Optimal Inference in Large-Scale Problems

Daniel Yekutieli and Asaf Weinstein

Statistical Learning Seminars
April 2021

Plan

1. High-dimensional logistic regression example
2. Background
3. Oracle-based optimal inference
   - Bayesian perspective
   - Frequentist perspective
4. Implementation by hierarchical Bayes modelling
5. Simulation results
6. Discussion

Illustrative example

High dimensional logistic regression example

- Fixed parameter vetor, $\vec{\beta} = (\beta_1, \cdots, \beta_m)$
- Fixed X matrix, $\mathbf{X}_{n \times m}$ generated by sampling iid $N(0, 1/n)$ entries
- Response vector, $\vec{Y} = (Y_1, \cdots, Y_n)$ with $Y_j \sim Bernoulli(q_j)$ for $q_j = \exp(\mu_j)/(1 + \exp(\mu_j))$ and $\vec{\mu} = \mathbf{X}\vec{\beta}$
- Candes and Sur (2019): $m = 800$ and $n = 4000$

Background

- Regularized estimation methods that correspond to eliciting a shrinkage prior distribution to model parameters (Ridge Regression; LASSO; Spike and Slab; ABSLOPE)
- Empirical Bayes (Robbins, 1956; James and Stein, 1961; Brown, 1966; Efron et al., 2001; Sun and Cai, 2007; Brown and Greenshtein, 2009; Efron, 2011)
- Compound decision approach (Robbins, 1951; Zhang, 2003; Weinstein et al., 2018): for sequence model likelihood and compound loss, Bayes rules with respect to the empirical distribution of the parameter vector minimize Risk for any fixed parameter vector.
- Our hierarchical Bayes modelling uses a finite Polya tree on dyadic partitions to define random distributions as shown in Ferguson (1974).

Bayesian perspective

- Prior distribution $\vec{\beta} \sim \pi(\vec{\beta})$
- Conditioning on $\vec{Y} = \vec{y}$ yields the posterior distribution

$$\pi(\vec{\beta}|\vec{y}) \propto \pi(\vec{\theta}) f(\vec{y}|\vec{\theta})$$

- For loss function $L(\hat{\beta}, \vec{\beta})$ the Bayes rule is given by

$$\hat{\beta}^{Bayes}(\vec{y}) = \underset{\hat{\beta}(\vec{y})}{\text{argmin}}\, E_{\vec{\beta} \sim \pi(\vec{\beta}|\vec{y})} L(\hat{\beta}(\vec{y}), \vec{\beta}).$$

Per definition, $\hat{\beta}^{Bayes}(\vec{y})$ minimizes the average risk

$$r(\hat{\beta}) = E_{\vec{y} \sim f(\vec{y})}\, E_{\vec{\beta} \sim \pi(\vec{\beta}|\vec{y})} L(\hat{\beta}(\vec{y}), \vec{\beta}) \qquad (1)$$

Oracle inferential framework

Suppose we have extra information – the parameter order statistic:

$$\vec{\beta}^{ord} = (\beta_{(1)} \leq \cdots \leq \beta_{(m)}) \quad \Leftrightarrow \quad \text{knowing empirical dist. of } \vec{\beta}$$

$\Rightarrow$ Conditioning on $\vec{Y} = \vec{y}$ and on $\vec{\beta}^{ord}$ yields better Bayes rules

- Given $\vec{y}$ and $\vec{\beta}^{ord}$ we may derive $\pi(\vec{\beta}|\vec{y}, \vec{\beta}^{ord})$
- A Bayes rule may now be derived

$$\hat{\beta}^{Bayes}(\vec{y}; \vec{\beta}^{ord}) = \underset{\hat{\beta}(\vec{y}, \vec{\beta}^{ord})}{\operatorname{argmin}} E_{\vec{\beta} \sim \pi(\vec{\beta}|\vec{y}, \vec{\beta}^{ord})} L(\hat{\beta}(\vec{y}, \vec{\beta}^{ord}), \vec{\beta})$$

And expressing the average risk in (1)

$$E_{\vec{y}, \vec{\beta}^{ord} \sim f(\vec{y}, \vec{\beta}^{ord})}\, E_{\vec{\beta} \sim \pi(\vec{\beta}|\vec{y}, \vec{\beta}^{ord})} L(\hat{\beta}(\vec{y}, \vec{\beta}^{ord}), \vec{\beta})$$

reveals that $\hat{\beta}^{Bayes}(\vec{y}; \vec{\beta}^{ord})$ yields smaller average risk then $\hat{\beta}^{Bayes}(\vec{y})$

Oracle inferential framework (cont.)

- Let $\mathcal{P}_m$ denote set of permutations on $\{1 \cdots m\}$.
  Then $\forall \vec{\beta}$, $\exists \tau' \in \mathcal{P}_m$ for which $\tau'(\vec{\beta}^{ord}) = \vec{\beta}$.
- Thus, π specifies distribution for $\vec{\beta}^{ord}$ and on $\mathcal{P}_m$

$$\pi(\vec{\beta}^{ord}) = \sum_{\tau \in \mathcal{P}_m} \pi(\tau(\vec{\beta}^{ord}))\,, \quad \tilde{\pi}(\tau|\vec{\beta}^{ord}) = \frac{\pi(\tau(\vec{\beta}^{ord}))}{\pi(\vec{\beta}^{ord})}$$

- We may then express

$$\begin{aligned} \pi(\vec{\beta}|\vec{y}, \vec{\beta}^{ord}) &= \frac{f(\vec{\beta}, \vec{y}, \vec{\beta}^{ord})}{f(\vec{y}, \vec{\beta}^{ord})} = \frac{f(\tau'(\vec{\beta}^{ord}), \vec{y})}{\sum_{\tau \in \mathcal{P}_m} f(\tau(\vec{\beta}^{ord}), \vec{y})} \\ &= \frac{f(\vec{y}|\tau'(\vec{\beta}^{ord}))\pi(\tau'|\vec{\beta}^{ord})}{\sum_{\tau \in \mathcal{P}_m} f(\vec{y}|\tau(\vec{\beta}^{ord}))\pi(\tau|\vec{\beta}^{ord}))} \end{aligned}$$

Oracle inferential framework – symmetric priors

- For cases in which all ordering $\vec{\beta}^{ord}$ are apriori equally probable (in shrinkage priors components of $\vec{\beta}$ iid; MLE = flat prior)

$$\tilde{\pi}(\tau'|\vec{\beta}^{ord}) = \frac{\pi(\tau'(\vec{\beta}^{ord}))}{\sum_{\tau \in \mathcal{P}_m} \pi(\tau(\vec{\beta}^{ord}))} = \frac{1}{m!}$$

for which we get

$$\pi(\vec{\beta}|\vec{y}, \vec{\beta}^{ord}) = \frac{f(\vec{y}|\tau'(\vec{\beta}^{ord}))}{\sum_{\tau \in \mathcal{P}_m} f(\vec{y}|\tau(\vec{\beta}^{ord}))} \qquad (2)$$

- e.g. Bayes rule for $L(\hat{\beta}, \vec{\beta}) = \|\hat{\beta} - \vec{\beta}\|^2$ is

$$\hat{\beta}^{Bayes}(\vec{y}; \vec{\beta}^{ord}) = \frac{\sum_{\tau \in \mathcal{P}_m} \tau(\vec{\beta}^{ord}) f(\vec{y}|\tau(\vec{\beta}^{ord}))}{\sum_{\tau \in \mathcal{P}_m} f(\vec{y}|\tau(\vec{\beta}^{ord}))}$$

Frequentist perspective on oracle inferential framework

- Fixed unknown $\vec{\beta}$ the goal is to find $\hat{\beta}$ minimizing the Risk

$$R(\hat{\beta}; \vec{\beta}) = E_{\vec{Y} \sim f(\vec{y}|\vec{\beta})} L(\hat{\beta}(\vec{Y}), \vec{\beta})$$

- To show that $\hat{\beta}^{Bayes}(\vec{y}; \vec{\beta}^{ord})$ yields small Risk we consider $(T, \vec{W})$:
  - $T \in \mathcal{P}_m$ is the parameter with $\Pr(T = \tau) = 1/m!$, $\vec{W}$ is the data with

$$\vec{W}|T = \tau \ \sim f(\vec{y}|\tau(\vec{\beta}^{ord})).$$

    It is easy to see that

$$\Pr(T = \tau'|\vec{W} = \vec{y}) = \frac{f(\vec{y}|\tau'(\vec{\beta}^{ord}))}{\sum_{\tau \in \mathcal{P}_m} f(\vec{y}|\tau(\vec{\beta}^{ord}))}$$

  - As this is same posterior distribution as (2), then also for $(T, \vec{W})$ $\hat{\beta}^{Bayes}(\vec{y}; \vec{\beta}^{ord})$ is Bayes rule for $L(\hat{\beta}(\vec{W}), T(\vec{\beta}^{ord}))$.

Frequentist perspective (cont.)

- Per construction, $\hat\beta^{Bayes}(\vec{y};\vec{\beta}^{ord})$ minimizes the average risk for $(T,\vec{W})$

$$
\begin{aligned}
E_{T,\vec{W}}\, L(\hat\beta(\vec{W}),T(\vec{\beta}^{ord})) &= E_T E_{\vec{W}|T}\, L(\hat\beta(\vec{W}),T(\vec{\beta}^{ord})) \\
&= \sum_{\tau\in\mathcal{P}_m} \frac{1}{m!}\, E_{\vec{W}\sim f(\vec{y}|\tau(\vec{\beta}^{ord}))}\, L(\hat\beta(\vec{W}),\tau(\vec{\beta}^{ord})) \\
&= \sum_{\tau\in\mathcal{P}_m} \frac{1}{m!}\, R(\hat\beta;\tau(\vec{\beta}^{ord})) \qquad (3)
\end{aligned}
$$

- Expression (3) implies that $\hat\beta^{Bayes}(\vec{y};\vec{\beta}^{ord})$ minimizes the mean Risk over all permutations of $\vec{\beta}^{ord}$ (VERY different than average risk $r(\hat\beta)$ in (1)).
- In particular, as in our example the $R(\hat\beta;\tau(\vec{\beta}^{ord}))$ is approximately the same for all $\tau\in\mathcal{P}_m$, then $\hat\beta^{Bayes}(\vec{y};\vec{\beta}^{ord})$ has small Risk for each $\tau(\vec{\beta}^{ord})$.

Hierarchical Bayes modeling for Large-Scale Inference

Implement hierarchical Bayes model that approximates $\pi(\vec{\beta}|\vec{y}, \vec{\beta}^{ord})$ in (2) and derives Bayes rules that approximate the oracle Bayes rules.

a. We imbed likelihood in (made up) generative model for the data:

1. Generate $f(\beta; \vec{a}, \vec{\pi})$ from hBeta model
2. For $i = 1 \cdots m$ generate iid $\beta_i \sim f(\beta; \vec{a}, \vec{\pi})$
3. Generate $\vec{Y} \sim f(\vec{y}|\vec{\beta})$

b. We use a Gibbs sampler to derive the posterior distribution of the hBeta model given $\vec{Y} = \vec{y}$, in which the Gibbs samples of $f(\beta; \vec{a}, \vec{\pi})$ are deconvolution estimates for distribution of $\vec{\beta}^{ord}$ and Gibbs samples of $\vec{\beta}$ approximate posterior samples from $\pi(\vec{\beta}|\vec{y}, \vec{\beta}^{ord})$ in (2)

c. Our inferences are Bayes rules for Gibbs sampling distribution of $\vec{\beta}$.

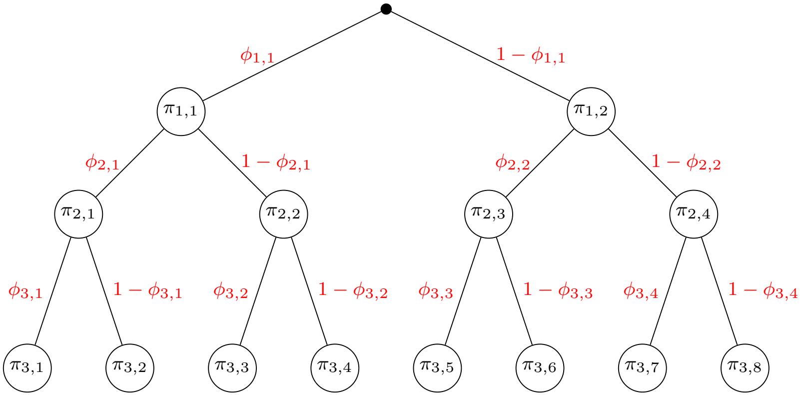
L level hierarchical Beta model

Finite Polya trees on dyadic partition of $\vec{a} = (a_0, \cdots, a_{2^L})$ that models random distributions with step function PDF's

- Model parameters: independent random variables $\phi_{l,j} \sim Beta(\alpha_{l,j}, \beta_{l,j})$ that specify the conditional subinterval probabilities for the dyadic partitions. In the generative model for the data $\phi_{l,j} \sim Beta(1,1)$
- $\pi_{1,1} \cdots \pi_{L,2^L}$ the probabilities of the subintervals in the dyadic partitions are products of the Beta random variables
- Step function PDF

$$f(\beta; \vec{a}, \vec{\pi}) = \pi_{L,1} \cdot \frac{I_{[a_0,a_1]}(\beta)}{a_1 - a_0} + \cdots + \pi_{L,2^L} \cdot \frac{I_{[a_{2^L-1},a_{2^L}]}(\beta)}{a_{2^L} - a_{2^L-1}}$$

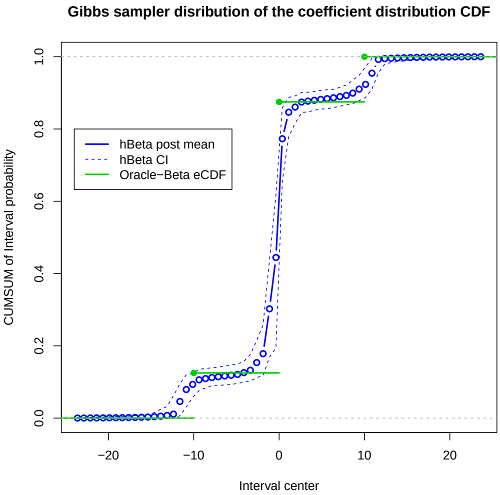
3 level hBeta model – highly regularized 7 parameter model



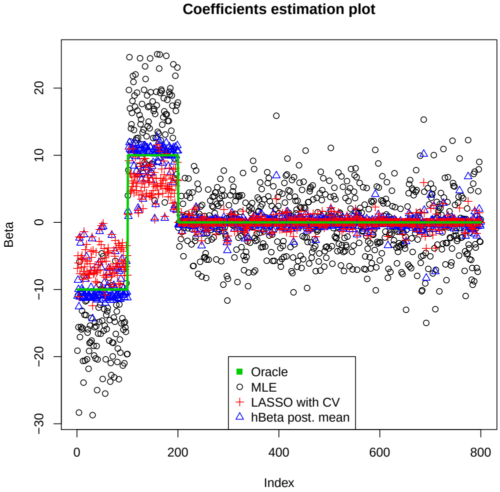
Candes and Sur (2019) simulation study

- Simulate High dimensional logistic regression example
  a. $\vec{\beta} = (-10, \cdots, -10, 10, \cdots, 10, 0, \cdots, 0)$
  b. $\beta_i \sim N(3, 4^2)$
  c. $\beta_i = 0$ or $\beta_i \sim N(3, 4^2)$ with probability 0.5
- We compare five estimates: MLE; “corrected” MLE of Candes and Sur (2019); LASSO and Ridge penalized likelihood estimates (R GLMNET); hBeta posterior means.
- Implement hBeta model with $L = 6$; $\vec{a}$ is a regular 65 point grid on $[-20, 20]$; in each simulated example we run 400 Gibbs sample iterations.

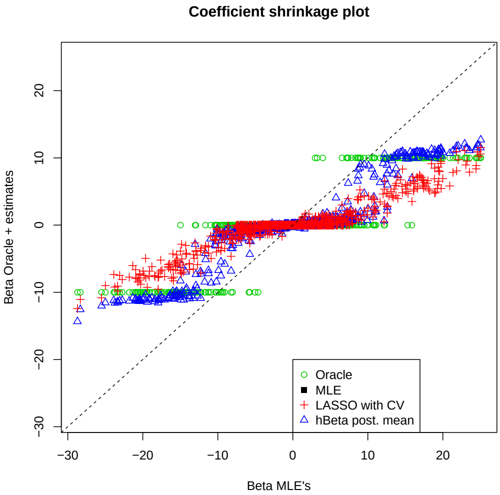
Simulated example a. results



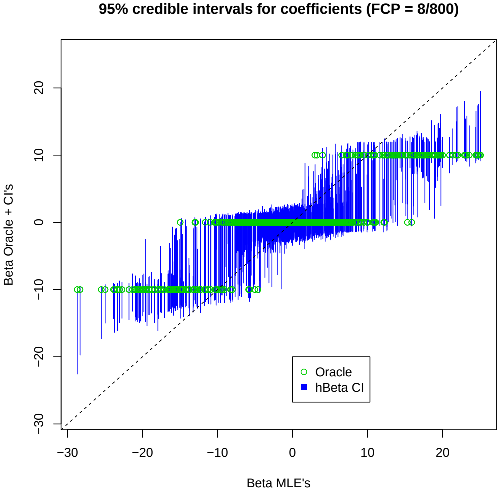
Simulated example a. results



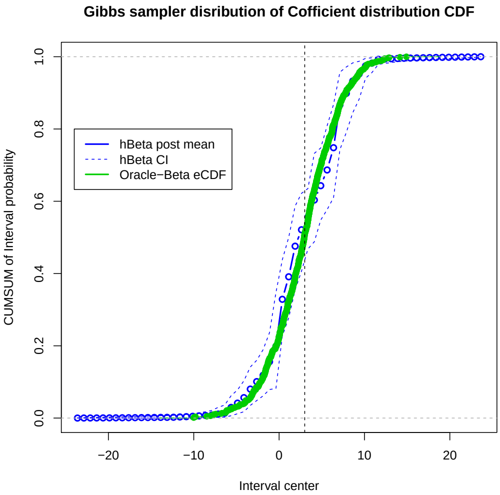
Simulated example a. results



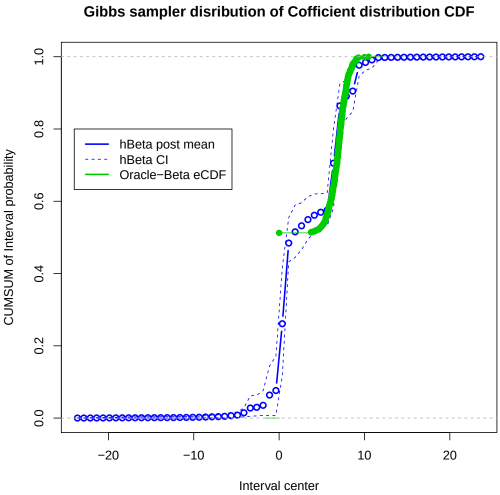
Simulated example a. results



Simulated example b. results



Simulated example c. results



Summary of results

		adj.MLE	LASSO	Ridge	hBeta
Example a	$\vec{\beta}$	0.33	0.19	0.19	0.10
	$\vec{\mu}$	0.31	0.21	0.20	0.11
	$\vec{p}$	0.75	0.49	0.61	0.34
Example b	$\vec{\beta}$	0.34	0.38	0.26	0.17
	$\vec{\mu}$	0.32	0.38	0.27	0.17
	$\vec{p}$	0.75	0.80	0.64	0.32
Example c	$\vec{\beta}$	0.36	0.27	0.25	0.18
	$\vec{\mu}$	0.34	0.26	0.26	0.19
	$\vec{p}$	0.76	0.67	0.63	0.48

Table: MSE for single realization displayed as fractions of the MSE for the MLE.

Discussion

- Propose GENERAL comprehensive eBayes approach for Large-Scale inference with explicit estimation target – the empirical distribution of $\vec{\beta}$.
- Scope of application is cases in which there is no previous information on problem (prior exchangeability in $\vec{\beta}$ excludes information from previous studies).
- Blessing of dimensionality: (1) distribution of $\vec{\beta}$ is easy to estimate in Large-Scale problems; (2) as the Risk tends to be similar for permutations of $\vec{\beta}$ our methods have good frequentist properties.
- Our methodology may also be used for diagnostics, specifying the difficulty of inferential problems and comparing and evaluating estimation methods.

Thank You!