu} \in \mathscr{L}_{\text{row}}(\boldsymbol{X})$$

- Under Model II (Bayesian linear regression) $\boldsymbol{\theta} = \boldsymbol{X}^T \boldsymbol{\beta} \quad \boldsymbol{\beta} \sim \mathcal{N}_k(\boldsymbol{\beta}_0, \lambda \boldsymbol{W})$

$$\hat{\boldsymbol{\theta}}^{\lambda,\boldsymbol{\beta}_0} = \boldsymbol{X}^T \hat{\boldsymbol{\beta}}^{\lambda,\boldsymbol{\beta}_0}, \quad \text{with } \hat{\boldsymbol{\beta}}^{\lambda,\boldsymbol{\beta}_0} = \lambda \boldsymbol{W}(\lambda \boldsymbol{W} + \boldsymbol{V})^{-1} \hat{\boldsymbol{\beta}}^{\text{WLS}} + \boldsymbol{V}(\lambda \boldsymbol{W} + \boldsymbol{V})^{-1} \boldsymbol{\beta}_0$$

where $\hat{\boldsymbol{\beta}}^{\text{WLS}} = \left(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right)^{-1} \boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{Y}$ is the weighted least square

$$\boldsymbol{V} = (\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T)^{-1}$$

Linear models

- Under Model **I** (hierarchical linear model) $\theta_i \overset{\text{indep.}}{\sim} \mathcal{N}\left(\boldsymbol{X}_i^T \boldsymbol{\beta}, \lambda\right)$

$$\hat{\theta}^{\lambda,\boldsymbol{\mu}} = \frac{\lambda}{\lambda + A_i} Y_i + \frac{A_i}{\lambda + A_i} \mu_i, \quad \text{with } \boldsymbol{\mu} \in \mathscr{L}_{\text{row}}(\boldsymbol{X})$$

- Semi-parametric estimators
  - Model I:

$$\hat{\theta}_i^{\boldsymbol{b},\boldsymbol{\mu}} = (1 - b_i) Y_i + b_i \mu_i, \quad \text{with } \boldsymbol{\mu} \in \mathscr{L}_{\text{row}}(\boldsymbol{X}),$$

where $\boldsymbol{b}$ satisfies the monotonic constraint

$$\text{MON}(\boldsymbol{A}) : b_i \in [0,1],\ b_i \leq b_j \text{ whenever } A_i \leq A_j.$$

Linear models

- Under Model **II** (Bayesian linear regression) $\boldsymbol{\theta} = \boldsymbol{X}^T\boldsymbol{\beta} \quad \boldsymbol{\beta} \sim \mathcal{N}_k(\boldsymbol{\beta}_0, \lambda \boldsymbol{W})$

$$\hat{\boldsymbol{\theta}}^{\lambda,\boldsymbol{\beta}_0} = \boldsymbol{X}^T\hat{\boldsymbol{\beta}}^{\lambda,\boldsymbol{\beta}_0}, \text{ with } \hat{\boldsymbol{\beta}}^{\lambda,\boldsymbol{\beta}_0} = \lambda \boldsymbol{W}(\lambda \boldsymbol{W} + \boldsymbol{V})^{-1}\hat{\boldsymbol{\beta}}^{\mathrm{WLS}} + \boldsymbol{V}(\lambda \boldsymbol{W} + \boldsymbol{V})^{-1}\boldsymbol{\beta}_0$$

where $\hat{\boldsymbol{\beta}}^{\mathrm{WLS}} = \left(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{Y}$ is the weighted least square

$$\boldsymbol{V} = (\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T)^{-1}$$

- Semi-parametric estimators
  - Model II: Let $\boldsymbol{W}^{-1/2}\boldsymbol{V}\boldsymbol{W}^{-1/2} = \boldsymbol{U}\bar{\boldsymbol{\Lambda}}\boldsymbol{U}^T \qquad \boldsymbol{Z} = \boldsymbol{U}^T\boldsymbol{W}^{1/2}\boldsymbol{X}$

    Then

$$\hat{\boldsymbol{\theta}}^{\lambda,\boldsymbol{\beta}_0} = \boldsymbol{Z}^T\left(\lambda(\lambda\boldsymbol{I}_k + \boldsymbol{\Lambda})^{-1}\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\hat{\boldsymbol{\beta}}^{\mathrm{WLS}} + \boldsymbol{\Lambda}(\lambda\boldsymbol{I}_k + \boldsymbol{\Lambda})^{-1}\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\boldsymbol{\beta}_0\right)$$

  - Suggest

$$\hat{\boldsymbol{\theta}}^{\boldsymbol{b},\boldsymbol{\beta}_0} = \boldsymbol{Z}^T\left((\boldsymbol{I}_k - \mathrm{diag}(\boldsymbol{b}))\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\hat{\boldsymbol{\beta}}^{\mathrm{WLS}} + \mathrm{diag}(\boldsymbol{b})\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\boldsymbol{\beta}_0\right)$$

where $\boldsymbol{b}$ satisfies the following monotonic constraint

$$\mathrm{MON}(\boldsymbol{D}) : b_i \in [0,1],\ b_i \le b_j \text{ whenever } d_i \le d_j.$$

Semi-parametric URE estimate

- Under Model I, an unbiased estimate of the risk is

$$\mathrm{URE}^{SP}(\boldsymbol{b},\boldsymbol{\mu}) = \frac{1}{p}\left\|\mathrm{diag}(\boldsymbol{b})(\boldsymbol{Y}-\boldsymbol{\mu})\right\|^2 + \frac{1}{p}\mathrm{tr}(\boldsymbol{A}-2\mathrm{diag}(\boldsymbol{b})\boldsymbol{A})$$

- Semi-parametric URE estimator

$$\hat{\boldsymbol{\theta}}_{SP}^{\mathrm{URE}} = (\boldsymbol{I}_p - \mathrm{diag}(\hat{\boldsymbol{b}}_{SP}^{\mathrm{URE}}))\boldsymbol{Y} + \mathrm{diag}(\hat{\boldsymbol{b}}_{SP}^{\mathrm{URE}})\hat{\boldsymbol{\mu}}_{SP}^{\mathrm{URE}},$$

where

$$\left(\hat{\boldsymbol{b}}_{SP}^{\mathrm{URE}}, \hat{\boldsymbol{\mu}}_{SP}^{\mathrm{URE}}\right) = \underset{\boldsymbol{b}\in\mathrm{MON}(\boldsymbol{A}),\ \boldsymbol{\mu}\in\mathscr{L}_{\mathrm{row}}(\boldsymbol{X})}{\mathrm{argmin}} \mathrm{URE}^{SP}(\boldsymbol{b},\boldsymbol{\mu})$$

Semi-parametric URE estimate

- Under Model II, an unbiased estimate of the risk is

$$\mathrm{URE}^{SP}(\boldsymbol{b},\boldsymbol{\beta}_0) = \frac{1}{p}\left\|\boldsymbol{Z}^T(\boldsymbol{I}_k - \mathrm{diag}(\boldsymbol{b}))\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{Y} + \boldsymbol{Z}^T\mathrm{diag}(\boldsymbol{b})\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\boldsymbol{\beta}_0 - \boldsymbol{Y}\right\|^2 + \frac{1}{p}\mathrm{tr}\left(2\boldsymbol{Z}^T(\boldsymbol{I}_k - \mathrm{diag}(\boldsymbol{b}))\boldsymbol{\Lambda}\boldsymbol{Z} - \boldsymbol{A}\right),$$

- Semi-parametric URE estimator

$$\hat{\boldsymbol{\theta}}^{\boldsymbol{b},\boldsymbol{\beta}_0} = \boldsymbol{Z}^T\left((\boldsymbol{I}_k - \mathrm{diag}(\boldsymbol{b}))\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\hat{\boldsymbol{\beta}}^{\mathrm{WLS}} + \mathrm{diag}(\boldsymbol{b})\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\boldsymbol{\beta}_0\right)$$

$$\left(\hat{\boldsymbol{b}}_{SP}^{\mathrm{URE}}, \left(\hat{\boldsymbol{\beta}}_0\right)_{SP}^{\mathrm{URE}}\right) = \underset{\boldsymbol{b}\in\mathrm{MON}(\boldsymbol{D}),\ \boldsymbol{\beta}_0\in\mathbb{R}^k}{\mathrm{argmin}} \mathrm{URE}^{SP}(\boldsymbol{b},\boldsymbol{\beta}_0)$$

Optimality (Model I)

Theorem 4. *Assume conditions (A)-(E). Then under Model I we have*

$$\sup_{\boldsymbol{b}\in \mathrm{MON}(\boldsymbol{A}),\, \boldsymbol{\mu}\in\mathscr{L}} \left| \mathrm{URE}^{SP}(\boldsymbol{b},\boldsymbol{\mu}) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{b},\boldsymbol{\mu}}\right)\right| \to 0 \text{ in } L^1 \text{ as } p\to\infty.$$

As a corollary, for any estimator $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p,\hat{\boldsymbol{\mu}}_p} = (\boldsymbol{I}_p - \mathrm{diag}(\hat{\boldsymbol{b}}_p))\boldsymbol{Y} + \mathrm{diag}(\hat{\boldsymbol{b}}_p)\hat{\boldsymbol{\mu}}_p$ *with* $\hat{\boldsymbol{b}}_p \in \mathrm{MON}(\boldsymbol{A})$ *and* $\hat{\boldsymbol{\mu}}_p \in \mathscr{L}$, *we always have*

$$\lim_{p\to\infty} \mathbb{P}\left(l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}_{SP}\right) \geq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p,\hat{\boldsymbol{\mu}}_p}\right) + \varepsilon\right) = 0 \quad \forall \varepsilon > 0,$$

$$\limsup_{p\to\infty} \left(R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}_{SP}\right) - R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p,\hat{\boldsymbol{\mu}}_p}\right)\right) \leq 0.$$

Optimality (Model II)

Theorem 5. *Assume conditions (A), (D)-(G). Then under Model II we have*

$$\sup_{\boldsymbol{b}\in \mathrm{MON}(\boldsymbol{D}),\, \boldsymbol{X}^T\boldsymbol{\beta}_0\in\mathscr{L}} \left|\mathrm{URE}^{SP}(\boldsymbol{b},\boldsymbol{\beta}_0) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{b},\boldsymbol{\beta}_0}\right)\right| \to 0 \text{ in } L^1 \text{ as } p\to\infty.$$

As a corollary, for any estimator $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p,\hat{\boldsymbol{\beta}}_{0,p}}$ *obtained from (16) with* $\hat{\boldsymbol{b}}_p \in \mathrm{MON}(\boldsymbol{D})$ *and* $\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_0 \in \mathscr{L}$, *we always have*

$$\lim_{p\to\infty} \mathbb{P}\left(l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}_{SP}\right) \geq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p,\hat{\boldsymbol{\beta}}_{0,p}}\right) + \varepsilon\right) = 0 \quad \forall \varepsilon > 0,$$

$$\limsup_{p\to\infty}\left(R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}_{SP}\right) - R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p,\hat{\boldsymbol{\beta}}_{0,p}}\right)\right) \leq 0.$$

Real data: Baseball

- Brown (2008) collects statistics of 567 MLB players of the 2005 season
- Task: Given first half season performance, predict the second half season batting average
- Bionomial $H_{ij} \sim Binomial(N_{ij}, p_j)$
- After arcsin transformation approximately normal

$$Y_{ij} = \arcsin \sqrt{\frac{H_{ij} + 1/4}{N_{ij} + 1/2}}$$

$$Y_{ij} \dot{\sim} N(\theta_j, (4N_{ij})^{-1}),\ \theta_j = \arcsin \sqrt{p_j}$$

Real data: Baseball

	All		Pitchers		Non-Pitchers	
p for estimation	567		81		486	
p for validation	499		64		435	
Covariates?	No	Yes	No	Yes	No	Yes
Naive	1	NA	1	NA	1	NA
Ordinary least squares (OLS)	0.852	0.242	0.127	0.115	0.378	0.333
Weighted least squares (WLS)	1.074	0.219	0.127	0.087	0.468	0.290
Parametric EBMOM	0.593	0.194	0.129	0.117	0.387	**0.256**
Parametric EBMLE	0.902	0.207	0.117	0.096	0.398	0.277
James-Stein	0.525	**0.184**	0.164	0.142	0.359	0.262
Parametric URE toward OLS	0.505	0.203	0.123	0.124	0.278	0.300
Parametric URE toward WLS	0.629	0.188	0.127	0.112	0.385	0.268
Parametric URE	0.422	0.215	0.123	0.130	0.282	0.310
Semiparametric URE toward OLS	0.409	0.197	0.081	0.097	0.261	0.299
Semiparametric URE toward WLS	0.499	**0.184**	0.098	**0.083**	0.336	**0.256**
Semiparametric URE	0.419	0.201	0.077	0.126	0.278	0.314

Real data: Baseball

	All		Pitchers		Non-Pitchers	
p for estimation	567		81		486	
p for validation	499		64		435	
Covariates?	No	Yes	No	Yes	No	Yes
Naive	1	NA	1	NA	1	NA
Ordinary least squares (OLS)	0.852	0.242	0.127	0.115	0.378	0.333
Weighted least squares (WLS)	1.074	0.219	0.127	0.087	0.468	0.290
Parametric EBMOM	0.593	0.194	0.129	0.117	0.387	**0.256**
Parametric EBMLE	0.902	0.207	0.117	0.096	0.398	0.277
James-Stein	0.525	**0.184**	0.164	0.142	0.359	0.262
Parametric URE toward OLS	0.505	0.203	0.123	0.124	0.278	0.300
Parametric URE toward WLS	0.629	0.188	0.127	0.112	0.385	0.268
Parametric URE	0.422	0.215	0.123	0.130	0.282	0.310
Semiparametric URE toward OLS	0.409	0.197	0.081	0.097	0.261	0.299
Semiparametric URE toward WLS	0.499	**0.184**	0.098	**0.083**	0.336	**0.256**
Semiparametric URE	0.419	0.201	0.077	0.126	0.278	0.314

Real data: Baseball

	All		Pitchers		Non-Pitchers	
p for estimation	567		81		486	
p for validation	499		64		435	
Covariates?	No	Yes	No	Yes	No	Yes
Naive	1	NA	1	NA	1	NA
Ordinary least squares (OLS)	0.852	0.242	0.127	0.115	0.378	0.333
Weighted least squares (WLS)	1.074	0.219	0.127	0.087	0.468	0.290
Parametric EBMOM	0.593	0.194	0.129	0.117	0.387	**0.256**
Parametric EBMLE	0.902	0.207	0.117	0.096	0.398	0.277
James-Stein	0.525	**0.184**	0.164	0.142	0.359	0.262
Parametric URE toward OLS	0.505	0.203	0.123	0.124	0.278	0.300
Parametric URE toward WLS	0.629	0.188	0.127	0.112	0.385	0.268
Parametric URE	0.422	0.215	0.123	0.130	0.282	0.310
Semiparametric URE toward OLS	0.409	0.197	0.081	0.097	0.261	0.299
Semiparametric URE toward WLS	0.499	**0.184**	0.098	**0.083**	0.336	**0.256**
Semiparametric URE	0.419	0.201	0.077	0.126	0.278	0.314

Conclusion

- We consider shrinkage estimate for heteroscedastic hierarchical Models (beyond Gaussian, linear models)
- Propose URE estimate: Par. and semi-par.
- Good theoretical properties of URE estimate
  - Oracle property
  - Asymptotically optimal
- Competitive numerical performance

Acknowledgement

- Larry Brown
- Xianchao Xie, Justin Yang
- NSF, NIH

Thank you!

References

- X. Xie, S. C. Kou and L. D. Brown (2012). SURE estimates for a heteroscedastic hierarchical model. *J. Amer. Statist. Assoc.*, 107, 1465-1479.

- X. Xie, S. C. Kou and L. D. Brown (2016). Optimal shrinkage estimation of mean parameters in family of distributions with quadratic variance. *Ann. Statist.*, 44, 564-597.

- S. C. Kou and J. J. Yang (2017). Optimal shrinkage estimation in heteroscedastic hierarchical linear models. In *Big and Complex Data Analysis: Methodologies and Applications*, (edited by S. Ejaz Ahmed), 249-284. Springer, New York.