A new double empirical Bayes approach for high-dimensional problems[1][2][3]

Ryan Martin
North Carolina State University
www4.stat.ncsu.edu/~rmartin

WHOA–PSI
St. Louis, MO
09/30/2016

[1] Joint work with Stephen Walker at UT–Austin
[2] Supported by NSF DMS–1507073.
[3] Paper is to appear in *Bernoulli*, see arXiv:1406.7718

Introduction

- For high-dim problems, e.g., $p \gg n$ regression, the Bayesian's choice of prior matters a lot.
- In particular, if the prior does not have sufficiently heavy tails, then posterior may behave sub-optimally.
- But, the priors that work in theory are non-conjugate, so the posterior computations can be difficult.
- *Question:* Can tail conditions and difficult computation be avoided without sacrificing optimality?
- *Answer:* YES, provided we center properly...

Double empirical Bayes

- To present the idea, consider the simpler normal mean model,[4] where $Y_i \sim \mathrm{N}(\theta_i, 1)$, $i = 1, \ldots, n$.
- Write the vector θ as (S, θ_S), where $S \subseteq \{1, 2, \ldots, n\}$ says which entries are non-zero, and θ_S is the non-zero value.
- Take prior $\pi(S)$ and a conditional prior for θ_S, given S.
- Posterior concentration requires that the conditional prior have tails at least as heavy as Laplace.
- Resulting posterior computation is more difficult compared to that based on a conjugate normal prior.
- To avoid the heavy-tailed prior, it is tempting to center a normal prior for θ_S at $\hat{\theta}_S = Y_S$...

[4] M. and Walker, *Electron. J. Stat.*, 2014, `arXiv:1304.7366`.

Double empirical Bayes (cont.)

- But centering prior for θ_S at $\hat{\theta}_S = Y_S$ is too greedy.
- Rather than abandon the centering idea, let's consider a second "empiricalization" step:

$$\textit{replace } L_n(\theta) \textit{ with } L_n(\theta)^\alpha, \textit{ for some } \alpha \in (0,1).$$

- Using a power likelihood is "empirical Bayes"

$$L_n(\theta)^\alpha \, \Pi(d\theta) = L_n(\theta) \, \frac{\Pi(d\theta)}{L_n(\theta)^{1-\alpha}}.$$

- So, we propose to combine the usual likelihood with a prior that uses data in two ways:

$$\text{Centering} + \text{Regularization} = \text{"double empirical Bayes"}$$

High-dim linear model

- Consider a standard linear model $Y = X\beta + \varepsilon$, where
  - Y is a n-vector of responses;
  - X is an $n \times p$ matrix of predictor variables;
  - β is a p-vector of coefficients;
  - and $\varepsilon \sim \mathrm{N}_n(0, \sigma^2 I_n)$ is the error, with σ known.
- Key point is that $p = p_n \gg n$, i.e., *high-dimensional!*
- Standard assumption: β is *sparse*, i.e., most entries are zero.
- Goal is to estimate the true sparse coefficient vector $\beta^\star$ and/or the true model $S^\star = \{j : \beta_j^\star \neq 0\}$.

Double EB for the linear model

- Need to specify prior for S and conditional prior for β_S.
- Restrict[5] prior to $|S| \leq R := \text{rank}(X)$.
- Since $|S| \leq R$, there is a least-squares estimator $\hat{\beta}_S$ to center on, and the classical distribution theory is available.
- Conditional prior for β_S, given S:

$$\beta_S \mid S \sim \mathrm{N}_{|S|}(\hat{\beta}_S, \gamma^{-1}(X_S^\top X_S)^{-1}), \quad \gamma > 0.$$

- Prior for $s = |S|$ and a conditional prior for S, given s:
  - Prior for S, given s, is uniform on the $\binom{p}{s}$ combinations.
  - Prior for s is restricted to $s \leq R$, but can take various forms.

[5] Not really a restriction: posterior should not be supported on models which are too complex for the given data!

Double EB for the linear model (cont.)

- Take a fraction $\alpha \in (0,1)$ to be used on the likelihood.
- Then the double empirical Bayes posterior for β is

$$\Pi^n(d\beta) \propto \sum_{S:|S|\leq R} \Big\{ \pi(S)\, \mathsf{N}_n(Y \mid X\beta, \sigma^2 I_n)^\alpha \times \mathsf{N}_{|S|}(d\beta_S \mid \hat{\beta}_S, \gamma^{-1}(X_S^\top X_S)^{-1})\, \delta_0(d\beta_{S^c}) \Big\}.$$

- We focus on the "complexity prior" for $s = |S|$:

$$f_n(s) \propto c^{-s} p^{-as}, \quad s = 0, 1, \ldots, R; \quad a, c > 0.$$

Double EB posterior concentration

- Assumptions:
  - Take complexity prior $f_n(s)$ supported on $s \leq R$.
  - X_S is full rank for all $|S| \leq R$.
- Let $B_n = \{\beta \in \mathbb{R}^{p_n} : \|\beta\|_0 = s_n\}$, for $s_n \leq n \ll p_n$.
- For prediction loss, the rate $\varepsilon_n = s_n \log(p_n/s_n)$ is *almost*[6] the optimal minimax rate on B_n.

Posterior concentration – prediction loss.

Under stated assumptions, there exists $M > 0$ such that

$$\sup_{\beta^\star \in B_n} \mathrm{E}_{\beta^\star} \Pi^n(\{\beta \in \mathbb{R}^{p_n} : \|X(\beta - \beta^\star)\|^2 > M\varepsilon_n\}) \to 0, \quad n \to \infty.$$

[6] Actual minimax rate is $\min\{R, s_n \log(p_n/s_n)\}$, which can be achieved by our double EB posterior with a slightly different prior...

Double EB posterior concentration (cont.)

- Does the posterior Π^n concentrate on sparse βs?
- For a given β, let $S_\beta = \{j : \beta_j \neq 0\}$ be its configuration.
- Ideally, Π^n concentrates on β such that $|S_\beta| = s_n$ $(= |S_{\beta^\star}|)$.

Posterior concentration – effective dimension.

Under the same assumptions, there exists $K > 0$ such that

$$\sup_{\beta^\star \in B_n} \mathrm{E}_{\beta^\star} \Pi^n(\{\beta \in \mathbb{R}^{p_n} : |S_\beta| > K s_n\}) \to 0, \quad n \to \infty.$$

Double EB posterior concentration (cont.)

- Other concentration rate results are available concerning estimation and model selection.
- These results are more technical and require some additional assumptions on, e.g., the condition number of X, etc.
- Here I give just a summary of the results:
  - Our concentration rate relative to ℓ_2-loss for β is optimal, better than the analogous results for lasso, etc.
  - Under the usual "beta-min" condition, we have model selection consistency, i.e., $\mathsf{E}_{\beta^\star} \Pi^n(\{\beta : S_\beta = S_{\beta^\star}\}) \to 1$.

Computational considerations

- Our model is conjugate, so a lot of analytical calculations can be done to simplify the Monte Carlo.
- In particular, the marginal posterior for the configuration S is

$$\pi^n(S) \propto \pi(S)(\gamma + \alpha/\sigma^2)^{-|S|/2} e^{-(\alpha/2\sigma^2)\|Y - \hat{Y}_S\|^2}.$$

- Now use Metropolis–Hastings to sample S.
- Moreover, if samples of β_S, given S, are also desired, these can be obtained via sampling from a suitable normal.
- R code available at `www4.stat.ncsu.edu/~rmartin`.
- What about tuning parameters?
  - Two (α, γ) settings: $(0.999, 0.001)$ and $(1, 0)$;
  - $a = 0.05$ and $c = 1$;
  - For σ^2, plug in a suitable estimator.

Double empirical Bayes model selection

- Inclusion probabilities $w_j = \Pi^n(S \ni j)$, $j = 1, \ldots, p$.
- To get $\hat{S}$, employ the *median probability rule:*

$$\text{select variable } j \text{ iff } w_j > 0.5, \quad j = 1, \ldots, p.$$

- Could also use a MAP rule...

Numerical results

- Normal linear model, with $\sigma^2 = 1$.
- Predictor variable matrix is multivariate normal, mean zero, unit variance, and pairwise correlation 0.25.
- Consider just the following setting:

$$n = 200, \quad p = 1000, \quad \beta_{S^\star} = (0.6, 1.2, 1.8, 2.4, 3.0)^\top.$$

- Compare double empirical Bayes with:
  - BASAD (Narisetty and He 2014);
  - BASAD.BIC (Narisetty and He 2014);
  - BCR.Joint (Bondell and Reich 2012);
  - SpikeSlab (Ishwaran and Rao 2005);
  - Lasso.BIC (Tibshirani 1996);
  - EN.BIC (Zou and Hastie 2005);
  - SCAD.BIC (Fan and Li 2001).

Numerical results (cont.)

Method	$\bar{w}_0$	$\bar{w}_1$	$P(\hat{S} = S^\star)$	$P(\hat{S} \supseteq S^\star)$	FDR
BASAD	0.000	0.986	0.930	0.950	0.000
BASAD.BIC	0.000	0.986	0.720	0.990	0.046
BCR.Joint			0.090	0.250	0.176
SpikeSlab			0.000	0.050	0.574
Lasso.BIC			0.020	1.000	0.430
EN.BIC			0.325	1.000	0.177
SCAD.BIC			0.650	1.000	0.091
$DEB_{\alpha=0.999}$	0.000	0.998	0.945	0.990	0.015
$DEB_{\alpha=1}$	0.000	0.999	0.950	0.995	0.011

(*DEB:* $\approx$ 25s to compute, using ordinary R)

Summary

- New and general *double empirical Bayes* framework based on using data in the prior in two ways:
  - centering
  - regularization.
- Posterior distribution has many desirable concentration results in the important $p \gg n$ linear model problem.
- Computation is fast and relatively easy due to conjugacy; the proper Bayesian methods with provable concentration rate results are much more difficult.
- Numerical results are very promising.

Discussion

- It is clear that the double empirical Bayes method described here would apply in other similar problems.
- But it turns out that the idea is more general.
- For example, in nonparametric problems, one often has in mind some kind of *sieve* that identifies relatively nice portions of the parameter space.
- This yields a sort of "$\theta = (\theta_S, S)$" decomposition like we had above, and we can put a prior on S and center a conditional prior for θ_S on, say, the sieve MLE $\hat{\theta}_S$.
- Very general results on adaptive posterior concentration rates with empirical priors have recently been worked out.[7]

[7] M. and Walker, `arXiv:1604.05734`

The end

Thank you!