NONPARAMETRIC SHRINKAGE ESTIMATION IN GENERALIZED LINEAR MODELS VIA POLYA TREES

Asaf Weinstein[1], Jonas Wallin[2], Daniel Yekutieli[3], and Malgorzata Bogdan[4]

[1]*Department of Statistics, Hebrew University of Jerusalem*

ORCID: 0000-0003-2237-2510

[2]*Department of Statistics, Lund University*

ORCID: 0000-0003-0381-6593

[3]*Department of Statistics, Tel Aviv University*

[4]*Institute of Mathematics, University of Wroclaw*

Supplementary Material

The supplement includes proof, and details on the Gibbs sampling algorithm.

S1 Proofs

Proof of Proposition 1. Let $\widehat{\boldsymbol{\beta}}(\mathbf{Z})$ be any PI rule under the (PI) model (3.12). Then we can proceed as in Weinstein (2021) and calculate the risk of $\widehat{\boldsymbol{\beta}}$ at $\boldsymbol{\beta}^*$ as

$$\begin{aligned} R(\boldsymbol{\beta}^*, \widehat{\boldsymbol{\beta}}) &= \mathbb{E}_{\boldsymbol{\beta}^*} L(\boldsymbol{\beta}^*, \widehat{\boldsymbol{\beta}}(\mathbf{Z})) \\ &= \mathbb{E}_{\boldsymbol{\beta}^*} L(\tau(\boldsymbol{\beta}^*), \tau(\widehat{\boldsymbol{\beta}}(\mathbf{Z}))) \\ &= \mathbb{E}_{\boldsymbol{\beta}^*} L(\tau(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}}(\tau(\mathbf{Z}))) \\ &= \mathbb{E}_{\tau(\boldsymbol{\beta}^*)} L(\tau(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}}(\mathbf{Z})) \\ &= R(\tau(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}}), \end{aligned} \tag{S1.1}$$

and we remind that the subscript on the expectation operator is the value of the parameter indexing the distribution of $\mathbf{Z}$ (not of $\tau(\mathbf{Z})$). Above, the second equality is because the *loss* is PI, the third equality is because the *rule* $\widehat{\boldsymbol{\beta}}$ is PI, and, crucially, the fourth inequality is because the *model* for $\mathbf{Z}$ is PI under (3.13). From (S1.1) it follows that

$$R(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}) = \frac{1}{p!} \sum_{\tau} R(\tau(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}}),$$

the sum taken over all $p!$ permutations τ. But this is precisely the Bayes risk of $\widehat{\boldsymbol{\beta}}$ under the prior $\widetilde{\Pi}_p^*$. The proof is complete because the oracle Bayes rule is *defined* to be the Bayes rule under the prior $\widetilde{\Pi}_p^*$. □

Proof of Proposition 2. Let $\widetilde{\Pi}$ be any exchangeable prior on $\boldsymbol{\beta}$. Per the technical modification in the statement of the proposition, the oracle Bayes rule $\widehat{\boldsymbol{\beta}}_{ol}$ is now also a function of the true (random) parameter vector $\boldsymbol{\beta}$

(through $\{\boldsymbol{\beta}\}$). Thus, first note that, quite trivially,

$$\min_{\widehat{\boldsymbol{\beta}}} \ \mathbb{E}_{\widetilde{\Pi}}[\ L(\ \boldsymbol{\beta},\ \widehat{\boldsymbol{\beta}}(\boldsymbol{Y},\{\boldsymbol{\beta}\})\ )\ ] \leq \min_{\widehat{\boldsymbol{\beta}}} \ \mathbb{E}_{\widetilde{\Pi}}[\ L(\ \boldsymbol{\beta},\ \widehat{\boldsymbol{\beta}}(\boldsymbol{Y})\ )\ ], \tag{S1.2}$$

where on the left hand side the minimum is over all functions $\widehat{\boldsymbol{\beta}}$ of $(\boldsymbol{Y},\{\boldsymbol{\beta}\})$, and on the right hand side the minimum is over all functions $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{Y}$ only; and where in both sides of the inequality the expectation is with respect to the joint distribution of $(\boldsymbol{Y},\boldsymbol{\beta})$ under the prior $\widetilde{\Pi}$. Therefore, it is enough to show that the oracle Bayes rule $\widehat{\boldsymbol{\beta}}_{ol}$ minimizes the left hand side of (S1.2), i.e., that

$$\arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \mathbb{E}_{\widetilde{\Pi}}[L(\boldsymbol{\beta},\boldsymbol{b})|\boldsymbol{Y},\{\boldsymbol{\beta}\}] = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \mathbb{E}_{\widetilde{\Pi}^*}[L(\boldsymbol{\beta},\boldsymbol{b})|\boldsymbol{Y}], \tag{S1.3}$$

Now, the posterior of $\boldsymbol{\beta}$ on the left hand side of (S1.3) is supported on the set of all possible orderings of the components of $\{\boldsymbol{\beta}\}$, i.e., on all permutations of $\boldsymbol{\beta}$. Calculating the posterior for $\boldsymbol{\beta}$ of $\boldsymbol{\beta}$, we have

$$\widetilde{\Pi}(\boldsymbol{\beta}|\boldsymbol{Y},\{\boldsymbol{\beta}\}) \propto \Pi(\boldsymbol{\beta}|\{\boldsymbol{\beta}\})f(\boldsymbol{Y}|\boldsymbol{\beta},\{\boldsymbol{\beta}\}) = \widetilde{\Pi}(\boldsymbol{\beta}|\{\boldsymbol{\beta}\})f(\boldsymbol{Y}|\boldsymbol{\beta}), \tag{S1.4}$$

and, since $\widetilde{\Pi}$ is exchangeable,

$$\widetilde{\Pi}(\boldsymbol{\beta}|\{\boldsymbol{\beta}\}) = \widetilde{\Pi}^*, \tag{S1.5}$$

the uniform distribution on all permutations of $\boldsymbol{\beta}$. From (S1.4) and (S1.5), we conclude that the posterior of $\boldsymbol{\beta}$ given $\boldsymbol{Y}$ and $\{\boldsymbol{\beta}\}$ is exactly the posterior with respect to which the minimum in (3.11) is taken. This completes the proof. $\square$

S2 Gibbs sampling of $\boldsymbol{\beta}$

To sample the vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ given $(\psi, \boldsymbol{\phi}, \boldsymbol{Y})$, we utilize a MH-within-Gibbs algorithm inspired by the coordinate descent algorithm of Friedman et al. (2007), that was shown to work very well for Lasso. Let $\boldsymbol{r} = (r_1, ..., r_p)$ denote the index of the FPT subintervals to which each component of $\boldsymbol{\beta}$ belongs, i.e. $r_j = k$ if $\beta_j \in \mathcal{I}_{L,k}$. We now detail how to generate a proposal for the MH algorithm for β_j:

1. Draw $r^* \sim r_j + Unif\{-K, \dots, K\}$, where K is a parameter which we estimate using an adaptive MCMC (AMCMC) scheme similar to Roberts and Rosenthal (2009). Specifically, we take an increasing batch size of MCMC samples; if the acceptance rate for the MH algorithm is above 0.3 we increase K, if it is below 0.3 we decrease it. The batch size is chosen so that updates are less frequent, in order to ensure convergence of the AMCMC algorithm.

2. Obtain a Taylor approximation of the log-likelihood $l(\beta_j) = \log f\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \psi\right)$,

$$l(\beta_j^*) \approx l(\beta_j) + l'(\beta_j)\left(\beta_j^* - \beta_j\right) + \frac{l''(\beta_j)}{2}\left(\beta_j^* - \beta_j\right)^2,$$

where

$$l'(\beta_j) = \frac{\partial}{\partial \beta_j} \log f\left(\boldsymbol{y}|\boldsymbol{\beta}, \psi\right) \qquad \text{and} \qquad l''(\beta_j) = \frac{\partial^2}{\partial \beta_j^2} \log f\left(\boldsymbol{Y}|\boldsymbol{\beta}, \psi\right).$$

3. Use the Taylor approximation to obtain a Normal approximation of the posterior as a proposal distribution, then generate a sample from the proposal given that proposal is contained in $\mathcal{I}_{L,r^*}$. That is, $\beta_j^* \sim \mathcal{N}_{\mathcal{I}_{L,r^*}}(\mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2)\,|\{\beta_j^* \in \mathcal{I}_{L,r^*}\}$, where $\mathcal{I}_{L,r^*} = (a_{r^*}, a_{r^*+1}]$, $\mu = \beta_j + \frac{l'(\beta_j)}{-l''(\beta_j)}$, and $\sigma^2 = \frac{1}{-l''(\beta_j)}$.

Thus, we first generate an interval $\mathcal{I}_{L,r^*}$ that includes β_j^*, then, given that the prior is constant on the interval, we use a quadratic approximation of the likelihood to sample β^* given it is in $\mathcal{I}_{L,r^*}$. Exactly as with coordinate descent algorithms, the main advantage of the algorithm is computational efficiency, for instance, one does not need to compute $\boldsymbol{X\beta}$ in each iteration, but instead store $\hat{\boldsymbol{Y}} = \boldsymbol{X\beta}$ and then compute $\hat{\boldsymbol{Y}}^* = \hat{\boldsymbol{Y}} + (\beta_j^* - \beta_j)\boldsymbol{X}^{(j)}$, where $\boldsymbol{X}^{(j)}$ is the jth column of $\boldsymbol{X}$. Hence, instead of computing a matrix-vector multiplication we only carry out a vector-scalar multiplication, and vector-addition. For the general likelihood (1.1), note that the log-likelihood, the gradient and the Hessian can be computed using only $\boldsymbol{X\beta}$, and $\boldsymbol{X}^{(j)}$.

S2.1 Sampling with oracle prior

Here we describe how to sample using MCMC when the prior is the oracle prior π_0 defined in (3.8). Since the vector of unique coefficients are fixed we simply permute the locations randomly. To generate proposal $\boldsymbol{\beta}^*$ given a previous sample $\boldsymbol{\beta}$ we do as follows: first set $\boldsymbol{\beta}^* = \boldsymbol{\beta}$, second generate two indices uniformly $i_1, i_2 \sim Unif\{1, 2, \ldots, p\}$ and set $\left(\beta^*_{i_1}, \beta^*_{i_2}\right) = (\beta_{i_2}, \beta_{i_1})$. Clearly this proposal is symmetric, and since oracle prior is constant over the permutation, the Metropolis-Hastings ratio is just $\frac{f(\boldsymbol{Y};\boldsymbol{\beta}^*,\psi)}{f(\boldsymbol{Y};\boldsymbol{\beta},\psi)}$.

Bibliography

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics 1*(2), 302 – 332.

Roberts, G. O. and J. S. Rosenthal (2009). Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics 18*(2), 349–367.

Weinstein, A. (2021). On permutation invariant problems in large-scale inference. *arXiv preprint arXiv:2110.06250*.