Beyond the Mean: From F-modeling to G-modeling

Joonhyuk Jung

March 5, 2024

THE UNIVERSITY OF CHICAGO

Introduction

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Irrational Exuberance: Correcting Bias in Probability Estimates

Gareth M. James, Peter Radchenko & Bradley Rava

To cite this article: Gareth M. James, Peter Radchenko & Bradley Rava (2022) Irrational Exuberance: Correcting Bias in Probability Estimates, Journal of the American Statistical Association, 117:537, 455-468, DOI: 10.1080/01621459.2020.1787175

To link to this article: https://doi.org/10.1080/01621459.2020.1787175

Introduction

Consider a statistical model given by

$$X_i \mid \theta_i \overset{ind}{\sim} \mathrm{N}\left(\theta_i, \sigma^2\right), \qquad \theta_i \in \mathbb{R},$$

$$L(\widehat{\theta}_i, \theta_i) = \left(\widehat{\theta}_i - \theta_i\right)^2, \qquad i = 1, \dots, n.$$

Assume $\sigma^2 > 0$ to be known.

Our goal is to minimize the compound loss

$$L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} L(\widehat{\theta}_i, \theta_i).$$

Introduction

Consider a statistical model given by

$$X_i \mid \theta_i \overset{ind}{\sim} \mathrm{Beta}\left(\frac{\theta_i}{\gamma}, \frac{1-\theta_i}{\gamma}\right), \qquad \theta_i \in (0,1),$$

$$L(\widehat{\theta}_i, \theta_i) = \left(\widehat{\theta}_i - \theta_i\right)^2, \qquad i = 1, \ldots, n.$$

Assume $\gamma > 0$ to be known.
Our goal is to minimize the compound loss

$$L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} L(\widehat{\theta}_i, \theta_i).$$

Introduction

Consider a statistical model given by

$$X_i \mid \theta_i \overset{ind}{\sim} \mathrm{Beta}\left(\frac{\theta_i}{\gamma}, \frac{1-\theta_i}{\gamma}\right), \qquad \theta_i \in (0,1),$$

$$L(\widehat{\theta}_i, \theta_i) = \left(\frac{\widehat{\theta}_i - \theta_i}{\min(\widehat{\theta}_i, 1-\widehat{\theta}_i)}\right)^2, \qquad i = 1, \ldots, n.$$

Assume $\gamma > 0$ to be known.
Our goal is to minimize the compound loss

$$L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} L(\widehat{\theta}_i, \theta_i).$$

Introduction - Remark

1. The dispersion parameter γ plays a similar role to σ^2.

$$\mathbb{E}(X_i \mid \theta_i) = \theta_i, \qquad \mathrm{Var}(X_i \mid \theta_i) = \frac{\gamma}{1+\gamma}\theta_i(1-\theta_i).$$

Introduction - Remark

1. The dispersion parameter γ plays a similar role to σ^2.

$$\mathbb{E}(X_i \mid \theta_i) = \theta_i, \qquad \mathrm{Var}(X_i \mid \theta_i) = \frac{\gamma}{1+\gamma}\theta_i(1-\theta_i).$$

2. We wish to heavily penalize settings where the estimate $\widehat{\theta}_i$ is closer to either zero or one than the truth.

$$L(\widehat{\theta}_i, \theta_i) = \left(\frac{\widehat{\theta}_i - \theta_i}{\min(\widehat{\theta}_i, 1-\widehat{\theta}_i)} \right)^2.$$

Bayes Estimator

Suppose a prior distribution G of θ_i is given.

$$\theta_i \overset{iid}{\sim} G.$$

Recall that the Bayes estimator of θ_i under the usual quadratic loss with respect to G is given by the posterior mean:

$$\widehat{\theta}_i^G(X_i) = \underset{a}{\operatorname{argmin}}\, \mathbb{E}_G\left[(a-\theta_i)^2 \mid X_i\right] = \mathbb{E}_G[\theta_i \mid X_i],$$

where the expectation is taken with respect to the posterior distribution of θ_i given X_i.

Bayes Estimator

Theorem (by the authors)

Consider the new loss function:

$$L(\widehat{\theta}_i, \theta_i) = \left(\frac{\widehat{\theta}_i - \theta_i}{\min(\widehat{\theta}_i, 1 - \widehat{\theta}_i)} \right)^2 .$$

Then, the Bayes estimator of θ_i with respect to G is given by

$$\widehat{\theta}_i^G(X_i) = \begin{cases} \min\left(\mathbb{E}_G(\theta_i \mid X_i) + \frac{\mathrm{Var}_G(\theta_i|X_i)}{\mathbb{E}_G(\theta_i|X_i)}, \frac{1}{2} \right), & \mathbb{E}_G(\theta_i \mid X_i) \leq \frac{1}{2}, \\ \max\left(\mathbb{E}_G(\theta_i \mid X_i) - \frac{\mathrm{Var}_G(\theta_i|X_i)}{1-\mathbb{E}_G(\theta_i|X_i)}, \frac{1}{2} \right), & \mathbb{E}_G(\theta_i \mid X_i) > \frac{1}{2}. \end{cases}$$

Bayes Estimator - Discussion

1. This result does not rely on the model $X_i \mid \theta_i$.

Bayes Estimator - Discussion

1. This result does not rely on the model $X_i \mid \theta_i$.
2. This new estimator $\widehat{\theta}_i^G$ depends solely on the first two posterior moments.

$$\widehat{\theta}_i^G(X_i) = \begin{cases} \min\left(\mathbb{E}_G(\theta_i \mid X_i) + \frac{\mathrm{Var}_G(\theta_i|X_i)}{\mathbb{E}_G(\theta_i|X_i)}, \frac{1}{2}\right), & \mathbb{E}_G(\theta_i \mid X_i) \leq \frac{1}{2}, \\ \max\left(\mathbb{E}_G(\theta_i \mid X_i) - \frac{\mathrm{Var}_G(\theta_i|X_i)}{1-\mathbb{E}_G(\theta_i|X_i)}, \frac{1}{2}\right), & \mathbb{E}_G(\theta_i \mid X_i) > \frac{1}{2}. \end{cases}$$

Bayes Estimator - Discussion

1. This result does not rely on the model $X_i \mid \theta_i$.
2. This new estimator $\widehat{\theta}_i^G$ depends solely on the first two posterior moments.

$$\widehat{\theta}_i^G(X_i) = \begin{cases} \min\left(\mathbb{E}_G(\theta_i \mid X_i) + \frac{\mathrm{Var}_G(\theta_i|X_i)}{\mathbb{E}_G(\theta_i|X_i)}, \frac{1}{2}\right), & \mathbb{E}_G(\theta_i \mid X_i) \leq \frac{1}{2}, \\ \max\left(\mathbb{E}_G(\theta_i \mid X_i) - \frac{\mathrm{Var}_G(\theta_i|X_i)}{1-\mathbb{E}_G(\theta_i|X_i)}, \frac{1}{2}\right), & \mathbb{E}_G(\theta_i \mid X_i) > \frac{1}{2}. \end{cases}$$

3. The estimator has the property of shifting the posterior mean toward 0.5.

Bayes Estimator - Discussion

1. This result does not rely on the model $X_i \mid \theta_i$.
2. This new estimator $\widehat{\theta}_i^G$ depends solely on the first two posterior moments.

$$\widehat{\theta}_i^G(X_i) = \begin{cases} \min\left(\mathbb{E}_G(\theta_i \mid X_i) + \frac{\mathrm{Var}_G(\theta_i|X_i)}{\mathbb{E}_G(\theta_i|X_i)}, \frac{1}{2}\right), & \mathbb{E}_G(\theta_i \mid X_i) \leq \frac{1}{2}, \\ \max\left(\mathbb{E}_G(\theta_i \mid X_i) - \frac{\mathrm{Var}_G(\theta_i|X_i)}{1-\mathbb{E}_G(\theta_i|X_i)}, \frac{1}{2}\right), & \mathbb{E}_G(\theta_i \mid X_i) > \frac{1}{2}. \end{cases}$$

3. The estimator has the property of shifting the posterior mean toward 0.5.
4. However, it never surpasses 0.5.

F-modeling

Theorem (by the authors)

Under the introduced Beta model where

$$X_i \mid \theta_i \overset{ind}{\sim} \mathrm{Beta}\left(\frac{\theta_i}{\gamma}, \frac{1-\theta_i}{\gamma}\right),$$

the first two posterior moments, $\mathbb{E}_G(\theta_i \mid X_i)$ *and* $\mathbb{E}_G(\theta_i^2 \mid X_i)$*, can be explicitly derived given knowledge of*

$$\frac{\partial}{\partial X_i} \log f_G(X_i) \quad and \quad \frac{\partial^2}{\partial X_i^2} \log f_G(X_i),$$

where $f_G(X_i)$ *denotes the marginal likelihood of* X_i*. [The explicit formula is given in the paper.]*

F-modeling - Discussion

1. One may call this “S-modeling” because they directly model the score function (and its partial derivative).

F-modeling - Discussion

1. One may call this "S-modeling" because they directly model the score function (and its partial derivative).
2. The authors use a natural cubic spline to compute the score function.

F-modeling - Discussion

1. One may call this “S-modeling” because they directly model the score function (and its partial derivative).
2. The authors use a natural cubic spline to compute the score function.
3. Like other F-modeling methods, the resulting $\widehat{\theta}_i(X_i)$ is not necessarily monotone in X_i, which is NOT desirable.

F-modeling - Discussion

1. One may call this “S-modeling” because they directly model the score function (and its partial derivative).
2. The authors use a natural cubic spline to compute the score function.
3. Like other F-modeling methods, the resulting $\widehat{\theta}_i(X_i)$ is not necessarily monotone in X_i, which is NOT desirable.
4. Their estimation is valid only under the assumption that G is symmetric about $1/2$. This constraint is too restrictive (in my opinion).

F-modeling - Discussion

1. One may call this "S-modeling" because they directly model the score function (and its partial derivative).
2. The authors use a natural cubic spline to compute the score function.
3. Like other F-modeling methods, the resulting $\widehat{\theta}_i(X_i)$ is not necessarily monotone in X_i, which is NOT desirable.
4. Their estimation is valid only under the assumption that G is symmetric about $1/2$. This constraint is too restrictive (in my opinion).
5. Hence, I propose an alternative approach.

G-modeling - Motivation

Theorem (by Joonhyuk Jung)

Under the introduced Beta model and loss function, the Bayes estimator

$$\widehat{\theta}_i^G(X_i) = \begin{cases} \min\left(\mathbb{E}_G(\theta_i \mid X_i) + \frac{\mathrm{Var}_G(\theta_i|X_i)}{\mathbb{E}_G(\theta_i|X_i)}, \frac{1}{2}\right), & \mathbb{E}_G(\theta_i \mid X_i) \le \frac{1}{2}, \\ \max\left(\mathbb{E}_G(\theta_i \mid X_i) - \frac{\mathrm{Var}_G(\theta_i|X_i)}{1-\mathbb{E}_G(\theta_i|X_i)}, \frac{1}{2}\right), & \mathbb{E}_G(\theta_i \mid X_i) > \frac{1}{2}. \end{cases}$$

is non-decreasing in X_i (for any prior distribution G).

Corollary

G-modeling necessarily results in a monotone estimator $\widehat{\theta}_i^{\widehat{G}}(X_i)$.

G-modeling - Proof

Proof of the Theorem (Sketch)

Rewrite the Bayes estimator as

$$\widehat{\theta}_i^G(X_i) = \begin{cases} \min\left(\frac{\mathbb{E}_G(\theta_i^2|X_i)}{\mathbb{E}_G(\theta_i|X_i)}, \frac{1}{2}\right), & \mathbb{E}_G(\theta_i \mid X_i) \le \frac{1}{2}, \\ 1 - \min\left(\frac{\mathbb{E}_G((1-\theta_i)^2|X_i)}{\mathbb{E}_G(1-\theta_i|X_i)}, \frac{1}{2}\right), & \mathbb{E}_G(\theta_i \mid X_i) > \frac{1}{2}. \end{cases}$$

Now, it suffices to prove that

$$\mathbb{E}_G(\theta_i \mid X_i) \quad and \quad \frac{\mathbb{E}_G(\theta_i^2 \mid X_i)}{\mathbb{E}_G(\theta_i \mid X_i)}$$

are non-decreasing in X_i, respectively. Here I will only handle the second one for brevity.

G-modeling - Proof

Proof of the Theorem (Sketch - Continued)

For simplicity, fix $\gamma = 1$. (The proof is essentially the same for general values of $\gamma > 0$.) Note that

$$\frac{\mathbb{E}_G(\theta_i^2 \mid X_i)}{\mathbb{E}_G(\theta_i \mid X_i)} = \frac{\int_0^1 \theta^2 X_i^{\theta-1}(1-X_i)^{-\theta} \frac{1}{\Gamma(\theta)\Gamma(1-\theta)} \, dG(\theta)}{\int_0^1 \theta X_i^{\theta-1}(1-X_i)^{-\theta} \frac{1}{\Gamma(\theta)\Gamma(1-\theta)} \, dG(\theta)}$$

$$= \frac{\int_0^1 e^{\theta Y_i} \frac{\theta^2}{\Gamma(\theta)\Gamma(1-\theta)} \, dG(\theta)}{\int_0^1 e^{\theta Y_i} \frac{\theta}{\Gamma(\theta)\Gamma(1-\theta)} \, dG(\theta)},$$

where $Y_i = \log \frac{X_i}{1-X_i} \in \mathbb{R}$.

G-modeling - Proof

Proof of the Theorem (Sketch - Continued)

By appealing to Cauchy-Schwarz inequality,

$$\frac{d}{dY_i}\frac{\mathbb{E}_G(\theta_i^2 \mid X_i)}{\mathbb{E}_G(\theta_i \mid X_i)} = \frac{J^3(Y_i)J(Y_i) - \left(J^2(Y_i)\right)^2}{\left(J(Y_i)\right)^2} \geq 0,$$

where we define

$$J^k(Y_i) := \int_0^1 e^{\theta Y_i} \frac{\theta^k}{\Gamma(\theta)\Gamma(1-\theta)} \, dG(\theta)$$

for $k = 1, 2, 3$. *Since* Y_i *is non-decreasing in* X_i, *we conclude the proof.*

Simulation - Setup

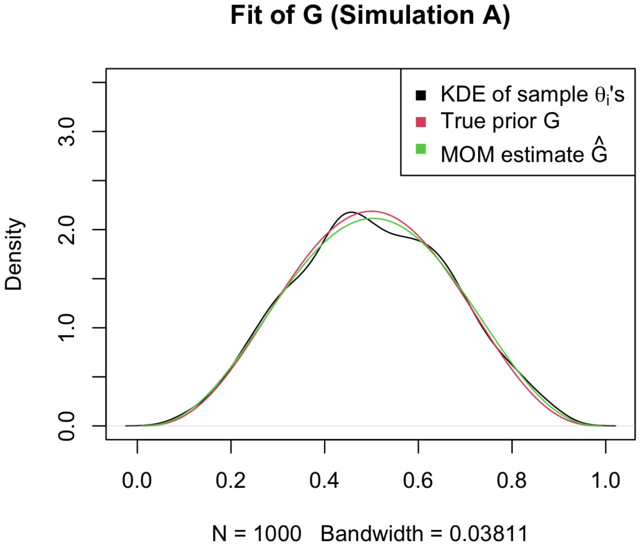
G	Beta	Non-Beta
Symmetric	$A = \mathrm{Beta}(4,4)$	$C = \frac{1}{2}\mathrm{Beta}(2,6) + \frac{1}{2}\mathrm{Beta}(6,2)$
Asymmetric	$B = \mathrm{Beta}(2,6)$	$D = \frac{1}{2}\mathrm{Beta}(2,6) + \frac{1}{2}\mathrm{Beta}(5,3)$

- Sample size $n = 1000$
- Number of iterations = 100 times
- Dispersion $\gamma = 0.03$

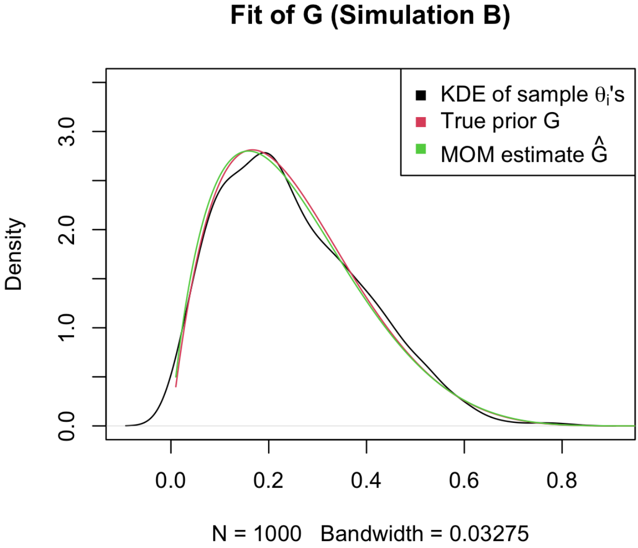
Simulation - Results

- Performance: ECAP (F-modeling) $\leq$ NP G-modeling $<$ Parametric G-modeling
  - Maybe the simulation setup is too simple.
- ECAP may result in very bad estimators if the true prior G is not symmetric about $1/2$.

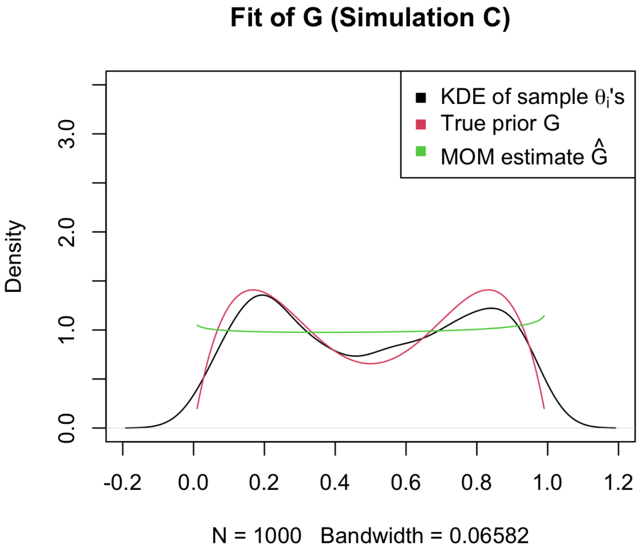
Simulation - Fit of G (Simulation A)



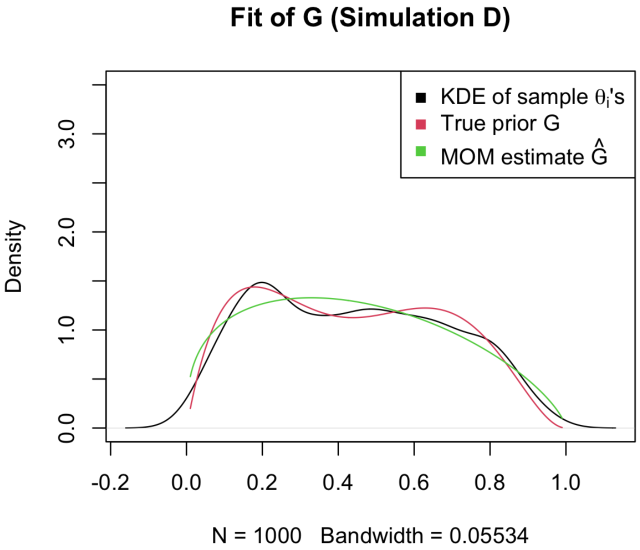
Simulation - Fit of G (Simulation B)



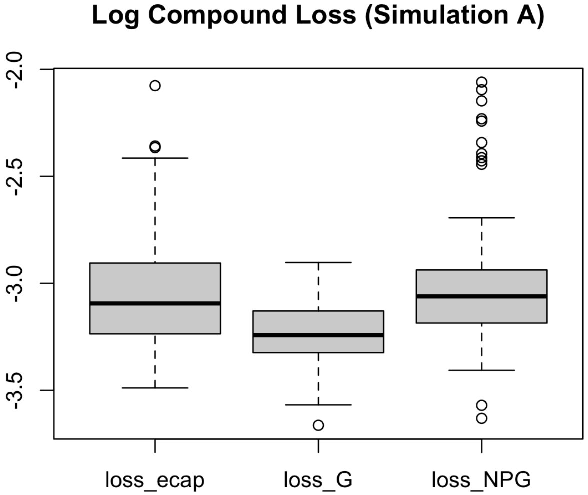
Simulation - Fit of G (Simulation C)



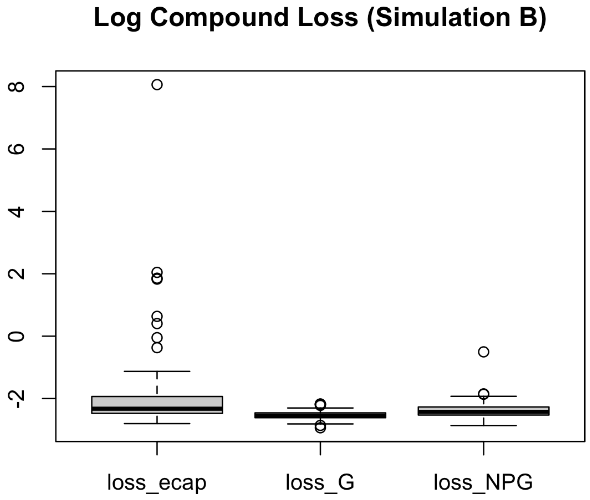
Simulation - Fit of G (Simulation D)



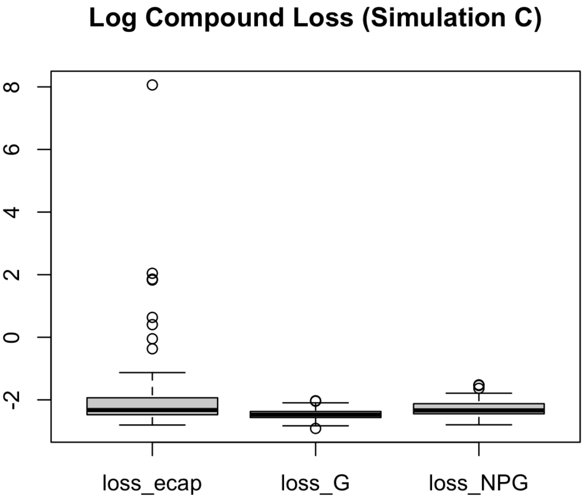
Simulation - Compound Loss (Simulation A)



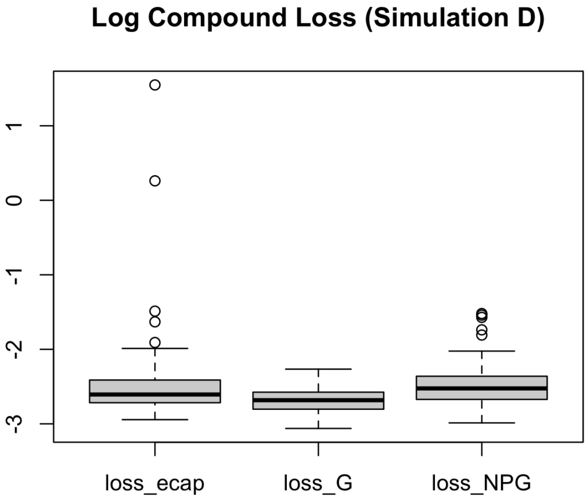
Simulation - Compound Loss (Simulation B)



Simulation - Compound Loss (Simulation C)



Simulation - Compound Loss (Simulation D)



Thank You

Contact: joonhyukjung (at) uchicago.edu