Empirical priors for high-dimensional problems

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

Jessie Jeng’s ST790 course
High-Dimensional Statistical Inference
March 27th, 2019
High-dimensional problems are everywhere now.

Hopeless without some structural assumptions.

*Sparsity* is a common type of structure, i.e., most signals are zero/small but a few are large.

An assumed structure is “prior information,” so Bayesians incorporate this structure via a prior.

But except for this vague notion of structure, we usually don’t know anything.

And since it’s high-dim, the choice of prior on the structure-specific parameters actually matters!
Priors that lead to posteriors with
- relatively easy computation tend to be thin-tailed and have sub-optimal theoretical properties
- good theoretical properties are heavier-tailed and tend to be more difficult to compute with.

*Do we have to choose between theory and computation?*

I say *NO!*

Take-away messages:
- *Tails don’t matter if centering is right.*
- To get the centering right, we need to use the data, i.e., an empirical prior.
This lecture

- Background on Bayesian stuff\(^1\)
- Sparse normal mean problem
  - empirical prior construction
  - posterior concentration results
  - uncertainty quantification
- Beyond the sparse normal mean problem
  - a piecewise constant structure
  - high-dim regression
- A general theory?
- Monotone density estimation
- Concluding remarks

\(^1\)My understanding is that you’ve talked about Bayesian large-scale testing, local fdr, etc. I worked on such things before and, if you’re interested, check out a recent review paper I wrote: arXiv:1812.02149
Basic Bayesian approach:
- statistical model with unknown $\theta \in \Theta$
- get prior for $\theta$
- combine prior with data/likelihood using Bayes’s formula
- use posterior distribution to make inference.

Selling points:
- conceptually easy to do
- gives answers to basically any question in one stroke.

But there are serious issues, so I’m not a Bayesian!²

The approach I’m going to talk about isn’t Bayesian.

But it’s better than Bayesian and the Bayesians don’t like it, so it’s a win–win for me!

²https://www.researchers.one/article/2019-02-1
Statistical model: $Y^n \sim P^n_\theta$, joint density $p^n_\theta$.

$\theta$ could be high- or infinite-dimensional.

Let $\Pi_n$ be a prior distribution on $\Theta$.

$L_n(\theta) = p^n_\theta(Y^n)$ is the likelihood.

Bayes’s formula defines the posterior distribution:

$$\Pi^n(d\theta) \propto L_n(\theta) \Pi_n(d\theta).$$

Answers to statistical questions based on the posterior:

- Estimators via posterior expectations
- confidence sets using posterior quantiles
- tests using posterior probabilities
- predict a new $\tilde{Y}$ by integrating $p_\theta(\tilde{y})$ wrt posterior
For low-dim problems, this is relatively easy:
- prior distribution for low-dim $\theta$ doesn’t matter too much
- MCMC computation is fast
- posterior has nice asymptotic properties almost automatically.

In high-dim problems, however:
- prior matters
- computation is hard(er)
- asymptotics are hard(er)

“High-dim” means more parameters than data points.

I discuss an important example below.
Sparse normal means problem

- $Y^n = (Y_1, \ldots, Y_n)$, independent, $Y_i \sim \text{N}(\theta_i, 1)$.
- High-dim, since each $Y_i$ has its own $\theta_i$.
- Assume vector $\theta$ is sparse, i.e., most $\theta_i$’s are 0.
- Non-Bayesian approach might do some kind of soft or hard thresholding, e.g., lasso, etc.
- How about a Bayesian?
- Express $\theta$ vector as a pair $(S, \theta_S)$:
  - $S \subseteq \{1, 2, \ldots, n\}$ denotes the location of non-zeros
  - $\theta_S$ the $|S|$-vector of non-zero values.
- Need a prior for $(S, \theta_S)$...\(^3\)

\(^3\)This discrete-$S$ approach is not the only Bayesian strategy, some adopt a “continuous shrinkage prior” like horseshoe, but I don’t really like this.
Can use sparsity assumption to motivate a marginal prior for $S$, more on this below.

But what about the conditional prior for $\theta_S$?

Since the model is normal, computational simplicity suggests a conjugate normal prior for $\theta_S$.

However, there are results\(^4\) that say a prior for $\theta_S$ with thin tails leads to sub-optimal posterior concentration.

Consequently, good theory requires tails to be heavier than normal, which makes computation difficult.

Apparently theory and computation are at odds.

\textit{Or maybe not...}

\(^4\) e.g., Castillo & van der Vaart, \textit{Ann. Statist.}, 2012
Idea: If sufficient information about $\theta_S$ is lacking to justify a prior, then why not let data help guide the choice?

Marginal prior $\pi_n(S)$ for $S$:
- $|S| \sim f(s) \propto n^{-as}$, $s = 0, \ldots, n$, where $a > 0$.
- Given the size, $S$ is uniform over all configs of that size.

Conditional prior for $\theta_S$, given $S$:

$$\pi_n(\theta_S \mid S) = N_{|S|}(Y_S, \gamma^{-1} I_{|S|}), \quad \gamma \in (0, 1).$$

Conditional prior for $\theta_S$, given $S$, times the marginal prior for $S$ gives an empirical prior for $\theta$ — call it $\Pi_n$.

Intuition:
- Informative prior on the thing we know about, $S$;
- “Non-informative” on the thing we don't know about, $\theta_S$
Combine prior and likelihood in *almost* the usual way:

$$\Pi_n(d\theta) \propto L_n(\theta)^\alpha \Pi_n(d\theta), \quad \alpha \in (0, 1).$$

The power $\alpha$ is unusual, a regularizer, more on this later.

Relatively simple computations thanks to:
- closed-form expression for posterior of $S$,

$$\pi^n(S) \propto \pi_n(S) e^{-\alpha\|Y_{sc}\|^2/2} (1 + \gamma^{-1} \alpha)^{-|S|/2}.$$

- “closed-form” marginal posterior for each $\theta_i$, more later.
- R code is on my website.

Good theoretical properties too...
Sparse normal means, cont.

- Summary of theoretical results spread over several papers.\(^5\)
- **Recovery.**
  - If true \(\theta^\star\) is \(s^\star\)-sparse, then \(\Pi_n\) concentrates around \(\theta^\star\) at the minimax rate \(s^\star \log(n/s^\star)\), i.e., for some \(M > 0\),

\[
\sup_{\theta^\star: |S_{\theta^\star}| = s^\star} \mathbb{E}_{\theta^\star} \Pi_n(\{\theta : \|\theta - \theta^\star\|^2 > Ms^\star \log(n/s^\star)\}) \to 0.
\]

- Prior doesn’t know \(s^\star\), so concentration rate is *adaptive*.
- Posterior mean is an asymptotically minimax estimator.
- \(\Pi_n\) supported on a set of dimension roughly \(s^\star\).

Sparse normal means, cont.

Average value of $||\hat{\theta} - \theta^*||^2$ in 100 replicates at $n = 200$.

<table>
<thead>
<tr>
<th>$s^*$</th>
<th>$10 = 0.05n$</th>
<th>$20 = 0.10n$</th>
<th>$40 = 0.20n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta^<em>_{S^</em>}$</td>
<td>7</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>BL</td>
<td>65</td>
<td>69</td>
<td>100</td>
</tr>
<tr>
<td>DL</td>
<td>16</td>
<td>14</td>
<td>33</td>
</tr>
<tr>
<td>HS</td>
<td>16</td>
<td>15</td>
<td>36</td>
</tr>
<tr>
<td>$EB$</td>
<td><strong>13</strong></td>
<td><strong>12</strong></td>
<td><strong>24</strong></td>
</tr>
</tbody>
</table>

(BL = Bayesian lasso; DL = Dirichlet-Laplace; HS = horseshoe)
Sparse normal means, cont.

- **Structure Learning.**
  - Inclusion probability satisfies

\[
p^n_i := \pi^n(S \ni i) \rightarrow \begin{cases} 
0 & \text{if } \theta^*_i = 0 \\
1 & \text{if } |\theta^*_i|^2 \gtrsim \log n.
\end{cases}
\]

- If all non-zero \(|\theta^*_i|\) are “large enough,” i.e., if

\[
\min_{j \in S^*} |\theta^*_j| \gtrsim (\log n)^{1/2},
\]

then \(\pi^n(S = S^*) \rightarrow 1\), where \(S^* = S_{\theta^*}\).
Sparse normal means, cont.

$n = 200$ and $\theta_1^* = \cdots = \theta_{10}^* = 7$, others zero.
Sparse normal means, cont.

- **Uncertainty quantification.**
  - Marginal posterior for $\theta_i$ is available in “closed-form,”

  $$(1 - p^n_i) \delta_0 + p^n_i N(Y_i, \nu_\alpha),$$

  where $p^n_i$ is the inclusion prob and $\nu_\alpha = (\alpha + \gamma)^{-1}$.

  - Note: if $\alpha \approx 1$ and $\gamma \approx 0$, then $\nu_\alpha \approx \text{var}(Y_i)$.

  - Credible interval for $\theta_i$ fails to hit the target coverage prob asymptotically only if $\theta^*_i \neq 0$ and $p^n_i \not\to 1$.

  - Conditions to avoid this, hopefully weaker than beta-min?

  - Specific results presented in arXiv:1812.02150, joint work with former NCSU student, Dr. Bo Ning.
$n = 200$, $\theta^* = (\text{five 7's, five 1.5's, } \theta^*_{11}, \text{rest equal 0})$.

Check coverage and length of 95% credible interval for $\theta_{11}$.

- **HS**: $\sigma^2 = 1$ fixed, $\tau$ via MMLE, horseshoe package
- **EB1**: $\alpha = 0.99$, $\gamma = 0.01$, from arXiv:1304.7366
- **EB2**: $\alpha = 0.99$, $\gamma = 0.01$, based on arXiv:1406.7718
Remarks about $\alpha$

- The power $\alpha$ might be a concern for some.
- Can absorb $\alpha$ into the empirical prior, regularization,
  \[
  \prod^n(d\theta) \propto L_n(\theta) \frac{\prod_n(d\theta)}{L_n(\theta)^{1-\alpha}}.
  \]
- Anyway, the above convergence theory holds for any $\alpha \in (0, 1)$ so there is no practical effect.
- $\alpha$ is only required for adaptive rates.
- It’s also possible to do the necessary regularization in the prior for $S$, but that’s more complicated.
- And there are situations where $\alpha < 1$ actually helps.\(^6\)

---
\(^6\)e.g., Peter Grünwald’s *SafeBayes.*
The above formulation seems especially tailored to the sparse normal means problem.

Can other problems/structures be handled similarly?

The approach is actually quite general — more later.

For now, two cases similar to that above:
- normal mean with piecewise constant structure
- high-dim linear regression
- Same normal mean model but, instead of sparsity, assume that there are blocks of means having the same value.
- This model is for change-point problems, regression trees, copy number variation, etc.
Makes sense to write mean vector $\theta$ as $(B, \theta_B)$, where
- $B$ is block configuration and
- $\theta_B$ are the block-specific parameters.

Sparsity is still relevant — not too many blocks.

My empirical Bayes approach\textsuperscript{7} proceeds as follows:
- informative prior on $|B|$, number of blocks
- uniform prior on block configuration, $B$, given the number
- non-informative empirical prior on block-specific means, $\theta_B$
- power $\alpha \in (0, 1)$ on likelihood.

R code is on my website.

\textsuperscript{7}arXiv:1712.03848
Let $\theta^*$ denote the true piecewise constant mean vector, and $B_{\theta^*}$ its “block configuration.”

If $\theta^*$ is such that $|B_{\theta^*}| > 2$, then the minimax rate is

$$\varepsilon_n(\theta^*) = |B_{\theta^*}| \log\left(\frac{en}{|B_{\theta^*}|}\right).$$

Our theorem says the empirical Bayes posterior adaptively attains this rate, i.e.,

$$\sup_{\theta^*} \mathbb{E}_{\theta^*} \Pi^n(\{\theta : \|\theta - \theta\|^2 > M\varepsilon_n(\theta^*)\}) \to 0, \quad n \to \infty.$$ 

A consequence is that the empirical Bayes posterior mean, $\hat{\theta}_n$, is an adaptive minimax estimator.
**Standard test example:**
- \( n = 497 \) means; shown in green
- \(|B_{\theta^*}| = 7\) blocks of unique means.

**Left plot** shows \( \hat{\theta}_n \) (black), 95% credible intervals (red)

**Right plot** shows posterior distribution for \(|B|\).
Not surprisingly, a similar analysis can be carried out for the linear regression setting, i.e.,

\[ Y = X\beta + \varepsilon, \]

where \( Y \) is a \( n \)-vector, \( X \) is a \( n \times p \) matrix, with \( p \gg n \).

If \( \beta \) is sparse, it makes sense to express \( \beta \) as \((S, \beta_S)\).

**Empirical prior \( \Pi_n \) for \( \beta \):**

- Similar prior encodes sparsity in \( S \),
- \((\beta_S \mid S) \sim N_{|S|}(\hat{\beta}_S, \gamma^{-1}(X_S^TX_S)^{-1})\).

**Posterior:** \( \Pi_n(d\beta) \propto L_n(\beta)^\alpha \Pi_n(d\beta) \).

As before, *is the posterior any good?*
Good computational and theoretical properties.\textsuperscript{8}

Computation:
- Conjugate normal prior, can integrate out $\beta_S$
- Marginal posterior for $S$, simple Metropolis.
- R code on my website.

Convergence theory:
- Minimax optimal rate with respect to $\|X\beta - X\beta^*\|$.
- Optimal rate for $\|\beta - \beta^*\|$.
- Model selection consistency.

Other things (with NCSU students):
- arXiv:1810.00739 is on empirical correlation-adaptative priors with Mr. Chang Liu and Ms. Yue Yang
- arXiv:1903.00961 is on prediction with Ms. Yiqi Tang.
High-dim linear regression, cont.

- Model selection results.\(^9\)
- Rows of \(X\): normal, pairwise correlation 0.25.
- \(n = 200, \ p = 1000, \ \sigma^2 = 1.\)
- True signal: \(\beta_{S^*}^\star = (0.6, 1.2, 1.8, 2.4, 3.0)^\top.\)

<table>
<thead>
<tr>
<th>Method</th>
<th>(P(\hat{S} = S^\star))</th>
<th>(P(\hat{S} \supseteq S^\star))</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASAD</td>
<td>0.930</td>
<td>0.950</td>
</tr>
<tr>
<td>BASAD.BIC</td>
<td>0.720</td>
<td>0.990</td>
</tr>
<tr>
<td>BCR.Joint</td>
<td>0.090</td>
<td>0.250</td>
</tr>
<tr>
<td>Lasso.BIC</td>
<td>0.020</td>
<td>1.000</td>
</tr>
<tr>
<td>EN.BIC</td>
<td>0.325</td>
<td>1.000</td>
</tr>
<tr>
<td>SCAD.BIC</td>
<td>0.650</td>
<td>1.000</td>
</tr>
<tr>
<td>(EB)</td>
<td>0.945</td>
<td>0.990</td>
</tr>
</tbody>
</table>

A general formulation/theory?

- Two examples above are still “normal.”
- **Question**: Is it possible to do the empirical prior construction for more general kinds of problems?
- **Yes!** New strategy\(^{10}\) based on an empirical KL condition can handle a lot, even nonparametric problems.
- This is messy, so my goal is a modest one:
  - share with you the basic idea
  - focus on intuition
  - one example: monotone density estimation\(^{11}\)

---

\(^{10}\) arXiv:1604.05734

\(^{11}\) arXiv:1706.08567
In a Bayesian setting, good asymptotic properties require that the prior put sufficient mass near the true $\theta^*$. Don’t know $\theta^*$ so it’s not obvious how to arrange this. This brings us back to the tails of the prior...

Remember:

- *tails don’t matter if the centering is right.*
- The “right” centering depends on data.
The KL condition is standard in the literature. That is, if the target rate is $\varepsilon_n$, then the prior should assign sufficiently large mass to the set

$$\{\theta \in \Theta_n : K(p^n_{\theta^*}, p^n_\theta) \leq n\varepsilon_n^2\}.$$

- If we knew $\theta^*$, then just put all the mass there.
- But we don’t know $\theta^*$, hence tails need to be chosen carefully.
- Or we could try an informative center...
- Idea: Replace KL condition with an *empirical version*?
For a suitable sieve, $\Theta_n$, let be $\hat{\theta}_n$ a sieve MLE, and

$$\mathcal{L}_n = \{ \theta \in \Theta_n : L_n(\theta) \geq e^{-cn_2}L_n(\hat{\theta}_n) \}.$$ 

Requiring the prior to put sufficient mass on $\mathcal{L}_n$ amounts to an empirical KL condition.

More specifically, I’ll require the prior $\Pi_n$ satisfy

$$\Pi_n(\mathcal{L}_n) > e^{-cn_2}, \quad \text{all large } n.$$ 

Key consequence:

- Let $D_n = \int_{\Theta_n} \{L_n(\theta)/L_n(\theta^*)\} \Pi_n(d\theta)$.
- If $n$ is large, then $D_n > e^{-(c-C)n_2}$. 
- Proof: $D_n > \int_{\mathcal{L}_n} \cdots \Pi_n(d\theta)$.
Lower-bounding the posterior denominator $D_n$ is key.

Comes basically for free under empirical KL condition.

Roughly, the general result\(^{12}\) assumes:
- sieve has good approximation properties
- empirical KL condition holds
- prior tails aren’t too heavy

If these conditions are satisfied for a particular sequence $\varepsilon_n$, then the posterior concentrates around $\theta^*$ at that rate.

This is for a non-adaptive rate result; adaptation requires some more complicated conditions, omitted here.

---

\(^{12}\)arXiv:1604.05734
Empirical priors are “non-informative” but not in the traditional sense.

That is, depending on the context, it may be that $\Pi_n$ is collapsing on $\hat{\theta}_n$ as $n \to \infty$, i.e., it’s not diffuse.

For example, in finite-dim problems, above procedure would suggest a prior with variance $O(n^{-1})$.

Seems counter-intuitive, but:

- No benefit to informative centering if the spread is wide!
- Ultimately, rate isn’t affected by prior concentration.

How to choose constants, etc, for good coverage??
Estimate a monotone non-increasing density $f$ with bounded support $[0, T]$; paper\textsuperscript{13} handles more general case.

Such densities have a mixture-of-uniforms representation

$$f(x) = \int k(x | \mu) \omega(d\mu), \quad k(x | \mu) = \mu^{-1}1(x \leq \mu).$$

Take a sieve of finite mixtures

$$f_\theta(x) = \sum_{j=1}^{J_n} \omega_j k(x | \mu_j), \quad \theta = (\omega, \mu).$$

Prior on $\theta$ leads to prior on $f$.

\textsuperscript{13}arXiv:1706.08567
Sieve: mixtures with $S$-many components, $\mu$'s in $[t, T]$.

Let $\hat{\theta} = (\hat{\omega}, \hat{\mu})$ be sieve MLE.

Prior for $\theta = (\omega, \mu)$:
- $\omega$ and $\mu$ independent;
- $\omega$ is $\text{Dir}(\hat{\alpha})$ with $\hat{\alpha}_j = 1 + c\hat{\omega}_j$;
- $\mu_j$ is $\text{Par}(\hat{\mu}_j, \delta)$.

Simple model, Pareto is conjugate to uniform kernel.

Existing Bayes results can’t use conjugate Pareto prior.
Need to choose $(J, t, T, c, \delta)$ carefully, theory helps.

R code on my website.
Target rate: $\varepsilon_n = (\log n)^{1/3} n^{-1/3}$, near minimax.

Rate doesn’t depend on true $f^*$, no adaptation needed.

Conditions:
- $J_n \propto \varepsilon_n^{-1}$;
- $c_n \propto n\varepsilon_n^{-2}$;
- $\delta_n \log(T_n/t_n) \propto \log n$.

Then there exists $M > 0$ such that

$$E_{f^*} \prod^n \left\{ f : H(f^*, f) > M\varepsilon_n \right\} \to 0,$$

for any $f^*$ monotone, supported on $[0, T^*]$, with $f^*(0) < \infty$. 
Figure 1: Exponential example. Plots of posterior samples (gray), the posterior mean (black), 95% credible band (dashed), the Grenander estimator (blue), and $f^*$ (red).
Monotone density, cont.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$x$</th>
<th>Coverage Prob.</th>
<th>Mean Length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EB</td>
<td>DPM</td>
</tr>
<tr>
<td>100</td>
<td>0.5</td>
<td>0.943</td>
<td>0.987</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.984</td>
<td>0.965</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>0.986</td>
<td>0.938</td>
</tr>
<tr>
<td></td>
<td>3.0</td>
<td>0.990</td>
<td>0.891</td>
</tr>
<tr>
<td>200</td>
<td>0.5</td>
<td>0.934</td>
<td>0.981</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.962</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>0.983</td>
<td>0.914</td>
</tr>
<tr>
<td></td>
<td>3.0</td>
<td>0.970</td>
<td>0.842</td>
</tr>
</tbody>
</table>

Table 1: Exponential example—coverage probability and mean length of the 95% posterior credible regions for the empirical Bayes (EB) and Dirichlet process mixture model (DPM), for several values of $x$ and $n$. 
For high-dim problems, incorporating structure is key.
- We often have relevant information about the structure.
- But a “good” prior for the structure-specific parameters can be difficult to come by.

Empirical priors have some nice features:
- good rates and model selection properties
- promising in uncertainty quantification
- fast to compute?!?!

Some general theory available, but more work to do.

Interesting applications I’m looking at now:
- sparse precision matrix estimation;
- general mixture models;
- high-dim GLMs;
- ...
Thank you!

rgmarti3@ncsu.edu
www4.stat.ncsu.edu/~rmartin

A new open-access publication platform is now open, featuring an author-driven peer review process.

www.researchers.one
www.twitter.com/@ResearchersOne