Appendix A Recommendations based on simulation study in Sec. 3

To specify our recommendations we assume the following, elaborated from Subec. 3.1. First, when $p > n$, the instability graphs are more important that the tables because the variable selection is so poor for all methods that we must rely solely on choosing the method that predicts the best (or least bad). Thus, for $n = 40$ we just consider the figures in choosing the best methods. For $n = 75$, unless one specifies a relative weighting among Tot, TP, and FP, there is usually no clearly preferred method. In these cases, we have used the instability curves to select from amongst the numerically best shrinkage methods.

Second, for larger sample sizes, and low to medium sparsity cases, we use both the instability curves and variable selection tables to make recommendations. That is, we choose the methods that are both selecting variables most appropriately and have the lowest instability generally. Again, there are cases where we have to weight Tot, TP, and FP.

Two limitations of our recommendations are that in practice i) the level of sparsity and ii) the variable selection tables cannot be known. It is only the instability curve that is available. Thus, our recommendations are based primarily on the instability curves but taking into account the variable selection tables. Moreover, there are other limitations that a potential user should take into account e.g., our restriction to linear models.

We begin our recommendations for small sample sizes with Table **??** (based on our computations from Subsec. 3.1). Recall that for $n = 40$ no method worked particularly well. However, the overall best performing of these poor methods was EN. RR was not too much worse than EN, and LASSO was better than the adaptive methods, but LASSO is unable to select more than n variables. Hence, in $n < p$ settings with little sparsity, LASSO cannot be expected to perform well.

Sparsity	.1	.5	.9
Ind, Li	EN	EN	EN
Ind, H	EN	EN	EN
Tri	EN	EN	EN
Toe	EN	EN	EN

Table A1 Best performing shrinkage methods for $n = 40$.

Our recommendations for $n = 40$ is the closest scenario to those considered in Wang et al (2020a). In fact, our recommendations for this case agree for the most part. They also conclude there is not a single best method that works best across many data generation scenarios. These authors suggest that often LASSO is a good compromise between other methods, and that suggestion may be reasonable. Although we never found LASSO performing best, it does often perform fairly close to other well performing methods, making it a 'safe' option.

Recommendations for the next sample size, $n = 75$, are given in Table **??** (based on the computations in Subsec. 3.2). EN was typically the preferred method for low to medium sparsity. For high sparsity, all methods besides RR performed well. Interestingly, we found SCAD2 or MCP performed best when there is strong dependence as in the Toeplitz structure.

Sparsity	.1	.5	.9
Ind, Li	EN	EN	**Not**: RR
Ind, H	EN	EN	**Not**: RR
Tri	EN	EN	**Not**: RR
Toe	EN	EN	EN/LASSO

Table A2 Best performing shrinkage methods for $n = 75$.

Table **??** summarizes our recommendations for $n = 150$ (based on the computations in Subsec. 3.3). For low sparsity, LM's were generally good. For medium sparsity, the conclusions did not follow a strong pattern. Several methods performed roughly equally well and tridiagonal was, as in the earlier cases, similar to the independence cases, except here for medium sparsity. With stronger dependence (Toeplitz), the biggest change from the other cases was that LM was the best for low to medium sparsity and SCAD/MCP performed best in the high sparsity case. We also see that generally, as sparsity increases, LM's do relatively worse.

Sparsity	.1	.5	.9
Ind, Li	LM	**Not**: RR	**Not**: RR/LM
Ind, H	LM	ALASSO/AEN	**Not**: RR/LM
Tri	LM/SCAD/MCP	SCAD/ MCP/AEN/ALASSO	**Not**: RR/LM
Toe	LM	LM	SCAD/ MCP

Table A3 Best performing shrinkage methods for $n = 150$.

Table **??** summarizes our recommendations for $n = 500$ (based on the computations in Subsec. 3.4) where it is safe to assume the asymptotic properties have begun to kick in. Thus, we see several methods performing well regardless of whether they have the OP. Indeed, it is very hard to choose a single best method with exception of the Toeplitz case where LM is the best across the board.

Sparsity	.1	.5	.9
Ind, Li	**Not**: RR	Not: RR	**Not**: RR/LM
Ind, H	LM/SCAD/MCP	**Not**: RR	**Not**: RR/LM
Tri	**Not**: RR	**Not**: RR	**Not**: RR/LM
Toe	LM	LM	LM

Table A4 Best performing of shrinkage methods for $n = 500$.

We conclude with some general observations. In the $n > p$ case, LM's often perform best except for high sparsity. Fortunately, at this sparsity level all shrinkage methods except for RR are essentially equivalent. LM's are sometimes are not very good in the heavy tailed case but, again, this mainly occurs for high sparsity as shrinkage methods set zero coefficients to zero faster than LM does. We observe that the higher the sparsity, the better the methods exclude variables that are not relevant and include only the relevant variables. This is generally true regardless of the sample size. That is, we are observing a sort of consistency under increasing sparsity that occurs even for sample sizes that are pre-asymptotic.

That said, as sample size increases, methods with the oracle property do indeed emerge as best, if not uniquely so, regardless of the heaviness of the tails – as long regularity conditions e.g., Conditions 2, 3 and 4 in Subsec. 2 are satisfied.

Appendix B Proofs from Sec. 2

B.1 Penalized Log-Likelihood Case, Sec. 2.1

Theorem B.1 *Suppose Conditions 1–6 are satisfied and suppose* $\sqrt{n}\lambda^*_{\max} \to 0$ *Then, there exists a local minimizer* $\hat{\beta}$ *of* $Q(\beta)$ *such that* $\|\hat{\beta} - \beta_0\| = O_p(n^{-\frac{1}{2}} + \lambda^*_{\max})$.

Proof **Step 1.** We want to show for any $\epsilon > 0$ there exists a large constant C such that

$$P\left\{\inf_{\|u\|=C} Q(\beta_0+\alpha_n u) > Q(\beta_0)\right\} \geq 1-\epsilon, \tag{B1}$$

where $\alpha_n = n^{-\frac{1}{2}} + \lambda^*_{\max}$. Denote

$$\begin{aligned}
D_n(u) &= Q(\beta_0+\alpha_n u) - Q(\beta_0)\\
&= L(\beta_0+\alpha_n u|x^n) + \sum_{j=1}^{p}\lambda_j f_j(\beta_{j0}+\alpha_n u_j) - L(\beta_0|x^n) - n\sum_{j=1}^{p}\lambda_j(f_j(\beta_{j0}))\\
&= L(\beta_0+\alpha_n u|x^n) - L(\beta_0|x^n) + n\sum_{j=1}^{p}\lambda_j[f_j(\beta_{j0}+\alpha_n u_j) - f_j(\beta_{j0})]\\
&= L(\beta_0+\alpha_n u|x^n) - L(\beta_0|x^n) + n\sum_{j=1}^{p_0}\lambda_j[f_j(\beta_{j0}+\alpha_n u_j) - f_j(\beta_{j0})] + n\sum_{j=p_0+1}^{p}\lambda_j[f_j(\beta_{j0}+\alpha_n u_j) - f_j(\beta_{j0})]\\
&= L(\beta_0+\alpha_n u|x^n) - L(\beta_0|x^n) + n\sum_{j=1}^{p_0}\lambda_j[f_j(\beta_{j0}+\alpha_n u_j) - f_j(\beta_{j0})] + n\sum_{j=p_0+1}^{p}\lambda_j f_j(\beta_{j0}+\alpha_n u_j)\\
&\geq L(\beta_0+\alpha_n u|x^n) - L(\beta_0|x^n) + n\sum_{j=1}^{p_0}\lambda_j[f_j(\beta_{j0}+\alpha_n u_j) - f_j(\beta_{j0})]\\
&= L(\beta_0+\alpha_n u|x^n) - L(\beta_0|x^n) + n\sum_{j=1}^{p_0}\lambda_j f_j'(\tilde{\beta}_j)\alpha_n u_j
\end{aligned}$$

where the fifth equality and the inequality are due to condition 5, and $f_j'(\tilde{\beta}_j)\alpha_n u_j$ is the Taylor expansion of the penalty term by Conditions 4 and 6 and where $\tilde{\beta}_j$ is on the line joining β_j to $\beta_j+\alpha_n u$. More formally, $\tilde{\beta}_j \in < \beta_j, \beta_j+\alpha_n u >$. Now define $\lambda^*_{\min} = \min\{\lambda_j | j=1,\ldots,p_0\}$. We have,

$$\begin{aligned}
D_n(u) &\geq L(\beta_0+\alpha_n u|x^n) - L(\beta_0|x^n) + n\alpha_n\sum_{j=1}^{p_0}\lambda_j f_j'(\tilde{\beta}_j)u_j\\
&= L(\beta_0+\alpha_n u|x^n) - L(\beta_0|x^n) + n(n^{-\frac{1}{2}}+\lambda^*_{\max})\sum_{j=1}^{p_0}\lambda_j f_j'(\tilde{\beta}_j)u_j\\
&= L(\beta_0+\alpha_n u|x^n) - L(\beta_0|x^n) + (\sqrt{n}+n\lambda^*_{\max})\sum_{j=1}^{p_0}\lambda_j f_j'(\tilde{\beta}_j)u_j\\
&\geq L(\beta_0+\alpha_n u|x^n) - L(\beta_0|x^n) + (\sqrt{n}+n\lambda^*_{\max})\lambda^*_{\min}\sum_{j=1}^{p_0} f_j'(\tilde{\beta}_j)u_j\\
&= L(\beta_0+\alpha_n u|x^n) - L(\beta_0|x^n) + (\sqrt{n}\lambda^*_{\min}+n\lambda^*_{\max}\lambda^*_{\min})\sum_{j=1}^{p_0} f_j'(\tilde{\beta}_j)u_j\\
&= L(\beta_0+\alpha_n u|x^n) - L(\beta_0|x^n) + \sqrt{n}\lambda^*_{\min}(1+\sqrt{n}\lambda^*_{\max})\sum_{j=1}^{p_0} f_j'(\tilde{\beta}_j)u_j
\end{aligned} \tag{B2}$$

where p_0 is the number of components in β_1, and inequality holds since $\lambda^*_{\min} \leq \lambda^*_{\max}$.

Now by Taylor expansion of the log likelihood at β_0,

$$\begin{aligned}
L(\beta_0+\alpha_n u|x^n) - L(\beta_0|x^n) &= \alpha_n u L'(\beta_0|x^n) + \frac{n}{2}(\alpha_n u)'\left(J(\beta_0|x^\infty)\right)(\alpha_n u)\\
&\quad + \frac{n}{2}(\alpha_n u)'\left(\frac{L''(\tilde{\beta}|x^n)}{n} - J(\beta_0|x^\infty)\right)(\alpha_n u)\\
&= \alpha_n L'(\beta_0|x^n)u + \frac{n}{2}u'J(\beta_0|x^\infty)u\alpha_n^2\{1+o_p(1)\}
\end{aligned} \tag{B3}$$

by Condition 2 and condition we have that

$$E\left[\sup_{\beta\in B(\beta_0,\eta)}\left|L''(\tilde{\beta}|x_i)-J(\beta_0|x_i)\right|\right]\to 0 \text{ as } \eta\to 0,$$

and the $1+\epsilon$ moment condition gives us uniform integrability. Thus, we have

$$E\left[\sup_{\beta\in B(\beta_0,\eta)}\left|\frac{L''(\tilde{\beta}|x^n)}{n}-J(\beta_0|x^\infty)\right|\right]\to 0 \text{ as } \eta\to 0.$$

Using (**??**) and (**??**) ,

$$\begin{aligned} D_n(u) &\geq \alpha_n L'(\beta_0|x^n)'u+\frac{n}{2}u'J(\beta_0|x^\infty)u\alpha_n^2\{1+o_p(1)\}+\sqrt{n}\lambda^*_{\min}(1+\sqrt{n}\lambda^*_{\max})\sum_{j=1}^{p_0}f_j'(\tilde{\beta}_j)u_j \\ &\geq \alpha_n L'(\beta_0|x^n)'u+\tau_{max}\frac{n}{2}u'u\alpha_n^2\{1+o_p(1)\}+\sqrt{n}\lambda^*_{\min}(1+\sqrt{n}\lambda^*_{\max})\sum_{j=1}^{p_0}f_j'(\tilde{\beta}_j)u_j \end{aligned} \tag{B4}$$

where τ_{max} is the largest eigenvalue of $J(\beta_0|x^\infty)$. Now we argue the second term on the RHS of (**??**) is dominant. To see this consider the first term on the RHS of (**??**) using Conditions 1 and 3 We multiply by $\frac{\sqrt{n}}{\sqrt{n}}$ and obtain $\sqrt{n}\alpha_n\frac{L'(\beta_0|x^n)}{\sqrt{n}}u$. Let $\bar{L}=\frac{L'(\beta_0|x^n)}{\sqrt{n}}$, and observe $\sqrt{n}\alpha_n\to 1$.

Then using the Cauchy-Schwarz inequality we have

$$C^2=\|u\|^2=u'u\geq\|\bar{L}\|\cdot\|u\|\geq|\bar{L}'u|. \tag{B5}$$

Thus we have that by choosing sufficiently large C for the first inequality in (**??**) to hold, the second term of (**??**) is larger in absolute value (with high probability) than the first term uniformly in $\|u\|=C$.

Also, by hypothesis $\sqrt{n}\lambda^*_{\max}\to 0$, which implies $\sqrt{n}\lambda^*_{\min}\to 0$ since $\lambda^*_{\min}\leq\lambda^*_{\max}$, so the whole third term in (**??**) goes to zero and the second term of (**??**) is also larger than the third term. Thus since the third term converges to 0, we have that the second term on the RHS of (**??**) is larger than all of the other terms.

Therefore, by choosing large enough C,

$$P\left\{\inf_{\|u\|=C}D_n(u)\geq\tau_{max}\frac{n}{2}u'u\alpha_n^2\{1+o_p(1)\}\right\}\geq 1-\epsilon.$$

Thus, (**??**) is true for sufficiently large C.

Step 2. We now show the conclusion of Step 1 holds for any $C^*>C$. Let $C^\star>C$. Then,

$$\begin{aligned} &\hat{\beta}\in\{\beta_0+\alpha_n u:\|u\|=C^*\} \\ &\iff \hat{\beta}\in\mathcal{B}\left(\beta_0,\alpha_n C^*\right) \\ &\iff \hat{\beta}-\beta_0\in B\left(0,\alpha_n C^*\right) \\ &\iff \alpha_n^{-1}(\hat{\beta}-\beta_0)\in B(0,C^*) \\ &\iff \|\alpha_n^{-1}(\hat{\beta}-\beta_0)\|<C^* \\ &\iff \|\hat{\beta}-\beta_0\|^2<\alpha_n^2C^{*2} \end{aligned}$$

It follows that $\sum_{j=1}^p(\hat{\beta}_j-\beta_{0j})^2\leq\alpha_n^2C^{*2}$ and $\forall j$, $(\hat{\beta}_j-\beta_{0j})^2\leq\alpha_n^2C^{*2}$ where β_{0j} is the j^{th} component of β_0. So,

$$-\alpha_nC^*\leq\hat{\beta}_j-\beta_{0j}\leq\alpha_nC^*$$

and we can absorb C^* into α_n because C^* is a constant. Thus we have $\hat{\beta}-\beta_0=O_p(\alpha_n)$. Note that

$$\alpha_n=\frac{1}{\sqrt{n}}+\lambda^*_{\max}=\frac{1+\sqrt{n}\lambda^*_{\max}}{\sqrt{n}}$$

and as $n\to\infty$ we have $\sqrt{n}\lambda^*_{\max}\to 0$, so $\hat{\beta}-\beta_0=O_p\left(\frac{1}{\sqrt{n}}\right)$. □

Lemma 1. *Assume conditions 1- 6, and the result from Theorem* **??** *holds. If* $\sqrt{n}\lambda_{\min} \to \infty$, $\liminf_{\beta_j \to 0^+} f_j'(\beta_j) > 0$ *and* $\liminf_{\beta_j \to 0^-} f_j'(\beta_j) < 0$ *, then with probability tending to 1, for any given* $\tilde{\beta}_1$ *satisfying* $\|\tilde{\beta}_1 - \beta_1\| = O_p(n^{-\frac{1}{2}})$ *and any constant* C*,*

$$Q\left\{\begin{pmatrix}\beta_1\\0\end{pmatrix}\right\} = \min_{\beta_2 \le Cn^{-\frac{1}{2}}} Q\left\{\begin{pmatrix}\beta_1\\\beta_2\end{pmatrix}\right\}. \tag{B6}$$

Proof Consider the objective function $Q(\beta) = L(\beta|x^n) + n\sum_{j=1}^{p}\lambda_j f_j(\beta_j)$. Note that

$$\frac{\partial Q(\beta)}{\partial \beta_j} = \frac{\partial L(\beta|x^n)}{\partial \beta_j} + n\lambda_j f_j'(\beta_j). \tag{B7}$$

Then by Taylor expanding at $\beta_0 = \begin{pmatrix}\beta_1\\0\end{pmatrix}$ we have

$$\frac{\partial Q(\beta)}{\partial \beta_j} = \frac{\partial L(\beta_0|x^n)}{\partial \beta_j} + \sum_{\ell=1}^{p}\frac{\partial^2 L(\beta_0|x^n)}{\partial\beta_j\partial\beta_\ell}(\beta_\ell - \beta_{\ell 0}) + \sum_{\ell=1}^{p}\sum_{k=1}^{p}\frac{\partial^3 L(\beta^*|x^n)}{\partial\beta_j\partial\beta_\ell\partial\beta_k}(\beta_\ell - \beta_{\ell 0})(\beta_k - \beta_{k0}) + n\lambda_j f_j'(\beta_j)$$

where β^* lies between $\hat{\beta}$ and β_0. Note that

$$\frac{1}{n}\frac{\partial L(\beta_0|x^n)}{\partial \beta_j} = O_p(n^{-\frac{1}{2}})$$

and due to Hoadley (1971) and by the Law of Large Numbers for non-identically distributed random variables we have

$$\frac{1}{n}\frac{\partial^2 L(\beta_0|x^n)}{\partial\beta_j\partial\beta_\ell} = E\left[\frac{\partial^2 L(\beta_0|x^n)}{\partial\beta_j\partial\beta_\ell}\right] + o_p(1).$$

Note that due to Conditions 1–6, and the result in Theorem 1, $\hat{\beta} - \beta_0 = O_p\left(\frac{1}{\sqrt{n}}\right)$. So, we have

$$\begin{aligned}
\frac{\partial Q(\beta)}{\partial \beta_j} &= n\frac{1}{n}\frac{\partial L(\beta_0|x^n)}{\partial \beta_j} + n\frac{1}{n}\sum_{\ell=1}^{p}\frac{\partial^2 L(\beta_0|x^n)}{\partial\beta_j\partial\beta_\ell}(\hat{\beta}_\ell - \beta_{\ell 0})\\
&\quad + n\frac{1}{n}\sum_{\ell=1}^{p}\sum_{k=1}^{p}\frac{\partial^3 L(\beta^*|x^n)}{\partial\beta_j\partial\beta_\ell\partial\beta_k}(\hat{\beta}_\ell - \beta_{\ell 0})(\hat{\beta}_k - \beta_{k0}) + n\lambda_j f_j'(\beta_j)\\
&= nO_p\left(\frac{1}{\sqrt{n}}\right) + n\sum_{\ell=1}^{p}\left(E\left[\frac{\partial^2 L(\beta_0|x^n)}{\partial\beta_j\partial\beta_\ell}\right] + o_p(1)\right)O_p\left(\frac{1}{\sqrt{n}}\right)\\
&\quad + n\sum_{\ell=1}^{p}\sum_{k=1}^{p}E\left[\sup_{\beta\in N}\left|\frac{\partial^3 L(\beta^*|x^n)}{\partial\beta_j\partial\beta_\ell\partial\beta_k}\right| + o_p(1)\right]O_p\left(\frac{1}{\sqrt{n}}\right)O_p\left(\frac{1}{\sqrt{n}}\right) + n\lambda_j f_j'(\beta_j)\\
&= n\left[O_p\left(\frac{1}{\sqrt{n}}\right) + O_p\left(\frac{1}{\sqrt{n}}\right) + O_p\left(\frac{1}{\sqrt{n}}\right)\right] + n\lambda_j f_j'(\beta_j)\\
&= n\left[O_p\left(\frac{1}{\sqrt{n}}\right) + \lambda_j f_j'(\beta_j)\right]\\
&\ge n\left[O_p\left(\frac{1}{\sqrt{n}}\right) + \lambda_{\min} f_j'(\beta_j)\right]\\
&= n\lambda_{\min}\left[O_p\left(\frac{1}{\sqrt{n}\lambda_{\min}}\right) + f_j'(\beta_j)\right]
\end{aligned}$$

because p is finite and using Slutsky's Theorem. Now since $n^{-1/2}\lambda_{\min}^{-1} \to 0$ the sign of

$$\frac{\partial Q(\beta)}{\partial \beta_j} = n\left[O_p\left(\frac{1}{\sqrt{n}}\right) + \lambda_j f_j'(\beta_j)\right]$$

is determined by the sign of β_j since $\liminf_{\beta_j\to 0^+} f_j'(\beta_j) > 0$ and $\liminf_{\beta_j\to 0^-} f_j'(\beta_j) < 0$. Therefore, we have a minimum at $\beta_j = 0$ and (**??**) is true. □

The proof of the OP for penalized log-likelihood Theorem 2.1 is as follows.

Proof First, we observe that under Conditions 1–6 and our assumptions on b_n, Lemma **??** holds. So, $\hat{\beta} = (\hat{\beta_1}', \hat{\beta_2}')'$ satisfies $P(\hat{\beta}_2 = 0) \to 1$.

Now to prove the asymptotic normality of $\hat{\beta}_1$, consider the gradient of the objective function, $\triangledown Q(\beta)$. Let $P_\lambda(\beta)$ denote the vector $(\lambda w_1 f_1(\beta_1), \ldots, \lambda w_{p_0} f_{p_0}(\beta_{p_0}))'$. It can be shown that there exists a $\hat{\beta}_1$ in Theorem 1 that is a $\sqrt{n}$-consistent minimizer of $Q\left\{\begin{pmatrix}\beta_1\\0\end{pmatrix}\right\}$ which is a function of β_1 and satisfies the likelihood equation

$$\triangledown Q(\beta)\Big|_{\beta=\hat{\beta}=\begin{pmatrix}\hat{\beta}_1\\0\end{pmatrix}} = 0,$$

for $j = 1, \ldots, p_0$. Now Taylor expanding the gradient of the log likelihood at β_0 gives

$$\begin{aligned} 0 = \triangledown Q(\beta)\Big|_{\beta=\begin{pmatrix}\hat{\beta}_1\\0\end{pmatrix}} &= \triangledown L(\hat{\beta}|x^n) + n\triangledown P_\lambda(\hat{\beta}) \\ &= \triangledown L(\beta_0|x^n) + \triangledown^2 L(\tilde{\beta}|x^n)(\hat{\beta} - \beta_0) + n\triangledown P_\lambda(\hat{\beta}) \end{aligned}$$

where $\tilde{\beta} \in < \beta_0, \hat{\beta}_1 >$.

Furthermore, observe

$$\begin{aligned} -\triangledown L(\beta_0|x^n) &= \triangledown^2 L(\tilde{\beta}|x^n)(\hat{\beta} - \beta_0) + n\triangledown P_\lambda(\hat{\beta}) \\ -\triangledown L(\beta_0|x^n) &= n\frac{1}{n}\triangledown^2 L(\tilde{\beta}|x^n)(\hat{\beta} - \beta_0) + n\triangledown P_\lambda(\hat{\beta}) \\ -\frac{1}{\sqrt{n}}\triangledown L(\beta_0|x^n) &= -\sqrt{n}\hat{J}(\tilde{\beta}|x^n)(\hat{\beta} - \beta_0) + \sqrt{n}\triangledown P_\lambda(\hat{\beta}) \\ \frac{1}{\sqrt{n}}\triangledown L(\beta_0|x^n) &= \sqrt{n}\hat{J}(\tilde{\beta}|x^n)(\hat{\beta} - \beta_0) - \sqrt{n}\triangledown P_\lambda(\hat{\beta}) \end{aligned} \tag{B8}$$

To finish the proof, we examine the terms in (**??**). By supposition, $\sqrt{n}\lambda^*_{\max} \to 0$ and $f_j'(\hat{\beta}_j)$ is uniformly bounded by Condition 4. Then for $p_0 < j \leq p$ we know $P_\lambda(0) = 0$ by Condition 5. Also for $j \leq p_0$ we have $\sqrt{n}\lambda_j f_j'(\hat{\beta}_j) \leq \sqrt{n}\lambda^*_{\max} f_j'(\hat{\beta}_j) \to 0$ by definition of $\lambda^*_{\max}$. Thus, we see since each component of $\sqrt{n}\triangledown P_\lambda(\hat{\beta})$ converges to zero as $n \to \infty$, the whole term converges to zero. For the LHS of (**??**), the Central Limit Theorem, Condition 3 and the fact that $E\left(\frac{\partial L(\beta_0|x^n)}{\partial \beta_j}\right) = 0$ gives

$$\sqrt{n}\left(\frac{1}{n}\triangledown L(\beta_0|x^n) - E\left(\triangledown L(\beta_0|x^n)\right)\right) \xrightarrow{D} \mathcal{N}(0, J(\beta_0)).$$

Thus,

$$\frac{1}{\sqrt{n}}\triangledown L(\beta_0|x^n) \xrightarrow{D} \mathcal{N}(0, J(\beta_0)).$$

By equality in (**??**), the fact that the second term on RHS of (**??**) converges in probability to zero, and $\tilde{\beta} \xrightarrow{p} \beta_0$, it follows that

$$\sqrt{n}\hat{J}(\beta_0|x^n)(\hat{\beta} - \beta_0) \xrightarrow{D} \mathcal{N}(0, J(\beta_0)),$$

where $J(\beta_0) = J(\beta_0|x^\infty)$. Since $P(\hat{\beta}_2 = 0) \to 1$, we get that

$$\sqrt{n}\hat{J}_1(\beta_1|x^n)(\hat{\beta}_1 - \beta_1) \xrightarrow{D} \mathcal{N}(0, J_1(\beta_1))$$

where $J_1(\beta_1)$ is the information matrix containing only β_1. Now observe

$$\begin{aligned} \sqrt{n}\hat{J}_1(\beta_1|x^n)(\hat{\beta}_1 - \beta_1) = \sqrt{n}\left(\hat{J}_1(\beta_1|x^n) - J_1(\beta_1|x^\infty)\right)(\hat{\beta}_1 - \beta_1) \\ + \sqrt{n}J_1(\beta_1|x^\infty)(\hat{\beta}_1 - \beta_1). \end{aligned} \tag{B9}$$

Using condition 2 and 3, we get the first term on RHS of (**??**) converges to zero which implies the second term on the RHS converges in distribution to $\mathcal{N}(0, J_1(\beta_1))$.

Hence we have the desired result

$$\sqrt{n}\left(J_1(\beta_1|x^\infty)\right)^{\frac{1}{2}}(\hat{\beta}_1 - \beta_1) \xrightarrow{D} \mathcal{N}(0, I_{p_0}).$$

□

B.1.1 Penalized Empirical Risk Case, Sec. 2.2

We omit the proof for Theorem 2.2 because it follows in the same manner as the proof for Theorem 2.1.