Laplace approximation for Bayesian variable selection via Le Cam's one-step procedure supplementary materials

AISTATS Spotlight Talk

Tianrui Hou   Yves Atchadé

Boston University

1. Problem setup: the marginal posterior over model supports

Define

$$\bar{\ell}^{\delta}(w;\mathcal{D}) := \ell^{\delta}(w;\mathcal{D}) - \frac{1}{2}\|w\|_2^2, \quad w \in \mathbb{R}^{\|\delta\|_0},$$

where

$$\ell^{\delta}(\vartheta;\mathcal{D}) := \ell((\vartheta,0)_{\delta};\mathcal{D}),$$

and, for $w \in \mathbb{R}^{\|\delta\|_0}$, the notation $(w,0)_\delta$ is the "padded vector of w", denoting the vector of $\mathbb{R}^p$ obtained from w by adding 0 at components where $\delta_j = 0$.

Bayesian variable selection target

For a support vector $\delta \in \{0,1\}^p$, let $\bar{\ell}^{\delta}(w;\mathcal{D})$ be the log-likelihood plus the log-prior over the active coefficients $w \in \mathbb{R}^{\|\delta\|_0}$. The posterior over models is

$$\Pi(\delta \mid \mathcal{D}) \propto p^{-u\|\delta\|_0} \underbrace{\int_{\mathbb{R}^{\|\delta\|_0}} \exp\left(\bar{\ell}^{\delta}(w;\mathcal{D})\right) \mathrm{d}w}_{\text{marginalization } m_{\delta}(\mathcal{D})} .$$

The computational problem

Sampling δ requires evaluating $m_\delta(\mathcal{D})$ for many neighboring models. The integral is not applicable beyond quadratic cases.

Standard Laplace is already an approximation

Around the submodel maximizer

$$\hat{\theta}^{\delta} \in \arg\max_{w \in \mathbb{R}^{\|\delta\|_0}} \bar{\ell}^{\delta}(w;\mathcal{D}), \qquad \hat{\mathcal{H}}_{\delta} = -\nabla^2 \bar{\ell}^{\delta}(\hat{\theta}^{\delta};\mathcal{D}),$$

$$m_\delta(\mathcal{D}) \approx \frac{(2\pi)^{\frac{\|\delta\|_0}{2}}}{\sqrt{\det(\hat{\mathcal{H}}^{\delta})}} \exp\left(\bar{\ell}^{\delta}(\hat{\theta}^{\delta};\mathcal{D})\right).$$

Why exact Laplace is costly

Every support δ has its own optimizer $\hat{\theta}^{\delta}$, so exact Laplace requires repeated local optimization in a large number of sub-models.

Goal: approximate the integral accurately, but without refitting each sub-model.

2. Method: replace each local MLE by a Le Cam one-step correction

Original Taylor/Laplace expansion around any v_0

For any $v_0 \in \mathbb{R}^{\|\delta\|_0}$, use a second-order Taylor approximation of $\bar{\ell}^{\delta}(\cdot;\mathcal{D})$ around v_0:

$$\int_{\mathbb{R}^{\|\delta\|_0}} e^{\bar{\ell}^{\delta}(w;\mathcal{D})}\mathrm{d}w \approx \frac{(2\pi)^{\frac{\|\delta\|_0}{2}}}{\sqrt{\det(\mathcal{H}^{v_0})}} \exp\left(\bar{\ell}^{\delta}(v_0;\mathcal{D}) + \frac{1}{2}(\mathcal{G}^{v_0})^{\mathsf{T}}(\mathcal{H}^{v_0})^{-1}\mathcal{G}^{v_0}\right).$$

$$\mathcal{G}^{v_0} := \nabla\bar{\ell}^{\delta}(w;\mathcal{D})|_{w=v_0}, \qquad \mathcal{H}^{v_0} := -\nabla^2\bar{\ell}^{\delta}(w;\mathcal{D})|_{w=v_0}.$$

One-step point estimate

We no longer fit $\hat{\theta}^{\delta}$ for every δ. Instead, we start with a global estimator $\tilde{\theta} \in \mathbb{R}^p$ and approximate $\hat{\theta}^{\delta}$ by

$$\check{\theta}^{\delta} = \tilde{\theta}^{\delta} + (\mathcal{H}^{\tilde{\theta}^{\delta}})^{-1}\mathcal{G}^{\tilde{\theta}^{\delta}}.$$

Using $\check{\theta}^{\delta}$ and Laplace approximation leads to the approximation of Π given by

$$\check{\Pi}(\delta|\mathcal{D}) \propto \left(\frac{1}{p^{\mathsf{u}}}\right)^{\|\delta\|_0} \frac{\exp\left(\bar{\ell}^{\delta}(\check{\theta}^{\delta};\mathcal{D}) + \frac{1}{2}(\check{\mathcal{G}}^{\delta})^{\mathsf{T}}(\check{\mathcal{H}}^{\delta})^{-1}\check{\mathcal{G}}^{\delta}\right)}{\sqrt{\det\left(\check{\mathcal{H}}^{\delta}\right)}},$$

OLAP: dropping vanishing Laplace corrections

Define $\check{\mathcal{G}}^{\delta} := \mathcal{G}^{\check{\theta}^{\delta}}$, $\check{\mathcal{H}}^{\delta} := \mathcal{H}^{\check{\theta}^{\delta}}$. If $\check{\theta}^{\delta}$ is iterated until convergence, then $\check{\theta}^{\delta} \approx \hat{\theta}^{\delta}$ and $\check{\mathcal{G}}^{\delta} \approx 0$. Hence we drop both

$$(\check{\mathcal{G}}^{\delta})^{\mathsf{T}}(\check{\mathcal{H}}^{\delta})^{-1}\check{\mathcal{G}}^{\delta} \quad \text{and} \quad \log\det(\check{\mathcal{H}}^{\delta}),$$

leading to

$$\check{\Pi}(\delta|\mathcal{D}) \propto p^{-\mathsf{u}\|\delta\|_0}\exp\left\{\bar{\ell}^{\delta}(\check{\theta}^{\delta};\mathcal{D})\right\}.$$

Gibbs update over δ

For coordinate j, compare $\delta^{(j,0)}$ and $\delta^{(j,1)}$:

$$\mathbb{P}(\delta_j = 1 \mid \delta_{-j},\mathcal{D}) = \frac{\check{\Pi}(\delta^{(j,1)} \mid \mathcal{D})}{\check{\Pi}(\delta^{(j,0)} \mid \mathcal{D}) + \check{\Pi}(\delta^{(j,1)} \mid \mathcal{D})}.$$

Main computational gain

No local MLEs inside the chain: each proposal uses a global estimator plus a one-step correction.

3. Empirical results: logistic and Poisson models, $p = 1000$, low correlation

Table 1: F1 scores for logistic model feature selections. $p = 1000$, low correlation.

	n=200	n=300	n=400	n=500	n=1000
OLAP	0.778	**1.000**	**1.000**	**1.000**	**1.000**
Exact	0.750	**1.000**	**1.000**	**1.000**	**1.000**
ALA-BB	0.462	0.824	0.947	1.000	**1.000**
ALA-Complex	0.000	0.182	0.333	0.667	**1.000**
Lasso	0.300	0.294	0.280	0.288	0.339
SCAD	0.526	0.625	0.656	0.785	0.976
MCP	0.800	0.909	0.931	0.952	**1.000**
ABESS	0.870	0.833	0.800	0.952	**1.000**
SparseVB	0.918	**1.000**	**1.000**	**1.000**	**1.000**
S-Gibbs	0.791	0.952	0.952	0.952	0.952

Table 2: Runtimes for logistic model feature selections. $p = 1000$, low correlation.

Method	n=200	n=300	n=400	n=500	n=1000
OLAP	1.11	1.77	1.09	0.81	1.38
Exact	0.89	1.74	2.19	2.40	8.56
ALA-BB	0.48	0.32	0.40	0.44	0.71
ALA-Complex	0.10	0.19	0.27	0.35	0.53
Lasso	0.23	0.34	0.48	0.63	1.69
SCAD	1.32	1.85	2.71	3.44	2.74
MCP	0.97	1.55	2.13	3.68	9.40
ABESS	0.06	0.14	0.26	0.43	2.29
SparseVB	1.20	2.20	2.88	3.68	8.20
S-Gibbs	12.36	12.82	13.38	13.50	15.35

Table 3: F1 scores for Poisson model feature selections. $p = 1000$, low correlation.

	n=200	n=300	n=400	n=500	n=1000
OLAP	0.429	0.789	0.900	0.952	**1.000**
Exact	0.533	0.833	0.947	**1.000**	**1.000**
ALA-BB	0.000	0.333	0.462	0.667	**1.000**
ALA-Complex	0.000	0.000	0.000	0.182	0.824
Lasso	0.462	0.606	0.597	0.714	0.817
SCAD	0.321	0.297	0.292	0.317	0.500
MCP	0.509	0.563	0.563	0.606	0.833
ABESS	**0.625**	**0.889**	**0.952**	**1.000**	**1.000**

Table 4: Runtimes for Poisson model feature selections. $p = 1000$, low correlation.

Method	n=200	n=300	n=400	n=500	n=1000
OLAP	0.43	0.37	0.37	0.36	0.88
Exact	0.45	4.03	6.6	8.46	14.35
ALA BB	2.06	2.94	3.89	4.69	4.85
ALA Complex	0.88	0.97	1.09	1.14	1.68
Lasso	0.41	0.63	0.89	1.21	5.45
SCAD	4.16	9.81	17.07	24.8	39.11
MCP	2.7	8.04	16.41	23.85	73.06
ABESS	0.05	0.12	0.23	0.41	2.51

Takeaway on simulation results

Set $p = 1000$ and vary n, we report medians of F1 scores over 50 replicated datasets. Against these benchmarks, OLAP attains top-tier F1 with markedly lower and more stable runtimes.