Application to the Linear Model

$$\begin{aligned} \boldsymbol{\theta} &= (\boldsymbol{\beta}, \sigma^2) \sim \pi(\boldsymbol{\theta}) \\ y_i &= x_i^\top \boldsymbol{\beta} + \epsilon_i, \text{ for } i = 1, \ldots, n \end{aligned}$$

- $T(\boldsymbol{y}) = (\boldsymbol{b}(X, \boldsymbol{y}), s(X, \boldsymbol{y}))$
  - $\boldsymbol{b}(X, \boldsymbol{y}) = (b_1(X, \boldsymbol{y}), \ldots, b_p(X, \boldsymbol{y}))^\top \in \mathbb{R}^p$
  - $s(X, \boldsymbol{y}) \in \{0\} \cup \mathbb{R}^+$
- E.g. M-estimators Huber (1964), Least Median Squares, Least Trimmed Squares

Computational Strategy

- Numerical integration (low dimension), Appeal to asymptotics.
- MCMC: Data augmented Gibbs Sampler targeting $f(\boldsymbol{\theta}, \boldsymbol{y} | T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs}))$
  1. $\pi(\boldsymbol{\theta} | \boldsymbol{y}, T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})) = \pi(\boldsymbol{\theta} | \boldsymbol{y})$ (full poserior)
  2. $f(\boldsymbol{y} | \boldsymbol{\theta}, T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs}))$
- For 2, prosose a Metropolis-Hastings sampler.
- accept/reject a sample full data $\boldsymbol{y} \in \mathcal{A} := \{\boldsymbol{y} \in \mathbb{R}^n | T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})\}$ from a well defined distribution with support $\mathcal{A}$.

Computational Strategy

- Difficult to sample from $\mathcal{A}$ directly
- Can sample $\boldsymbol{z}^* \in \mathbb{R}^n$ and transform:

$$\boldsymbol{y} = h(\boldsymbol{z}^*) := \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)} \boldsymbol{z}^* + X \left(\boldsymbol{b}(X, \boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)} \boldsymbol{z}^*) \right)$$

- With regression/scale equivariance/invariance properties (C3-C8 in the paper) $T(y) = T(y_{obs})$
- Idea:
  - sample $\boldsymbol{z}^* \sim p(\boldsymbol{z}^*)$, transform via h
  - proposal $p(\boldsymbol{y}|\boldsymbol{\theta})$ is then a change-of-variables adjustment on $p(\boldsymbol{z}^*)$.

Computational Strategy

- h is not 1-1/onto. (Change of variables difficult)
- Can restrict sample space of z^*, so that it is
  - $\mathcal{A}$ is an $n - p - 1$ space
  - Sample space for z^*: $\mathbb{S} := \{\boldsymbol{z}^* \in \mathcal{C}^{\perp}(X) \mid ||\boldsymbol{z}^*|| = 1\}$
    - i.e. the unit space in the orthogonal complement of the column space of the design matrix.
  - $h : \mathbb{S} \to \mathcal{A}$ is then 1-1/onto.
  - easier to figure out the Jacobian of the transformation from $p(\boldsymbol{z}^*)$ to $p(\boldsymbol{y}|\boldsymbol{\theta})$

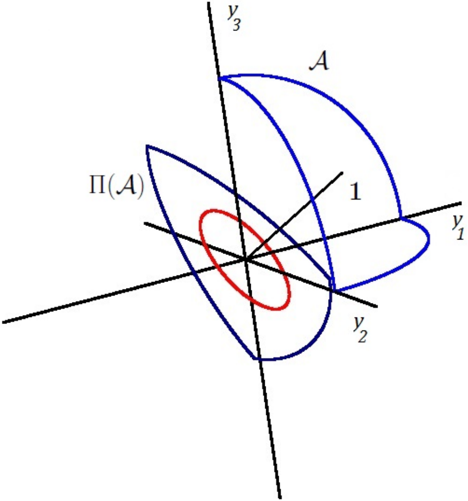
Computational Strategy

For the proposal distribution.

1. Sample $\boldsymbol{z}^*$ from a distribution with known density with support $\mathbb{S}$.
2. Set $\boldsymbol{y} = h(\boldsymbol{z}^*)$
3. Jacobian broken into steps
   - $\boldsymbol{z} = \frac{s(X,\boldsymbol{y}_{obs})}{s(X,\boldsymbol{z}^*)}\boldsymbol{z}^*$.
     - Scale from $\mathbb{S}$ to the set $\Pi(\mathcal{A}) := \{\boldsymbol{z} \in \mathbb{R}^n | \exists\ \boldsymbol{y} \in \mathcal{A}\ s.t.\ \boldsymbol{z} = Q\boldsymbol{y}\}$ with $Q = I - XX^\top$.
   - $\boldsymbol{y} = \boldsymbol{z} + X\left(\boldsymbol{b}(X, \boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \boldsymbol{z})\right)$.
     - Shift from $\Pi(\mathcal{A})$ to $\mathcal{A}$ along $\mathcal{C}(X)$

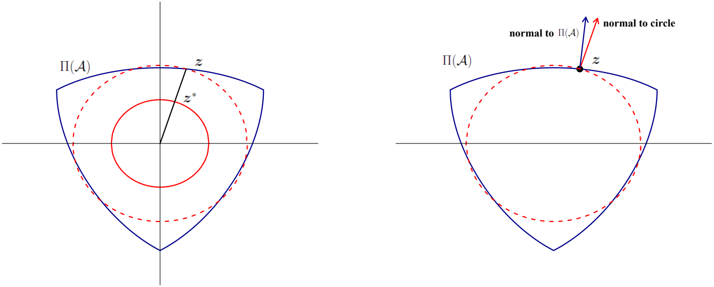
Computational Strategy (Visualization, $n = 3, p = 1$)

$$T(\boldsymbol{y}) = (\min(\boldsymbol{y}), \sum(y_i - \min(\boldsymbol{y}))^2), \; T(\boldsymbol{y}_{obs}) = (0, 1)$$



Computational Strategy (Visualization, $n = 3, p = 1$)

Scaling step



- resize sphere: $r^{-(n-p-1)}$
- deformation onto $\Pi(\mathcal{A})$: $\cos(\gamma) = \frac{\nabla s(X,\boldsymbol{y})^\top \boldsymbol{z}}{\|\nabla s(X,\boldsymbol{y})\|\|\boldsymbol{z}\|}$

Computational Strategy (Visualization, $n = 3, p = 1$)

Shifting step of $\boldsymbol{z}$ to $\boldsymbol{y}$ along the column space of X

- Contribution is the ratio of the infinitesimal volumes along $\Pi(\mathcal{A})$ at $\boldsymbol{z}$ to the corresponding volume along $\mathcal{A}$ at $\boldsymbol{y}$.
- $\mathrm{Vol}(P) := \sqrt{\det(P^\top P)} = \prod_{i=1}^{r} \sigma_i$
  - $P = QA$,
  - columns of A: an orthonormal basis for the tangent space to $\mathcal{A}$ at $\boldsymbol{y}$.
  - $\nabla s(X, \boldsymbol{y}), \nabla b_1(X, \boldsymbol{y}), \ldots, \nabla b_p(X, \boldsymbol{y})$ form basis for the orthogonal complement.
  - σ_i are the singular values of P

Full Jacobian: $p(\boldsymbol{y}) = p(\boldsymbol{z}^*) r^{-(n-p-1)} \cos(\gamma) \mathrm{Vol}(P)$

Real Data: Nationwide Insurance Data

- Nationwide Insurance sells many polices through insurance agencies.
- Agencies provide direct service to policy holders.
- Contractual agreements between Nationwide and the agencies vary.
- Interested in future performance of agencies.

Real Data: Nationwide Insurance Data

- Data grouped by states, varying number of agencies by state.

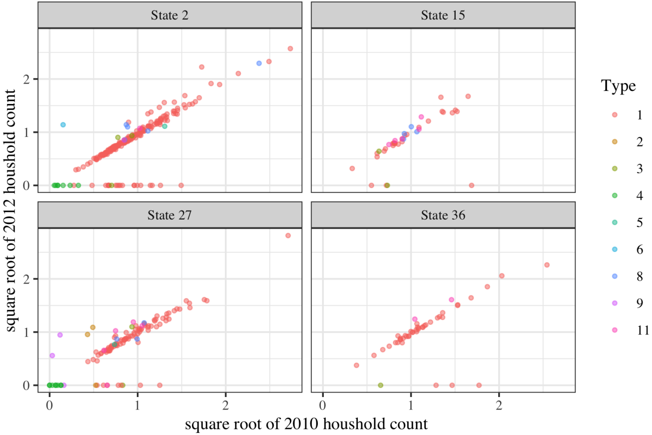


Figure: The square root of (scaled) count in 2012 versus that in 2010 for four states.

Real Data: State Level Regressions

Regression fit seperately within each state

$$\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_0, \Sigma_0); \quad \sigma^2 \sim IG(a_0, b_0); \quad y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

- Covariates: square root of household count in 2010, two different measures of size and experience.
- Response: square root of household count in 2012.
- model misspecification: omitting contract type, closure information.
- many cases to appear "outlying" due to misspecification.

Method of Model Comparison

- Both training and validation data will contain outliers
- Trimming approach with log predictive density (Jung et al., 2014)
- Case i in holdout set: $log(f(y_i))$
- Scoring a procedure:
  - Choose a base method (e.g. Student-t model) and trimming fractions α
  - Order holdout sample by $log(f_b(y_i))$
  - Denote ordering by: $y^b_{(1)}, y^b_{(2)}, \ldots, y^b_{(M)}$
  - score each method with "mean trimmed log margainal psuedo likelihood"

$$TLM_b(A) = (M - [\alpha M])^{-1} \sum_{i=[\alpha M]+1}^{M} \log(f_A(y^b_{(i)})),$$

- f_A - predictive distribution under the method "A" being scored.

Predictive Performance: 50 training/holdouts sets

Evaluation of 'Type 1' agencies (of special interest to the company)

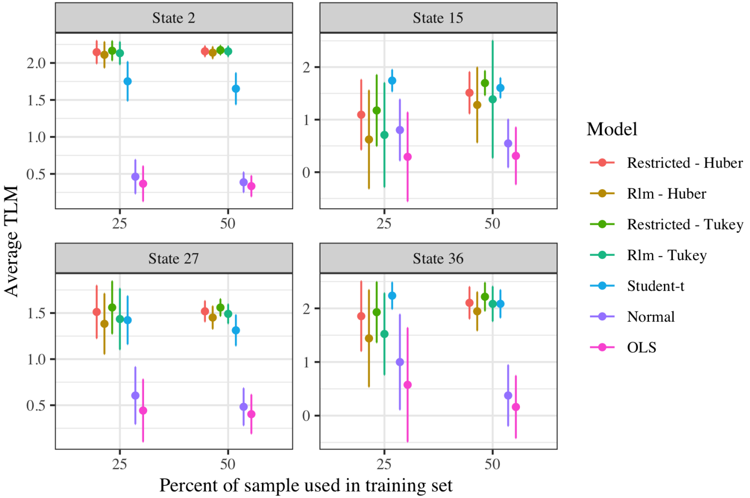


Figure: $\alpha = 0.3$. States 2, 15, 27, and 36, have $n = 222, 40, 117$, and 46

Predictive Performance

- Normal Theory/OLS perform poorly due to not accounting for misspecification
- Small, consistent improvement over classical methods
- variance reduction
- diminishing of effect of prior - similar performance in larger states.
- heavy-tailed model performs worse in the larger states - more outliers appear.

Real Data: Hierarchical Regression

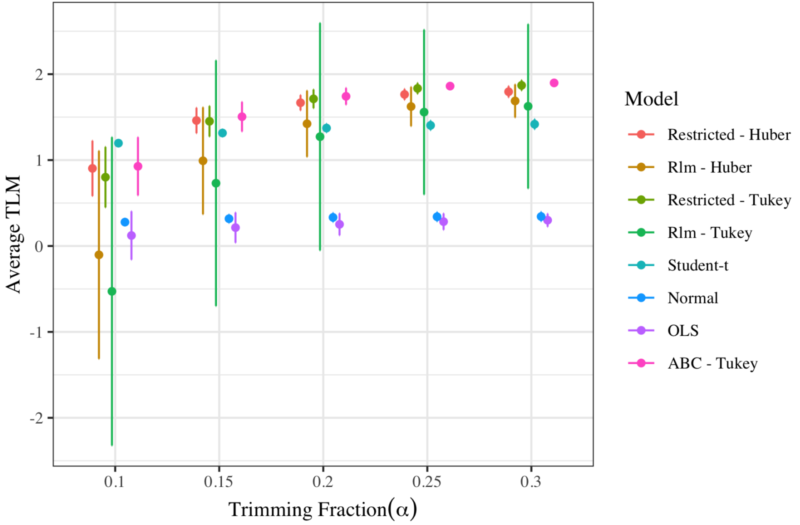
$$\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}_0, a\Sigma_0); \quad \boldsymbol{\beta}_j \stackrel{iid}{\sim} N_p(\boldsymbol{\beta}, b\Sigma_0); \quad \sigma_j^2 \sim IG(a_0, b_0);$$

$$y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j + \epsilon_{ij}, \quad \epsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_j^2), \; i = 1, \ldots, n_j, \; j = 1, \ldots, J$$

Predictive Performance

Average of State-level performance

$$\overline{TLM}_b(A)_{\cdot} = \frac{1}{J}\sum_{j=1}^{J} TLM_b(A)_j$$

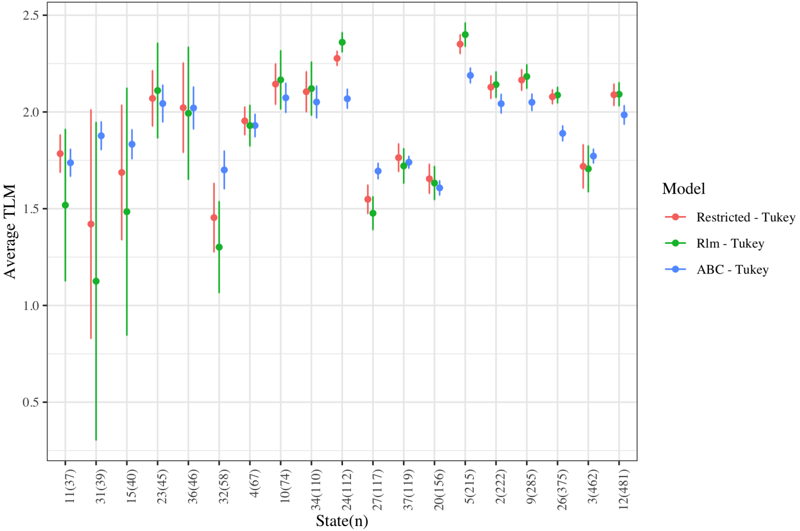


Predictive Performance - Notes

- t-model's poorer performance attributed to the heavier tails - no natural mechanism for prediction.
- Bayesian versions outperform classical robust counterparts. Also reduction in variance.
- ABC also performs well - perhaps better, but this can be attributed to a single state.

Predictive Performance by State

- restricted likelihood average TLM is larger than ABC in 14 of the 20 states, median difference of 0.04



Predictive Performance by State - Notes

- Our method tends to perform well in the smaller states in comparison to classical counter part
- Similar performance in larger states - as expected
- Better than ABC in 14/20 states. Better overall performance of ABC attributed to a single state (31).