Web-based Supplementary Materials for "Sparse Kernel Machine Regression for Ordinal Outcomes"

by

Yuanyuan Shen, Katherine P. Liao and Tianxi Cai

This supplementary material contains the algorithm details and proof of the main theorem along with necessary lemmas.

Web Appendix A *Algorithm details*

To numerically obtain $\widehat{\boldsymbol{\theta}}$ in (9), we first perform a variable transformation by letting $\boldsymbol{\delta}^{(c)}$ to represent the differences between adjacent categories: $\boldsymbol{\delta}^{(c)} = \boldsymbol{\beta}^{(c+1)}_{(r_n)} - \boldsymbol{\beta}^{(c)}_{(r_n)}$, for $c = 1, \ldots, C-2$. Let $\boldsymbol{\Theta} = (\gamma_0^{(1)}, \cdots, \gamma_0^{(C-1)}, \boldsymbol{\beta}^{(1)}_{(r_n)}, \boldsymbol{\delta}^{(1)}, \ldots, \boldsymbol{\delta}^{(C-2)})$ be our new parameters after transformation, which relates to the original parameter vector $\boldsymbol{\theta} = (\gamma_0^{(1)}, ..., \gamma_0^{(C-1)}, \boldsymbol{\beta}^{(1)}_{(r_n)}, \boldsymbol{\beta}^{(2)}_{(r_n)}, ..., \boldsymbol{\beta}^{(C-2)}_{(r_n)}, \boldsymbol{\beta}^{(C-1)}_{(r_n)})^\intercal$ through $\boldsymbol{\theta} = \mathbb{M}\boldsymbol{\Theta}$, where

$$\mathbb{M} = \begin{bmatrix} \mathbb{I}_{C-1} & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \mathbb{I}_r & 0 & 0 & 0 & \cdots & 0 \\ 0 & \mathbb{I}_r & \mathbb{I}_r & 0 & 0 & \cdots & 0 \\ 0 & \mathbb{I}_r & \mathbb{I}_r & \mathbb{I}_r & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \mathbb{I}_r & \mathbb{I}_r & \mathbb{I}_r & \mathbb{I}_r & \cdots & \mathbb{I}_r \end{bmatrix}$$

Let $\widetilde{\mathbf{X}} = \widetilde{A}^\intercal \mathbb{M}$ and $\widetilde{\mathbf{Y}} = \widetilde{A}^\intercal \widetilde{\boldsymbol{\theta}}$, where $\widetilde{\mathbb{A}} = \widetilde{A}\widetilde{A}^\intercal$. Therefore, (9) is transformed into a linear adaptive group LASSO (gLASSO) problem:

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left[\frac{1}{2} \|\widetilde{\mathbf{X}}\boldsymbol{\Theta} - \widetilde{\mathbf{Y}}\|_2^2 + \tau_1 \sum_{c=1}^{C-2} \frac{\|\boldsymbol{\Theta}^{(c)}\|_2}{\|\widetilde{\boldsymbol{\Theta}}^{(c)}\|_2} \right] \tag{1}$$

where $\widetilde{\mathbf{\Theta}} = \mathbb{M}^{-1}\widetilde{\boldsymbol{\theta}}$.

Web Appendix B *Parameter tuning*

There are three tuning parameters involved in our proposed procedure, ρ, τ_1 and τ_2, where ρ is the parameter for kernel $k(\cdot,\cdot;\rho)$, τ_2 is the tuning parameter for the ridge penalty, and τ_1 is the gLASSO penalty parameter controlling the amount of penalty for the differences between adjacent categories. Commonly used methods for selecting tuning parameters for ridge regression and gLASSO penalties include AIC, BIC, cross-validation, and generalized cross validation (GCV) (**????**). For each given ρ, we obtain an optimal τ_2 based on the GCV criterion (**?**), denoted by $\tau_2(\rho)$. Then with each given ρ and $\tau_2(\rho)$, we obtain the corresponding synthetic data $\{\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}, \widetilde{\delta}^{(c)}\}$ for fitting the gLASSO penalized least square in (1). The tuning parameters τ_1 and ρ are then selected via the AIC criterion. The degree of freedom in the AIC criterion is estimated analogous to those proposed in **?**) and **?**). Specifically, we define $\mathrm{DF}(\rho,\tau_1) = \sum_{c=1}^{C-1} I\{\|\widehat{\delta}^{(c)}(\rho,\tau_1)\| > 0\} + \sum_{c=1}^{C-1} \frac{\|\widehat{\delta}^{(c)}(\rho,\tau_1)\|_2}{\|\widetilde{\delta}^{(c)}(\rho)\|_2}(d_c(\rho)-1)$, where $d_c(\rho)$ is the effective number of parameters in the c^{th} group from the ridge regression, calculated as the sum of the diagonal elements of Hessian matrix $\widetilde{\mathbb{A}}(\rho)$ that correspond to the c^{th} group. We then select the optimal (ρ,τ_1) as the minimizer of $\mathrm{AIC}(\rho,\tau_1) = -2\mathrm{loglik}(\rho,\tau_1) + 2\mathrm{DF}(\rho,\tau_1)$.

Web Appendix C *Asymptotic Properties of* $\widehat{h}^{(c)}(\cdot)$

Here, when $\mathcal{H}_k$ is finite dimensional, we aim to establish the root-n convergence rate of $\widehat{h}^{(c)}(\mathbf{x})$ and model selection consistency in the sense that $P\{\widehat{h}^{(c)}(\mathbf{x}) = \widehat{h}^{(c+1)}(\mathbf{x})\} \to 1$ when $h^{(c)} = h^{(c+1)}$. To this end, we first note that we can write our penalized likelihood (7) in the same form as in **?**). It is the summation of $C-1$ independent terms, each of which takes the form:

$$\sum_{i=1}^{n} I(y_i \geqslant c)\left[D_i^{(c)}\log\{g(\gamma_0^{(c)} + \widetilde{\boldsymbol{\psi}}_i^{\mathsf{T}}\boldsymbol{\beta}_{(r_n)}^{(c)})\} + (1-D_i^{(c)})\log\{1-g(\gamma_0^{(c)} + \widetilde{\boldsymbol{\psi}}_i^{\mathsf{T}}\boldsymbol{\beta}_{(r_n)}^{(c)})\}\right] - \tau_2\|\boldsymbol{\beta}_{(r_n)}^{(c)}\|_2^2$$

Therefore, using the same arguments as given in **?**), we have

Lemma 1: *$P(r_n = r) \to 1$ and* $\|\widetilde{\mathbf{\Psi}}(\mathbf{x}) - \mathbf{\Psi}(\mathbf{x})\|_2 + n^{-\frac{1}{2}}\|\widetilde{\mathbf{\Psi}} - \mathbf{\Psi}\|_F + \|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 = O_p(n^{-\frac{1}{2}})$.

It also directly implies that $\widetilde{h}^{(c)}(\mathbf{x}) - h^{(c)}(\mathbf{x}) = O_p(n^{-\frac{1}{2}})$ and we may need establish the convergences conditioning on $r_n = r$. In view of this together with the parametrization in (1), it suffices to show that $\widehat{\boldsymbol{\delta}}^{(c)} - \boldsymbol{\delta}^{(c)} = O_p(n^{-\frac{1}{2}})$, if $c \in \mathcal{A}$; and $P(\widehat{\boldsymbol{\delta}}^{(c)} = 0) \to 1$, if $c \notin \mathcal{A}$, where $\mathcal{A} = \{c : \boldsymbol{\delta}^{(c)} \neq 0\}$. These are parallel to Theorem 1 and 2 in (**?**) where they show the estimation consistency and selection consistency of the adaptive group lasso estimator. The main difference between our problem and the setting considered in **?**) is that our $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ are not directly observed data but are estimated quantities with $\widetilde{\mathbf{X}} = \widetilde{A}^{\intercal}\mathbb{M}$, $\widetilde{\mathbf{Y}} = \widetilde{A}^{\intercal}\widetilde{\boldsymbol{\theta}}$, where $\widetilde{\mathbb{A}} = \widetilde{A}\widetilde{A}^{\intercal}$, so we need to take into account the randomness in $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$. In their proof, the main arguments rely on two convergences: $n^{-1}\mathbf{X}^{\intercal}\mathbf{X} \to E(\mathbf{X}_i\mathbf{X}_i^{\intercal})$ in probability and $n^{-\frac{1}{2}}\mathbf{X}^{\intercal}(\mathbf{Y} - \mathbf{X}\boldsymbol{\delta}) = O_p(1)$. In our case, the corresponding convergences we need to establish are the probability convergence of $\mathbb{M}^{\intercal}\widetilde{\mathbb{A}}\mathbb{M}$ and $\mathbb{M}^{\intercal}\widetilde{\mathbb{A}}[n^{\frac{1}{2}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})] = O_p(1)$. By the Lemma, $n^{\frac{1}{2}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = O_P(1)$, and since $\mathbb{M}$ is a constant, it is suffice to show that $\widetilde{\mathbb{A}} = \text{diag}\{\widetilde{\mathbb{A}}^{(1)}, ..., \widetilde{\mathbb{A}}^{(C-1)}\}$ converges to $\mathbb{A} = \text{diag}\{\mathbb{A}^{(1)}, ..., \mathbb{A}^{(C-1)}\}$ in probability, where

$$\begin{aligned}
\widetilde{\mathbb{A}}^{(c)} =& n^{-1}\sum_{i=1}^{n} I(y_i \geqslant c)\left[\underline{\widetilde{\boldsymbol{\psi}}}_i\underline{\widetilde{\boldsymbol{\psi}}}_i^{\intercal} g(\underline{\widetilde{\boldsymbol{\psi}}}_i^{\intercal}\underline{\widetilde{\boldsymbol{\beta}}}_{(r)}^{(c)})(1 - g(\underline{\widetilde{\boldsymbol{\psi}}}_i^{\intercal}\underline{\widetilde{\boldsymbol{\beta}}}_{(r)}^{(c)}))\right] \\
\mathbb{A}^{(c)} =& E\left\{I(y_i \geqslant c)\left[\underline{\boldsymbol{\psi}}_i\underline{\boldsymbol{\psi}}_i^{\intercal} g(\underline{\boldsymbol{\psi}}_i^{\intercal}\underline{\boldsymbol{\beta}}_{(r)}^{(c)})(1 - g(\underline{\boldsymbol{\psi}}_i^{\intercal}\underline{\boldsymbol{\beta}}_{(r)}^{(c)}))\right]\right\}
\end{aligned}$$

$\underline{\boldsymbol{\psi}}_i = [1, \psi_i^{\intercal}]^{\intercal}$, $\underline{\boldsymbol{\beta}}_{(r)}^{(c)} = [\boldsymbol{\gamma}_0^{(c)}, \boldsymbol{\beta}_{(r)}^{(c)\intercal}]^{\intercal}$; $\underline{\widetilde{\boldsymbol{\psi}}}_i = [1, \widetilde{\psi}_i^{\intercal}]^{\intercal}$, and $\underline{\widetilde{\boldsymbol{\beta}}}_{(r)}^{(c)} = [\widetilde{\boldsymbol{\gamma}}_0^{(c)}, \boldsymbol{\beta}_{(r)}^{(c)\intercal}]^{\intercal}$.

Since $\|\widetilde{\mathbb{A}} - \mathbb{A}\|_F^2 = \sum_{c=1}^{C-1}\|\widetilde{\mathbb{A}}^{(c)} - \mathbb{A}^{(c)}\|_F^2$, so if we can show the convergence of each of the $C-1$ blocks, we will have convergence for the entire matrix $\widetilde{\mathbb{A}}$. Let $\widetilde{\mathbb{A}}^{\star(c)} = n^{-1}\sum_{i=1}^{n} I(y_i \geqslant c)\left[\underline{\boldsymbol{\psi}}_i\underline{\boldsymbol{\psi}}_i^{\intercal} g(\underline{\boldsymbol{\psi}}_i^{\intercal}\underline{\boldsymbol{\beta}}_{(r)}^{(c)})(1 - g(\underline{\boldsymbol{\psi}}_i^{\intercal}\underline{\boldsymbol{\beta}}_{(r)}^{(c)}))\right]$, we have $\|\widetilde{\mathbb{A}}^{(c)} - \mathbb{A}^{(c)}\|_F^2 = \|\widetilde{\mathbb{A}}^{(c)} - \widetilde{\mathbb{A}}^{\star(c)} + \widetilde{\mathbb{A}}^{\star(c)} - \mathbb{A}^{(c)}\|_F^2 \leqslant \|\widetilde{\mathbb{A}}^{(c)} - \widetilde{\mathbb{A}}^{\star(c)}\|_F^2 + \|\widetilde{\mathbb{A}}^{\star(c)} - \mathbb{A}^{(c)}\|_F^2$. Note that since $\widetilde{\mathbb{A}}^{\star(c)} \to \mathbb{A}^{(c)}$ with probability 1 by Law of Large Numbers, so we only need to show $\|\widetilde{\mathbb{A}}^{(c)} - \mathbb{A}^{\star(c)}\|_F^2 \to 0$. To simplify notation, we drop (c)

superscripts and the (r) subscripts. We first split $\widetilde{\mathbb{A}} - \widetilde{\mathbb{A}}^{\star}$ into summation of three parts:

$$\widetilde{\mathbb{A}} - \widetilde{\mathbb{A}}^{\star} = n^{-1} \sum_{i=1}^{n} I(y_i \geqslant c) \left[(\widetilde{\underline{\boldsymbol{\psi}}}_i - \psi_i) \widetilde{\underline{\boldsymbol{\psi}}}_i^{\intercal} g(\widetilde{\underline{\boldsymbol{\psi}}}_i^{\intercal} \widetilde{\underline{\boldsymbol{\beta}}})(1 - g(\widetilde{\underline{\boldsymbol{\psi}}}_i^{\intercal} \widetilde{\underline{\boldsymbol{\beta}}})) \right] \tag{P1}$$

$$+ n^{-1} \sum_{i=1}^{n} I(y_i \geqslant c) \left[\underline{\boldsymbol{\psi}}_i (\widetilde{\underline{\boldsymbol{\psi}}}_i - \underline{\boldsymbol{\psi}}_i)^{\intercal} g(\widetilde{\underline{\boldsymbol{\psi}}}_i^{\intercal} \widetilde{\underline{\boldsymbol{\beta}}})(1 - g(\widetilde{\underline{\boldsymbol{\psi}}}_i^{\intercal} \widetilde{\underline{\boldsymbol{\beta}}})) \right] \tag{P2}$$

$$+ n^{-1} \sum_{i=1}^{n} I(y_i \geqslant c) \left[\underline{\boldsymbol{\psi}}_i \underline{\boldsymbol{\psi}}_i^{\intercal} (g(\widetilde{\underline{\boldsymbol{\psi}}}_i^{\intercal} \widetilde{\underline{\boldsymbol{\beta}}})(1 - g(\widetilde{\underline{\boldsymbol{\psi}}}_i^{\intercal} \widetilde{\underline{\boldsymbol{\beta}}})) - g(\underline{\boldsymbol{\psi}}_i^{\intercal} \underline{\boldsymbol{\beta}})(1 - g(\underline{\boldsymbol{\psi}}_i^{\intercal} \underline{\boldsymbol{\beta}}))) \right] \tag{P3}$$

Assume that $\|\underline{\boldsymbol{\psi}}_i\|_2 \leqslant R$ and apply the Lemma, (P1) can be bounded since

$$\left\| n^{-1} \sum_{i=1}^{n} I(y_i \geqslant c) \left[(\widetilde{\underline{\boldsymbol{\psi}}}_i - \underline{\boldsymbol{\psi}}_i) \widetilde{\underline{\boldsymbol{\psi}}}_i^{\intercal} g(\widetilde{\underline{\boldsymbol{\psi}}}_i^{\intercal} \widetilde{\underline{\boldsymbol{\beta}}})(1 - g(\widetilde{\underline{\boldsymbol{\psi}}}_i^{\intercal} \widetilde{\underline{\boldsymbol{\beta}}})) \right] \right\|_F$$

$$\leqslant n^{-1} \left[R n^{\frac{1}{2}} \|\widetilde{\underline{\boldsymbol{\Psi}}} - \underline{\boldsymbol{\Psi}}\|_F + \|\widetilde{\underline{\boldsymbol{\Psi}}} - \underline{\boldsymbol{\Psi}}\|_F^2 \right] = R \cdot O_p(n^{-1}) + O_p(n^{-2})$$

The term (P2) can also be bounded similarly with

$$n^{-1} \left\| \sum_{i=1}^{n} I(y_i \geqslant c) \left[\underline{\boldsymbol{\psi}}_i (\widetilde{\underline{\boldsymbol{\psi}}}_i - \underline{\boldsymbol{\psi}}_i)^{\intercal} g(\widetilde{\underline{\boldsymbol{\psi}}}_i^{\intercal} \widetilde{\underline{\boldsymbol{\beta}}})(1 - g(\widetilde{\underline{\boldsymbol{\psi}}}_i^{\intercal} \widetilde{\underline{\boldsymbol{\beta}}})) \right] \right\|_F \leqslant n^{-1/2} R \|\widetilde{\underline{\boldsymbol{\Psi}}} - \underline{\boldsymbol{\Psi}}\|_F = R \cdot O_p(n^{-1})$$

Since $\|\widetilde{\underline{\boldsymbol{\beta}}} - \underline{\boldsymbol{\beta}}\|_2 = O_p(n^{-1/2})$, $\|\widetilde{\boldsymbol{\Psi}} - \boldsymbol{\Psi}\|_F = O_p(1)$, $\|\underline{\boldsymbol{\psi}}_i\|_2 \leqslant R$ and $\|\underline{\boldsymbol{\beta}}\|_2 < \infty$, we can easily obtain $(P3) = O_p(n^{-\frac{1}{2}})$. Therefore $\|\widetilde{\mathbb{A}} - \widetilde{\mathbb{A}}^{*}\|_F \to 0$ in probability and hence $\|\widetilde{\mathbb{A}} - \mathbb{A}\|_F \to 0$ in probability. This, together with the same arguments as given in (?), implies that $\|\widehat{\boldsymbol{\delta}}^{(c)} - \boldsymbol{\delta}^{(c)}\|_2 = O_p(n^{-\frac{1}{2}})$ when $c \in \mathcal{A}$ and $P(\widehat{\boldsymbol{\delta}}^{(c)} = 0) \to 1$ when $c \notin \mathcal{A}$.