The Blended Paradigm: Robust Bayesian Modeling using Non-Sufficient Statistics

John Lewis
Joint work with Yoonkyung Lee, Steven MacEachern

The Ohio State University With Support from:
The Nationwide Center for Advanced Customer Insights
NSF

Tuesday, April 8, 2014

Data analysis and Bayesian methods

- The Bayesian paradigm is optimal, and unquestionably the correct way to make inference–except for a few troublesome details
  - Choice of prior distribution
  - Choice of loss function
  - Choice of likelihood
- For each of these pieces, we should be concerned about the impact of misspecification on understanding and inference
- The likelihood is the focus of this talk
  - Particularly problematic – doesn't 'wash out' with large samples

Example: Large Corporate Data Bases

- Nationwide Insurance Company offers a broad array of finance and insurance products. These products have historically been sold and serviced through a network of agents
- Understanding drivers of agency growth and forecasting performance is of great interest
- Focus here is on the number of households having policies through a given agency
- Salient features of the data include the closure of some agencies (e.g., through retirement of the principal agent), varying contractual details with agents, and differing behavior by state

Household count by agency

- Handling outliers: discard, model directly
- Model misspecification: outliers arise from transitory phenomena
- Also:
  - Imperfect data quality, from various sources; sliding definitions of categories/terms; constantly changing systems (government regulations, pricing policies, competition)

The Blended Paradigm

- Strategy: Decide when an attempt to model and include a phenomenon will result in an overall deterioration of inference. Trim this bit of the model or this part of the data
- Framework for restricted likelihood

$$f(y|\theta) = f(T(y)|\theta) f(y|T(y), \theta)$$

- Look for a good $T(y)$ (non-sufficient)
- Drop the term $f(y|T(y), \theta)$
  - What is a good statistic? Often, one for which the final inference is relatively insensitive to plausible variations in the likelihood $f(y|\theta)$
- Connected to other restricted likelihood methods (Hoff, rank likelihood; Clarke, mean likelihood; ABC; many more)

The Blended Paradigm

- Base Model: $\theta \sim \pi(\theta),\ y \sim f(y|\theta)$
  - Full posterior:
$$\pi(\theta|y) \propto \pi(\theta) f(y|\theta)$$
  - Restricted (blended) posterior:
$$\pi(\theta|T(y)) \propto \pi(\theta) f(T(y)|\theta)$$

Implementation

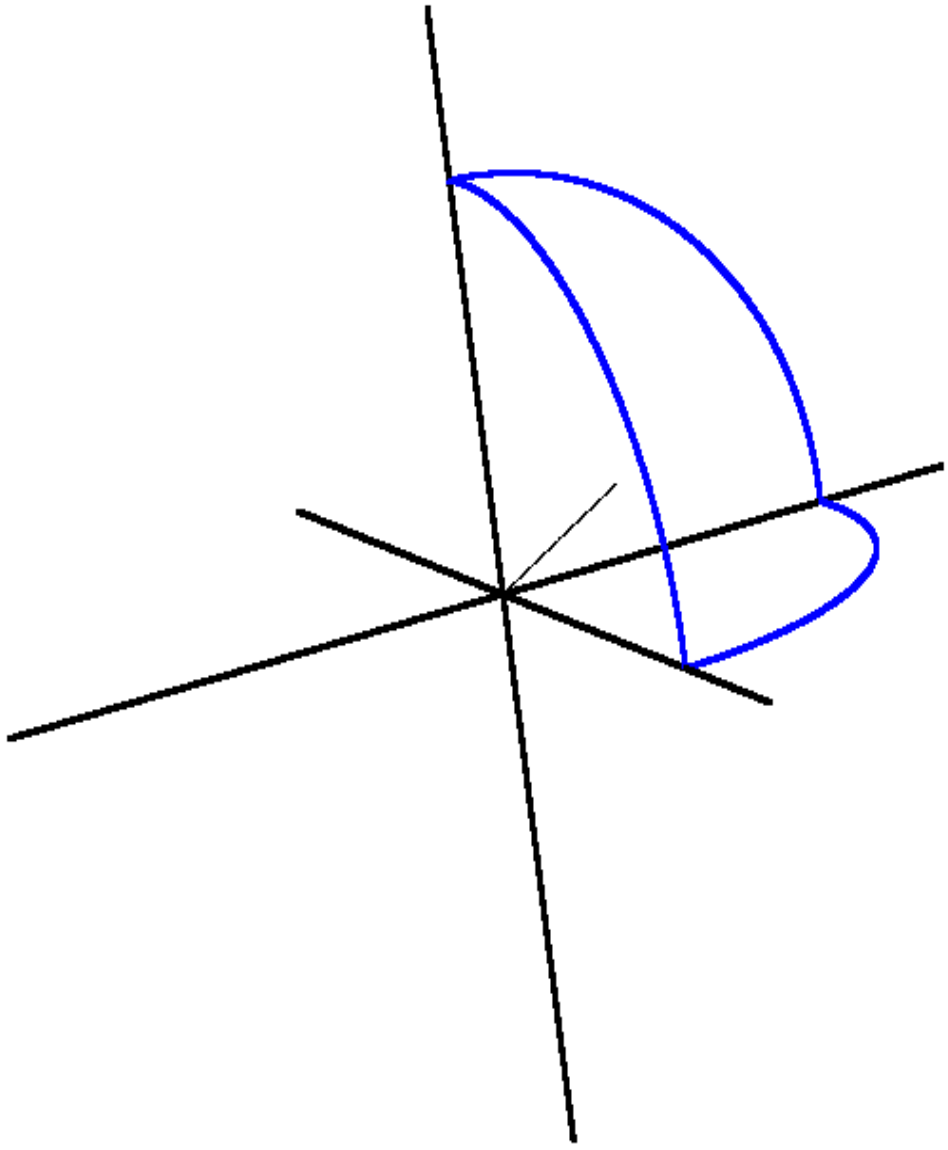
- The blended paradigm models are rarely conjugate, computational methods are needed to fit them
  - **Low dimensional problems**-grid estimation techniques
  - **High dimensional problems**-MCMC
    - Any standard algorithm as the base
    - Data augmentation $[\mathbf{y}|T(\mathbf{y}), \theta]$
    - Step may be simple (order statistics, trimming) or difficult (M-estimators)

MCMC

- Data augmentation – fill in $\mathbf{y}$ from the appropriate conditional distribution
  - Metropolis-Hastings step to move from one augmentation to another
  - Conditioning statistic in regression setting: $T(\mathbf{y}) = (\hat{\beta}, \hat{\sigma}^2)$
- Evaluation of the proposal density is tricky as one must match observed estimates of β and σ^2
  - Using initial proposal: rescale and recenter to match the observed estimates
  - Need density of the resulting proposal, accounting for Jacobian

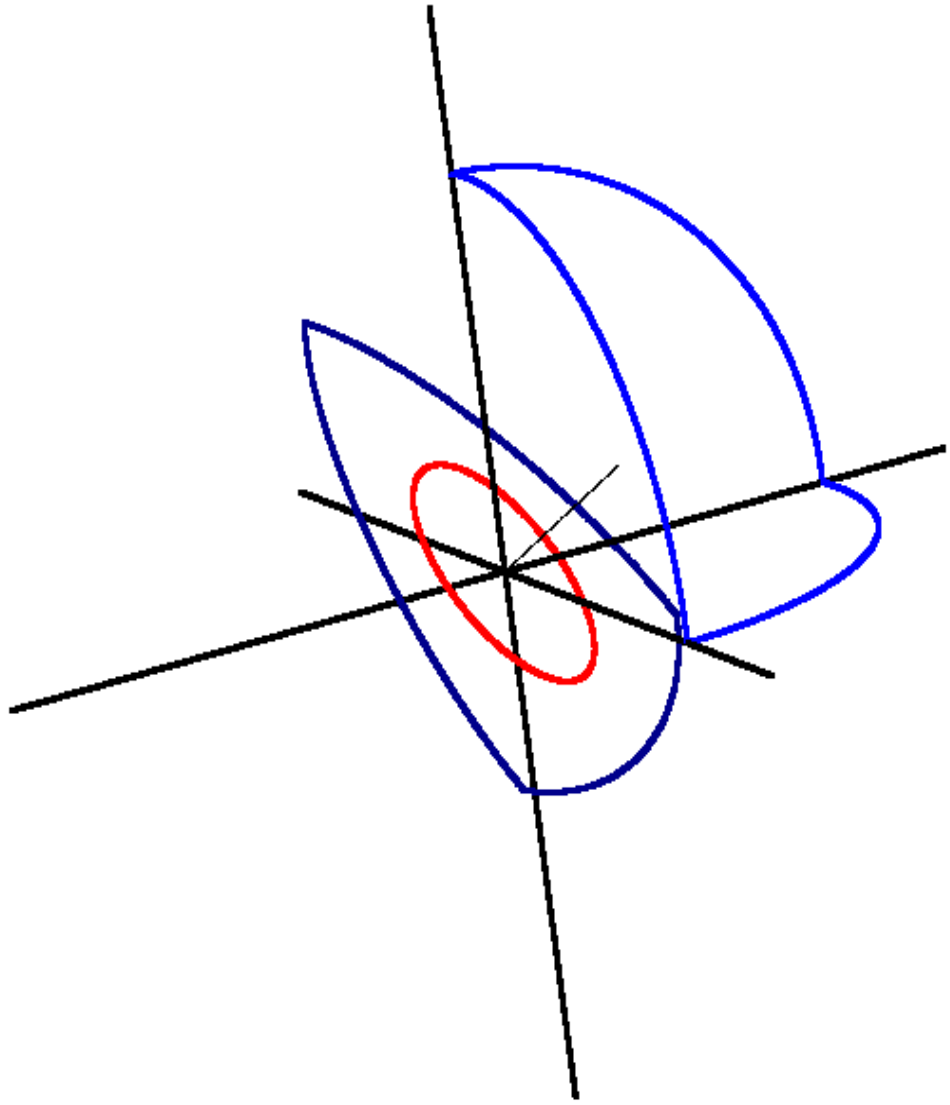
Example – $T(\mathbf{y}) : \min \mathbf{y} = 0, \sum (y_i - \min \mathbf{y})^2 = 1$

- n=3



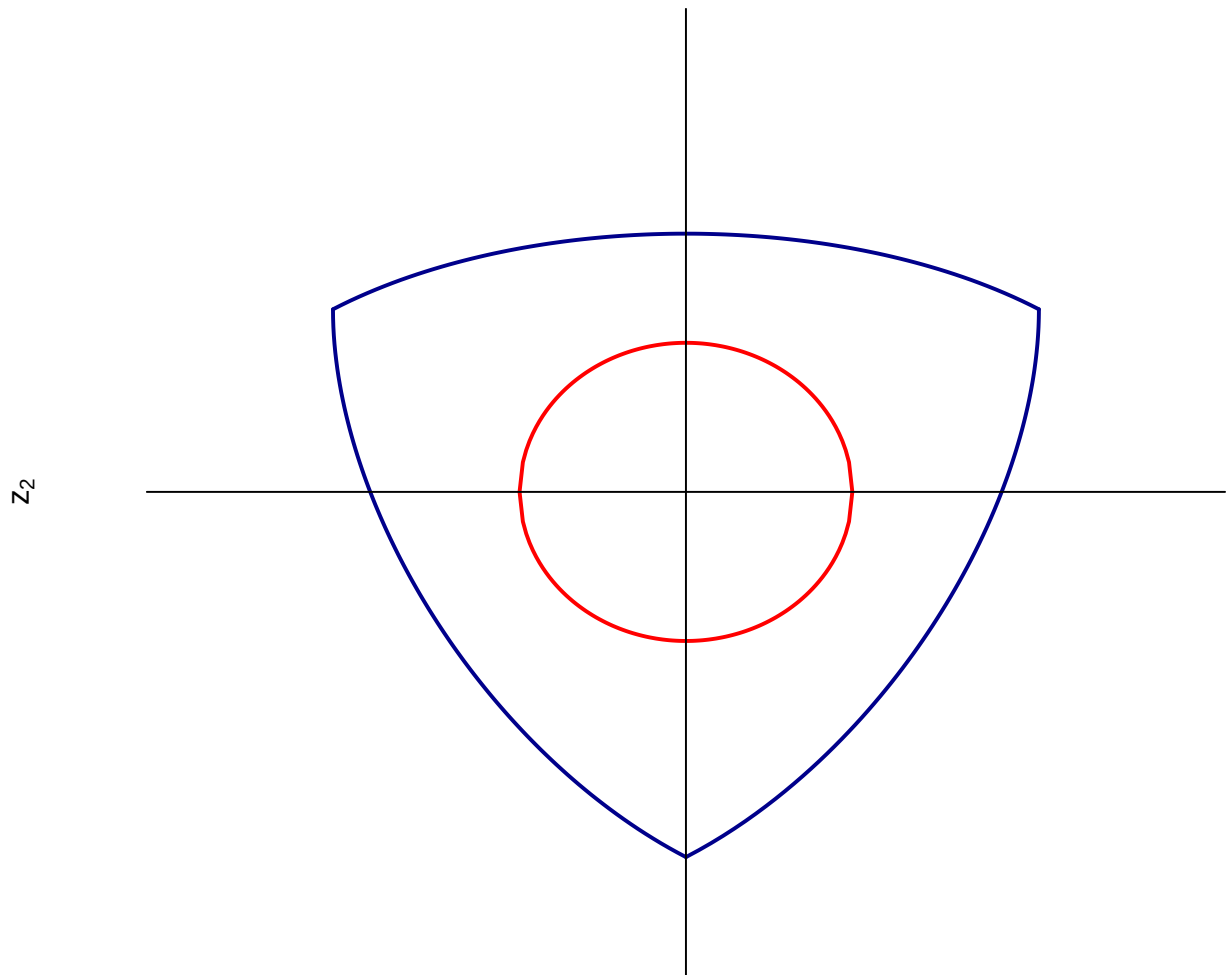
Example – $T(\mathbf{y}) : \min \mathbf{y} = 0, \sum (y_i - \min \mathbf{y})^2 = 1$

- n=3



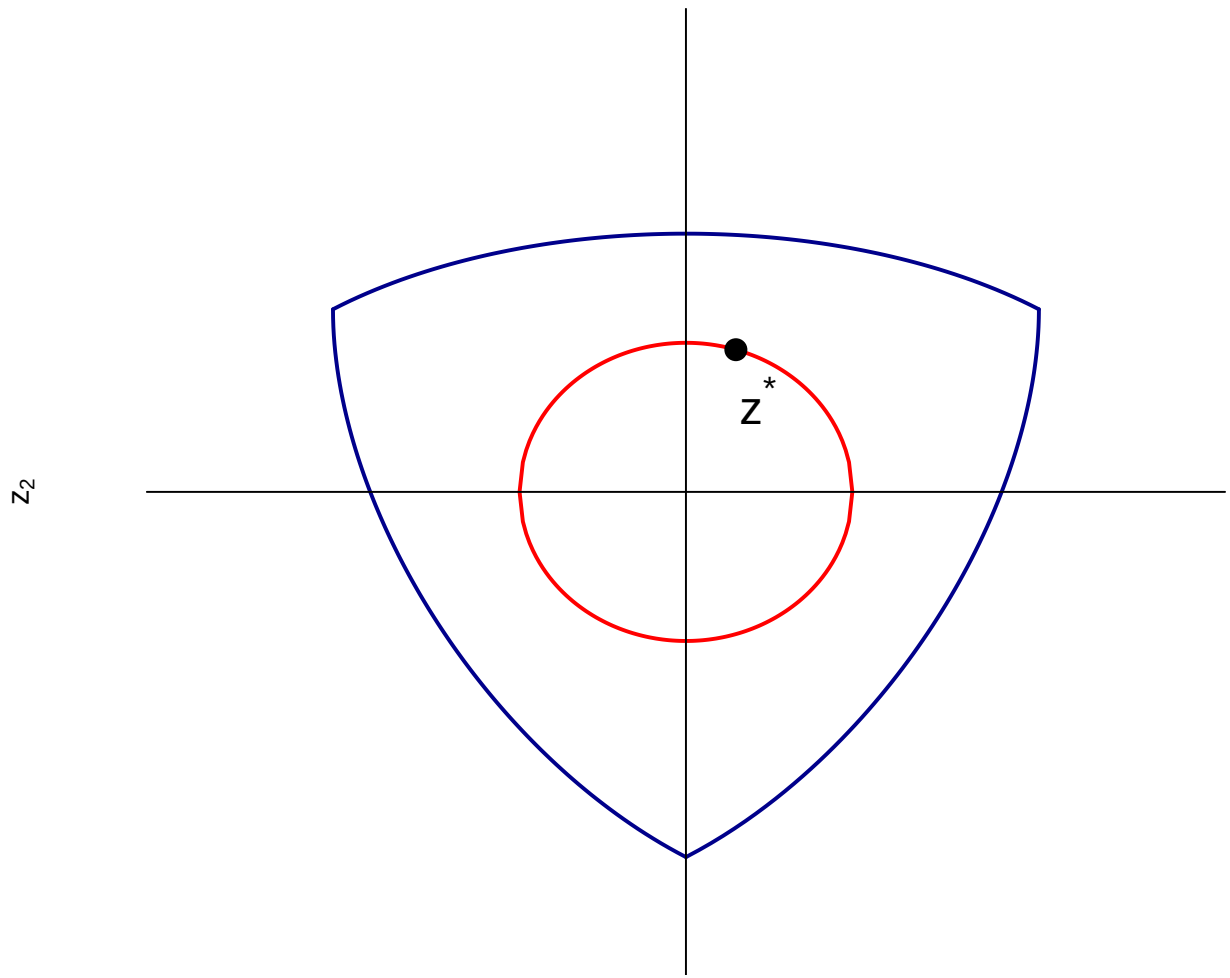
Example – $T(\mathbf{y}) : \min \mathbf{y} = 0, \sum (y_i - \min \mathbf{y})^2 = 1$

- A two dimensional view of the plane



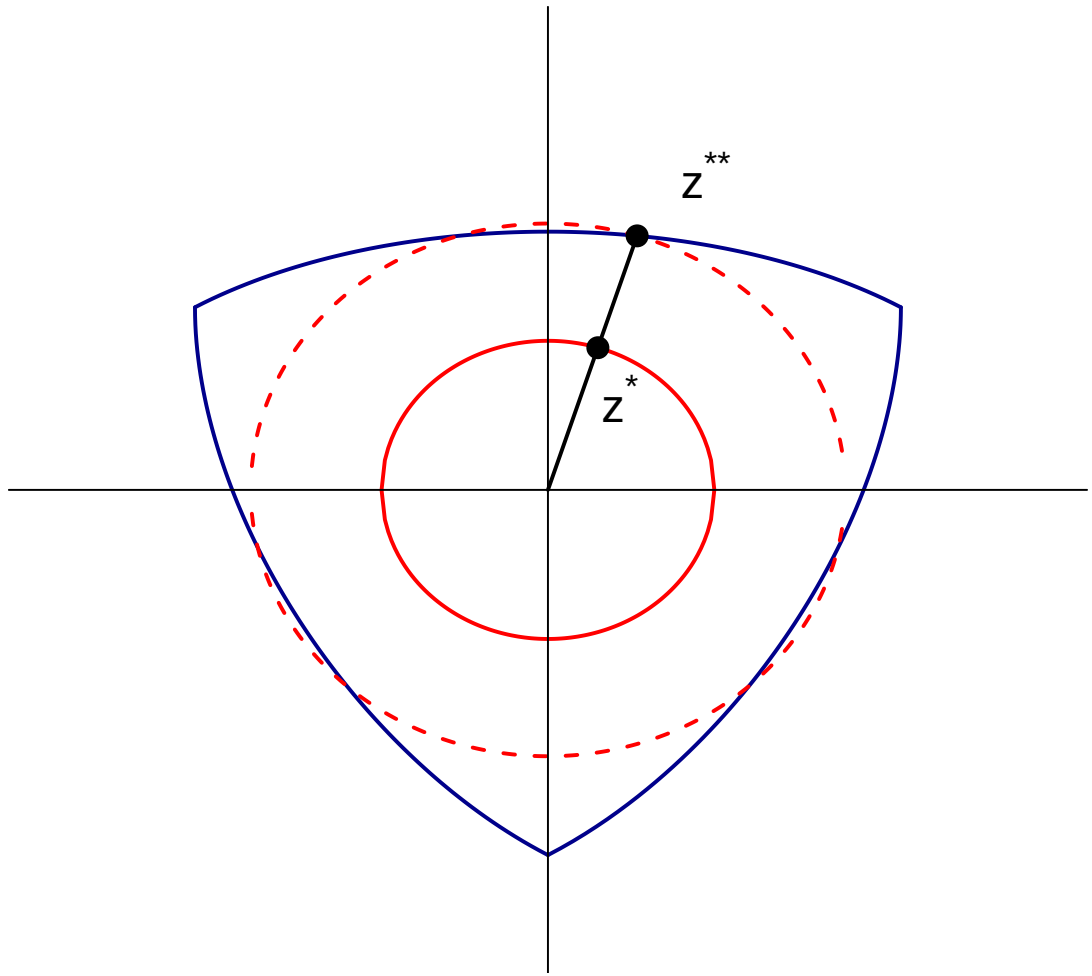
Example – $T(\mathbf{y}) : \min \mathbf{y} = 0, \sum (y_i - \min \mathbf{y})^2 = 1$

- A two dimensional view of the plane



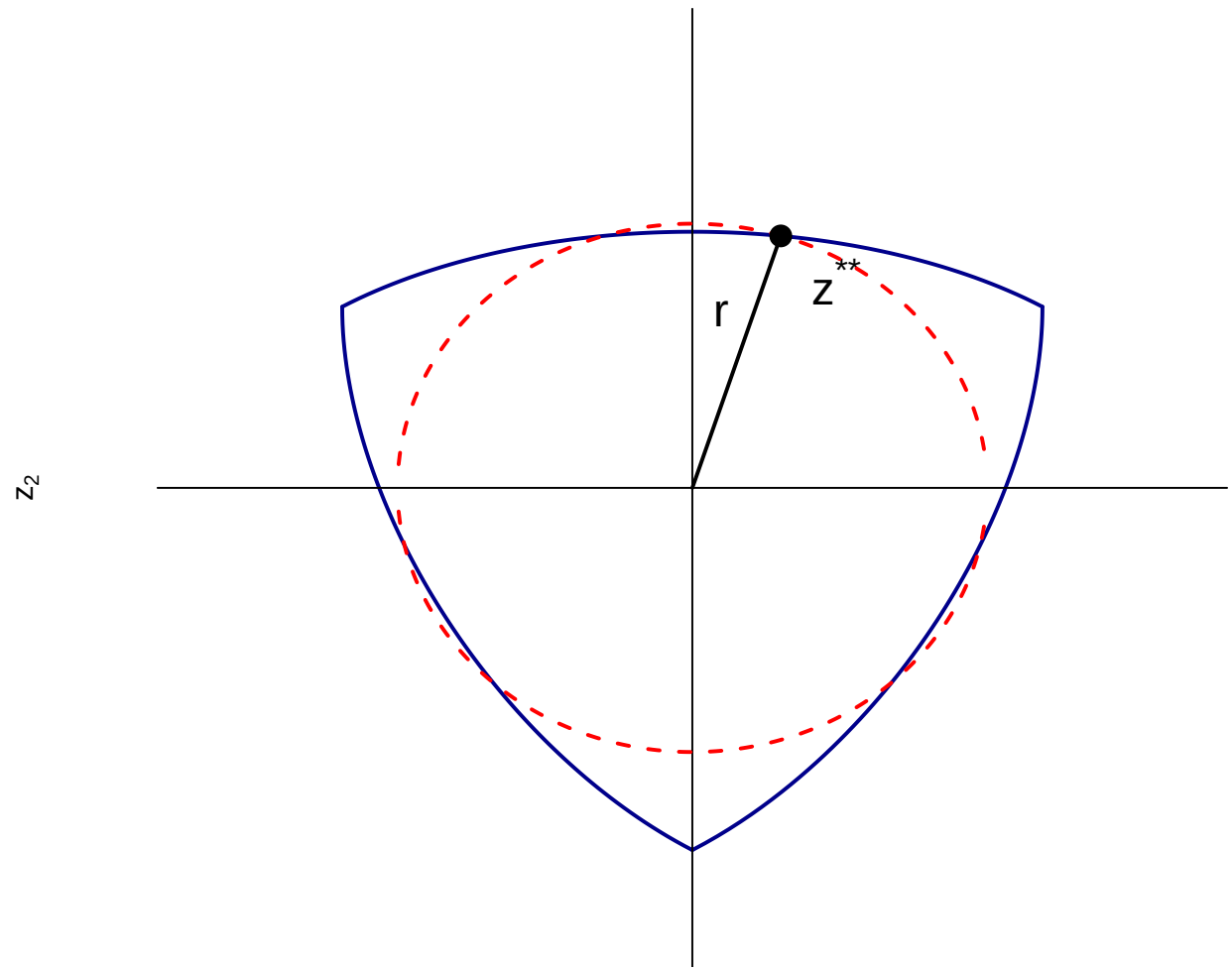
Example – $T(\mathbf{y}) : \min \mathbf{y} = 0, \sum (y_i - \min \mathbf{y})^2 = 1$

- A two dimensional view of the plane



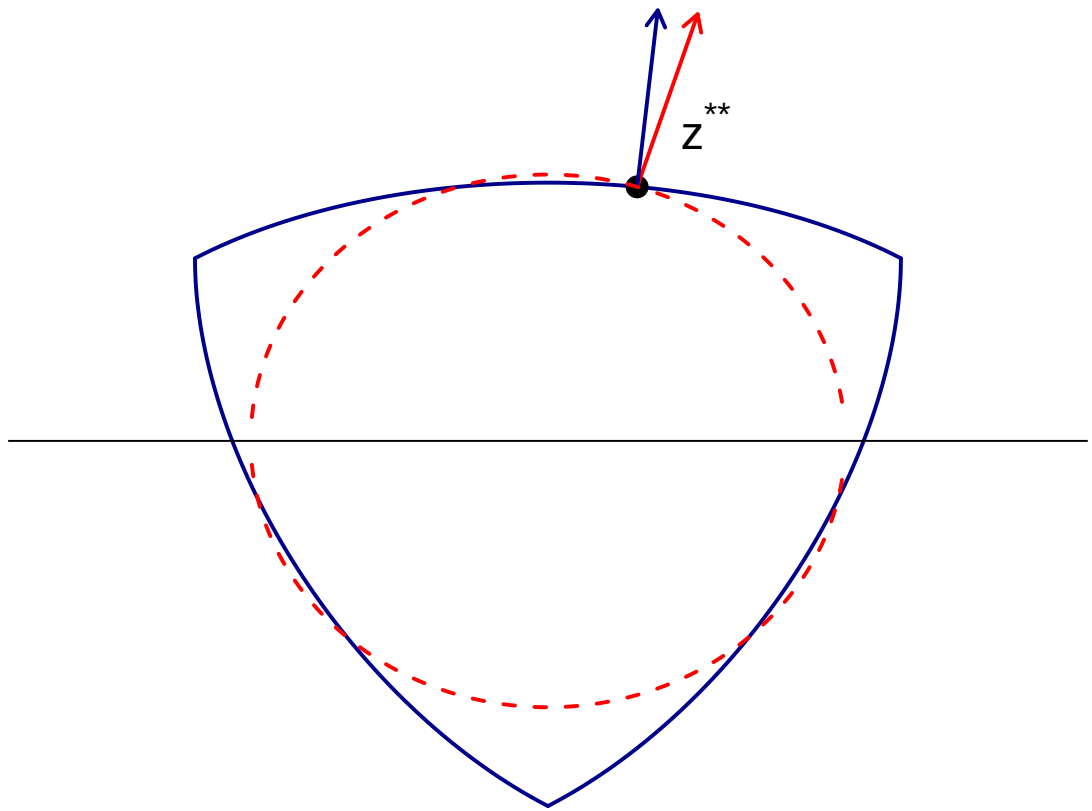
Example – $T(\mathbf{y}) : \min \mathbf{y} = 0, \sum (y_i - \min \mathbf{y})^2 = 1$

- A two dimensional view of the plane



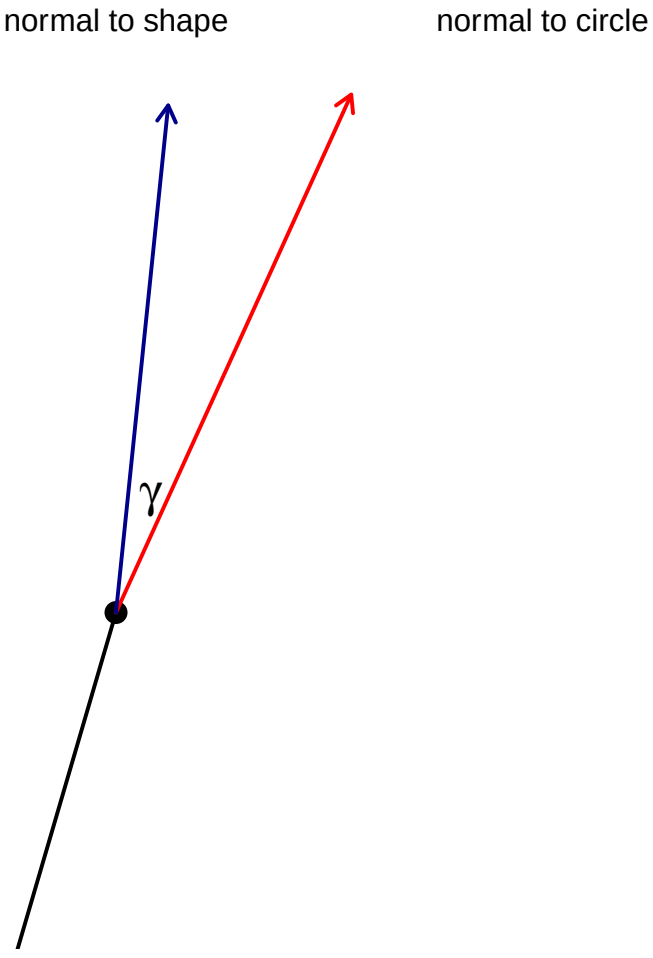
Example – $T(\mathbf{y}) : \min \mathbf{y} = 0, \sum (y_i - \min \mathbf{y})^2 = 1$

- A two dimensional view of the plane



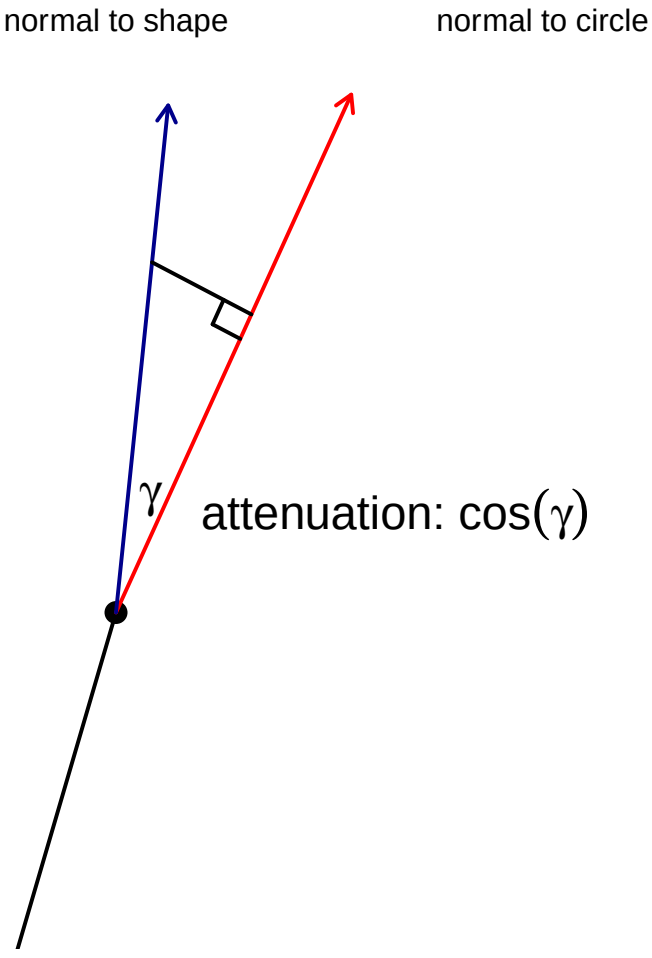
Example – $T(\mathbf{y}) : \min \mathbf{y} = 0, \sum (y_i - \min \mathbf{y})^2 = 1$

- A two dimensional view of the plane



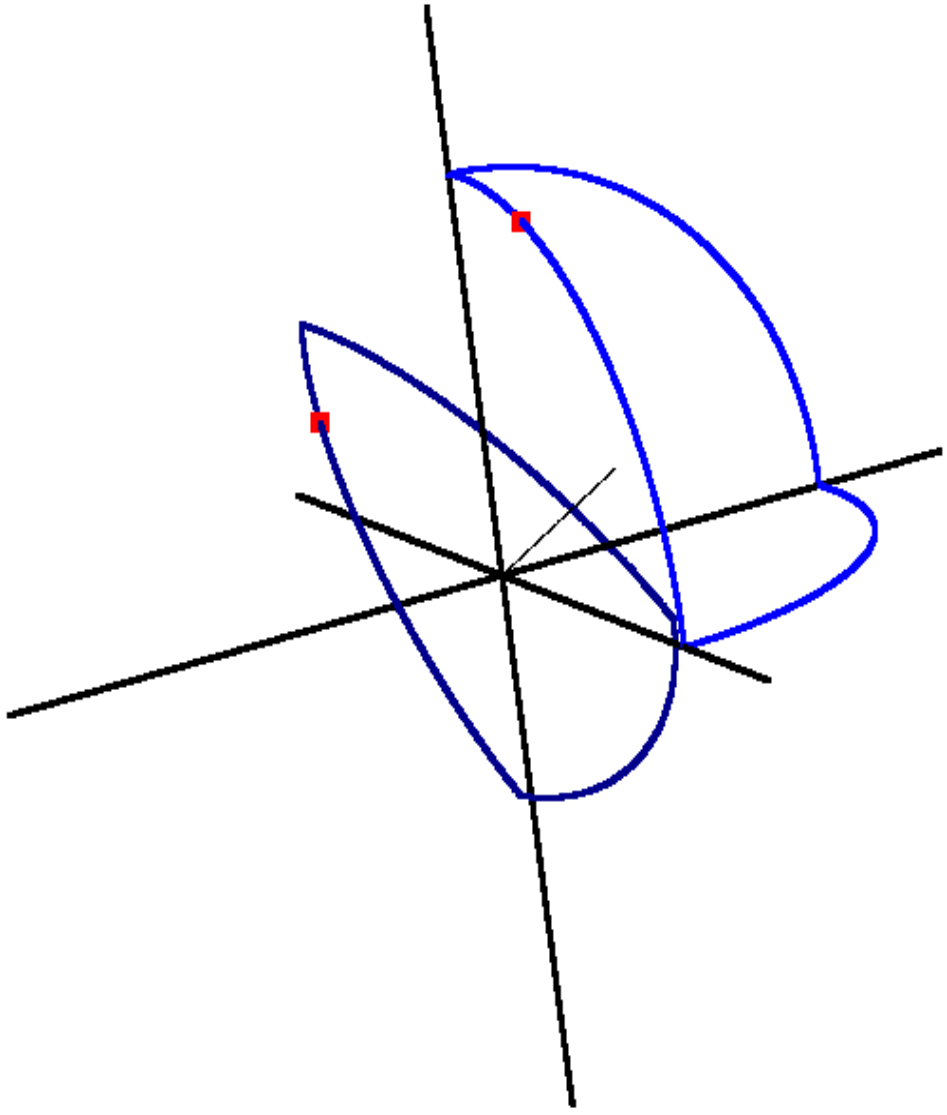
Example – $T(\mathbf{y}) : \min \mathbf{y} = 0, \sum (y_i - \min \mathbf{y})^2 = 1$

- A two dimensional view of the plane



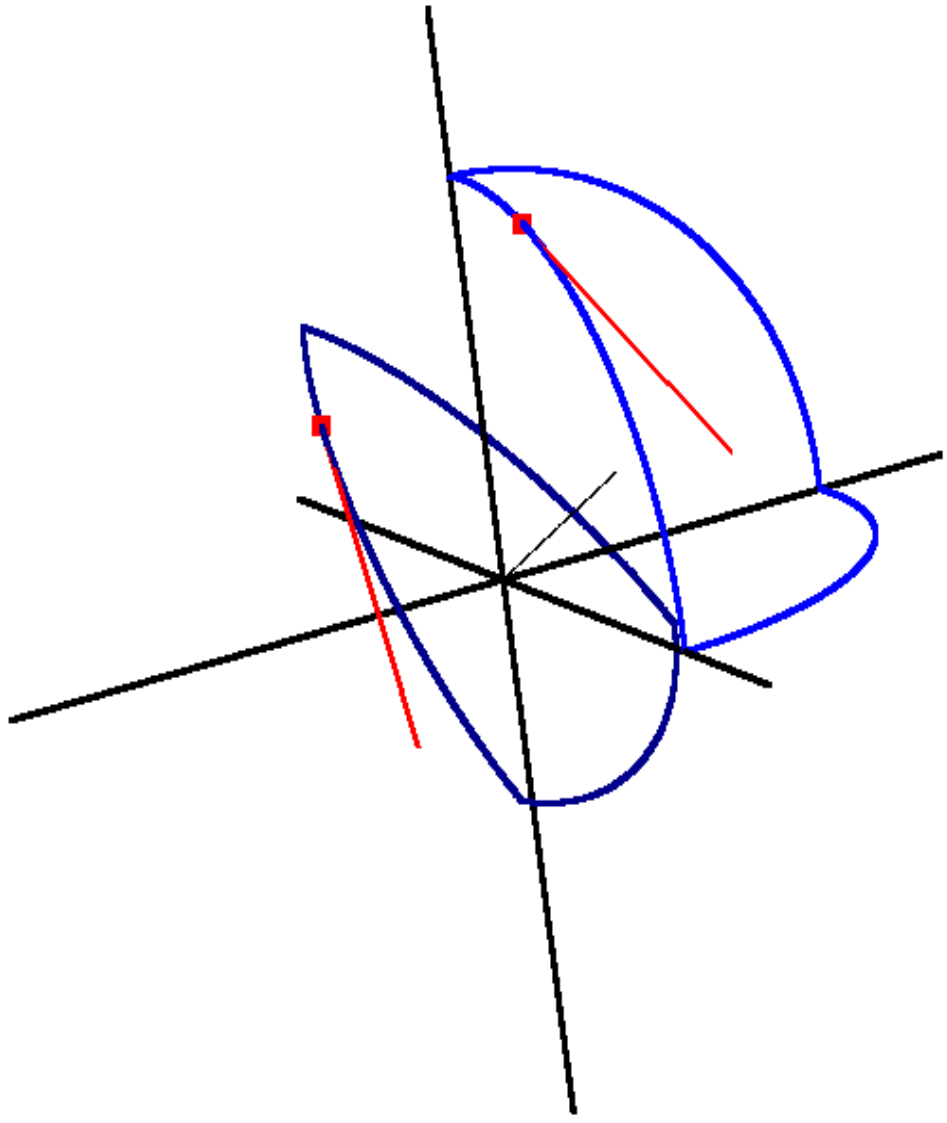
Example – $T(\mathbf{y}) : \min \mathbf{y} = 0, \sum (y_i - \min \mathbf{y})^2 = 1$

- A view in the full space



Example – $T(\mathbf{y}) : \min \mathbf{y} = 0, \sum (y_i - \min \mathbf{y})^2 = 1$

- A view in the full space



Modeling household count by agency

- A standard, normal-theory regression model as the base

$$\beta \sim \mathsf{N}(\mu, \sigma^2\Sigma_0); \qquad \sigma^2 \sim \mathsf{IG}(a_0, b_0)$$
$$\mathbf{y} = X\beta + \epsilon; \qquad \epsilon \sim \mathsf{N}(0, \sigma^2 I)$$

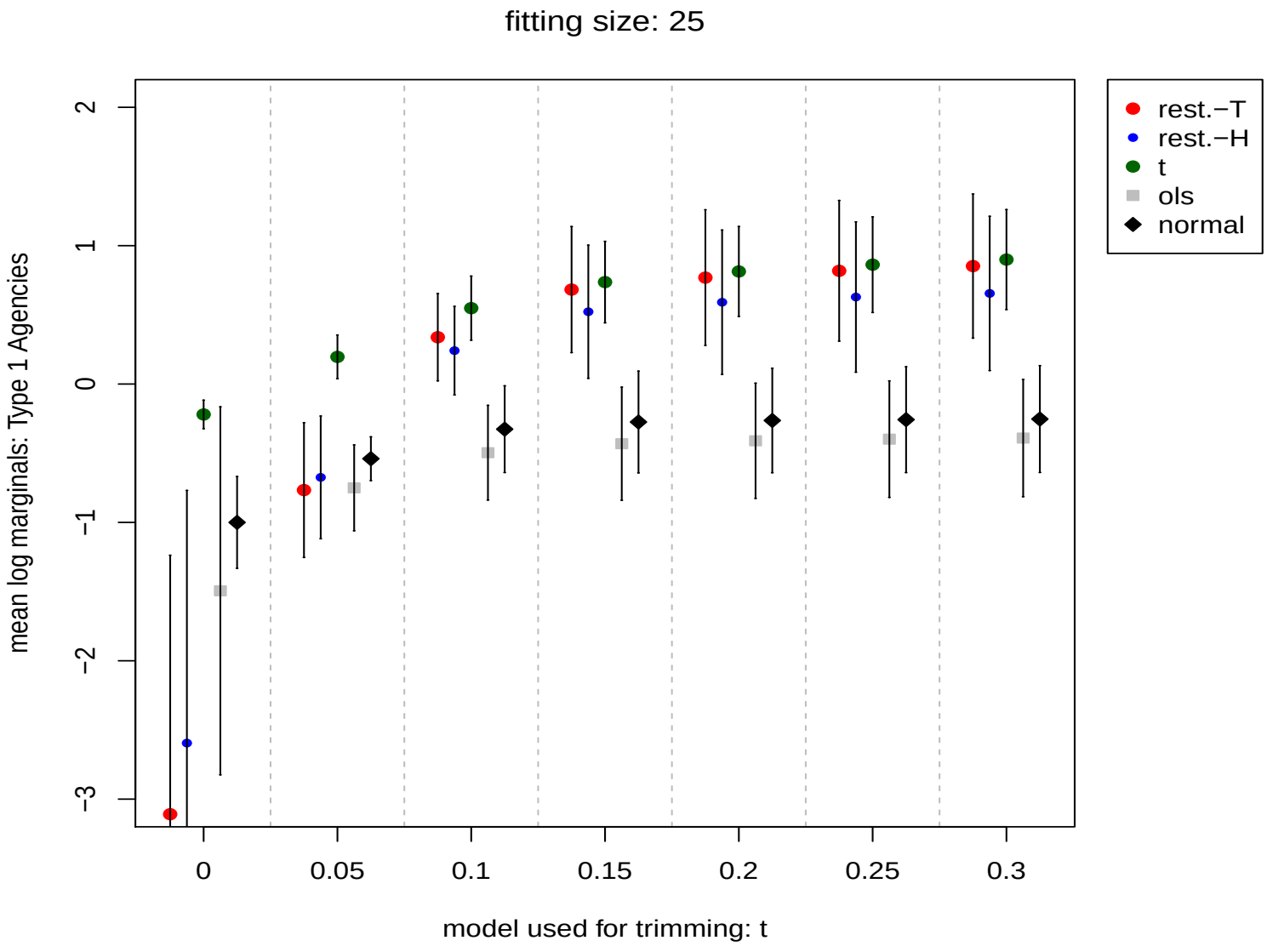
- The data
  - Household count square-rooted to stabilize variance
  - Covariates consist of three measures of agency size
  - Covariates and response centered and scaled to anonymize data
- Analyses include
  - full-likelihood analysis, restricted-likelihood analysis (Huber and Tukey M-estimators), thick-tailed model

Cross-validation study

- Fit on a random sample of data
- Predict on the holdout set; compute mean of log marginals across holdout set as a measure of model fit
- Repeat 100 times, collect and average the model fit measures from each fold
- Model evaluation is tricky
  - Interest primarily lies in Type 1 agencies (most numerous)
  - Holdout sets also contain outliers; do we want to include these in the evaluation?
  - Trim lowest log marginals (according to a single model) before calculating average (use several trimming proportions)

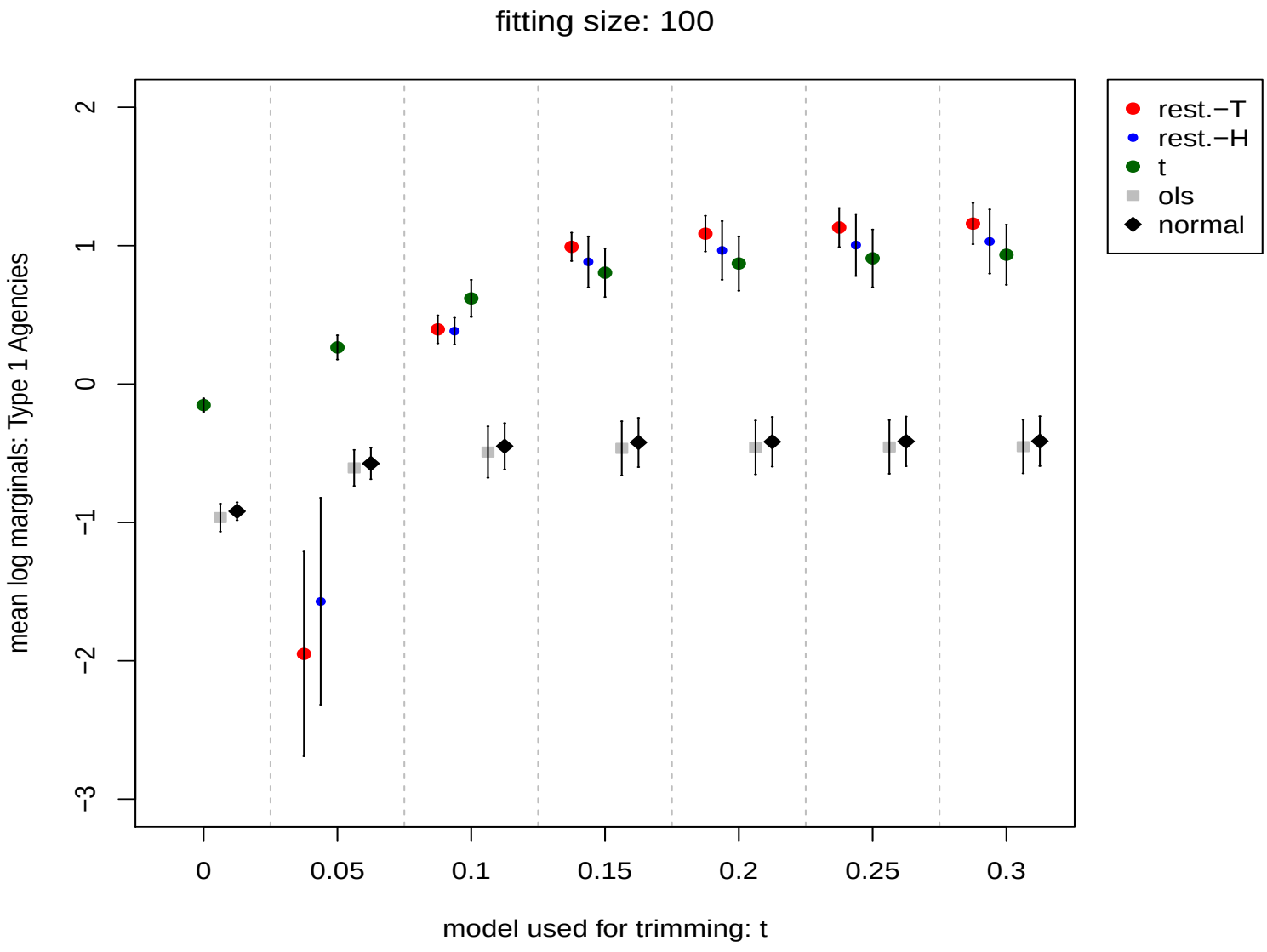
log Marginals-Type 1 Agencies

Blended-Tukey; Blended-Huber; thick tailed-t; normal; ols



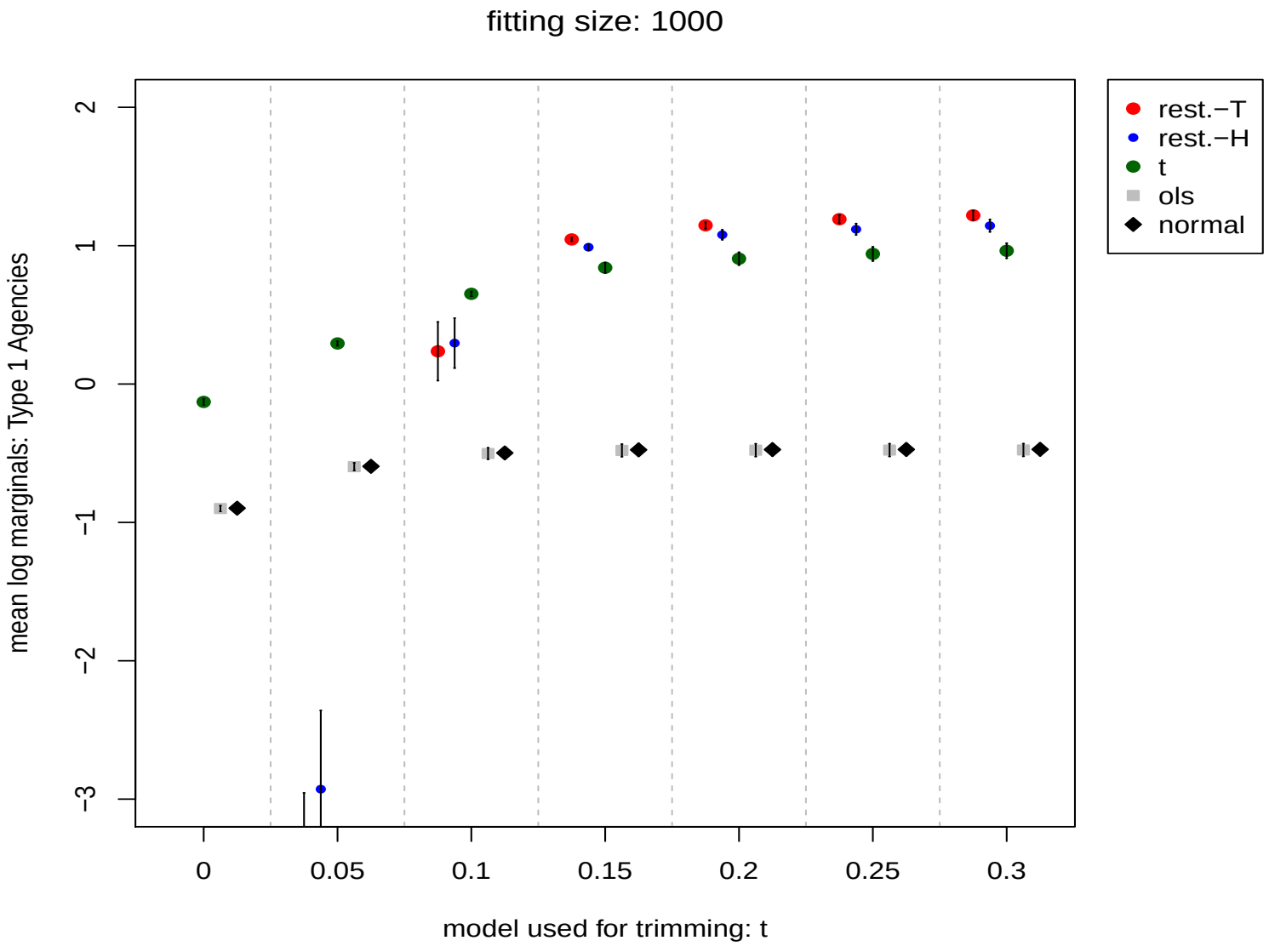
log Marginals-Type 1 Agencies

Blended-Tukey; Blended-Huber; thick tailed-t; normal; ols



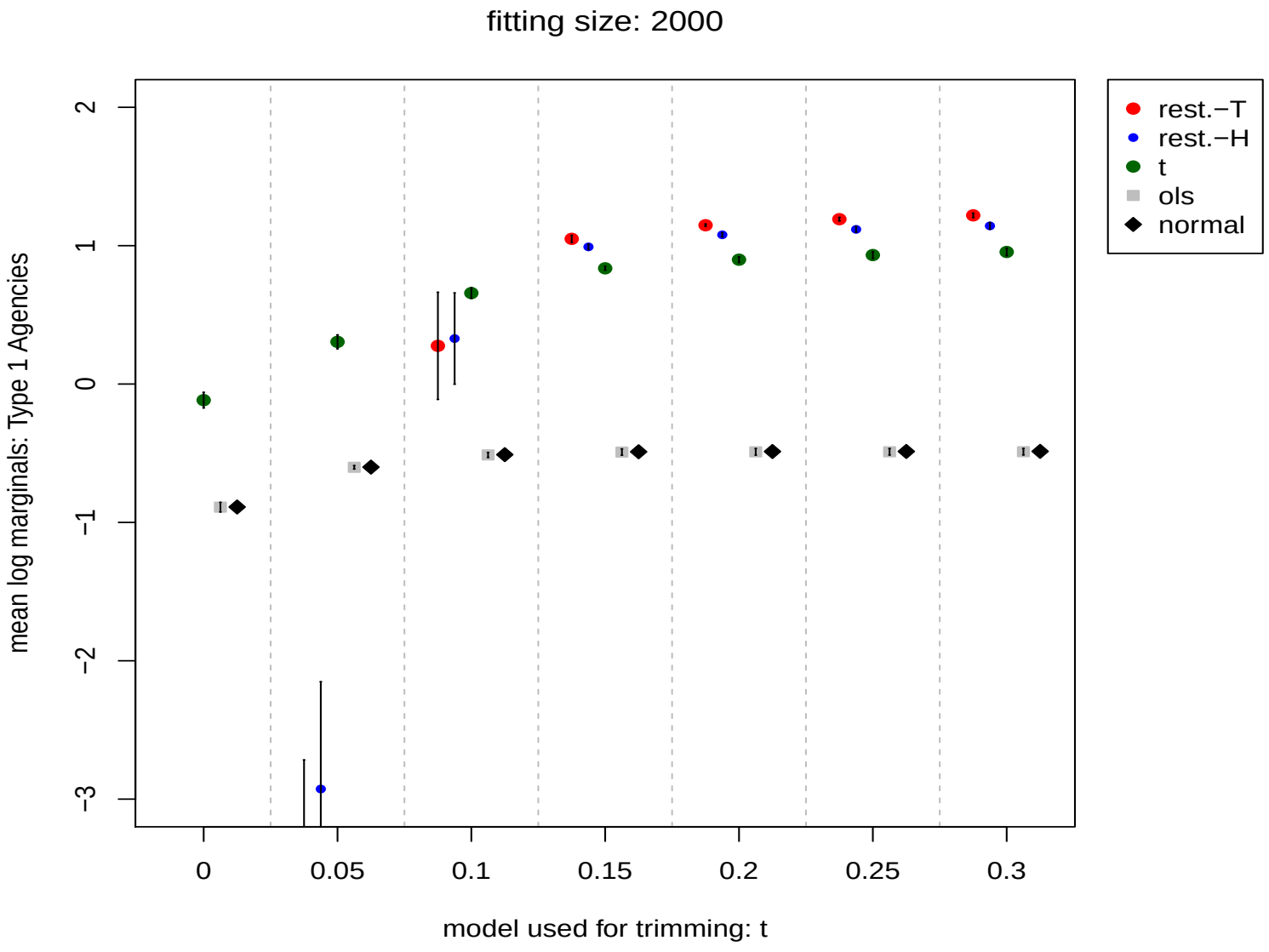
log Marginals-Type 1 Agencies

Blended-Tukey; Blended-Huber; thick tailed-t; normal; ols



log Marginals-Type 1 Agencies

Blended-Tukey; Blended-Huber; thick tailed-t; normal; ols



Summary

- A major question in applied work is when to stop modeling
  - Common practice is to exclude covariates, to reduce dimensionality of response, to preprocess measurements, etc.
  - Decision is based on whether, by stretching to include more in the model, overall model suffers
  - Outliers often removed before formal analysis
- Bayesian traditions differ (a little)
  - Model everything, including outliers
  - A complete model can be used to generate data
- Robust estimation traditions differ (a lot)
  - Model as little as possible; omit case analyses
  - Create estimator that is insensitive to certain facets of the data

Summary

- Blend the two approaches
  - Avoid attempting to model transitory phenomena
  - Use likelihood from robust estimators for Bayesian update
  - Get benefits of posterior for use in making decisions
- Consider these methods any time you worry about the likelihood,

Summary

- Blend the two approaches
  - Avoid attempting to model transitory phenomena
  - Use likelihood from robust estimators for Bayesian update
  - Get benefits of posterior for use in making decisions
- Consider these methods any time you worry about the likelihood,

  and you should always worry about the likelihood . . .