Appendices for "A Unified Approach to Sparse Tweedie Modeling of Multi-Source Insurance Claim Data"

Simon Fontaine[*], **Yi Yang**[†], **Wei Qian**[‡], **Yuwen Gu**[§], **Bo Fan**[¶]

July 19, 2019

Appendix A. Projection onto the L_1-Ball

Proof of Lemma 1. Note that (16) can be written as

$$\mathrm{prox}_{\tau h}(\mathbf{u}) = \arg\min_{\boldsymbol{\beta}_j} \frac{1}{2}\|\boldsymbol{\beta}_j - (\tilde{\boldsymbol{\beta}}_j - t_j \nabla_j \ell(\tilde{\boldsymbol{\beta}}_j; \tilde{\mathbf{b}}_{-j}))\|_2^2 + \lambda v_j t_j \|\boldsymbol{\beta}_j\|_\infty,$$

where $\tau = \lambda v_j t_j$, $h = ||\cdot||_\infty$ and $\mathbf{u} = \tilde{\boldsymbol{\beta}}_j - t_j \nabla_j \ell(\tilde{\boldsymbol{\beta}}_j; \tilde{\mathbf{b}}_{-j})$. By the Moreau decomposition (Parikh et al., 2014), we have

$$\mathrm{prox}_h(\mathbf{u}) = \mathbf{u} - \mathrm{prox}_{h^*}(\mathbf{u}),$$

where h^* denotes the convex conjugate of h. We want to derive a similar identity for τh, $\tau > 0$. The convex conjugate of τh is

$$(\tau h)^*(\mathbf{u}) = \sup_{\mathbf{v}} \left(\mathbf{u}^\top \mathbf{v} - \tau h(\mathbf{v})\right) = \tau \sup_{\mathbf{v}} \left(\frac{1}{\tau}\mathbf{u}^\top \mathbf{v} - h(\mathbf{v})\right) = \tau h^* \left(\frac{\mathbf{u}}{\tau}\right).$$

[*] Department of Mathematics and Statistics, University of Montreal (fontaines@dms.umontreal.ca)

[†] Corresponding author, Department of Mathematics and Statistics, McGill University (yi.yang6@mcgill.ca)

[‡] Department of Applied Economics and Statistics, University of Delaware (weiqian@udel.edu)

[§] Department of Statistics, University of Connecticut (yuwen.gu@uconn.edu)

[¶] Department of Statistics, University of Oxford (bo.fan@lmh.ox.ac.uk)

Then, we get

$$\begin{aligned}
\operatorname{prox}_{(\tau h)^*}(\mathbf{u}) &= \arg\min_{\mathbf{v}} (\tau h)^*(\mathbf{v}) + \frac{1}{2}\left\|\mathbf{v}-\mathbf{u}\right\|_2^2 \\
&= \arg\min_{\mathbf{v}} \tau h^*\left(\frac{\mathbf{v}}{\tau}\right) + \frac{1}{2}\left\|\mathbf{v}-\mathbf{u}\right\|_2^2 \\
&= \arg\min_{\mathbf{v}} h^*\left(\frac{\mathbf{v}}{\tau}\right) + \frac{1}{2\tau}\left\|\mathbf{v}-\mathbf{u}\right\|_2^2 \qquad (\mathbf{v}=\tau\mathbf{z}) \\
&= \arg\min_{\tau\mathbf{z}} h^*\left(\mathbf{z}\right) + \frac{1}{2\tau}\left\|\tau\mathbf{z}-\mathbf{u}\right\|_2^2 \\
&= \tau \arg\min_{\mathbf{z}} h^*\left(\mathbf{z}\right) + \frac{1}{2\frac{1}{\tau}}\left\|\mathbf{z}-\frac{\mathbf{u}}{\tau}\right\|_2^2 \\
&= \tau \operatorname{prox}_{\frac{1}{\tau}h^*}\left(\frac{\mathbf{u}}{\tau}\right),
\end{aligned}$$

so we have the identity

$$\operatorname{prox}_{\tau h}(\mathbf{u}) = \mathbf{u} - \operatorname{prox}_{(\tau h)^*}(\mathbf{u}) = \mathbf{u} - \tau \operatorname{prox}_{\frac{1}{\tau}h^*}\left(\frac{\mathbf{u}}{\tau}\right).$$

For $h = ||\cdot||_\infty$, it can be shown that $\tau \operatorname{prox}_{\frac{1}{\tau}h^*}\left(\frac{\mathbf{u}}{\tau}\right)$ is equivalent to the L_2-projection of $\mathbf{u}$ onto an L_1-ball $B_1(\tau)$,

$$\tau \operatorname{prox}_{\frac{1}{\tau}h^*}\left(\frac{\mathbf{u}}{\tau}\right) = \operatorname{Proj}_{B_1(\tau)}(\mathbf{u}).$$

To see this, note that the convex conjugate h^* of $h = ||\cdot||_\infty$ is

$$h^*(\mathbf{u}) = I_{\{\mathbf{u}:||\mathbf{u}||_1\le 1\}} = \begin{cases} 0, & ||\mathbf{u}||_1 \le 1, \\ +\infty, & ||\mathbf{u}||_1 > 1, \end{cases}$$

and

$$2\tau h^*\left(\frac{\mathbf{z}}{\tau}\right) = \begin{cases} 0, & ||\mathbf{z}||_1 \le \tau, \\ +\infty, & ||\mathbf{z}||_1 > \tau. \end{cases}$$

Then

$$
\begin{aligned}
\tau \operatorname{prox}_{\frac{1}{\tau}h^*}\left(\frac{\mathbf{u}}{\tau}\right) &= \tau \arg\min_{\mathbf{v}} h^*(\mathbf{v}) + \frac{\tau}{2}\left\|\mathbf{v} - \frac{\mathbf{u}}{\tau}\right\|_2^2 \\
&= \arg\min_{\mathbf{z}} h^*\left(\frac{\mathbf{z}}{\tau}\right) + \frac{\tau}{2}\left\|\frac{\mathbf{z}}{\tau} - \frac{\mathbf{u}}{\tau}\right\|_2^2 \qquad (\mathbf{z} = \tau\mathbf{v}) \\
&= \arg\min_{\mathbf{z}} h^*\left(\frac{\mathbf{z}}{\tau}\right) + \frac{1}{2\tau}\left\|\mathbf{z} - \mathbf{u}\right\|_2^2 \\
&= \arg\min_{\mathbf{z}} 2\tau h^*\left(\frac{\mathbf{z}}{\tau}\right) + \left\|\mathbf{z} - \mathbf{u}\right\|_2^2 .
\end{aligned}
$$

The objective function is minimized at where $2\tau h^*\left(\frac{\mathbf{z}}{\tau}\right)$ is finite, i.e., $||\mathbf{z}||_1 \leq \tau$. Hence, we get

$$
\tau \operatorname{prox}_{\frac{1}{\tau}h^*}\left(\frac{\mathbf{u}}{\tau}\right) = \arg\min_{\mathbf{z}:||\mathbf{z}||_1 \leq \tau} \left\|\mathbf{z} - \mathbf{u}\right\|_2^2 = \operatorname{Proj}_{B_1(\tau)}(\mathbf{u}).
$$

If $||\mathbf{u}||_1 \leq \tau$, we obviously have $\operatorname{Proj}_{B_1(\tau)}(\mathbf{u}) = \mathbf{u}$. Otherwise, we have to solve

$$
\sum_{k=1}^{K} \left(|u_k| - \xi\right)_+ = \tau
$$

for ξ and compute

$$
\left[\operatorname{Proj}_{B_1(\tau)}(\mathbf{u})\right]_k = \operatorname{sgn}(u_k)\left(|u_k| - \xi\right)_+ .
$$

□

Duchi et al. (2008) suggest a linear time algorithm to perform projection onto the simplex that can be easily extended to projection onto the L_1-ball. Algorithm 5 summarizes the procedure.

Appendix B. KKT Conditions

Denote $\mathbf{u} = \tilde{\boldsymbol{\beta}}_j - t_j \nabla_j \ell(\tilde{\boldsymbol{\beta}}_j | \tilde{\mathbf{b}}_{-j})$. Note that

$$
||\mathbf{u}||_\infty = \max_k |u_k| = \max_k |\mathbf{e}_k^\top \mathbf{u}|,
$$

where $\mathbf{e}_k = (I(j = k), 1 \leq j \leq K)^\top$. For each individual $|\mathbf{e}_k^\top \mathbf{u}|$, we have

$$
\partial |\mathbf{e}_k^\top \mathbf{u}| = \mathbf{e}_k \partial |\mathbf{e}_k^\top \mathbf{u}| = \mathbf{e}_k \cdot s_k,
$$

Algorithm 5: Linear time projection of $\mathbf{y} \in \mathbb{R}^n$ onto the L_1-ball of radius $z > 0$ (Duchi et al., 2008)

```
1. Consider $\mathbf{v} = (|y_1|, \ldots, |y_n|)^\top$;
2. Project $\mathbf{v}$ onto the simplex:
    (a) Initialize $U = \{1, \ldots, n\}$, $s = 0$, $\rho = 0$;
    (b) While $U \neq \emptyset$, do:
        i. Pick $k \in U$ at random;
        ii. Partition $U = G \cup L$, where $G = \{j \in U | v_j \geq v_k\}$ and $L = U \setminus G$;
        iii. Compute $\Delta\rho = |G|$ and $\Delta s = \sum_{j \in G} v_j$;
        iv. If $(s + \Delta s) - (\rho + \Delta\rho)v_k < z$, then set $s \leftarrow s + \Delta s$, $\rho \leftarrow \rho + \Delta\rho$ and $U \leftarrow L$.
            Otherwise, set $U \leftarrow G \setminus \{k\}$;
    (c) Set $\theta = (s - z)/\rho$;
    (d) Compute the projection onto the simplex $\mathbf{w} = (w_1, \ldots, w_n)^\top$, where
        $w_i = \max(v_i - \theta, 0)$;
3. Output $\mathbf{x} = (x_1, \ldots, x_n)^\top$, the projection onto the $L_1$-Ball, where $x_i = w_i \cdot \mathrm{sgn}(y_i)$.
```

where

$$s_k = \begin{cases} \{1\} & \mathbf{e}_k^\top \mathbf{u} > 0, \\ \{-1\} & \mathbf{e}_k^\top \mathbf{u} < 0, \\ [-1, 1] & \mathbf{e}_k^\top \mathbf{u} = 0. \end{cases}$$

Thus we can obtain the sub-differential for $||\mathbf{u}||_\infty$

$$\partial ||\mathbf{u}||_\infty = \text{conv} \bigcup_{k \in M(\mathbf{u})} \{\mathbf{e}_k \cdot s_k\},$$

where $M(\mathbf{u}) = \left\{k : \ |\mathbf{e}_k^\top \mathbf{u}| = ||\mathbf{u}||_\infty\right\}$ is the maximizing indices set and conv denotes the convex hull. This implies that an optimal solution needs to satisfy the condition: $\mathbf{0} \in \nabla_j \ell(\tilde{\boldsymbol{\beta}}_j; \tilde{\mathbf{b}}_{-j}) + t_j^{-1}(\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j) + \lambda v_j \partial ||\boldsymbol{\beta}_j||_\infty$, i.e.,

$$\frac{1}{\lambda v_j t_j} \left(\tilde{\boldsymbol{\beta}}_j - t_j \nabla_j \ell(\tilde{\boldsymbol{\beta}}_j; \tilde{\mathbf{b}}_{-j}) \right) - \frac{1}{\lambda v_j t_j} \boldsymbol{\beta}_j \in \text{conv} \bigcup_{k \in M(\boldsymbol{\beta}_j)} \{\mathbf{e}_k \cdot s_k\}. \tag{24}$$

If $\boldsymbol{\beta}_j = \mathbf{0}$, then $M(\boldsymbol{\beta}_j) = \{1, \ldots, K\}$ resulting in a convex hull equal to the L_1 unit ball formed by $\{\mathbf{e}_k \cdot s\}_{k=1}^{K}$. Thus, from (24), we require $\|\tilde{\boldsymbol{\beta}}_j - t_j \nabla_j \ell(\tilde{\boldsymbol{\beta}}_j; \tilde{\mathbf{b}}_{-j})\|_1 \leq \lambda v_j t_j$. In practice, our algorithm builds the model upwards: it will never exclude a feature from the model (i.e., by setting $\boldsymbol{\beta}_j = \mathbf{0}$) once it is already included (i.e., $\tilde{\boldsymbol{\beta}}_j \neq \mathbf{0}$ for some previous iteration) so that these two inequalities will be equivalent.

For $\boldsymbol{\beta}_j \neq \mathbf{0}$, we need to verify the above inclusion directly. If (24) holds, then we must have

$$\frac{1}{\lambda v_j t_j} \left(\tilde{\beta}_j^{(k)} - t_j \nabla_j \ell(\tilde{\boldsymbol{\beta}}_j; \tilde{\mathbf{b}}_{-j})^{(k)} \right) - \frac{1}{\lambda v_j t_j} \beta_j^{(k)} = 0$$

for all $k \notin M(\boldsymbol{\beta}_j)$, i.e., $|\beta_j^{(k)}| \neq \|\boldsymbol{\beta}_j\|_\infty$, while $\|t_j^{-1}\tilde{\boldsymbol{\beta}}_j - \nabla_j \ell(\tilde{\boldsymbol{\beta}}_j; \tilde{\mathbf{b}}_{-j}) - t_j^{-1}\boldsymbol{\beta}_j\|_1 = \lambda v_j$ since the convex hull must be a subset of the boundary of the L_1 ball of radius λv_j. These two conditions are also sufficient for (24) to hold.

Appendix C. Algorithm Verification

To check the validity of our algorithm, we consider the modeling under L_1/L_∞ regularization of simulated data with $K = 5$, $p = 20$, $n_k = 200$ and 4 true variables in setting 1.

In Section 3.1, we have seen that the inner loop of the algorithm (the MStweedie-GPG algorithm) should feature the strict descent property. We can plot the difference in the objective function $\ell_Q(\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\beta}}) - \ell_Q(\boldsymbol{\beta}_0, \boldsymbol{\beta})$ and check whether this value is positive for every cycle of the MStweedie-GPG algorithm. The theoretical solution should always exhibit the descent property where a numerical solution will possibly violate that check. Figure A1 displays this verification for the current example. Except minor violations, we can see that this property is satisfied by our implementation.

The KKT conditions are at the heart of minimizing the penalized likelihood $\ell(\boldsymbol{\beta}_0, \boldsymbol{\beta}) + \lambda P_\alpha(\boldsymbol{\beta})$. Along the solution path, the KKT conditions in (18) should always be verified by the theoretical solution. However, a numerical solution could only approach this analytical value within certain precision and therefore may fail the KKT check. Thus, we can plot the values of these conditions for both zero and non-zero estimates and check how far they deviate from their theoretical values. Figure A2 shows these conditions for every $j = 1, \ldots, p$ along the sequence of λ values. There are exactly no violations of the condition on excluded variables and the condition on included variables is never violated by a large value.

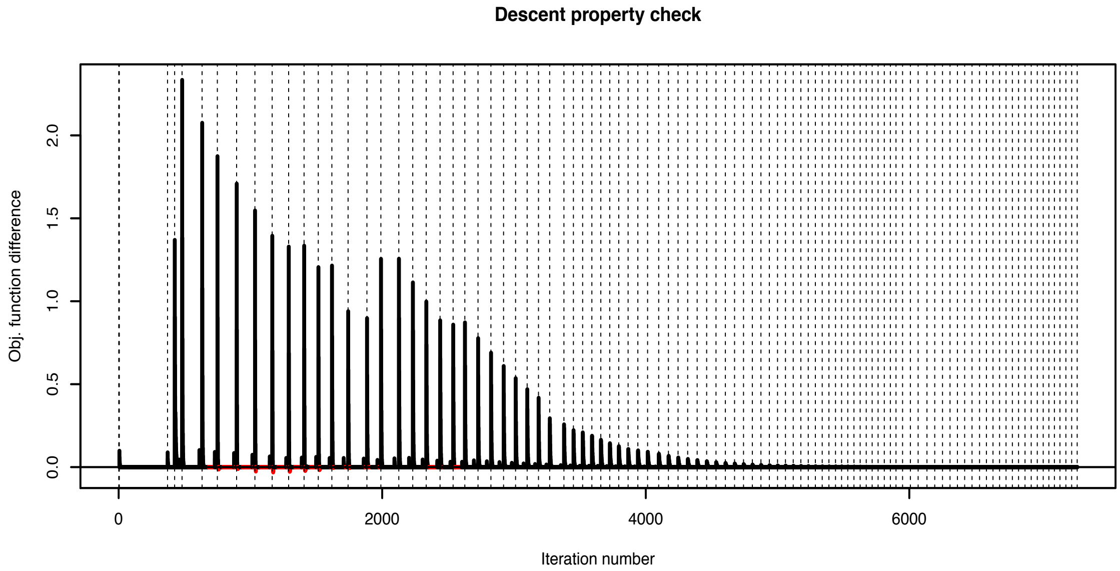

Figure A1: *Verification of the descent property in the MStweedie-GPG algorithm with synthetic data: the difference in objective function is plotted versus the iteration number (representing one MStweedie-GPG cycle). The vertical dotted lines represent new λ values in the solution path.*

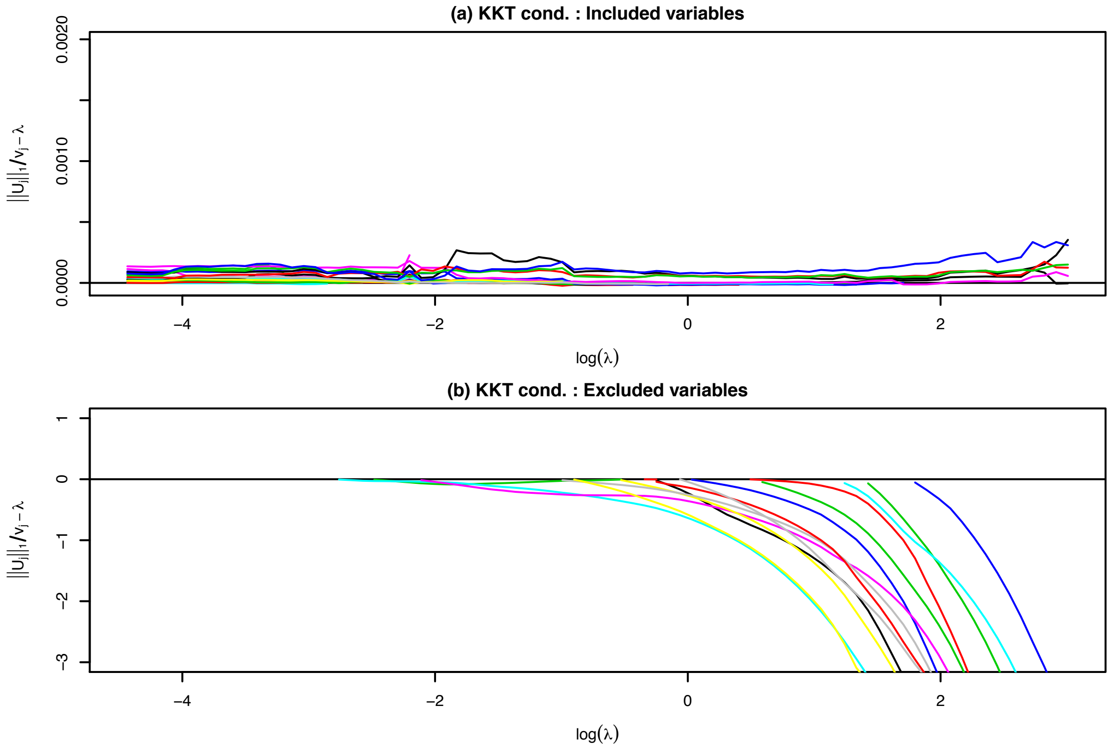

Figure A2: *Verification of the KKT conditions with synthetic data. The curves in each panel trace the path of the value $||\boldsymbol{\beta}_j||_1/v_j - \lambda$ for one j. In part (a), we verify the condition on non-zero estimates, i.e. variables included in the model for a given λ, where we expect the value to be 0. In part (b), we verify the condition on zero estimates, i.e. variables excluded from the model, where we expect the value to be below 0.*

Appendix D. Convergence of MStweedie-GPG with Line Search

Lemma 2. *For each $j \in \{0,1,\ldots,p\}$, $\nabla_j \ell(\boldsymbol{\beta}_j; \tilde{\mathbf{b}}_{-j})$ is uniformly Lipschitz continuous in the sublevel set $\mathcal{L}_0 = \{(\boldsymbol{\beta}_0, \boldsymbol{\beta}) : \ f(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \le f(\mathbf{0}, \mathbf{0})\}$, where $f(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \ell(\boldsymbol{\beta}_0, \boldsymbol{\beta}) + \lambda P_\alpha(\boldsymbol{\beta})$. In other words, there exists $M_j \in (0, \infty)$ such that the inequality*

$$\|\nabla_j \ell(\boldsymbol{\beta}_j; \tilde{\mathbf{b}}_{-j}) - \nabla_j \ell(\boldsymbol{\beta}'_j; \tilde{\mathbf{b}}_{-j})\|_2 \le M_j \|\boldsymbol{\beta}_j - \boldsymbol{\beta}'_j\|_2$$

holds for any $\boldsymbol{\beta}_j, \boldsymbol{\beta}'_j$ and $\tilde{\mathbf{b}}_{-j}$ such that $(\boldsymbol{\beta}_j, \tilde{\mathbf{b}}_{-j}) \in \mathcal{L}_0$ and $(\boldsymbol{\beta}'_j, \tilde{\mathbf{b}}_{-j}) \in \mathcal{L}_0$. Moreover, $\nabla \ell(\boldsymbol{\beta}_0, \boldsymbol{\beta})$ is uniformly Lipschitz continuous with constant $M \in (0, \infty)$, i.e., for all $(\boldsymbol{\beta}_0, \boldsymbol{\beta}),\ (\boldsymbol{\beta}'_0, \boldsymbol{\beta}') \in \mathcal{L}_0$,

$$\|\nabla \ell(\boldsymbol{\beta}_0, \boldsymbol{\beta}) - \nabla \ell(\boldsymbol{\beta}'_0, \boldsymbol{\beta}')\|_2 \le M \|(\boldsymbol{\beta}_0, \boldsymbol{\beta}) - (\boldsymbol{\beta}'_0, \boldsymbol{\beta}')\|_2.$$

Proof of Lemma 2

Proof. As will be shown in the proof of Theorem 1, the MStweedie-GPG algorithm is descending along its iterations and we can thus restrict the domain of $(\boldsymbol{\beta}_0, \boldsymbol{\beta})$ to the sublevel set $\mathcal{L}_0$. Without loss of generality, assume not all $y_i^{(k)}$'s are zero. Define $\eta_i^{(k)} = \beta_0^{(k)} + \mathbf{x}_i^{(k)\top} \boldsymbol{\beta}^{(k)}$, $i = 1, \ldots, n_k, k = 1, \ldots, K$. It follows that the set

$$\mathcal{C}_0 = \{\boldsymbol{\eta} = (\eta_i^{(k)}, 1 \le i \le n_k, 1 \le k \le K) : (\boldsymbol{\beta}_0, \boldsymbol{\beta}) \in \mathcal{L}_0\}$$

is convex compact. Therefore, for all $(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \in \mathcal{L}_0$, $\eta_i^{(k)}$ is bounded by $\eta_{\max}$, where

$$\eta_{\max} = \max_{1 \le i \le n_k, 1 \le k \le K} \sup_{(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \in \mathcal{L}_0} |\eta_i^{(k)}| < \infty.$$

Also, $w_i^{(k)}$ and $y_i^{(k)}$ are bounded, respectively, by

$$w_{\max} = \max_{1 \le i \le n_k, 1 \le k \le K} w_i^{(k)} \quad \text{and} \quad y_{\max} = \max_{1 \le i \le n_k, 1 \le k \le K} y_i^{(k)}.$$

Let

$$\overline{w}_i^{(k)} = w_i^{(k)} \big((\rho - 1) y_i^{(k)} e^{(1-\rho)\eta_i^{(k)}} + (2 - \rho) e^{(2-\rho)\eta_i^{(k)}} \big).$$

Note that $\overline{w}_i^{(k)}$ is bounded by

$$\max_{1\leq i\leq n_k, 1\leq k\leq K} \sup_{(\boldsymbol{\beta}_0,\boldsymbol{\beta})\in\mathcal{L}_0} |\overline{w}_i^{(k)}| \leq w_{\max}\big(y_{\max}(\rho-1)e^{(\rho-1)\eta_{\max}} + (2-\rho)e^{(2-\rho)\eta_{\max}}\big) \equiv C.$$

Let $M_j = C\max_{1\leq k\leq K}\|X_j^{(k)}\|_2^2$. We can see that

$$\begin{aligned}\nabla_j^2\ell(\boldsymbol{\beta}_j;\tilde{\mathbf{b}}_{-j}) &= \frac{\partial^2}{\partial\boldsymbol{\beta}_j\partial\boldsymbol{\beta}_j^\top}\ell(\boldsymbol{\beta}_j;\tilde{\mathbf{b}}_{-j})\\ &= \mathrm{diag}\Big(X_j^{(k)\top}[\mathrm{diag}(\overline{w}_1^{(k)},\ldots,\overline{w}_{n_k}^{(k)})]X_j^{(k)}, k=1,\ldots,K\Big)\\ &\preceq M_j\mathbf{I}_K, \qquad \forall(\boldsymbol{\beta}_j;\tilde{\mathbf{b}}_{-j})\in\mathcal{L}_0.\end{aligned}$$

It follows from the mean-value theorem that $\nabla_j\ell(\boldsymbol{\beta}_j;\tilde{\mathbf{b}}_{-j})$ is uniformly Lipschitz continuous on the sublevel set $\mathcal{L}_0$. Indeed, the inequality

$$\|\nabla_j\ell(\boldsymbol{\beta}_j;\tilde{\mathbf{b}}_{-j}) - \nabla_j\ell(\boldsymbol{\beta}_j';\tilde{\mathbf{b}}_{-j})\|_2 \leq M_j\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j'\|_2$$

holds for any $\boldsymbol{\beta}_j, \boldsymbol{\beta}_j'$ and $\tilde{\mathbf{b}}_{-j}$ satisfying $(\boldsymbol{\beta}_j,\tilde{\mathbf{b}}_{-j})\in\mathcal{L}_0$ and $(\boldsymbol{\beta}_j',\tilde{\mathbf{b}}_{-j})\in\mathcal{L}_0$. Now let

$$M = \max_{1\leq k\leq K} C\Lambda_{\max}(\hat{\mathbf{X}}^{(k)\top}\hat{\mathbf{X}}^{(k)}),$$

where $\hat{\mathbf{X}}^{(k)} = (\mathbf{1}_{n_k}, \mathbf{X}^{(k)})$ and $\Lambda_{\max}(\cdot)$ denotes the largest eigenvalue of the enclosed matrix. We can similarly show that $\nabla\ell(\boldsymbol{\beta}_0,\boldsymbol{\beta})$ is uniformly Lipschitz continuous with constant M for all $(\boldsymbol{\beta}_0,\boldsymbol{\beta})\in\mathcal{L}_0$. □

Proof of Theorem 1

Proof. To simplify notation, let $\mathbf{b} = (\boldsymbol{\beta}_0,\boldsymbol{\beta})$ such that $\mathbf{b}_j = \boldsymbol{\beta}_j, 0\leq j\leq p$. Also, let $\ell(\mathbf{b}) = \ell(\boldsymbol{\beta}_0,\boldsymbol{\beta})$, $h(\mathbf{b}) = \lambda P_\alpha(\boldsymbol{\beta})$ and $f(\mathbf{b}) = \ell(\mathbf{b}) + h(\mathbf{b})$. Since h is separable in $\mathbf{b}$, we let $h_j(\mathbf{b}_j) = \lambda P_{\alpha,j}(\mathbf{b}_j), 0\leq j\leq p$. Denote by $\nabla\ell(\mathbf{b}) = \partial\ell(\mathbf{b})/\partial\mathbf{b}$ the gradient of ℓ and by $\nabla_j\ell(\mathbf{b}) = \partial\ell(\mathbf{b})/\partial\mathbf{b}_j$ the groupwise gradient of ℓ. Let $\nabla_j^2\ell(\mathbf{b}) = \partial^2\ell(\mathbf{b})/(\partial\mathbf{b}_j\partial\mathbf{b}_j^\top)$ be the Hessian matrix of $\ell(\cdot)$ for group j. In Lemma 2, we have shown that $\nabla\ell(\cdot)$ is uniformly Lipschitz continuous on the sublevel set $\mathcal{L}_0$ with constant M and $\nabla_j\ell(\cdot)$ is uniformly Lipschitz continuous on the sublevel set $\mathcal{L}_0$ with constant $M_j,\ 0\leq j\leq p$. Moreover, from (10), it can be shown that $\overline{w}_i^{(k)}$ is lower-bounded in the sublevel set $\mathcal{L}_0$. First, we have

$$\overline{w}_i^{(k)} \geq \Big(\frac{\rho-1}{2-\rho}\Big)^{3-2\rho} w_i^{(k)} (y_i^{(k)})^{2-\rho} I(y_i^{(k)} > 0) + (2-\rho) e^{-(2-\rho)\eta_{\max}} I(y_i^{(k)} = 0) > 0$$

for all $\mathbf{b} \in \mathcal{L}_0$ and $1 \leq i \leq n_k, 1 \leq k \leq K$. Let

$$w_{\min} = \min\bigg\{ \Big(\frac{\rho-1}{2-\rho}\Big)^{3-2\rho} \min_{i,k: y_i^{(k)}>0} w_i^{(k)} (y_i^{(k)})^{2-\rho},\ (2-\rho) e^{-(2-\rho)\eta_{\max}} \bigg\}.$$

Then we can see that $\overline{w}_i^{(k)} \geq w_{\min} > 0$. Therefore

$$\begin{aligned} \nabla_j^2 \ell(\mathbf{b}) &\succeq \operatorname{diag}\Big(X_j^{(k)\top} [\operatorname{diag}(\overline{w}_1^{(k)}, \ldots, \overline{w}_{n_k}^{(k)})] X_j^{(k)}, k = 1, \ldots, K \Big) \\ &\succeq w_{\min} \operatorname{diag}\Big(\|X_j^{(k)}\|_2^2, k = 1, \ldots, K \Big). \end{aligned}$$

As long as none of $\hat{\mathbf{X}}^{(k)}$'s columns are zero (otherwise we simply remove that column and the corresponding group variable), this implies that $\ell(\cdot)$ is groupwise strongly convex in $\mathcal{L}_0$.

Let t_j^{r+1} be the first step size that satisfies (13) when updating group $\mathbf{b}_j$ in the $(r+1)$-st cycle of MStweedie-GPG. We claim that

$$\frac{\delta}{M_j} \leq t_j^{r+1} \leq t_{\max},\ 0 \leq j \leq p. \tag{25}$$

Indeed, recall that in the line search, t_j starts with $t_{\max}$. The search then continues by scaling t_j down with the factor $\delta \in (0,1)$. Therefore, the last inequality holds in (25). Denote

$$G_{t_j}(\tilde{\mathbf{b}}) = G_{t_j}(\tilde{\boldsymbol{\beta}}_j; \tilde{\mathbf{b}}_{-j}) = \frac{\tilde{\boldsymbol{\beta}}_j - \operatorname{prox}_{\lambda v_j t_j h}(\tilde{\boldsymbol{\beta}}_j - t_j \nabla_j \ell(\tilde{\boldsymbol{\beta}}_j; \tilde{\mathbf{b}}_{-j}))}{t_j} = \frac{\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^+}{t_j}.$$

By the definition of M_j, we can see that

$$\begin{aligned} \ell(\boldsymbol{\beta}_j^+; \tilde{\mathbf{b}}_{-j}) &\leq \ell(\tilde{\mathbf{b}}) + \nabla_j \ell(\tilde{\boldsymbol{\beta}}_j; \tilde{\mathbf{b}}_{-j})^\top (\boldsymbol{\beta}_j^+ - \tilde{\boldsymbol{\beta}}_j) + \frac{M_j}{2} \|\boldsymbol{\beta}_j^+ - \tilde{\boldsymbol{\beta}}_j\|_2^2 \\ &= \ell(\tilde{\mathbf{b}}) - t_j \nabla_j \ell(\tilde{\boldsymbol{\beta}}_j; \tilde{\mathbf{b}}_{-j})^\top G_{t_j}(\tilde{\mathbf{b}}) + \frac{M_j t_j^2}{2} \|G_{t_j}(\tilde{\mathbf{b}})\|_2^2 \end{aligned}$$

holds for any t_j. Compared to (13), the above inequality implies that (13) can be satisfied by all $t_j \in [0, M_j^{-1}]$. Consequently, the first inequality holds in (25). Now let $t_{\min} = \delta / (\max_{0 \leq j \leq p} M_j)$, we conclude that $t_j^{r+1} \in [t_{\min},\ t_{\max}]$ for all j and r.

In the cyclic MStweedie-GPG algorithm, let $\mathbf{b}^r$ be the update of $\mathbf{b}$ after the r-th cycle. For notational convenience, define the following auxiliary variables

$$\mathbf{B}_j^{r+1} \equiv (\mathbf{b}_0^{r+1}, \dots, \mathbf{b}_{j-1}^{r+1}, \mathbf{b}_j^r, \mathbf{b}_{j+1}^r, \dots, \mathbf{b}_p^r)^\top, j = 0, \dots, p,$$
$$\mathbf{B}_{-j}^{r+1} \equiv (\mathbf{b}_0^{r+1}, \dots, \mathbf{b}_{j-1}^{r+1}, \mathbf{b}_{j+1}^r, \dots, \mathbf{b}_p^r)^\top, j = 0, \dots, p,$$

For $\mathbf{z} \in \mathbb{R}^K$, let

$$(\mathbf{z}; \mathbf{B}_{-j}^{r+1}) \equiv (\mathbf{b}_0^{r+1}, \dots, \mathbf{b}_{j-1}^{r+1}, \mathbf{z}, \mathbf{b}_{j+1}^r, \dots, \mathbf{b}_p^r)^\top.$$

Clearly we have $\mathbf{B}_0^{r+1} = \mathbf{b}^r$ and $\mathbf{B}_{p+1}^{r+1} = \mathbf{b}^{r+1}$, and we have

$$\mathbf{B}_j^{r+1} = (\mathbf{b}_j^r; \mathbf{B}_{-j}^{r+1}), \qquad \mathbf{B}_{j+1}^{r+1} = (\mathbf{b}_j^{r+1}; \mathbf{B}_{-j}^{r+1}).$$

Under the new notation, (13) can be rewritten as

$$\ell(\mathbf{B}_{j+1}^{r+1}) = \ell(\mathbf{b}_j^{r+1}; \mathbf{B}_{-j}^{r+1}) \leq \ell(\mathbf{B}_j^{r+1}) - t_j^{r+1} \nabla_j \ell(\mathbf{B}_j^{r+1})^\top G_{t_j^{r+1}}(\mathbf{B}_j^{r+1}) + \frac{t_j^{r+1}}{2} \|G_{t_j^{r+1}}(\mathbf{B}_j^{r+1})\|_2^2, \quad (26)$$

where

$$G_{t_j^{r+1}}(\mathbf{B}_j^{r+1}) \equiv G_{t_j^{r+1}}(\mathbf{b}_j^r; \mathbf{B}_{-j}^{r+1}) = -\frac{\mathbf{b}_j^{r+1} - \mathbf{b}_j^r}{t_j^{r+1}}. \quad (27)$$

Next, we show that for any $\mathbf{z} \in \mathbb{R}^K$,

$$f(\mathbf{B}_{j+1}^{r+1}) \leq f(\mathbf{z}; \mathbf{B}_{-j}^{r+1}) + G_{t_j^{r+1}}(\mathbf{B}_j^{r+1})^\top (\mathbf{b}_j^r - \mathbf{z}) - \frac{t_j^{r+1}}{2} \|G_{t_j^{r+1}}(\mathbf{B}_j^{r+1})\|_2^2. \quad (28)$$

Let

$$\ell_{Q_j}(\mathbf{B}_{j+1}^{r+1}) = \ell_{Q_j}(\mathbf{b}_j^{r+1}; \mathbf{B}_{-j}^{r+1}) = \ell(\mathbf{B}_j^{r+1}) + \nabla_j \ell(\mathbf{B}_j^{r+1})^\top (\mathbf{b}_j^{r+1} - \mathbf{b}_j^r) + \frac{1}{2t_j^{r+1}} \|\mathbf{b}_j^{r+1} - \mathbf{b}_j^r\|_2^2.$$

The gradient of ℓ_{Q_j} is

$$\nabla_j \ell_{Q_j}(\mathbf{B}_{j+1}^{r+1}) = \nabla_j \ell(\mathbf{B}_j^{r+1}) + \frac{\mathbf{b}_j^{r+1} - \mathbf{b}_j^r}{t_j} = \nabla_j \ell(\mathbf{B}_j^{r+1}) - G_{t_j^{r+1}}(\mathbf{B}_j^{r+1}). \quad (29)$$

By subgradient optimality condition, we have

$$\mathbf{0} \in \nabla_j \ell_{Q_j}(\mathbf{B}_{j+1}^{r+1}) + \partial h_j(\mathbf{b}_j^{r+1}),$$

thus

$$G_{t_j^{r+1}}(\mathbf{B}_j^{r+1}) - \nabla_j \ell(\mathbf{B}_j^{r+1}) \in \partial h_j(\mathbf{b}_j^{r+1}). \tag{30}$$

Now by convexity of ℓ

$$\ell(\mathbf{z}; \mathbf{B}_{-j}^{r+1}) \geq \ell(\mathbf{b}_j^r; \mathbf{B}_{-j}^{r+1}) + \nabla_j \ell(\mathbf{B}_j^{r+1})^\top (\mathbf{z} - \mathbf{b}_j^r), \tag{31}$$

and the convexity of h

$$h(\mathbf{z}; \mathbf{B}_{-j}^{r+1}) = h_j(\mathbf{z}) + \sum_{0 \leq m \leq p, m \neq j} h_m(\mathbf{b}_m^{r+I(m<j)}) \geq h(\mathbf{b}_j^{r+1}; \mathbf{B}_{-j}^{r+1}) + \partial h_j(\mathbf{b}_j^{r+1})^\top (\mathbf{z} - \mathbf{b}_j^{r+1}) \tag{32}$$

and (13), we have that for any $\mathbf{z} \in \mathbb{R}^K$,

$$\begin{aligned}
& f(\mathbf{B}_{j+1}^{r+1}) = f(\mathbf{b}_j^{r+1}; \mathbf{B}_{-j}^{r+1}) = \ell(\mathbf{b}_j^{r+1}; \mathbf{B}_{-j}^{r+1}) + h(\mathbf{b}_j^{r+1}; \mathbf{B}_{-j}^{r+1}) \\
& \overset{(26)}{\leq} \ell(\mathbf{B}_j^{r+1}) - t_j^{r+1} \nabla_j \ell(\mathbf{B}_j^{r+1})^\top G_{t_j^{r+1}}(\mathbf{B}_j^{r+1}) + \frac{t_j^{r+1}}{2} \|G_{t_j^{r+1}}(\mathbf{B}_j^{r+1})\|_2^2 + h(\mathbf{b}_j^{r+1}; \mathbf{B}_{-j}^{r+1}) \\
& \overset{(31)(32)}{\leq} \ell(\mathbf{z}; \mathbf{B}_{-j}^{r+1}) + \nabla_j \ell(\mathbf{B}_j^{r+1})^\top (\mathbf{b}_j^r - \mathbf{z}) - t_j^{r+1} \nabla_j \ell(\mathbf{B}_j^{r+1})^\top G_{t_j^{r+1}}(\mathbf{B}_j^{r+1}) \\
& \quad + \frac{t_j^{r+1}}{2} \|G_{t_j^{r+1}}(\mathbf{B}_j^{r+1})\|_2^2 + h_j(\mathbf{z}) + \partial h_j(\mathbf{b}_j^{r+1})^\top (\mathbf{b}_j^{r+1} - \mathbf{z}) + \sum_{0 \leq m \leq p, m \neq j} h_m(\mathbf{b}_m^{r+I(m<j)}) \\
& \overset{(30)}{=} \ell(\mathbf{z}; \mathbf{B}_{-j}^{r+1}) + \nabla_j \ell(\mathbf{B}_j^{r+1})^\top (\mathbf{b}_j^r - \mathbf{z}) - t_j^{r+1} \nabla_j \ell(\mathbf{B}_j^{r+1})^\top G_{t_j^{r+1}}(\mathbf{B}_j^{r+1}) \\
& \quad + \frac{t_j^{r+1}}{2} \|G_{t_j^{r+1}}(\mathbf{B}_j^{r+1})\|_2^2 + h_j(\mathbf{z}) + (G_{t_j^{r+1}}(\mathbf{B}_j^{r+1}) - \nabla_j \ell(\mathbf{B}_j^{r+1}))^\top (\mathbf{b}_j^{r+1} - \mathbf{z}) \\
& \quad + \sum_{0 \leq m \leq p, m \neq j} h_m(\mathbf{b}_m^{r+I(m<j)}) \\
& = \ell(\mathbf{z}; \mathbf{B}_{-j}^{r+1}) + h(\mathbf{z}; \mathbf{B}_{-j}^{r+1}) + \nabla_j \ell(\mathbf{B}_j^{r+1})^\top (\mathbf{b}_j^r - \mathbf{b}_j^{r+1}) - t_j^{r+1} \nabla_j \ell(\mathbf{B}_j^{r+1})^\top G_{t_j^{r+1}}(\mathbf{B}_j^{r+1}) \\
& \quad + \frac{t_j^{r+1}}{2} \|G_{t_j^{r+1}}(\mathbf{B}_j^{r+1})\|_2^2 + G_{t_j^{r+1}}(\mathbf{B}_j^{r+1})^\top (\mathbf{b}_j^{r+1} - \mathbf{b}_j^r + \mathbf{b}_j^r - \mathbf{z}) \\
& \overset{(27)}{=} f(\mathbf{z}; \mathbf{B}_{-j}^{r+1}) + G_{t_j^{r+1}}(\mathbf{B}_j^{r+1})^\top (\mathbf{b}_j^r - \mathbf{z}) - \frac{t_j^{r+1}}{2} \|G_{t_j^{r+1}}(\mathbf{B}_j^{r+1})\|_2^2,
\end{aligned}$$

which proves (28).

Now taking $\mathbf{z} = \mathbf{b}_j^r$ in (28), we have

$$f(\mathbf{B}_j^{r+1}) - f(\mathbf{B}_{j+1}^{r+1}) \geq \frac{t_j^{r+1}}{2}\|G_{t_j^{r+1}}(\mathbf{B}_j^{r+1})\|_2^2 = \frac{1}{2t_j^{r+1}}\|\mathbf{b}_j^r - \mathbf{b}_j^{r+1}\|_2^2 \geq \frac{1}{2t_{\max}}\|\mathbf{b}_j^r - \mathbf{b}_j^{r+1}\|_2^2,$$

which implies that the MStweedie-GPG algorithm is descending. Moreover, we have the descent property of MStweedie-GPG over the cycles

$$f(\mathbf{b}^r) - f(\mathbf{b}^{r+1}) = \sum_{j=0}^{p}[f(\mathbf{B}_j^{r+1}) - f(\mathbf{B}_{j+1}^{r+1})] \geq (2t_{\max})^{-1}\|\mathbf{b}^r - \mathbf{b}^{r+1}\|_2^2. \tag{33}$$

Now let $\mathcal{X}^* := \{\mathbf{b}^* \in \mathcal{L}_0 : f(\mathbf{b}^*) = \min_{\mathbf{b}\in\mathcal{L}_0} f(\mathbf{b})\}$ be the optimal solution set of problem (6) and define $\mathrm{d}_{\mathcal{X}^*}(\mathbf{b}) := \min_{\mathbf{b}^*\in\mathcal{X}^*}\|\mathbf{b} - \mathbf{b}^*\|_2$ to be the minimum distance from $\mathbf{b}$ to $\mathcal{X}^*$. Let $\mathbf{b}^{r*}$ be the point in $\mathcal{X}^*$ such that $\|\mathbf{b}^r - \mathbf{b}^{r*}\|_2 = \mathrm{d}_{\mathcal{X}^*}(\mathbf{b}^r)$. We also have $f(\mathbf{b}^{r*}) = f^* := \min_{\mathbf{b}\in\mathcal{L}_0} f(\mathbf{b})$. By the mean value theorem, there exists $\mu \in [0,1]$ and $\boldsymbol{\zeta}^r = \mu\mathbf{b}^{r+1} + (1-\mu)\mathbf{b}^{r*}$ such that

$$\ell(\mathbf{b}^{r+1}) - \ell(\mathbf{b}^{r*}) = (\nabla\ell(\boldsymbol{\zeta}^r))^\top(\mathbf{b}^{r+1} - \mathbf{b}^{r*}).$$

It follows that

$$\begin{aligned}
&f(\mathbf{b}^{r+1}) - f^* = f(\mathbf{b}^{r+1}) - f(\mathbf{b}^{r*})\\
&= \ell(\mathbf{b}^{r+1}) - \ell(\mathbf{b}^{r*}) + \sum_{j=0}^{p}[h_j(\mathbf{b}_j^{r+1}) - h_j(\mathbf{b}_j^{r*})]\\
&= \sum_{j=0}^{p}[\nabla_j\ell(\boldsymbol{\zeta}^r)^\top(\mathbf{b}_j^{r+1} - \mathbf{b}_j^{r*}) + h_j(\mathbf{b}_j^{r+1}) - h_j(\mathbf{b}_j^{r*})]\\
&= \sum_{j=0}^{p}[\nabla_j\ell(\mathbf{B}_j^{r+1})^\top(\mathbf{b}_j^{r+1} - \mathbf{b}_j^{r*}) + h_j(\mathbf{b}_j^{r+1}) - h_j(\mathbf{b}_j^{r*})\\
&\qquad + (\nabla_j\ell(\boldsymbol{\zeta}^r) - \nabla_j\ell(\mathbf{B}_j^{r+1}))^\top(\mathbf{b}_j^{r+1} - \mathbf{b}_j^{r*})].
\end{aligned}$$

By convexity of h, we have

$$\begin{aligned}
&\nabla_j \ell(\mathbf{B}_j^{r+1})^\top (\mathbf{b}_j^{r+1} - \mathbf{b}_j^{r*}) + h_j(\mathbf{b}_j^{r+1}) - h_j(\mathbf{b}_j^{r*}) \\
&\le \nabla_j \ell(\mathbf{B}_j^{r+1})^\top (\mathbf{b}_j^{r+1} - \mathbf{b}_j^{r*}) - \partial h_j(\mathbf{b}_j^{r+1})^\top (\mathbf{b}_j^{r*} - \mathbf{b}_j^{r+1}) \\
&\overset{(30)}{=} \nabla_j \ell(\mathbf{B}_j^{r+1})^\top (\mathbf{b}_j^{r+1} - \mathbf{b}_j^{r*}) - (G_{t_j^{r+1}}(\mathbf{B}_j^{r+1}) - \nabla_j \ell(\mathbf{B}_j^{r+1}))^\top (\mathbf{b}_j^{r*} - \mathbf{b}_j^{r+1}) \\
&= -G_{t_j^{r+1}}(\mathbf{B}_j^{r+1})(\mathbf{b}_j^{r*} - \mathbf{b}_j^{r+1}) \\
&= \frac{1}{t_j^{r+1}}(\mathbf{b}_j^{r+1} - \mathbf{b}_j^{r})(\mathbf{b}_j^{r*} - \mathbf{b}_j^{r} + \mathbf{b}_j^{r} - \mathbf{b}_j^{r+1}) \\
&\le \frac{1}{t_j^{r+1}}[(\mathbf{b}_j^{r+1} - \mathbf{b}_j^{r})^\top (\mathbf{b}_j^{r*} - \mathbf{b}_j^{r}) - \|\mathbf{b}_j^{r+1} - \mathbf{b}_j^{r}\|_2^2] \\
&\le \frac{1}{2t_j^{r+1}}[\|\mathbf{b}_j^{r*} - \mathbf{b}_j^{r}\|_2^2 + \|\mathbf{b}_j^{r} - \mathbf{b}_j^{r+1}\|_2^2] \\
&\le \frac{1}{2t_{\min}}[\|\mathbf{b}_j^{r*} - \mathbf{b}_j^{r}\|_2^2 + \|\mathbf{b}_j^{r} - \mathbf{b}_j^{r+1}\|_2^2].
\end{aligned}$$

Moreover, by the Lipschitz continuity of $\nabla \ell(\cdot)$ and the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
&\left(\sum_{j=0}^{p} (\nabla_j \ell(\boldsymbol{\zeta}^r) - \nabla_j \ell(\mathbf{B}_j^{r+1}))^\top (\mathbf{b}_j^{r+1} - \mathbf{b}_j^{r*})\right)^2 \\
&\le \left(\sum_{j=0}^{p} \|\nabla \ell(\boldsymbol{\zeta}^r) - \nabla \ell(\mathbf{B}_j^{r+1})\|_2^2\right)\left(\sum_{j=0}^{p} \|\mathbf{b}_j^{r+1} - \mathbf{b}_j^{r*}\|_2^2\right) \\
&\le \left(\sum_{j=0}^{p} M^2 \|\boldsymbol{\zeta}^r - \mathbf{B}_j^{r+1}\|_2^2\right) \|\mathbf{b}^{r+1} - \mathbf{b}^{r*}\|_2^2 \\
&= \left(\sum_{j=0}^{p} M^2 \sum_{j'=0}^{p} \|\mu(\mathbf{b}_{j'}^{r+1} - \mathbf{b}_{j'}^{r}) + (1-\mu)(\mathbf{b}_{j'}^{r*} - \mathbf{b}_{j'}^{r}) + \mathbf{b}_{j'}^{r} - \mathbf{b}_{j'}^{r+I(j' \le j)}\|_2^2\right) \\
&\quad \cdot 2(\|\mathbf{b}^{r+1} - \mathbf{b}^{r}\|_2^2 + \|\mathbf{b}^{r*} - \mathbf{b}^{r}\|_2^2) \\
&\le \left(2\sum_{j=0}^{p} M^2 \|\mathbf{b}^{r+1} - \mathbf{b}^{r}\|_2^2 + \|\mathbf{b}^{r*} - \mathbf{b}^{r}\|_2^2\right) \cdot 2(\|\mathbf{b}^{r+1} - \mathbf{b}^{r}\|_2^2 + \|\mathbf{b}^{r*} - \mathbf{b}^{r}\|_2^2) \\
&\le 4(p+1)M^2 \big(\|\mathbf{b}^{r+1} - \mathbf{b}^{r}\|_2^2 + \|\mathbf{b}^{r*} - \mathbf{b}^{r}\|_2^2\big)^2.
\end{aligned}$$

Altogether these imply

$$\begin{aligned} f(\mathbf{b}^{r+1}) - f^* &\le \sum_{j=0}^{p} \frac{1}{2t_{\min}} [\|\mathbf{b}_j^{r*} - \mathbf{b}_j^r\|_2^2 + \|\mathbf{b}_j^r - \mathbf{b}_j^{r+1}\|_2^2] \\ &\quad + 2M\sqrt{p+1}\big(\|\mathbf{b}^{r+1} - \mathbf{b}^r\|_2^2 + \mathrm{d}_{\mathcal{X}^*}^2(\mathbf{b}^r)\big) \\ &\le \Big(\frac{1}{2t_{\min}} + 2M\sqrt{p+1}\Big)\big(\|\mathbf{b}^{r+1} - \mathbf{b}^r\|_2^2 + \mathrm{d}_{\mathcal{X}^*}^2(\mathbf{b}^r)\big). \end{aligned} \tag{34}$$

According to our algorithm,

$$\begin{aligned} \mathbf{b}_j^{r+1} &= \arg\min_{\mathbf{z}\in\mathbb{R}^K} \ell_{Q_j}(\mathbf{z}; \mathbf{B}_j^{r+1}) + h_j(\mathbf{z}) \\ &= \arg\min_{\mathbf{z}\in\mathbb{R}^K} \ell(\mathbf{B}_j^{r+1}) + \nabla_j \ell(\mathbf{B}_j^{r+1})^\top(\mathbf{z} - \mathbf{b}_j^r) + \frac{1}{2t_j^{r+1}}\|\mathbf{z} - \mathbf{b}_j^r\|_2^2 + h_j(\mathbf{z}). \end{aligned} \tag{35}$$

By the optimality condition of $\mathbf{b}_j^{r+1}$ in (35), we have

$$\mathbf{b}_j^{r+1} = \mathrm{prox}_{t_j^{r+1} h_j}(\mathbf{b}_j^{r+1} - t_j^{r+1}\nabla_j \ell_{Q_j}(\mathbf{b}_j^{r+1}; \mathbf{B}_j^{r+1})).$$

Now let $c_0 = \min(1, t_{\max})$. It follows from Lemma 4.3 of Kadkhodaie et al. (2014) that

$$\begin{aligned}
&\|\mathbf{b}_j^r - \operatorname{prox}_{h_j}(\mathbf{b}_j^r - \nabla_j \ell(\mathbf{b}^r))\|_2 \\
&\le \frac{1}{\max(1, 1/t_j^{r+1})} \|\mathbf{b}_j^r - \operatorname{prox}_{t_j^{r+1}h_j}(\mathbf{b}_j^r - t_j^{r+1}\nabla_j \ell(\mathbf{b}^r))\|_2 \\
&= \min(1, t_j^{r+1}) \|\mathbf{b}_j^r - \operatorname{prox}_{t_j^{r+1}h_j}(\mathbf{b}_j^r - t_j^{r+1}\nabla_j \ell(\mathbf{b}^r))\|_2 \\
&\le c_0 \|\mathbf{b}_j^r - \operatorname{prox}_{t_j^{r+1}h_j}(\mathbf{b}_j^r - t_j^{r+1}\nabla_j \ell(\mathbf{b}^r)) + \mathbf{b}_j^{r+1} - \mathbf{b}_j^{r+1}\|_2 \\
&\le c_0 [\|\mathbf{b}_j^{r+1} - \operatorname{prox}_{t_j^{r+1}h_j}(\mathbf{b}_j^r - t_j^{r+1}\nabla_j \ell(\mathbf{b}^r))\|_2 + \|\mathbf{b}_j^{r+1} - \mathbf{b}_j^r\|_2] \\
&\le c_0 [\| \operatorname{prox}_{t_j^{r+1}h_j}(\mathbf{b}_j^{r+1} - t_j^{r+1}\nabla_j \ell_{Q_j}(\mathbf{b}_j^{r+1}; \mathbf{B}_j^{r+1})) \\
&\qquad\qquad - \operatorname{prox}_{t_j^{r+1}h_j}(\mathbf{b}_j^r - t_j^{r+1}\nabla_j \ell(\mathbf{b}^r))\|_2 + \|\mathbf{b}_j^{r+1} - \mathbf{b}_j^r\|_2] \\
&\le 2c_0 \|\mathbf{b}_j^{r+1} - \mathbf{b}_j^r\|_2 + c_0 t_j^{r+1} \|\nabla_j \ell_{Q_j}(\mathbf{b}_j^{r+1}; \mathbf{B}_j^{r+1}) - \nabla_j \ell(\mathbf{b}^r)\|_2 \\
&\overset{(29)}{=} 2c_0 \|\mathbf{b}_j^{r+1} - \mathbf{b}_j^r\|_2 + c_0 t_j^{r+1} \|\nabla_j \ell(\mathbf{B}_j^{r+1}) + \frac{1}{t_j^{r+1}}(\mathbf{b}_j^{r+1} - \mathbf{b}_j^r) - \nabla_j \ell(\mathbf{b}^r)\|_2 \\
&\le 3c_0 \|\mathbf{b}_j^{r+1} - \mathbf{b}_j^r\|_2 + c_0 t_{\max} \|\nabla_j \ell(\mathbf{B}_j^{r+1}) - \nabla_j \ell(\mathbf{b}^r)\|_2 \\
&\le 3c_0 \|\mathbf{b}_j^{r+1} - \mathbf{b}_j^r\|_2 + c_0 t_{\max} \|\nabla \ell(\mathbf{B}_j^{r+1}) - \nabla \ell(\mathbf{b}^r)\|_2 \\
&\le 3c_0 \|\mathbf{b}_j^{r+1} - \mathbf{b}_j^r\|_2 + c_0 t_{\max} M \|\mathbf{B}_j^{r+1} - \mathbf{b}^r\|_2.
\end{aligned}$$

It follows that

$$\|\mathbf{b}^r - \operatorname{prox}_h(\mathbf{b}^r - \nabla \ell(\mathbf{b}^r))\|_2 \le (3c_0 + c_0 t_{\max} M \sqrt{p+1}) \|\mathbf{b}^{r+1} - \mathbf{b}^r\|_2. \tag{36}$$

Note that

$$\ell(\boldsymbol{\eta}) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} w_i^{(k)} \left\{ -\frac{y_i^{(k)} e^{(1-\rho)\eta_i^{(k)}}}{1-\rho} + \frac{e^{(2-\rho)\eta_i^{(k)}}}{2-\rho} \right\}$$

is strongly convex in $\boldsymbol{\eta} \in \mathcal{C}_0$ and $\boldsymbol{\eta}$ is an affine transformation of $(\boldsymbol{\beta}_0, \boldsymbol{\beta})$, i.e., $\eta_i^{(k)} = \beta_0^{(k)} + \mathbf{x}_i^{(k)\top} \boldsymbol{\beta}^{(k)}$. It follows from Zhang et al. (2013) that for any given $\xi \ge f^* = \min_{\mathbf{b} \in \mathcal{L}_0} f(\mathbf{b})$, there exists $\kappa, \epsilon > 0$ such that, for all $\mathbf{b} \in \mathcal{L}_0$ satisfying $f(\mathbf{b}) \le \xi$ and $\|\mathbf{b} - \operatorname{prox}_h(\mathbf{b} - \nabla \ell(\mathbf{b}))\|_2 \le \epsilon$, we have

$$\mathrm{d}_{\mathcal{X}^*}(\mathbf{b}) \le \kappa \|\mathbf{b} - \operatorname{prox}_h(\mathbf{b} - \nabla \ell(\mathbf{b}))\|_2. \tag{37}$$

From (33), we can see that

$$\sum_{i=0}^{r} \|\mathbf{b}^i - \mathbf{b}^{i+1}\|_2^2 \leq 2t_{\max} \sum_{i=0}^{r} \left[f(\mathbf{b}^i) - f(\mathbf{b}^{i+1})\right] = 2t_{\max} \left[f(\mathbf{b}^0) - f(\mathbf{b}^{r+1})\right] \leq 2t_{\max} f(\mathbf{b}^0) < \infty,$$

then we must have $\|\mathbf{b}^{r+1} - \mathbf{b}^r\|_2 \to 0$ as $r \to \infty$. Thus, it follows from (36) that as $r \to \infty$, $\|\mathbf{b}^r - \mathrm{prox}_h(\mathbf{b}^r - \nabla \ell(\mathbf{b}^r))\|_2 \to 0$, and further by (37), this implies that $\mathrm{d}_{\mathcal{X}^*}(\mathbf{b}^r) \to 0$ as $r \to \infty$. Consequently, from (34) it follows that $f(\mathbf{b}^r) \to f^*$, which proves that the MStweedie-GPG algorithm converges to the global minimum. Let $\Delta^r = f(\mathbf{b}^r) - f^*$, $c_1 = \frac{1}{2t_{\min}} + 2M\sqrt{p+1}$. By (37) and (34) again, we have for large enough r,

$$\begin{aligned}
\Delta^{r+1} = f(\mathbf{b}^{r+1}) - f^* &\leq c_1[\mathrm{d}^2_{\mathcal{X}^*}(\mathbf{b}^r) + \|\mathbf{b}^{r+1} - \mathbf{b}^r\|_2^2] \\
&\leq c_1\kappa^2 \|\mathbf{b}^r - \mathrm{prox}_h(\mathbf{b}^r - \nabla \ell(\mathbf{b}^r))\|_2^2 + c_1\|\mathbf{b}^{r+1} - \mathbf{b}^r\|_2^2 \\
&\leq (c_1\kappa^2(3c_0 + c_0 t_{\max} M\sqrt{p+1})^2 + c_1)\|\mathbf{b}^{r+1} - \mathbf{b}^r\|_2^2 \\
&\leq (c_1\kappa^2(3c_0 + c_0 t_{\max} M\sqrt{p+1})^2 + c_1) \cdot 2t_{\max}[f(\mathbf{b}^r) - f(\mathbf{b}^{r+1})] \\
&= 2t_{\max}(c_1\kappa^2(3c_0 + c_0 t_{\max} M\sqrt{p+1})^2 + c_1)(\Delta^r - \Delta^{r+1}).
\end{aligned}$$

This implies that

$$\Delta^{r+1} \leq \frac{c_2}{1+c_2}\Delta^r, \tag{38}$$

where $c_2 = 2t_{\max}(c_1\kappa^2(3c_0 + c_0 t_{\max} M\sqrt{p+1})^2 + c_1)$. Let $c_3 = c_2/(1+c_2)$. From (38), we can see that $f(\mathbf{b}^r)$ approaches f^* with linear rate $O(c_3^r)$. By (33) this further implies that $\{\mathbf{b}^r, r \geq 0\}$ converges at least linearly. □

Appendix E. Numerical Studies on Correlated Responses

Setting 6 – Correlated responses

In this simulation setting, we study the impact of having correlated responses on the performance of our proposed algorithm. Correlation is introduced using two compounded techniques. We consider a simultaneous setting, i.e. an observation consists of a vector of features $\mathbf{x}$ which is used to predict all K responses. When the coefficients are similar across tasks, then there will be correlation induced from the fact that the means $\mu^{(k)} = \exp\left(\mathbf{x}\boldsymbol{\beta}^{(k)}\right)$, $k = 1, \ldots, K$, will be related. If we simply generate K Tweedie variables from these means, then the random variables will be independent, conditionally on the vector of means. To introduce additional correlation, we consider the following setup inspired from a claim count decomposition suggested by Bermúdez and Karlis (2011). We generate $K' > K$ independent Tweedie variables with means $\mu^{(k)} = \exp\left(\mathbf{x}\boldsymbol{\beta}^{(k)}\right)$, $k = 1, \ldots, K'$, for some choice of coefficients $\boldsymbol{\beta}^{(k)}$ and produce the responses by taking a linear combination of these independent Tweedie random variables. In particular, we consider $\widetilde{Y}^{(k)}$, $k = 1, \ldots, 6$, the independent Tweedie random variables generating the $K = 3$ observed responses as follows:

$$\begin{bmatrix} Y^{(1)} \\ Y^{(2)} \\ Y^{(3)} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}}_{=:A} \begin{bmatrix} \widetilde{Y}^{(1)} & \widetilde{Y}^{(2)} & \widetilde{Y}^{(3)} & \widetilde{Y}^{(4)} & \widetilde{Y}^{(5)} & \widetilde{Y}^{(6)} \end{bmatrix}^{\top}.$$

The correlation depends on the mean of each independent Tweedie, but it is clear there will be correlation introduced in that way. Indeed, if $\widetilde{Y}^{(4)} > 0$, then both $Y^{(1)}$ and $Y^{(2)}$ will be non-zero.

This construction actually has a real interpretation. Suppose the $Y^{(k)}$ represent different aspect of a car insurance policy (1: personal injury, 2: property damage, 3: third party). Then, the random variable $\widetilde{Y}^{(4)}$ can be seen as the total claim amount that is common to personal injuries and property damages but without third party damages, while the difference between those aspects is captured by $\widetilde{Y}^{(1)}$ and $\widetilde{Y}^{(2)}$, which are independent.

We consider three experiments under this setting. In the first two cases, we set $\rho = 1.5$, $\phi = 40$ and $n^{(k)} = 1000$ and consider $p = 50$ features of which only the first 5 are truly generating the data. Each $x_{ij}^{(k)}$ is produced from a standard normal distribution. In the experiment 6A, we consider equal contribution of the features across the sources so that the Lasso on the full dataset should be sufficient:

$$\begin{bmatrix} \boldsymbol{\beta}^{(1)} & \boldsymbol{\beta}^{(2)} & \boldsymbol{\beta}^{(3)} & \boldsymbol{\beta}^{(4)} & \boldsymbol{\beta}^{(5)} & \boldsymbol{\beta}^{(6)} \end{bmatrix} = \begin{bmatrix} 0.2 & 0.2 & 0.2 & 0.8 & 0.8 & 0.8 \\ 0.2 & 0.2 & 0.2 & 0.8 & 0.8 & 0.8 \\ 0.2 & 0.2 & 0.2 & 0.8 & 0.8 & 0.8 \\ -0.2 & -0.2 & -0.2 & -0.8 & -0.8 & -0.8 \\ -0.2 & -0.2 & -0.2 & -0.8 & -0.8 & -0.8 \\ \mathbf{0}_{45} & \mathbf{0}_{45} & \mathbf{0}_{45} & \mathbf{0}_{45} & \mathbf{0}_{45} & \mathbf{0}_{45} \end{bmatrix}.$$

Upon generating 100 replications of the experiment, we obtain the following empirical correlation

(a) Setting 6A: Mean (standard error)

	Full Lasso	Ind. Lasso	L_1/L_∞	a-L_1/L_∞	L_1/L_2	a-L_1/L_2
Test dev.	48.13 (0.28)	52.11 (0.43)	49.82 (0.33)	48.63 (0.29)	49.85 (0.35)	48.32 (0.29)
Size	11.47 (0.48)	10.02 (0.35)	6.74 (0.20)	5.54 (0.13)	7.40 (0.21)	5.44 (0.10)
Accuracy	87.06 (0.96)	89.92 (0.70)	96.52 (0.39)	98.92 (0.25)	95.20 (0.43)	99.08 (0.21)
Precision	50.31 (1.82)	55.07 (1.67)	78.96 (1.77)	93.26 (1.37)	72.76 (1.89)	94.02 (1.23)
Recall	100.00 (0.00)	99.80 (0.20)	100.00 (0.00)	100.00 (0.00)	100.00 (0.00)	99.80 (0.20)

(b) Setting 6A: Mean rank (nb. times best)

	Full Lasso	Ind. Lasso	L_1/L_∞	a-L_1/L_∞	L_1/L_2	a-L_1/L_2
Test dev.	1.75 (50)	5.89 (0)	4.39 (1)	2.58 (20)	4.28 (0)	2.11 (29)
Size	5.21 (1)	4.83 (3)	2.50 (29)	1.26 (83)	3.20 (21)	1.18 (87)
Accuracy	5.20 (1)	4.84 (3)	2.49 (30)	1.25 (84)	3.19 (22)	1.21 (86)
Precision	5.20 (1)	4.84 (3)	2.49 (30)	1.25 (84)	3.19 (22)	1.18 (87)
Recall	1.00 (100)	1.04 (99)	1.00 (100)	1.00 (100)	1.00 (100)	1.04 (99)

Table A1: *Results from Setting 6A with 100 replications. Part (a) shows the mean values of the statistics and their standard errors in parentheses. Part (b) shows the mean rank across the six models and, in parentheses, the number of times the model is best.*

matrix,

$$\begin{bmatrix} 1.00 & 0.76 & 0.73 \\ 0.76 & 1.00 & 0.74 \\ 0.73 & 0.74 & 1.00 \end{bmatrix},$$

and the three sources respectively have 73.6%, 74.5% and 74.9% of zeroes.

Since the mean is exponential in the coefficients, we cannot compute the true coefficients generating the real responses so that it is impossible to produce the L_2-loss measure of performance. However, the selection accuracy measures (accuracy, precision and recall) are still relevant. The results of training and testing the usual six models are contained in Table A1. The two adaptive versions of our algorithm (especially a-L_1/L_2) achieve test deviance values similar to that of Full Lasso, but using far less features. The accuracy and precision are therefore much better with a similar fit to the test data. This suggests that our proposal is quite robust to correlated data and can actually benefit from it.

In experiment 6B, we consider unequal contribution of the coefficients to the means of the independent random variables:

$$\begin{bmatrix} \boldsymbol{\beta}^{(1)} & \boldsymbol{\beta}^{(2)} & \boldsymbol{\beta}^{(3)} & \boldsymbol{\beta}^{(4)} & \boldsymbol{\beta}^{(5)} & \boldsymbol{\beta}^{(6)} \end{bmatrix} = \begin{bmatrix} 0.5 & 0.5 & 0 & 1.5 & 3.0 & 1.5 \\ 0.5 & 0 & 0.5 & 3.0 & 1.5 & 1.5 \\ 0 & 0.5 & 0.5 & 1.5 & 1.5 & 3.0 \\ 0 & 0 & 0 & 0 & -3.0 & 0 \\ 0 & 0 & 0 & -3.0 & 0 & 0 \\ \mathbf{0}_{45} & \mathbf{0}_{45} & \mathbf{0}_{45} & \mathbf{0}_{45} & \mathbf{0}_{45} & \mathbf{0}_{45} \end{bmatrix}.$$

Hence, the Full Lasso should not perform well in this case, but the individual Lasso should be able to capture the differences between sources. Upon generating 100 replications of the experiment, we obtain the following empirical correlation matrix,

$$\begin{bmatrix} 1.00 & 0.83 & 0.21 \\ 0.83 & 1.00 & 0.48 \\ 0.21 & 0.48 & 1.00 \end{bmatrix},$$

and the three sources respectively have 73.0%, 69.3% and 73.3% of zeroes. The results of training and testing the usual six models are contained in Table A2. We get that all our models systematically out-performs both the Full Lasso and independent Lasso in term of test deviance. While the independent Lasso can assign different parameter values in each sources, it does not benefit from the sharing of information between sources and is thus more prone to over-fit. This is what we observe through the poor model fit to test data and larger number of variables in the model.

In experiment 6C, rather than considering a linear combination of independent random variables, we consider a product:

$$\begin{aligned} Y^{(1)} &= \widetilde{Y}^{(1)}\widetilde{Y}^{(4)}\widetilde{Y}^{(6)}, \\ Y^{(2)} &= \widetilde{Y}^{(2)}\widetilde{Y}^{(4)}\widetilde{Y}^{(5)}, \\ Y^{(3)} &= \widetilde{Y}^{(3)}\widetilde{Y}^{(5)}\widetilde{Y}^{(6)}. \end{aligned}$$

This new construction allows us to compute the true generating coefficients in each task and to produce the L_2-loss measure of performance. Indeed, we find that they are given by the sub of the

(a) Setting 6B: Mean (standard error)

	Full Lasso	Ind. Lasso	L_1/L_∞	a-L_1/L_∞	L_1/L_2	a-L_1/L_2
Test dev.	39.58 (0.80)	43.13 (1.81)	36.22 (0.78)	31.69 (0.53)	36.55 (0.86)	31.51 (0.60)
Size	7.48 (0.23)	16.55 (0.49)	10.63 (0.37)	5.62 (0.13)	10.29 (0.38)	5.57 (0.13)
Accuracy	94.80 (0.44)	76.90 (0.98)	88.70 (0.73)	98.72 (0.27)	89.34 (0.76)	98.74 (0.25)
Precision	71.28 (1.83)	32.72 (0.96)	51.53 (1.47)	91.96 (1.42)	53.85 (1.67)	92.23 (1.38)
Recall	98.80 (0.48)	100.00 (0.00)	99.80 (0.20)	99.80 (0.20)	99.60 (0.28)	99.40 (0.34)

(b) Setting 6B: Mean rank (nb. times best)

	Full Lasso	Ind. Lasso	L_1/L_∞	a-L_1/L_∞	L_1/L_2	a-L_1/L_2
Test dev.	5.10 (0)	5.33 (0)	3.67 (2)	1.67 (39)	3.76 (0)	1.47 (59)
Size	2.79 (21)	5.86 (0)	4.21 (0)	1.32 (78)	3.99 (2)	1.25 (82)
Accuracy	2.82 (20)	5.86 (0)	4.20 (0)	1.28 (82)	4.00 (2)	1.25 (82)
Precision	2.82 (20)	5.86 (0)	4.21 (0)	1.29 (81)	4.01 (2)	1.24 (83)
Recall	1.25 (94)	1.00 (100)	1.03 (99)	1.04 (99)	1.06 (98)	1.11 (97)

Table A2: *Results from Setting 6B with 100 replications. Part (a) shows the mean values of the statistics and their standard errors in parentheses. Part (b) shows the mean rank across the six models and, in parentheses, the number of times the model is best.*

coefficients of the independent random variables of the product. For example,

$$\mathbb{E}\left\{Y^{(1)}\right\} = \mu^{(1)}\mu^{(4)}\mu^{(6)} = \exp\left\{\mathbf{x}\left(\boldsymbol{\beta}^{(1)} + \boldsymbol{\beta}^{(4)} + \boldsymbol{\beta}^{(6)}\right)\right\}.$$

Hence the true coefficients is the product between the matrix of independent coefficients and the structure matrix A:

$$\boldsymbol{\beta}^{\text{true}} = \boldsymbol{\beta}A^\top = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ \mathbf{0}_{45} & \mathbf{0}_{45} & \mathbf{0}_{45} & \mathbf{0}_{45} & \mathbf{0}_{45} & \mathbf{0}_{45} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 3 & 2 \\ 3 & 2 & 3 \\ -2 & -3 & -3 \\ 0 & -1 & -1 \\ -1 & -1 & 0 \\ \mathbf{0}_{45} & \mathbf{0}_{45} & \mathbf{0}_{45} \end{bmatrix}.$$

In this setting, pairs of responses are often zero at the same: e.g. $\widetilde{Y}^{(4)} = 0$ implies $Y^{(1)} = 0$ and $Y^{(2)} = 0$. Upon generating 100 replications of the experiment with $\phi = 10$, we obtain the following

(a) Setting 6C: Mean (standard error)

	Full Lasso	Ind. Lasso	L_1/L_∞	a-L_1/L_∞	L_1/L_2	a-L_1/L_2
Test dev.	7.44 (3.86)	8.10 (1.29)	6.13 (1.84)	2.99 (0.88)	7.11 (2.00)	3.45 (1.24)
Size	7.91 (0.31)	17.50 (0.64)	9.97 (0.54)	5.60 (0.32)	9.99 (0.46)	5.08 (0.24)
Accuracy	89.34 (0.51)	72.80 (1.16)	86.38 (0.92)	94.88 (0.50)	86.50 (0.73)	95.80 (0.36)
Precision	54.26 (1.93)	28.55 (1.01)	49.90 (2.01)	83.26 (2.20)	48.59 (1.90)	86.32 (1.87)
Recall	75.80 (1.49)	89.00 (1.25)	81.60 (1.63)	80.40 (1.63)	82.40 (1.74)	79.80 (1.65)
L2 loss	4.32 (0.08)	5.09 (0.10)	4.32 (0.08)	2.86 (0.08)	4.38 (0.10)	2.86 (0.09)

(b) Setting 6C: Mean rank (nb. times best)

	Full Lasso	Ind. Lasso	L_1/L_∞	a-L_1/L_∞	L_1/L_2	a-L_1/L_2
Test dev.	3.68 (6)	5.39 (1)	4.21 (2)	1.77 (39)	4.27 (3)	1.68 (49)
Size	3.19 (17)	5.66 (0)	3.84 (9)	1.69 (60)	3.88 (6)	1.42 (68)
Accuracy	3.43 (9)	5.68 (0)	3.89 (4)	1.52 (69)	3.84 (5)	1.33 (75)
Precision	3.42 (13)	5.71 (0)	3.93 (6)	1.54 (70)	3.82 (7)	1.29 (77)
Recall	2.95 (36)	1.32 (84)	1.99 (57)	2.16 (54)	1.88 (60)	2.13 (50)
L2 loss	4.05 (0)	5.42 (0)	4.12 (1)	1.60 (54)	4.19 (1)	1.62 (44)

Table A3: *Results from Setting 6C with 100 replications. Part (a) shows the mean values of the statistics and their standard errors in parentheses. Part (b) shows the mean rank across the six models and, in parentheses, the number of times the model is best.*

empirical correlation matrix,

$$\begin{bmatrix} 1.00 & 0.35 & 0.38 \\ 0.35 & 1.00 & 0.41 \\ 0.38 & 0.41 & 1.00 \end{bmatrix},$$

and the three sources all have 92.9% of zeroes. Table A3 contain the results for the six models. All versions of our algorithm significantly produce better test deviance and the two adaptive versions clearly beats all other models. Also, the adaptive versions have smaller models and therefore much improved selection accuracy. Finally, the estimated coefficients by the adaptive algorithms are much closer to the truth.