Modern Statistical Methods
Example sheet 3

1. $Y = X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbb{1}$ $\qquad Q_\lambda(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$

As $\hat\beta$ minimises $Q_\lambda(\beta)$,

$$Q_\lambda(\hat\beta) \le Q_\lambda((1-t)\hat\beta + t\beta^0) \qquad \text{for } t \in [0,1]$$

$$\Rightarrow \frac{1}{2n}\|X(\beta^0-\hat\beta)\|_2^2 + \frac{1}{n}\varepsilon^T X(\beta^0-\hat\beta) + \frac{1}{2n}\|\varepsilon - \bar\varepsilon\mathbb{1}\|_2^2 + \lambda\|\hat\beta\|_1$$

$$\le \frac{1}{2n}\|(1-t)[X(\beta^0-\hat\beta) \cancel{+(\varepsilon - \bar\varepsilon\mathbb{1})}] + t(\varepsilon-\bar\varepsilon\mathbb{1})\|_2^2 + \lambda\|(1-t)\hat\beta + t\beta^0\|_1$$

$$= \frac{1}{2n}(1-t)^2\|X(\beta^0-\hat\beta)\|_2^2 + \frac{1}{n}(1-t)\varepsilon^T X(\beta^0-\hat\beta) + \frac{1}{2n}\|\varepsilon - \bar\varepsilon\mathbb{1}\|_2^2 + \lambda\|(1-t)\hat\beta + t\beta^0\|_1$$

$$\Rightarrow \frac{1-(1-t)^2}{2n}\|X(\beta^0-\hat\beta)\|_2^2 \le t\frac{1}{n}\varepsilon^T X(\hat\beta-\beta^0) + \lambda\underbrace{\|(1-t)\hat\beta + t\beta^0\|_1}_{\le (1-t)\|\hat\beta\|_1 + t\|\beta^0\|_1} - \lambda\|\hat\beta\|_1$$

$$\le t\left[\frac{1}{n}\varepsilon^T X(\hat\beta-\beta^0) + \lambda\|\beta^0\|_1 - \lambda\|\hat\beta\|_1\right]$$

$$\Rightarrow \frac{1}{n}\|X(\beta^0-\hat\beta)\|_2^2 \le \left[\frac{1}{n}\varepsilon^T X(\hat\beta-\beta^0) + \lambda\|\beta^0\|_1 - \lambda\|\hat\beta\|_1\right]\underbrace{\frac{2t}{1-(1-t)^2}}_{=\frac{2t}{2t-t^2} = \frac{2}{2-t}} \qquad \text{for } t\in(0,1]$$

Hence, taking $\lim_{t\to 0}$ gives the expected result. □

2. Start with improved inequality

$$\frac{1}{n}\|X(\hat\beta-\beta^0)\|_2^2 + \lambda\|\hat\beta\|_1 \le \frac{1}{n}\varepsilon^T X(\hat\beta-\beta^0) + \lambda\|\beta^0\|_1$$

$$\le \frac{1}{n}\|\varepsilon^T X\|_\infty\|\hat\beta-\beta^0\|_1 + \lambda\|\beta^0\|_1 \qquad \text{Hölder's inequality}$$

$$\Omega = \{2\|X^T\varepsilon\|_\infty/n \le \lambda\}, \quad P(\Omega) = 1 - 2p^{-(A^2/8-1)}$$

$$\Rightarrow \frac{1}{n}\|X(\hat\beta-\beta^0)\|_2^2 + \lambda\|\hat\beta\|_1 \le \frac{\lambda}{2}\|\hat\beta-\beta^0\|_1 + \lambda\|\beta^0\|_1 \quad -(*)$$

$$\frac{2}{n}\|X(\hat\beta-\beta^0)\|_2^2 \le 3\lambda\|\hat\beta-\beta^0\|_1 \qquad \text{as } \|\beta^0\|_1 - \|\hat\beta\|_1 \le \|\beta^0-\hat\beta\|_1$$

Write $a \equiv \frac{2}{n\lambda}\|X(\hat\beta-\beta^0)\|_2^2$, from $(*)$

$$a + 2(\|\hat\beta_N\|_1 + \|\hat\beta_S\|_1) \le \|\hat\beta_S - \beta^0_S\|_1 + \|\hat\beta_N - \cancelto{0}{\beta_N}\|_1 + 2\|\beta^0_S\|_1 + 2\|\cancelto{0}{\beta^0_N}\|_1 \quad \text{by definition}$$

$$\le \|\hat\beta_S - \beta^0_S\|_1 + \|\hat\beta_N\|_1 + 2\|\beta^0_S\|_1$$

$$a + \|\hat\beta_N\|_1 \le \|\hat\beta_S - \beta^0_S\|_1 + 2\|\beta^0_S\|_1 - 2\|\hat\beta_S\|_1 \le 3\|\beta^0_S - \hat\beta_S\|_1$$

$$\Rightarrow \frac{2}{n}\|X(\hat\beta-\beta^0)\|_2^2 \le 3\lambda\|\beta^0_S - \hat\beta_S\|_1 \quad \text{as } \|\hat\beta_N\|_1 \ge 0$$

Compatibility condition for $\beta = \beta^0 - \hat{\beta}$,

$$\frac{1}{n}\|X(\hat{\beta}-\beta^0)\|_2^2 \geq \frac{\phi^2}{s}\|\hat{\beta}_S - \beta^0_S\|_1^2$$

$$\Rightarrow \frac{2}{n}\|X(\hat{\beta}-\beta^0)\|_2^2 \leq 3\lambda\|\hat{\beta}_S - \beta^0_S\|_1 \leq \frac{3\lambda}{\phi}\sqrt{\frac{s}{n}}\|X(\hat{\beta}-\beta^0)\|_2$$

$$\therefore \frac{1}{n}\|X(\hat{\beta}-\beta^0)\|_2^2 \leq \frac{9}{4}\frac{\lambda^2 s}{\phi^2} = \frac{9A^2\log(p)}{4\phi^2}\frac{\sigma^2 s}{n} \qquad \square$$

3. $Y = X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbb{1}$ $\quad S = \{k : \beta^0_k \neq 0\}$ $\quad N = \{1, \dots, p\} \setminus S$

$|S| = s$ $\quad S = \{1, \dots, s\}$ $\quad \text{rank}(X_S) = s$

Write $\hat{\beta}^L_\lambda$ as $\hat{\beta}$. KKT conditions

$$\frac{1}{n}X^T(Y - X\hat{\beta}) = \lambda\hat{\nu} \qquad \text{for } \|\hat{\nu}\|_\infty \leq 1, \ \hat{\nu}_{\hat{S}} = \text{sgn}(\hat{\beta}_{\hat{S}}) \text{ with } \hat{S} = \{k : \hat{\beta}_k \neq 0\}$$

$$\Rightarrow \frac{1}{n}X^TX(\beta^0 - \hat{\beta}) + \frac{1}{n}X^T(\varepsilon - \bar{\varepsilon}\mathbb{1}) = \frac{1}{n}X^TX(\beta^0 - \hat{\beta}) + \frac{1}{n}X^T\varepsilon = \lambda\hat{\nu}$$

Expand this

$$\frac{1}{n}\begin{pmatrix} X_S^T X_S & X_S^T X_N \\ X_N^T X_S & X_N^T X_N \end{pmatrix}\begin{pmatrix} \beta^0_S - \hat{\beta}_S \\ -\hat{\beta}_N \end{pmatrix} = \begin{pmatrix} \lambda\hat{\nu}_S - \frac{1}{n}X_S^T\varepsilon \\ \lambda\hat{\nu}_N - \frac{1}{n}X_N^T\varepsilon \end{pmatrix}$$

⌈ If $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^0)$, $\hat{\nu}_S = \text{sgn}(\beta^0_S)$ and $\hat{\beta}_N = 0$. Hence,

$$\frac{1}{n}X_S^TX_S(\beta^0_S - \hat{\beta}_S) = \lambda\,\text{sgn}(\beta^0_S) - \frac{1}{n}X_S^T\varepsilon \qquad \text{for top}$$

$$\Rightarrow \beta^0_S - \hat{\beta}_S = \left(\frac{1}{n}X_S^TX_S\right)^{-1}\left(\lambda\,\text{sgn}(\beta^0_S) - \frac{1}{n}X_S^T\varepsilon\right)$$

$$\frac{1}{n}X_N^TX_S\left(\frac{1}{n}X_S^TX_S\right)^{-1}\left(\lambda\,\text{sgn}(\beta^0_S) - \frac{1}{n}X_S^T\varepsilon\right) = \lambda\hat{\nu}_N - \frac{1}{n}X_N^T\varepsilon$$ ⌋

Claim:

$$(\hat{\beta}_S, \hat{\beta}_N) = \left(\beta^0_S - \left(\tfrac{1}{n}X_S^TX_S\right)^{-1}\left(\lambda\,\text{sgn}(\beta^0_S) - \tfrac{1}{n}X_S^T\varepsilon\right), 0\right)$$

$$(\hat{\nu}_S, \hat{\nu}_N) = \left(\text{sgn}(\beta^0_S), \frac{1}{\lambda}X_N^T\left\{\tfrac{1}{n}\varepsilon + X_S(X_S^TX_S)^{-1}\left(\lambda\,\text{sgn}(\beta^0_S) - \tfrac{1}{n}X_S^T\varepsilon\right)\right\}\right)$$

is a Lasso solution.

By above, it satisfies KKT conditions. But need to check $\|\hat{\nu}\|_\infty \leq 1$,

$$\hat{\nu}_N = X_N^T\left\{\frac{1}{n\lambda}\varepsilon + X_S(X_S^TX_S)^{-1}\left(\text{sgn}(\beta^0_S) - \frac{1}{n\lambda}X_S^T\varepsilon\right)\right\}$$

Can check that $\|\hat{\nu}_N\|_\infty \leq 1$ provided the condition.

$$\cancel{\|\hat{\nu}_N\|_\infty \leq \|X_N^TX_S(X_S^TX_S)^{-1}\text{sgn}(\beta^0_S)\|_\infty + \frac{1}{\lambda}\|X_N^T(I - X_S(X_S^TX_S)^{-1}X_S^T)\varepsilon\|_\infty/n}$$

$$\cancel{\leq \frac{\lambda - \lambda_0}{\lambda + \lambda_0} + \frac{1}{\lambda}\|X_N^T(I - X_S(X_S^TX_S)^{-1}X_S^T)\varepsilon\|_\infty/n}$$ ~~by condition~~

Finally, $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^0) \Leftarrow \|\{(\frac{1}{n}X_S^TX_S)^{-1}\}_k(\lambda\,\text{sgn}(\beta^0_S) - \frac{1}{n}X_S^T\varepsilon)\|_1 \leq \|\{(\frac{1}{n}X_S^TX_S)^{-1}\}_k\|_1\|\lambda\,\text{sgn}(\beta^0_S) - \frac{1}{n}X_S^T\varepsilon\|_\infty$

$$\leq \|\{(\tfrac{1}{n}X_S^TX_S)^{-1}\}_k\|_1(\lambda + \underbrace{\|X_S^T\varepsilon\|_\infty/n}_{\leq\lambda}) \leq |\beta^0_k| \quad \text{by condition}$$

4. Group Lasso

$$Q_\lambda(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{q} m_j \|\beta_{G_j}\|_2$$

$\bigcup_{k=1}^{q} G_k = \{1, \dots, p\}$

$G_j \cap G_k = \phi$ for $j \neq k$

For $i \in G_j$,

$$(\partial \|\beta_{G_j}\|_2)_i = \frac{\beta_i}{\|\beta_{G_j}\|_2}$$

$$\Rightarrow \partial Q_\lambda(\beta) = -\frac{1}{n} X^T(Y - X\beta) + \lambda \tilde{\beta}$$ where $\tilde{\beta}_i = m_j \beta_i / \|\beta_{G_j}\|_2$ for $i \in G_j$

$\therefore \frac{1}{n} X^T(Y - X\beta) = \lambda \tilde{\beta}$ by KKT conditions.

5. (a) Lasso:

$$Q(\beta, z) = \frac{1}{2n}\|Y - z\|_2^2 + \lambda\|\beta\|_1 \quad \text{subject to} \quad q(\beta, z) = z - X\beta = 0$$

Lagrangian for this problem

$$L(\beta, z, \theta) = Q(\beta, z) + \theta^T q(\beta, z) = \frac{1}{2n}\|Y - z\|_2^2 + \lambda\|\beta\|_1 + \theta^T(z - X\beta)$$

Note that

$$\inf_{(\beta,z)\in\mathbb{R}^p\times\mathbb{R}^n} L(\beta, z, \theta) \le \inf_{(\beta,z)\in\mathbb{R}^p\times\mathbb{R}^n : q(\beta,z)=0} L(\beta, z, \theta) = c^*$$

where c^* is the optimal value of the original problem

$$c^* = Q(\hat{\beta}_\lambda^L, \hat{z}) = \frac{1}{2n}\|Y - X\hat{\beta}_\lambda^L\|_2^2 + \lambda\|\hat{\beta}_\lambda^L\|_1$$

$\forall\, \theta \in \mathbb{R}^n$

Consider

$$\max_{\theta : \|X^T\theta\|_\infty \le \lambda} \inf_{(\beta,z)\in\mathbb{R}^p\times\mathbb{R}^n} \frac{1}{2n}\|Y - z\|_2^2 + \lambda\|\beta\|_1 + \theta^T(z - X\beta)$$

$$\cancel{\max_{\theta : \|X^T\theta\|_\infty \le \lambda} \inf_{(\beta,z):q(\beta,z)=0} \frac{1}{2n}\|Y - z\|_2^2 + \lambda\|\beta\|_1 + \theta^T(z - X\beta) \le \max_{\theta:}}$$

$$\cancel{\le \max_{\theta : \|X^T\theta\|_\infty \le \lambda} \inf_{(\beta,z):q(\beta,z)=0} \frac{1}{2n}\|Y - z\|_2^2 + \lambda\|\beta\|_1}$$

$$\ge \max_{\theta : \|X^T\theta\|_\infty \le \lambda} \inf_{(\beta,z)} \frac{1}{2n}\|Y - z\|_2^2 + \lambda\|\beta\|_1 + \theta^T z - \|X^T\theta\|_\infty \|\beta\|_1$$ by Hölder

$$\ge \max_{\theta : \|X^T\theta\|_\infty \le \lambda} \inf_{(\beta,z)} \frac{1}{2n}\|Y - z\|_2^2 + \theta^T z$$ using the condition $\|X^T\theta\|_\infty \le \lambda$.

$$= \max_{\theta : \|X^T\theta\|_\infty \le \lambda} \inf_{z\in\mathbb{R}^n} \frac{1}{2n}\|Y - z\|_2^2 + \theta^T z$$

Now

$$\inf_{z\in\mathbb{R}^n}\left(\frac{1}{2n}\|Y-z\|^2+\theta^T z\right)=\inf_{z\in\mathbb{R}^n}\left(\frac{1}{2n}\|Y\|_2^2-\frac{1}{n}Y^Tz+\frac{1}{2n}\|z\|_2^2+\theta^Tz\right)$$

$$=\frac{1}{2n}\|Y\|_2^2+\inf_{z\in\mathbb{R}^n}\left(\frac{1}{2n}\|z\|_2^2-\frac{1}{n}(Y-n\theta)^Tz\right)$$

$$=\frac{1}{2n}\|Y\|_2^2+\inf_{z\in\mathbb{R}^n}\left(\frac{1}{2n}\|z\|_2^2-\frac{1}{n}(Y-n\theta)^Tz+\frac{1}{2n}\|Y-n\theta\|_2^2\right)-\frac{1}{2n}\|Y-n\theta\|_2^2$$

$$=\frac{1}{2n}\|Y\|_2^2-\frac{1}{2n}\|Y-n\theta\|_2^2+\inf_{z\in\mathbb{R}^n}\frac{1}{2n}\|Y-n\theta-z\|_2^2$$

$$=\frac{1}{2n}\|Y\|_2^2-\frac{1}{2n}\|Y-n\theta\|_2^2 \qquad \text{by choosing } z=Y-n\theta$$

Hence,

$$\max_{\theta:\|X^T\theta\|_\infty\le\lambda}\frac{1}{2n}\|Y\|_2^2-\frac{1}{2n}\|Y-n\theta\|_2^2\le c^*$$

or equivalently

$$\max_{\theta:\|X^T\theta\|_\infty\le\lambda} G(\theta)\le c^*=\frac{1}{2n}\|Y-X\hat\beta_\lambda^L\|_2^2+\lambda\|\hat\beta_\lambda^L\|_1 \qquad \text{where } G(\theta)=\frac{1}{2n}\|Y\|^2-\frac{1}{2n}\|Y-n\theta\|_2^2$$

Consider $\theta^*=\frac{1}{n}(Y-X\hat\beta_\lambda^L)$,

$\|X^T\theta^*\|_\infty=\|\frac{1}{n}X^T(Y-X\hat\beta_\lambda^L)\|_\infty=\lambda\|\hat\nu\|_\infty\le\lambda$ by KKT conditions on the Lasso soln.

For the equality to hold above, $\max_{\theta:\|X^T\theta\|_\infty\le\lambda}G(\theta)=c^*$, must have $z-X\beta=0$ (Lagrangian method).

$\Rightarrow \theta=\frac{1}{n}(Y-X\beta)$ i.e. $\theta^*=\frac{1}{n}(Y-X\hat\beta_\lambda^L)$ and this satisfies $\|X^T\theta\|_\infty\le\lambda$ as shown above.

Hence, this θ^* is the unique solution because $X\hat\beta_\lambda^L$ is unique.

(b) $\tilde{\theta} : \|X^T\tilde{\theta}\|_\infty \leq \lambda$

$$\max_{\theta : G(\theta) \geq G(\tilde{\theta})} |X_k^T\theta| < \lambda \qquad \Longleftrightarrow \qquad |X_k^T\theta| < \lambda \quad \forall\ \theta : G(\theta) \geq G(\tilde{\theta})$$

Now $\max\limits_{\theta : \|X^T\theta\|_\infty \leq \lambda} G(\theta) \geq G(\tilde{\theta})$ as $\tilde{\theta} : \|X^T\tilde{\theta}\|_\infty \leq \lambda$.

Hence, $|X_k^T\theta^*| < \lambda$ where $\theta^* = \frac{1}{n}(Y - X\hat{\beta}_\lambda^L)$ as shown in part (a).

$\Rightarrow |X_k^T\theta^*| = |\frac{1}{n}X_k^T(Y - X\hat{\beta}_\lambda^L)| = |\lambda\hat{\nu}_k| < \lambda$ by KKT conditions.

Know that $\hat{\nu}_{\hat{S}} = \text{sgn}(\hat{\beta}_{\lambda,\hat{S}}^L)$ where $\hat{S} = \{k : \hat{\beta}_{\lambda,k}^L \neq 0\}$. Thus, if $\hat{\beta}_{\lambda,k}^\lambda \neq 0 \Rightarrow |\lambda\hat{\nu}_k| = \lambda$.

$\therefore \hat{\beta}_{\lambda,k}^L = 0 \quad \square$

$\tilde{\theta} = \frac{1}{n}Y\frac{\lambda}{\lambda_{max}}$, $\lambda_{max} = \frac{1}{n}\|X^TY\|_\infty$

$\Rightarrow \|X^T\tilde{\theta}\|_\infty = \frac{1}{n}\frac{\lambda}{\lambda_{max}}\|X^TY\|_\infty = \lambda \leq \lambda$

Hence, $G(\tilde{\theta}) \leq \max\limits_{\theta : \|X^T\theta\|_\infty \leq \lambda} G(\theta)$.

If $G(\theta) \geq G(\tilde{\theta})$,

$$\frac{1}{2n}\|Y\|_2^2 - \frac{1}{2n}\|Y - n\theta\|_2^2 \geq \frac{1}{2n}\|Y\|_2^2 - \frac{1}{2n}\|Y\|_2^2\left(1 - \frac{\lambda}{\lambda_{max}}\right)^2$$

$$\Rightarrow \frac{1}{n}\|Y\|_2^2\left(1 - \frac{\lambda}{\lambda_{max}}\right)^2 \geq \frac{1}{n}\|Y - n\theta\|_2^2$$

In the case $\theta = \theta^* = \frac{1}{n}(Y - X\hat{\beta}_\lambda^L)$,

$$\frac{1}{\sqrt{n}}\|Y\|_2\left(1 - \frac{\lambda}{\lambda_{max}}\right) \geq \frac{1}{\sqrt{n}}\|X\hat{\beta}_\lambda^L\|_2 \qquad \text{if} \quad \lambda \leq \lambda_{max}.$$

Now consider the LHS of KKT conditions

$$\begin{aligned}\frac{1}{n}|X_k^T(Y - X\hat{\beta}_\lambda^L)| &\leq \frac{1}{n}|X_k^TY| + \frac{1}{n}|X_k^TX\hat{\beta}_\lambda^L| \\ &\leq \frac{1}{n}|X_k^TY| + \frac{1}{n}\underbrace{\|X_k\|_2}_{=\sqrt{n} \text{ by scaling}}\|X\hat{\beta}_\lambda^L\|_2 \quad \text{by Hölder's inequality} \\ &\leq \frac{1}{n}|X_k^TY| + \frac{1}{\sqrt{n}}\|Y\|_2\left(1 - \frac{\lambda}{\lambda_{max}}\right) \quad \text{from above.}\end{aligned}$$

Hence, if $\frac{1}{n}|X_k^TY| < \lambda - \frac{1}{\sqrt{n}}\|Y\|_2\left(1 - \frac{\lambda}{\lambda_{max}}\right)$,

$\frac{1}{n}|X_k^T(Y - X\hat{\beta}_\lambda^L)| < \lambda \Rightarrow \hat{\beta}_{\lambda,k}^L = 0$.

Finally, if $\lambda > \lambda_{max}$, $\hat{\beta}_\lambda^L = 0$. $\square$

6. $\hat{E}_\lambda = \{k : \frac{1}{n}|X_k^T(Y - X\hat\beta^L_\lambda)| = \lambda\}$ $\operatorname{rank}(X_{\hat E_\lambda}) = |\hat E_\lambda|$ $\forall\ \lambda > 0$

$\operatorname{sgn}(\hat\beta^L_{\lambda_1}) = \operatorname{sgn}(\hat\beta^L_{\lambda_2})$

KKT conditions:

$$\frac{1}{n}|X_k^T(Y - X\hat\beta^L_{\lambda_1})| \begin{cases} = \lambda_1 & \text{if } k \in \hat E_{\lambda_1} \\ \le \lambda_1 & \text{otherwise} \end{cases}$$

similarly for λ_2.

Consider

$$\frac{1}{n} X_k^T(Y - X(t\hat\beta^L_{\lambda_1} + (1-t)\hat\beta^L_{\lambda_2})) = \frac{1}{n}\left[t X_k^T(Y - X\hat\beta^L_{\lambda_1}) + (1-t)X_k^T(Y - X\hat\beta^L_{\lambda_2})\right]$$

$$= t\lambda_1 \hat v_{\lambda_1,k} + (1-t)\lambda_2 \hat v_{\lambda_2,k}$$

If $k \in \hat E_{\lambda_1} \cap \hat E_{\lambda_2}$

$$\frac{1}{n} X_k^T(Y - X(t\hat\beta^L_{\lambda_1} + (1-t)\hat\beta^L_{\lambda_2})) = t\lambda_1 \operatorname{sgn}(\hat\beta^L_{\lambda_1,k}) + (1-t)\lambda_2 \operatorname{sgn}(\hat\beta^L_{\lambda_2,k})$$

$$= [t\lambda_1 + (1+t)\lambda_2]\operatorname{sgn}(\hat\beta^L_k) \qquad (k \in \hat E_{t\lambda_1 + (1+t)\lambda_2})$$

Otherwise,

$$\frac{1}{n}|X_k^T(Y - X(t\hat\beta^L_{\lambda_1} + (1-t)\hat\beta^L_{\lambda_2}))| = |t\lambda_1 \hat v_{\lambda_1,k} + (1-t)\lambda_2 \hat v_{\lambda_2,k}|$$

$$\le t\lambda_1|\hat v_{\lambda_1,k}| + (1-t)\lambda_2|\hat v_{\lambda_2,k}| < t\lambda_1 + (1-t)\lambda_2 \qquad (k \notin \hat E_{t\lambda_1 + (1+t)\lambda_2})$$

Hence

$t\hat\beta^L_{\lambda_1} + (1-t)\hat\beta^L_{\lambda_2} = \hat\beta^L_{t\lambda_1 + (1-t)\lambda_2}$ as it satisfies the corresponding KKT conditions. □

Thus, $\operatorname{sgn}(\hat\beta^L_\lambda)$ constant along $\lambda_1 \to \lambda_2$, whereas $\hat E_\lambda$ may change at either end point. The solution path $\lambda \mapsto \hat\beta^L_\lambda$ is then piecewise linear between points where $\operatorname{sgn}(\hat\beta^L_\lambda)$ changes, the number of knots is finite as $\operatorname{sgn}(\hat\beta^L_\lambda)$ can only take finitely many values and any two points with an equal value must be part of the same linear piece.

7. $Q_{\lambda,\alpha}(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2/2)$ $\alpha \in [0,1]$

(a) Suppose $\beta^{*(1)}$ and $\beta^{*(2)}$ are two minimisers of $Q_{\lambda,\alpha}$. Note that $\|\cdot\|_2^2$ is strictly convex, $\|\cdot\|_1$ is convex. $c^* = Q_{\lambda\alpha}(\beta^{*(1)}) = Q_{\lambda,\alpha}(\beta^{*(2)})$

$\|Y - tX\beta^{*(1)} - (1+t)X\beta^{*(2)}\|_2^2 \le t\|Y - X\beta^{*(1)}\|_2^2 + (1+t)\|Y - X\beta^{*(2)}\|_2^2$

with equality iff $X\beta^{*(1)} = X\beta^{*(2)}$,

$\|t\beta^{*(1)} + (1-t)\beta^{*(2)}\|_1 \le t\|\beta^{*(1)}\|_1 + (1-t)\|\beta^{*(2)}\|_1$,

$\|t\beta^{*(1)} + (1-t)\beta^{*(2)}\|_2^2 \le t\|\beta^{*(1)}\|_2^2 + (1-t)\|\beta^{*(2)}\|_2^2$,

with equality iff $\beta^{*(1)} = \beta^{*(2)}$

$\Rightarrow c^* \le Q_{\lambda\alpha}(t\beta^{*(1)} + (1-t)\beta^{*(2)}) = \frac{1}{2n}\|Y - tX\beta^{*(1)} - (1-t)X\beta^{*(2)}\|_2^2 + \lambda(\alpha\|t\beta^{*(1)} + (1-t)\beta^{*(2)}\|_1 + (1-\alpha)\|t\beta^{*(1)} + (1-t)\beta^{*(2)}\|_2^2/2) \le t\{\frac{1}{2n}\|Y - X\beta^{*(1)}\|_2^2 + \lambda(\alpha\|\beta^{*(1)}\|_1 + (1-\alpha)\|\beta^{*(1)}\|_2^2/2)\} + (1-t)\{\frac{1}{2n}\|Y - X\beta^{*(2)}\|_2^2 + \lambda(\alpha\|\beta^{*(2)}\|_1 + (1-\alpha)\|\beta^{*(2)}\|_2^2/2)\}$

$$= t\, Q_{\lambda,\alpha}(\beta^{*(1)}) + (1-t)\, Q_{\lambda,\alpha}(\beta^{*(2)}) = c^*$$

Hence, equality must prevail at each step and as $\alpha < 1$ then from the $\|\beta\|_2^2$ term get $\beta^{*(1)} = \beta^{*(2)}$. β^* is unique. □

If X has two columns X_j and X_k identical, suppose $\beta_k^* \neq \beta_j^*$. Then by switching ~~columns~~ labels $j \leftrightarrow k$ get the same X matrix but $\beta_j^{*\prime} = \beta_k^*$ and $\beta_k^{*\prime} = \beta_j^*$ where prime denotes β^* values after switching. However, $\beta_j^{*\prime} \neq \beta_j^*$ and hence have two different solutions β^* but β^* was shown to be unique. #

(b) $$\frac{\partial Q_{\lambda,\alpha}(\beta)}{\partial \beta_k} = -\frac{1}{n} X_k^T (Y - X\beta) + \lambda\alpha\, \mathrm{sgn}(\beta_k) + \lambda(1-\alpha)\beta_k$$

Coordinate descent

$$\hat\beta_k^{(m)} = \underset{\beta_k \in \mathbb{R}}{\arg\min}\; Q_{\lambda,\alpha}(\hat\beta_1^{(m)}, \dots, \hat\beta_{k-1}^{(m)}, \beta_k, \hat\beta_{k+1}^{(m-1)}, \dots, \hat\beta_p^{(m-1)})$$

$$= \underset{\beta_k \in \mathbb{R}}{\arg\min}\; Q_{\lambda,\alpha}(\hat\beta_A^{(m)}, \beta_k, \hat\beta_B^{(m-1)})$$

At the minimum must have $\partial_{\beta_k} Q_{\lambda,\alpha} = 0$ and $X\beta \to X_A \hat\beta_A^{(m)} + X_B \hat\beta_B^{(m-1)}$

$$-\frac{1}{n} X_k^T (Y - X_A \hat\beta_A^{(m)} - X_B \hat\beta_B^{(m-1)}) + \lambda\alpha\, \mathrm{sgn}(\beta_k) + \lambda(1-\alpha)\beta_k = 0$$

$- X_k \hat\beta_k^{(m)}$

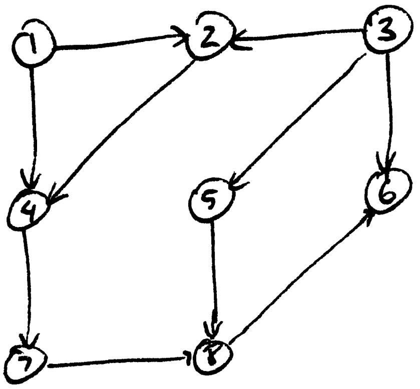
$$\Rightarrow \frac{1}{n} \underbrace{X_k^T X_k}_{\|X_k\|_2^2 = n \text{ by scaling}} \hat\beta_k^{(m)} + \lambda(1-\alpha)\hat\beta_k^{(m)} = \frac{1}{n} X_k^T (Y - X_A \hat\beta_A^{(m)} - X_B \hat\beta_B^{(m-1)}) - \lambda\alpha\, \mathrm{sgn}(\hat\beta_k^{(m)})$$

$$\Rightarrow \hat\beta_k^{(m)} = \frac{n^{-1} X_k^T (Y - X_A \hat\beta_A^{(m)} - X_B \hat\beta_B^{(m-1)}) - \lambda\alpha\, \mathrm{sgn}(\beta_k)}{1 + \lambda(1-\alpha)}$$

By considering $\hat\beta_k^{(m)} > 0$ and $\hat\beta_k^{(m)} < 0$,

$$\hat\beta_k^{(m)} = \frac{S_{\lambda\alpha}(n^{-1} X_k^T (Y - X_A \hat\beta_A^{(m)} - X_B \hat\beta_B^{(m-1)}))}{1 + \lambda(1-\alpha)} \quad \text{with } S_t(u) = \mathrm{sgn}(u)(|u| - t)_+ .$$ □

8.



DAG $\mathcal{G}$

(a) $de(3) = \{2, 4, 7, 8, 5, 6\} = \{1, \dots, 8\} \setminus \{1, 3\}$

(b) Sets of variables that d-separate 1 and 3:

Paths from 1 to 3:

i) $1 \to 2 \leftarrow 3$

ii) $1 \to 2 \to 4 \to 7 \to 8 \leftarrow 5 \leftarrow 3$

iii) $1 \to 2 \to 4 \to 7 \to 8 \to 6 \leftarrow 3$

iv) $1 \to 4 \to 7 \to 8 \leftarrow 5 \leftarrow 3$

v) $1 \to 4 \to 7 \to 8 \to 6 \leftarrow 3$

Sets that block path (i):

$\{5\}, \phi$ as $1 \to 2 \leftarrow 3$ is a collider and hence any set that contains 2 or any $de(2)$ would not block

$\{5\}$ also blocks all paths containing 5 as 5 is not a collider, ϕ blocks these as 8 is a collider

They also block the other two as 6 is a collider.

$\therefore S = \{5\}, \phi$

(c) d-separate $\{1, 4\}$ and 6:

Again, any path containing $1 \to 2 \leftarrow 3$ is only blocked by $\{5\}, \phi$.

However, $(1 \to) 4 \to 7 \to 8 \to 6$ is blocked by nothing.

$\therefore S \in \phi$, no S d-separates $\{1, 4\}$ and 6.

(d) All v-structures:

$\{1, 2, 3\}$, $\{7, 8, 5\}$, $\{3, 6, 8\}$

9. $Z = (Z_1, \dots, Z_p)^T \in \{0,1\}^p$

$$\mathbb{P}(Z_1 = z_1, \dots, Z_p = z_p) = \exp\left(\Theta_{00} + \sum_{k=1}^{p} \Theta_{0k} z_k + \sum_{k=1}^{p} \sum_{j=1}^{k-1} \Theta_{jk} z_j z_k - \Phi(\Theta)\right)$$

$$\begin{aligned} \text{logit}(\mathbb{P}(Z_k = 1 \mid Z_{-k} = z_{-k})) &= \log(\mathbb{P}(Z_k = 1 \mid Z_{-k} = z_{-k})) - \log(\mathbb{P}(Z_k = 0 \mid Z_{-k} = z_{-k})) \\ &= \log(\mathbb{P}(Z_k = 1, Z_{-k} = z_{-k})) - \log(\mathbb{P}(Z_k = 0, Z_{-k} = z_{-k})) \\ &= \Theta_{0k} + \sum_{j : j<k} \Theta_{jk} z_j + \sum_{j : j>k} \Theta_{kj} z_j \ . \end{aligned}$$

For $j < k$,

$$Z_j \perp\!\!\!\perp Z_k \mid Z_{-jk} \iff \mathbb{P}(Z_j = 1 \mid Z_{-j}) = \mathbb{P}(Z_j = 1 \mid Z_{-jk})$$

$$\iff \text{logit}(\mathbb{P}(Z_j = 1 \mid Z_{-j})) = \text{logit}(\mathbb{P}(Z_j = 1 \mid Z_{-jk})) \quad \text{as logit is 1-1 in } (0,1).$$

$$\begin{aligned} \text{logit}(\mathbb{P}(Z_j = 1 \mid Z_{-jk})) &= \log(\mathbb{P}(Z_j = 1 \mid Z_{-jk})) - \log(\mathbb{P}(Z_j = 0 \mid Z_{-jk})) \\ &= \log(\mathbb{P}(Z_j = 1, Z_{-jk} = z_{-jk})) - \log(\mathbb{P}(Z_j = 0, Z_{-jk} = z_{-jk})) \\ &= \log\Big(\mathbb{P} \bigcap_{z_k \in \{0,1\}} (Z_j = 1, Z_{-jk} = z_{-jk}, Z_k = z_k)\Big) - \log\Big(\mathbb{P} \bigcap_{z_k \in \{0,1\}} (Z_j = 0, Z_{-jk} = z_{-jk}, Z_k = z_k)\Big) \\ &= \Theta_{0j} + \log\Big(\sum_{z_k \in \{0,1\}} \exp\Big(\sum_{m=1}^{p} \sum_{n=1}^{m-1} \Theta_{nm} z_n z_m\Big)\Big|_{\substack{z_j = 1 \\ z_k = z_k}}\Big) - \log\Big(\sum_{z_k \in \{0,1\}} \exp\Big(\sum_{m=1}^{p} \sum_{n=1}^{m-1} \Theta_{nm} z_n z_m\Big)\Big|_{\substack{z_j = 0 \\ z_k = z_k}}\Big) \\ &= \Theta_{0j} + \sum_{i : i<j} \Theta_{ij} z_i + \sum_{\substack{i : i>j \\ i \neq k}} \Theta_{ji} z_i + \log(1 + \exp \Theta_{jk}) - \log 2 \end{aligned}$$

$$\therefore \ Z_j \perp\!\!\!\perp Z_k \mid Z_{-jk} \iff \Theta_{0j} + \sum_{i : i<j} \Theta_{ij} z_i + \sum_{i : i>j} \Theta_{ji} z_i$$

$$= \Theta_{0j} + \sum_{i : i<j} \Theta_{ij} z_i + \sum_{\substack{i : i>j \\ i \neq k}} \Theta_{ji} z_i + \log(1 + \exp(\Theta_{jk})) - \log 2$$

$$\iff \Theta_{jk} z_k = \log \frac{1 + \exp \Theta_{jk}}{2} \iff \Theta_{jk} = 0 \quad \text{as it must hold for } z_k \in \{0,1\}.$$

Estimate Θ_{jk} from

$$\text{logit}(\mathbb{P}(Z_j = 1 \mid Z_{-j})) - \text{logit}(\mathbb{P}(Z_j = 1 \mid Z_{-jk})) = \Theta_{jk} z_k - \log \frac{1 + \exp \Theta_{jk}}{2}$$

once have probability mass function estimates.

10. $Z \sim N_p(\mu, \Sigma)$ pairwise Markov w.r.t. G

$j, k \in V$, $j \neq k$, $(j,k), (k,j) \notin E \Rightarrow Z_j \perp\!\!\!\perp Z_k \mid Z_{-jk} \Leftrightarrow \Omega_{jk} = 0$