Design Matrix Uncertainty: Robust Optimization and Approximate MLE Approaches

Richard J. Clancy and Stephen Becker

University of Colorado at Boulder
Department of Applied Mathematics

July 21st
SIAM OPT 2021

University of Colorado
Boulder

Outline

Table of Contents

Problem Setup

We consider the generative model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\eta}$

- $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, design/data matrix
- $\boldsymbol{\eta} \in \mathbb{R}^{m}$, additive noise
- $\boldsymbol{x} \in \mathbb{R}^{n}$, model parameters
- $\boldsymbol{y} \in \mathbb{R}^{m}$, measurements

Given measurements $\boldsymbol{y}$ and design $\boldsymbol{A}$, estimate $\boldsymbol{x}$. Consider over-determined case where $m > n$.

Example: Estimating home prices

Generative model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\eta}$

Example

- $\boldsymbol{A} \in \mathbb{R}^{m \times 2}$
  - Column 1: Home size (1366 sqft)
  - Column 2: Lot size (5036 sqft)
- $\boldsymbol{y} \in \mathbb{R}^m$: selling price for corresponding home ($\$207k$)
- Each row is data for a particular home
- GOAL: Estimate parameters $\boldsymbol{x}$ to predict home price given info about home and lot size.

Ordinary least squares

Mismatch between modeled and true data given by $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}$.

Ordinary least squares (OLS): minimize residual sum of squares

$$\min_{\boldsymbol{x}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2.$$

Closed form solution of $\boldsymbol{x}_{\text{OLS}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}$.

Implicitly assumes Gaussian noise.

Uncertain design matrix

CONCERN: Assumes $\boldsymbol{A}$ is known precisely!!!

In practice, it is uncommon to know $\boldsymbol{A}$ with certainty. Causes are

- precision limits in measurement (dynamic range of sensor)
- truncation for memory savings (fixed number of sig. figs.)
- subjective features (survey responses)
- modeling error (features are approximate functions of data)

Example: Estimating home prices with uncertain data

Given the generative model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\eta}$, where $\boldsymbol{A}$ and $\boldsymbol{\eta}$ are both uncertain, estimate $\boldsymbol{x}$.

Example

- $\boldsymbol{A} \in \mathbb{R}^{m \times 2}$,
  - Column 1: *Rounded* home size (~~1366~~ 1400 sqft)
  - Column 2: *Rounded* lot size (~~5036~~ 5000 sqft).
- $\boldsymbol{y} \in \mathbb{R}^m$ selling price for corresponding home ($\$207k$)
- Each row is data for a particular home
- GOAL: Estimate parameters $\boldsymbol{x}$ to predict home price given uncertain info about home and lot size.

Total least squares

Total least squares considers uncertainty in $\boldsymbol{A}$ as well and solves

$$\begin{aligned} \min_{\boldsymbol{U},\boldsymbol{\eta},\boldsymbol{x}} \quad & \left\|[\boldsymbol{U},\boldsymbol{\eta}]\right\|_F \\ \text{subject to} \quad & (\boldsymbol{A}+\boldsymbol{U})\boldsymbol{x} = \boldsymbol{y}+\boldsymbol{\eta}. \end{aligned}$$

- Closed form solution of $\boldsymbol{x}_{\text{TLS}} = (\boldsymbol{A}^T\boldsymbol{A} - s_{n+1}^2\boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{y}$ where s_{n+1} is the smallest singular value of $[\boldsymbol{A},\boldsymbol{y}] \in \mathbb{R}^{m\times(n+1)}$ [H. Golub and F. van Loan, 1980].
- TLS is *always* worse conditioned than OLS (which is known to be poorly conditioned)
- Often performs poorly in practice...need better method

Our contribution

To address limitations of existing methods for regression with uncertain design matrices, we present two methods:

1. using box-constrained robust optimization, and
2. an approximate MLE framework based on the saddle point approximation from complex analysis.

Both approaches allow us to move away from implicit/explicit assumptions of Gaussianity and provides data scientists with more tools for handling model uncertainty.

Table of Contents

Robust optimization

Robust optimization (RO) seeks optima over all realizations of uncertain data

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{U} \in \mathcal{U}} f(\boldsymbol{x}, \boldsymbol{U})$$

where $\mathcal{U}$ is an uncertainty set and f is a real-valued function.

- Noise comes from a bounded set
- Uncertainty has geometric rather than distributional interpretation (can be viewed as uniform noise over set)
- Considers "worst-case" data

Robust least squares

Let $f(\boldsymbol{x}, \boldsymbol{U}) = \|(\boldsymbol{A} + \boldsymbol{U})\boldsymbol{x} - \boldsymbol{y}\|^2$ and solve the robust least squares problem

$$\min_{\boldsymbol{x}} \underbrace{\left\{ \max_{\|\boldsymbol{U}\|_\infty \leq \delta} \left\|(\boldsymbol{A} + \boldsymbol{U})\boldsymbol{x} - \boldsymbol{y}\right\|^2 \right\}}_{F(\boldsymbol{x})}$$

where $\|\boldsymbol{U}\|_\infty$ is the largest magnitude element of matrix $\boldsymbol{U}$. Parameter δ can be inferred from observed data.

Other works in robust least squares:

- constrained Frobenius norm [El Ghaoui and Lebret, 1997],
- constrained 1-norm [Xu et al., 2010],
- and more [Bertsimas et al., 2011].

Reformulate objective

Must optimize

$$\min_{\boldsymbol{x}} \Bigg\{ \underbrace{\max_{\|\boldsymbol{U}\|_\infty \leq \delta} \left\| (\boldsymbol{A}+\boldsymbol{U})\boldsymbol{x} - \boldsymbol{y} \right\|^2}_{F(\boldsymbol{x})} \Bigg\}.$$

Theorem

The inner maximization can be written as

$$F(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 + 2\,\delta \|\boldsymbol{x}\|_1 \, \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_1 + m\,\delta^2 \, \|\boldsymbol{x}\|_1^2.$$

with the argmax given by $\boldsymbol{U}_{\boldsymbol{x}}^* = \delta \cdot sign\left(\boldsymbol{x}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})^T\right)$ *for a fixed* $\boldsymbol{x}$.

Tractability

Rewrite robust least squares as

$$\min_{\boldsymbol{x}} \underbrace{\left\{ \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 + 2\,\delta \|\boldsymbol{x}\|_1 \, \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_1 + m\,\delta^2 \, \|\boldsymbol{x}\|_1^2 \right\}}_{F(\boldsymbol{x})}.$$

Not differentiable, but convex since max of convex functions.

Lemma

Element of subdifferential given by

$$F'(\boldsymbol{x}) = 2(\boldsymbol{A} + \boldsymbol{U}_{\boldsymbol{x}}^*)^T[(\boldsymbol{A} + \boldsymbol{U}_{\boldsymbol{x}}^*)\boldsymbol{x} - \boldsymbol{y}] \in \partial F(\boldsymbol{x}).$$

Access to subdifferential $\rightarrow$ tractable

Algorithms

For non-smooth problems, appropriate to use:

- Subgradient method
- Bundle method

In practice, slow to converge.

Quasi-Newton methods performed well with considerable speed-up and were used for numerical experiments.

Numerical Experiments

$$\min_{x} \left\{ \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 + 2\,\delta \|\boldsymbol{x}\|_1 \, \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_1 + m\,\delta^2 \, \|\boldsymbol{x}\|_1^2 \right\}$$

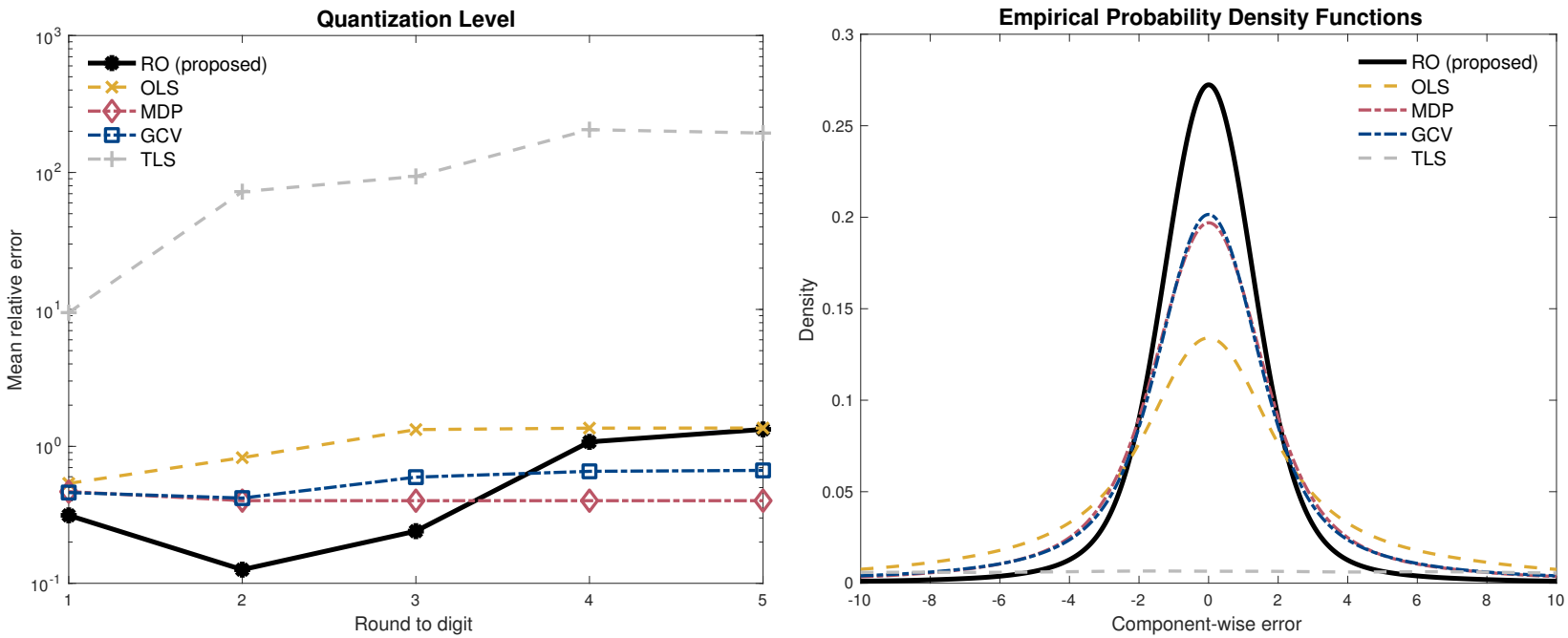

Figure: Left: mean relative error $e = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|/\|\boldsymbol{x}\|$ over 10k simulations for different estimators as a function of quantization level. Right: empirical PDF for errors of different methods.

Table of Contents

Maximum likelihood estimation

Once again, start with the generative model

$$\boldsymbol{y} = \boldsymbol{G}\boldsymbol{x} + \boldsymbol{\eta}$$

- $\boldsymbol{G} \in \mathbb{R}^{m \times n}$, random/uncertain design matrix
- $\boldsymbol{\eta} \in \mathbb{R}^{m}$, additive noise
- $\boldsymbol{x} \in \mathbb{R}^{n}$, model parameters
- $\boldsymbol{y} \in \mathbb{R}^{m}$, measurements

Unlike robust LS, we assume distributional knowledge of $\boldsymbol{G}$ and $\boldsymbol{\eta}$, use to find PDF of $\boldsymbol{y}$

Maximum likelihood estimation

Maximum likelihood estimation (MLE) works as follows:

- Observe measurements $\boldsymbol{y}$ from some distribution $\mathcal{P}_{\boldsymbol{y}}$ determined by distributions $\mathcal{P}_{\boldsymbol{G}}$ and $\mathcal{P}_{\boldsymbol{\eta}}$
- Form likelihood function from PDF of $\boldsymbol{y}$, i.e. $L(\boldsymbol{x}) = f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{x})$, where $f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{x})$ is the PDF of $\boldsymbol{y}$
- What parameters $\boldsymbol{x}$ best explain observed data $\boldsymbol{y}$?
- Find out by maximizing the likelihood function

$$\underset{\boldsymbol{x}}{\operatorname{argmax}} \ \{L(\boldsymbol{x}) = f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{x})\}$$

Maximum likelihood estimation

- We assume independence of G_{ij}'s and η_i's throughout, implies independence of y_i's
- When components of $\boldsymbol{y}$ are independent, we can split PDF

$$f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{x}) = \prod_{i=1}^{m} f_{Y_i}(y_i;\boldsymbol{x})$$

- Focus on maximizing the log-likelihood function

$$\hat{\boldsymbol{x}}_{\text{MLE}} = \underset{\boldsymbol{x}}{\text{argmax}} \left\{ \sum_{i=1}^{m} \ln\left[f_{Y_i}(y_i;\boldsymbol{x})\right] \right\}$$

Justification of MLE for regression problem

For additive noise models $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\eta}$ where $\boldsymbol{A}$ is known precisely

- $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2$, negative log-likelihood for Gaussian noise (OLS)
- $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_1$, negative log-likelihood for Laplacian noise
- $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_\infty$, negative log-likelihood for uniform noise
- For uncertainty in operator, solving TLS problem

$$\begin{aligned} \min_{\boldsymbol{x}, \boldsymbol{U}, \boldsymbol{\eta}} \quad & \|[\boldsymbol{U}, \boldsymbol{\eta}]\|_F \\ \text{Subject to} \quad & (\boldsymbol{A} + \boldsymbol{U})\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{\eta} \end{aligned}$$

yields MLE for i.i.d. Gaussian in $\boldsymbol{U}$ and $\boldsymbol{\eta}$

MLE for uncertainty in design matrix

- GOAL: find PDF for $\boldsymbol{y} = \boldsymbol{G}\boldsymbol{x} + \boldsymbol{\eta}$ to form a likelihood function
- Each component can be rewritten as a sum of random variables, i.e., $y_i = \boldsymbol{g}_i^T \boldsymbol{x} + \eta_i = \sum_{j=1}^{n} G_{ij} x_j + \eta_i$
- PDFs for sums of random variables are challenging to derive since they require convolutions

Example

Let $Z \sim \mathcal{N}(0, 1)$ and $U \sim \text{Uniform}(0, 1)$. The PDF for $U + Z$ is

$$f_{U+Z}(t) = \frac{1}{\sqrt{2\pi}} \int_0^1 e^{-(t-s)^2/2} \, ds.$$

No analytic form!

Moment generating functions

- The moment generating function (MGF) is a bilateral Laplace transform of PDF
$$M_Y(t) = \mathbb{E}(e^{tY})$$

- When MGF exists, it uniquely characterizes the distribution
- Useful properties: for independent random variables U and Z and $a \in \mathbb{R}$
$$M_{aU+Z}(t) = M_U(at)M_Z(t),$$
with MGFs, convolution $\rightarrow$ multiplication

Example

Let $Z \sim \mathcal{N}(0,1)$ and $U \sim \text{Uniform}(0,1)$. The MGF for $U + Z$ is

$$M_{U+Z}(t) = \frac{(e^t - 1)e^{-t^2/2}}{t}.$$

No need for quadrature

Moment generating functions

- Using properties of MGFs,

$$M_{Y_i}(t) = M_{\boldsymbol{g}_i^T \boldsymbol{x} + \eta_i}(t) = M_{\eta_i}(t) \prod_{j=1}^{n} M_{G_{ij}}(t x_j)$$

  Express complicated MGF for Y_i as the product of univariate MGFs for G_{ij} and η_i.
- By inverting transform for M_{Y_i}, we can recover density, but difficult in practice
- Approximate PDF instead

Density approximation

Options to approximate PDFs:

- Kernel density estimation : data intensive
- Edgeworth series: poor tail behavior (poly. series)
- Saddle point approximation [Daniels, 1954, Strawderman et al., 1996]: uses complex analysis, works well in practice

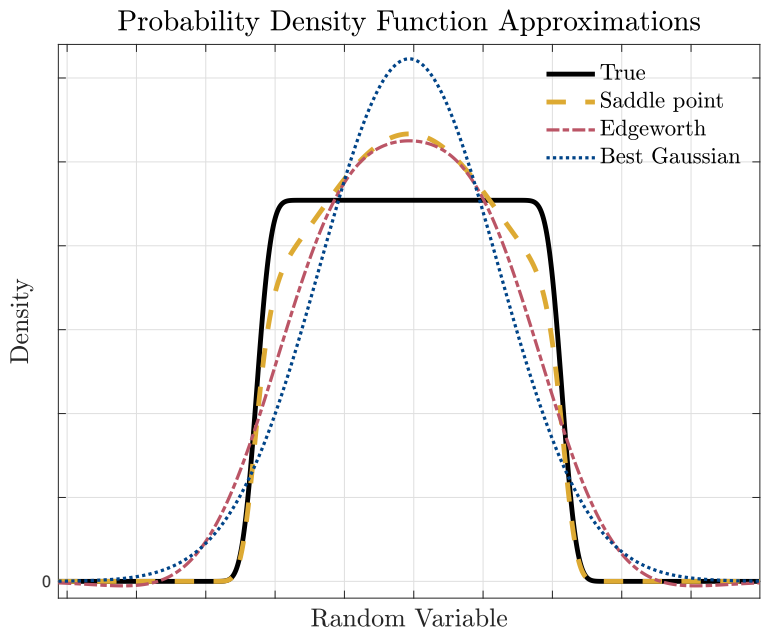

Figure: True density and several approximations for $y = \boldsymbol{g}^T \boldsymbol{x} + \eta$

Saddle point approximation

Saddle point approximation for PDF of Y is

$$f_Y(y) \approx \sqrt{\frac{1}{2\pi K_Y''(t_0)}} e^{K_Y(t_0) - y t_0}$$

- $K_Y(t) = \ln M_Y(t)$ is *Cumulant Generating Function* (CGF)
- t_0 is the solution to $K_Y'(t) - y = 0$ (use Newton's method)

Approximate MLE

Using the saddle point approximation and eliminating constants, write the approximate log-likelihood function, $\ell(\boldsymbol{x}) \approx \ln L(\boldsymbol{x})$, as

$$\begin{aligned}
\ell(\boldsymbol{x}) &= \sum_{i=1}^{m} \ln \underbrace{\left\{ \sqrt{\frac{1}{K''_{Y_i}(t_i)}} e^{K_{Y_i}(t_i) - y_i t_i} \right\}}_{\approx \text{ constant } \cdot \ f_{Y_i}(y_i)} \\
&= \sum_{i=1}^{m} \left\{ K_{Y_i}(t_i) - t_i y_i - \frac{1}{2} \ln \left(K''_{Y_i}(t_i) \right) \right\} \\
&= \sum_{i=1}^{m} \left[K_{\boldsymbol{g}_i^T \boldsymbol{x} + \eta_i}(t_i(\boldsymbol{x})) - \frac{1}{2} \ln \left(K''_{\boldsymbol{g}_i^T \boldsymbol{x} + \eta_i}(t_i(\boldsymbol{x})) \right) - t_i(\boldsymbol{x}) y_i \right].
\end{aligned}$$

where $\boldsymbol{g}_i^T$ is i^{th} row of $\boldsymbol{G}$ and $t_i(\boldsymbol{x})$ is solution to $K'_{\boldsymbol{g}_i^T \boldsymbol{x} + \eta_i}(t) = y_i$.

Optimization problem

The approximate MLE can be cast generically in vector form

$$\begin{aligned} \underset{\boldsymbol{x},\boldsymbol{t}}{\text{argmax}} \quad & \mathbb{1}^T \left(K_{\boldsymbol{G}\boldsymbol{x}+\boldsymbol{\eta}}(\boldsymbol{t}) - \frac{1}{2} \ln \left(K''_{\boldsymbol{G}\boldsymbol{x}+\boldsymbol{\eta}}(\boldsymbol{t}) \right) \right) - \boldsymbol{t}^T \boldsymbol{y} \\ \text{subject to} \quad & K'_{\boldsymbol{G}\boldsymbol{x}+\boldsymbol{\eta}}(\boldsymbol{t}) = \boldsymbol{y} \end{aligned}$$

Example

When $\boldsymbol{G} \sim \text{Uniform}(\boldsymbol{H} - \delta \mathbb{1}\mathbb{1}^T, \boldsymbol{H} + \delta \mathbb{1}\mathbb{1}^T)$ and $\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$,

$$\begin{aligned} \underset{\boldsymbol{x},\boldsymbol{t}}{\text{argmax}} \quad & \boldsymbol{t}^T \left(\frac{\sigma^2}{2} \boldsymbol{t} + \boldsymbol{H}\boldsymbol{x} - \boldsymbol{y} \right) + \mathbb{1}^T \ln \left[\sinh \left(\delta \boldsymbol{t}\boldsymbol{x}^T \right) \oslash \left(\delta \boldsymbol{t}\boldsymbol{x}^T \right) \right] \mathbb{1} \\ & \qquad - \frac{1}{2} \mathbb{1}^T \ln \left[\sigma^2 \mathbb{1} - \delta^2 \operatorname{csch}^2 \left(\delta \boldsymbol{t}\boldsymbol{x}^T \right) \boldsymbol{x}^2 \right] \\ \text{Subject to} \quad & \sigma^2 \boldsymbol{t} + \boldsymbol{H}\boldsymbol{x} + \delta \coth \left(\delta \boldsymbol{t}\boldsymbol{x}^T \right) \boldsymbol{x} - n(\mathbb{1} \oslash \boldsymbol{t}) = \boldsymbol{y}. \end{aligned}$$

Gradients

Despite constraint requiring a numeric solve, have gradients.

Letting $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{t}) = K'_{\boldsymbol{G}\boldsymbol{x}+\boldsymbol{\eta}}(\boldsymbol{t}) - \boldsymbol{y}$ be our constraint, using the chain rule and implicit differentiation we have

$$\nabla_{\boldsymbol{x}} \ell = \frac{\partial \ell}{\partial \boldsymbol{x}} - \left(\frac{\partial \ell}{\partial \boldsymbol{t}}\right) \left(\frac{\partial \boldsymbol{q}}{\partial \boldsymbol{t}}\right)^{-1} \left(\frac{\partial \boldsymbol{q}}{\partial \boldsymbol{x}}\right).$$

Unimportant, but for completeness, each factor is given by:

$$\begin{aligned}
\frac{\partial \ell}{\partial \boldsymbol{x}} &= \mathbb{1}^T \left(\frac{\partial}{\partial \boldsymbol{x}} K_{\boldsymbol{G}\boldsymbol{x}+\boldsymbol{\eta}}(\boldsymbol{t}) - \frac{1}{2} \left\{ \text{diag}\left(K''_{\boldsymbol{G}\boldsymbol{x}+\boldsymbol{\eta}}(\boldsymbol{t})\right) \right\}^{-1} \frac{\partial}{\partial \boldsymbol{x}} K''_{\boldsymbol{G}\boldsymbol{x}+\boldsymbol{\eta}}(\boldsymbol{t}) \right) \\
\frac{\partial \ell}{\partial \boldsymbol{t}} &= \left(K'_{\boldsymbol{G}\boldsymbol{x}+\boldsymbol{\eta}}(\boldsymbol{t}) - \frac{1}{2} \left(K'''_{\boldsymbol{G}\boldsymbol{x}+\boldsymbol{\eta}}(\boldsymbol{t}) \oslash K''_{\boldsymbol{G}\boldsymbol{x}+\boldsymbol{\eta}}(\boldsymbol{t}) \right) - \boldsymbol{y} \right)^T \\
\frac{\partial \boldsymbol{q}}{\partial \boldsymbol{t}} &= \text{diag}\left(K''_{\boldsymbol{G}\boldsymbol{x}+\boldsymbol{\eta}}(\boldsymbol{t})\right), \\
\frac{\partial \boldsymbol{q}}{\partial \boldsymbol{x}} &= \frac{\partial}{\partial \boldsymbol{x}} K'_{\boldsymbol{G}\boldsymbol{x}+\boldsymbol{\eta}}(\boldsymbol{t}).
\end{aligned}$$

Numerical experiments

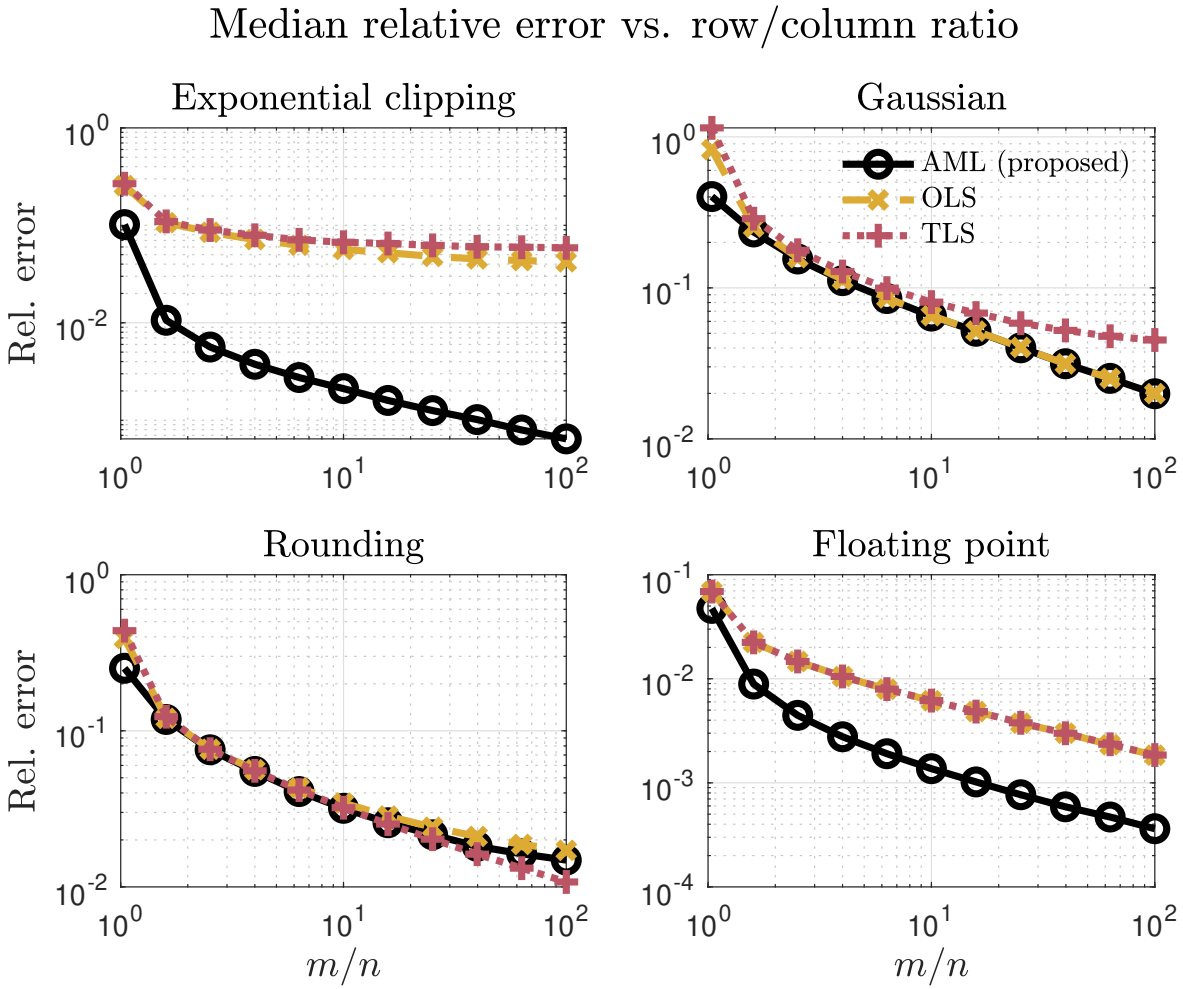

Figure: Median relative error over 1,000 simulations as number of rows increase for fixed number of colums in design matrix.

Numerical Experiments

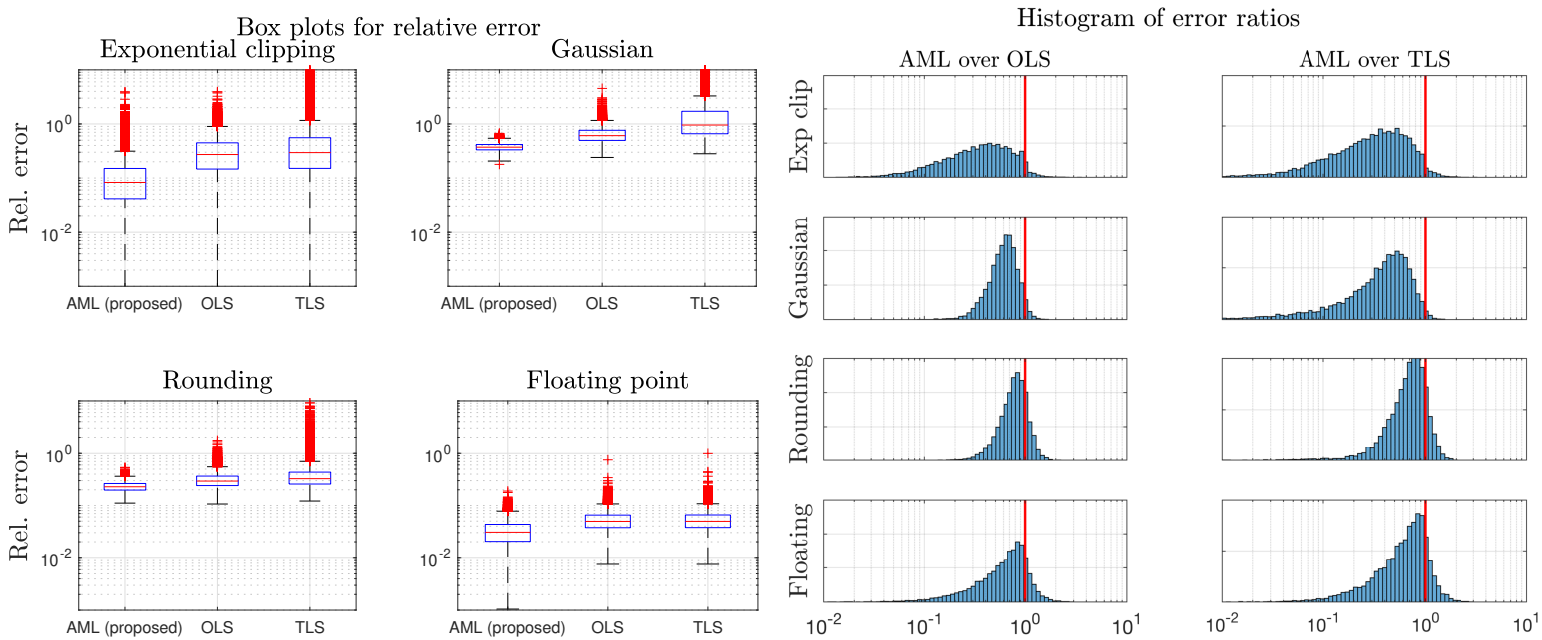

Figure: Error metrics for simulations $\boldsymbol{G} \in \mathbb{R}^{55\times 50}$. Left: box-plot of relative error for different methods. Right: histogram of error ratio $\|\boldsymbol{x}_{\mathsf{AML}} - \boldsymbol{x}_{\mathsf{TRU}}\|/\|\boldsymbol{x}_{\mathsf{OLS}} - \boldsymbol{x}_{\mathsf{TRU}}\|$. Values less than 1 indicate AML outperformed competing method for identical data.

Summary

- Motivated the need for solving regression problems with uncertain design matrices.
- Formulated and solved a robust least squares problem that is well suited for regression problems with data subject to quantization error.
- Derived an approximate MLE function and provided its gradient. Ideal for regression problems with different noise models across features.

References

D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. SIAM Review, 53 (3):464–501, 2011. doi: 10.1137/080734510. URL https://doi.org/10.1137/080734510.

Henry E Daniels. Saddlepoint approximations in statistics. The annals of mathematical statistics, pages 631–650, 1954.

Laurent El Ghaoui and Herve Lebret. Robust solutions to least-squares problems with uncertain data. SIAM J. Matrix Anal. Appl., 18(4):1035 – 1064, 10 1997.

Gene H. Golub and Charles F. van Loan. An analysis of the total least squares problem. SIAM J. Numer. Anal., 17 (6):883 – 893, 12 1980.

Robert L Strawderman, George Casella, and Martin T Wells. Practical small-sample asymptotics for regression problems. Journal of the American Statistical Association, 91(434):643–654, 1996.

H. Xu, C. Caramanis, and S. Mannor. Robust regression and lasso. IEEE Tran. Info. Theory, 56(7):3561–3574, July 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2048503.

Thank you for your time!

- Stephen Becker, and Richard J. Clancy. "Robust least squares for quantized data matrices." Signal Processing 176 (2020)
- Richard J. Clancy and Stephen Becker. "Approximate maximum likelihood estimators for linear regression with design matrix uncertainty." arXiv preprint arXiv:2104.03307 (2021).

Questions?