Supplementary proofs for ”Consistent and Conservative Model Selection with the adaptive Lasso in Stationary and Nonstationary Autoregressions”

Anders Bredahl Kock

March 2, 2015

1 Supplementary proofs

This document contains the proofs of Theorems 1, 2, 6, and 7 of ”Consistent and Conservative Model Selection with the adaptive Lasso in Stationary and Nonstationary Autoregressions”. Please consult the main paper for notation.

Proof of Theorem 1. For the proof of this theorem we will need the following results which can be found in e.g. Hamilton (1994), Chapter 17.

$$S_T^{-1}X_T'X_TS_T^{-1}\overset{\sim}{\to}\begin{pmatrix} \frac{\sigma^2}{(1-\sum_{j=1}^p\beta_j^*)^2}\int_0^1 W_r^2dr & 0 \\ 0 & \Sigma \end{pmatrix}=:A, \tag{1}$$

$$S_T^{-1}X_T'\epsilon\overset{\sim}{\to}\begin{pmatrix} \frac{\sigma^2}{(1-\sum_{j=1}^p\beta_j)}\int_0^1 W_rdW_r \\ Z \end{pmatrix}=:B. \tag{2}$$

We shall also make use of the fact that the least squares estimator, $(\hat{\rho}_I, \hat{\beta}_I')$, of $(\rho^*, \beta^{*\prime})$ in (1) of the main paper satisfies that $\left\|S_T\left[\left(\hat{\rho}_I, \hat{\beta}_I'\right)' - \left(\rho^*, \beta^{*\prime}\right)'\right]\right\|_{\ell_2} \in O_p(1)$

The idea of the proof is as in the proof of Theorem 2 in Zou (2006). Alternatively, one could follow the route of Wang and Leng (2008), which is very different from the one here. First, let $u = (u_1, u_2')'$ where u_1 is a scalar and u_2 a $p \times 1$ vector. Set $\rho = u_1/T$ and $\beta_j = \beta_j^* + u_{2j}/\sqrt{T}$ which implies that (2) in the main paper as a function of u can be written as

$$\Psi_T(u) = \left\|\Delta y - \frac{u_1}{T}y_{-1} - \sum_{j=1}^p\left(\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right)\Delta y_{-j}\right\|_{\ell_2}^2$$
$$+ \lambda_T w_1^{\gamma_1}\left|\frac{u_1}{T}\right| + \lambda_T\sum_{j=1}^p w_{2j}^{\gamma_2}\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right|.$$

Let $\hat{u} = (\hat{u}_1, \hat{u}_2')' = \arg\min \Psi_T(u)$ and notice that $\hat{u}_1 = T\hat{\rho}$ and $\hat{u}_{2j} = \sqrt{T}(\hat{\beta}_j - \beta_j^*)$ for $j = 1, ..., p$. Define

$$\begin{aligned} V_T(u) &= \Psi_T(u) - \Psi_T(0) \\ &= u'S_T^{-1}X_T'X_TS_T^{-1}u - 2u'S_T^{-1}X_T'\epsilon + \lambda_T w_1^{\gamma_1}\left|\frac{u_1}{T}\right| + \lambda_T \sum_{j=1}^{p} w_{2j}^{\gamma_2}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right). \end{aligned}$$

Consider the first two terms in the above display. It follows from (1) and (2) that

$$u'S_T^{-1}X_T'X_TS_T^{-1}u - 2u'S_T^{-1}X_T'\epsilon \tilde{\to} u'Au - 2u'B \tag{3}$$

for all $u \in \mathbb{R}^{p+1}$. Furthermore,

$$\lambda_T w_1^{\gamma_1}\left|\frac{u_1}{T}\right| = \lambda_T \frac{1}{|\hat{\rho}_I|^{\gamma_1}}\left|\frac{u_1}{T}\right| = |u_1|\frac{\lambda_T}{T^{1-\gamma_1}}\frac{1}{|T\hat{\rho}_I|^{\gamma_1}} \to \begin{cases} \infty \text{ in probability if } u_1 \neq 0 \\ 0 \text{ in probability if } u_1 = 0 \end{cases} \tag{4}$$

since $T\hat{\rho}_I$ is tight. Also, if $\beta_j^* \neq 0$

$$\begin{aligned} \lambda_T w_{2j}^{\gamma_2}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) &= \lambda_T \left|\frac{1}{\hat{\beta}_{I,j}}\right|^{\gamma_2} \frac{u_{2j}}{\sqrt{T}}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) / \left(\frac{u_{2j}}{\sqrt{T}}\right) \\ &= \frac{\lambda_T}{T^{1/2}}\left|\frac{1}{\hat{\beta}_{I,j}}\right|^{\gamma_2} u_{2j}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) / \left(\frac{u_{2j}}{\sqrt{T}}\right) \\ &\to 0 \text{ in probability} \end{aligned} \tag{5}$$

since (i): $\lambda_T/T^{1/2} \to 0$, (ii): $\left|1/\hat{\beta}_{I,j}\right|^{\gamma_2} \to \left|1/\beta_j^*\right|^{\gamma_2} < \infty$ in probability and (iii): $u_{2j}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) / \left(\frac{u_{2j}}{\sqrt{T}}\right) \to u_{2j}\text{sign}(\beta_j^*)$.

Finally, if $\beta_j^* = 0$,

$$\begin{aligned} \lambda_T w_{2j}^{\gamma_2}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) &= \frac{\lambda_T}{T^{1/2}}\left|\frac{1}{\hat{\beta}_{I,j}}\right|^{\gamma_2}|u_{2j}| = \frac{\lambda_T}{T^{1/2-\gamma_2/2}}\left|\frac{1}{\sqrt{T}\hat{\beta}_{I,j}}\right|^{\gamma_2}|u_{2j}| \\ &\to \begin{cases} \infty \text{ in probability if } u_{2j} \neq 0 \\ 0 \text{ in probability if } u_{2j} = 0 \end{cases} \end{aligned} \tag{6}$$

since (i): $\frac{\lambda_T}{T^{1/2-\gamma_2/2}} \to \infty$ and (ii): $\sqrt{T}\hat{\beta}_{I,j}$ is tight.

Putting together (3)-(6) one concludes:

$$V_T(u) \tilde{\to} \Psi(u) = \begin{cases} u'Au - 2u'B \text{ if } u_1 = 0 \text{ and } u_{2j} = 0 \text{ for all } j \in \mathcal{A}^c \\ \infty \ \text{ if } u_1 \neq 0 \text{ or } u_{2j} \neq 0 \text{ for some } j \in \mathcal{A}^c \end{cases}$$

Since $V_T(u)$ is convex and $\Psi(u)$ has a unique minimum it follows from Knight (1999) that $\arg\min V_T(u) \tilde{\to} \arg\min \Psi(u)$. Hence,

$$\hat{u}_1 \tilde{\to} \delta_0 \tag{7}$$

$$\hat{u}_{2\mathcal{A}^c} \tilde{\to} \delta_0^{|\mathcal{A}^c|} \tag{8}$$

$$\hat{u}_{2\mathcal{A}} \tilde{\to} N(0, \sigma^2[\Sigma_{\mathcal{A}}]^{-1}) \tag{9}$$

where δ_0 is the Dirac measure at 0 and $|\mathcal{A}^c|$ is the cardinality of $\mathcal{A}^c$ (hence, $\delta_0^{|\mathcal{A}^c|}$ is the $|\mathcal{A}^c|$-dimensional Dirac measure at 0). Notice that (7) and (8) imply that $\hat{u}_1 \to 0$ in probability and $\hat{u}_{2\mathcal{A}^c} \to 0$ in probability. An equivalent formulation of (7)-(9) is

$$T\hat{\rho} \overset{\sim}{\to} \delta_0 \tag{10}$$

$$\sqrt{T}(\hat{\beta}_{\mathcal{A}^c} - \beta^*_{\mathcal{A}^c}) \overset{\sim}{\to} \delta_0^{|\mathcal{A}^c|} \tag{11}$$

$$\sqrt{T}(\hat{\beta}_{\mathcal{A}} - \beta^*_{\mathcal{A}}) \overset{\sim}{\to} N(0, \sigma^2[\Sigma_{\mathcal{A}}]^{-1}) \tag{12}$$

(10)-(12) yield the consistency part of the theorem at the rate of T for $\hat{\rho}$ and $\sqrt{T}$ for $\hat{\beta}$. Notice that this also implies that no $\hat{\beta}_j,\ j \in \mathcal{A}$ will be set equal to 0 since for all $j \in \mathcal{A},\ \hat{\beta}_j$ converges in probability to $\beta^*_j \neq 0$. (12) also yields the oracle efficient asymptotic distribution for $\hat{\beta}_{\mathcal{A}}$, i.e. part (3) of the theorem. It remains to show part (2) of the theorem; $P(\hat{\rho} = 0) \to 1$ and $P(\hat{\beta}_{T,\mathcal{A}^c} = 0) \to 1$.

First, assume $\hat{\rho} \neq 0$. Then the first order conditions for a minimum read:

$$2y'_{-1}\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right) + \lambda_T w_1^{\gamma_1}\text{sign}(\hat{\rho}) = 0$$

which is equivalent to

$$\frac{2y'_{-1}\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T} + \frac{\lambda_T w_1^{\gamma_1}\text{sign}(\hat{\rho})}{T} = 0$$

Consider first the second term:

$$\left|\frac{\lambda_T w_1^{\gamma_1}\text{sign}(\hat{\rho})}{T}\right| = \frac{\lambda_T}{T^{1-\gamma_1}}\frac{1}{|T\hat{\rho}_I|^{\gamma_1}} \to \infty \text{ in probability}$$

since $T\hat{\rho}_I$ is tight. For the first term one has:

$$\frac{2y'_{-1}\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T} = \frac{2y'_{-1}\left(\epsilon - X_T S_T^{-1} S_T[\hat{\rho}, \hat{\beta}' - \beta^{*\prime}]'\right)}{T}$$
$$= \frac{2y'_{-1}\epsilon}{T} - \frac{2y'_{-1}X_T S_T^{-1} S_T[\hat{\rho}, \hat{\beta}' - \beta^{*\prime}]'}{T}$$

By (2), $\frac{y'_{-1}\epsilon}{T} \overset{\sim}{\to} \frac{\sigma^2}{1-\sum_{j=1}^p \beta^*_j}\int_0^1 W_r dW_r$. Furthermore, $\frac{y'_{-1}X_T S_T^{-1}}{T} \overset{\sim}{\to} \left(\left(\frac{\sigma}{1-\sum_{j=1}^p \beta^*_j}\right)^2 \int_0^1 W_r^2 dr, 0, ..., 0\right)$ by (1). Hence, $\frac{y'_{-1}\epsilon}{T}$ and $\frac{y'_{-1}X_T S_T^{-1}}{T}$ are tight. We also know that $S_T[\hat{\rho}, \hat{\beta}' - \beta^{*\prime}]'$ converges weakly by (10)-(12) which implies it is tight as well. Taken together, $\frac{2y'_{-1}\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T}$ is tight and so

$$P(\hat{\rho} \neq 0) \leq P\left(\frac{2y'_{-1}\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T} + \frac{\lambda_T w_1^{\gamma_1}\text{sign}(\hat{\rho})}{T} = 0\right) \to 0$$

Next, assume $\hat{\beta}_j \neq 0$ for $j \in \mathcal{A}^c$. From the first order conditions

$$\Delta y'_{-j}(\Delta y - X_T(\hat{\rho}, \hat{\beta}')') + \lambda_T w_{2j}^{\gamma_2} \text{sign}(\hat{\beta}_j) = 0$$

or equivalently,

$$\frac{2\Delta y'_{-j}\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T^{1/2}} + \frac{\lambda_T w_{2j}^{\gamma_2} \text{sign}(\hat{\beta}_j)}{T^{1/2}} = 0$$

First, consider the second term

$$\left| \frac{\lambda_T w_{2j}^{\gamma_2} \text{sign}(\hat{\beta}_j)}{T^{1/2}} \right| = \frac{\lambda_T w_{2j}^{\gamma_2}}{T^{1/2}} = \frac{\lambda_T}{T^{1/2-\gamma_2/2} \left|T^{1/2}\hat{\beta}_{I,j}\right|^{\gamma_2}} \to \infty$$

since $\sqrt{T}\hat{\beta}_{I,j}$ is tight. Regarding the first term,

$$\frac{2\Delta y'_{-j}\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T^{1/2}} = \frac{2\Delta y'_{-j}\left(\epsilon - X_T S_T^{-1} S_T[\hat{\rho}, \hat{\beta}' - \beta^{*\prime}]'\right)}{T^{1/2}}$$
$$= \frac{2\Delta y'_{-j}\epsilon}{T^{1/2}} - \frac{2\Delta y'_{-j} X_T S_T^{-1} S_T[\hat{\rho}, \hat{\beta}' - \beta^{*\prime}]'}{T^{1/2}}$$

By (2) $\frac{\Delta y'_{-j}\epsilon}{T^{1/2}} \tilde{\rightarrow} N(0, \sigma^2\Sigma_j)$ where in accordance with previous notation Σ_j is the jth diagonal element of Σ. $\frac{\Delta y'_{-j} X_T S_T^{-1}}{T^{1/2}} \tilde{\rightarrow} (0, \Sigma_{(j,1)}, ..., \Sigma_{(j,p)})$ by (1). Hence, $\frac{\Delta y'_{-j}\epsilon}{T^{1/2}}$ and $\frac{\Delta y'_{-j} X_T S_T^{-1}}{T^{1/2}}$ are tight. The same is the case for $S_T[\hat{\rho}, \hat{\beta}' - \beta^{*\prime}]$ since it converges weakly by (10)-(12). Taken together, $\frac{2\Delta y'_{-j}\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T^{1/2}}$ is tight and so

$$P(\hat{\beta}_j \neq 0) \leq P\left(\frac{2\Delta y'_{-j}\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T^{1/2}} + \frac{\lambda_T w_{2j}^{\gamma_2} \text{sign}(\hat{\beta}_j)}{T^{1/2}} = 0\right) \to 0$$

We next turn to proving part b). The proof runs along the same lines as the proof part a). For the proof we will need (13) and (14) below which can be found in e.g. Hamilton (1994), Chapter 8. Notice that by definition of $x_t = (y_{t-1}, z_t')'$ the lower right hand $(p \times p)$ block of Q is Σ.

We shall make use of the following limit results:

$$\frac{1}{T} X_T' X_T \xrightarrow{p} Q \tag{13}$$

$$\frac{1}{\sqrt{T}} X_T' \epsilon \tilde{\rightarrow} N_{p+1}(0, \sigma^2 Q) =: \tilde{B} \tag{14}$$

where the definition of $\tilde{B}$ means that $\tilde{B}$ is a random vector distributed as $N_{p+1}(0, \sigma^2 Q)$ We shall also make use of the fact that the least squares estimator is $\sqrt{T}$ consistent under stationarity, i.e. $\left\| \sqrt{T}\left[\left(\hat{\rho}_I, \hat{\beta}_I'\right)' - \left(\rho^*, \beta^{*\prime}\right)' \right] \right\|_{\ell_2} \in O_p(1)$

First, let $u = (u_1, u_2')'$ where u_1 is a scalar and u_2 a $p \times 1$ vector. Set $\rho = \rho^* + u_1/\sqrt{T}$ and $\beta_j = \beta_j^* + u_{2j}/\sqrt{T}$ and

$$\Psi_T(u) = \left\| \Delta y - \left(\rho^* + \frac{u_1}{\sqrt{T}}\right) y_{-1} - \sum_{j=1}^{p} \left(\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right) \Delta y_{-j} \right\|_{\ell_2}^2$$
$$+ \lambda_T w_1^{\gamma_1} \left| \rho^* + \frac{u_1}{\sqrt{T}} \right| + \lambda_T \sum_{j=1}^{p} w_{2j}^{\gamma_2} \left| \beta_j^* + \frac{u_{2j}}{\sqrt{T}} \right|$$

Let $\hat{u} = (\hat{u}_1, \hat{u}_2')' = \arg\min \Psi_T(u)$ and notice that $\hat{u}_1 = \sqrt{T}(\hat{\rho} - \rho^*)$ and $\hat{u}_{2j} = \sqrt{T}(\hat{\beta}_j - \beta_j^*)$ for $j = 1, ..., p$. Define

$$\begin{aligned}\tilde{V}_T(u) &= \Psi_T(u) - \Psi_T(0) \\ &= \frac{1}{T} u' X_T' X_T u - 2 \frac{1}{\sqrt{T}} u' X_T' \epsilon + \lambda_T w_1^{\gamma_1} \left(\left| \rho^* + \frac{u_1}{\sqrt{T}} \right| - |\rho^*| \right) + \lambda_T \sum_{j=1}^{p} w_{2j}^{\gamma_2} \left(\left| \beta_j^* + \frac{u_{2j}}{\sqrt{T}} \right| - \left| \beta_j^* \right| \right)\end{aligned}$$

Consider the first two terms in the above display. It follows from (13) and (14) that

$$\frac{1}{T} u' X_T' X_T u - 2 \frac{1}{\sqrt{T}} u' X_T' \epsilon \tilde{\to} u' Q u - 2 u' \tilde{B} \tag{15}$$

for all $u \in \mathbb{R}^{p+1}$. Furthermore, since $\rho^* \neq 0$

$$\begin{aligned}\lambda_T w_1^{\gamma_1} \left(\left| \rho^* + \frac{u_1}{\sqrt{T}} \right| - |\rho^*| \right) &= \lambda_T \left| \frac{1}{\hat{\rho}_I} \right|^{\gamma_1} \frac{u_1}{\sqrt{T}} \left(\left| \rho^* + \frac{u_1}{\sqrt{T}} \right| - |\rho^*| \right) / \left(\frac{u_1}{\sqrt{T}} \right) \\ &= \frac{\lambda_T}{T^{1/2}} \left| \frac{1}{\hat{\rho}_I} \right|^{\gamma_1} u_1 \left(\left| \rho^* + \frac{u_1}{\sqrt{T}} \right| - |\rho^*| \right) / \left(\frac{u_1}{\sqrt{T}} \right) \\ &\to 0 \text{ in probability}\end{aligned} \tag{16}$$

since (i): $\lambda_T / T^{1/2} \to 0$, (ii): $\left| 1/\hat{\rho}_I \right|^{\gamma_1} \to \left| 1/\rho^* \right|^{\gamma_1} < \infty$ in probability and (iii): $u_1 \left(\left| \rho^* + \frac{u_1}{\sqrt{T}} \right| - |\rho^*| \right) / \left(\frac{u_1}{\sqrt{T}} \right) \to u_1 \text{sign}(\rho^*)$.

Similarly, if $\beta_j^* \neq 0$

$$\begin{aligned}\lambda_T w_{2j}^{\gamma_2} \left(\left| \beta_j^* + \frac{u_{2j}}{\sqrt{T}} \right| - \left| \beta_j^* \right| \right) &= \lambda_T \left| \frac{1}{\hat{\beta}_{I,j}} \right|^{\gamma_2} \frac{u_{2j}}{\sqrt{T}} \left(\left| \beta_j^* + \frac{u_{2j}}{\sqrt{T}} \right| - \left| \beta_j^* \right| \right) / \left(\frac{u_{2j}}{\sqrt{T}} \right) \\ &= \frac{\lambda_T}{T^{1/2}} \left| \frac{1}{\hat{\beta}_{I,j}} \right|^{\gamma_2} u_{2j} \left(\left| \beta_j^* + \frac{u_{2j}}{\sqrt{T}} \right| - \left| \beta_j^* \right| \right) / \left(\frac{u_{2j}}{\sqrt{T}} \right) \\ &\to 0 \text{ in probability}\end{aligned} \tag{17}$$

since (i): $\lambda_T / T^{1/2} \to 0$, (ii): $\left| 1/\hat{\beta}_{I,j} \right|^{\gamma_2} \to \left| 1/\beta_j^* \right|^{\gamma_2} < \infty$ in probability and (iii): $u_{2j} \left(\left| \beta_j^* + \frac{u_{2j}}{\sqrt{T}} \right| - \left| \beta_j^* \right| \right) / \left(\frac{u_{2j}}{\sqrt{T}} \right) \to u_{2j} \text{sign}(\beta_j^*)$.

Finally, if $\beta_j^* = 0$,

$$\lambda_T w_{2j}^{\gamma_2} \left(\left| \beta_j^* + \frac{u_{2j}}{\sqrt{T}} \right| - \left| \beta_j^* \right| \right) = \frac{\lambda_T}{T^{1/2}} \left| \frac{1}{\hat{\beta}_{I,j}} \right|^{\gamma_2} |u_{2j}| = \frac{\lambda_T}{T^{1/2-\gamma_2/2}} \left| \frac{1}{\sqrt{T}\hat{\beta}_{I,j}} \right|^{\gamma_2} |u_{2j}|$$

$$\to \begin{cases} \infty \text{ in probability if } u_{2j} \neq 0 \\ 0 \text{ in probability if } u_{2j} = 0 \end{cases} \tag{18}$$

since (i): $\frac{\lambda_T}{T^{1/2-\gamma_2/2}} \to \infty$ and (ii) $\sqrt{T}\hat{\beta}_{I,j}$ is tight.

Putting (15)-(18) together one concludes:

$$\tilde{V}_T(u) \tilde{\to} \Psi(u) = \begin{cases} u'Qu - 2u'\tilde{B} \text{ if } u_{2j} = 0 \text{ for all } j \in \mathcal{A}^c \\ \infty \;\; \text{if } u_{2j} \neq 0 \text{ for some } j \in \mathcal{A}^c \end{cases}$$

Since $\tilde{V}_T(u)$ is convex and $\Psi(u)$ has a unique minimum it follows from Knight (1999) that $\arg\min \tilde{V}_T(u) \tilde{\to} \arg\min \Psi(u)$. Hence,

$$(\hat{u}_1, \hat{u}_{2\mathcal{A}}')' \tilde{\to} N\left(0, \sigma^2 [Q_{\mathcal{B}}]^{-1}\right) \tag{19}$$

$$\hat{u}_{2\mathcal{A}^c} \tilde{\to} \delta_0^{|\mathcal{A}^c|} \tag{20}$$

where δ_0 is the Dirac measure at 0 and $|\mathcal{A}^c|$ is the cardinality of $\mathcal{A}^c$ (hence, $\delta_0^{|\mathcal{A}^c|}$ is the $|\mathcal{A}^c|$-dimensional Dirac measure at 0). Notice that (20) implies that $\hat{u}_{2\mathcal{A}^c} \to 0$ in probability. An equivalent formulation of (19) and (20) is

$$\begin{pmatrix} \sqrt{T}(\hat{\rho} - \rho^*) \\ \sqrt{T}(\hat{\beta}_{\mathcal{A}} - \beta_{\mathcal{A}}^*) \end{pmatrix} \tilde{\to} N\left(0, \sigma^2 [Q_{\mathcal{B}}]^{-1}\right) \tag{21}$$

$$\sqrt{T}(\hat{\beta}_{\mathcal{A}^c} - \beta_{\mathcal{A}^c}^*) \tilde{\to} \delta_0^{|\mathcal{A}^c|} \tag{22}$$

(21) and (22) establish the consistency part of the theorem at the oracle rate of $\sqrt{T}$. Note that this also implies that for no $j \in \mathcal{A}$ will $\hat{\beta}_j$ be set equal to 0 since for each $j \in \mathcal{A}$, $\hat{\beta}_j$ converges in probability to $\beta_j^* \neq 0$. The same is true for $\hat{\rho}$. (21) also yields the oracle efficient asymptotic distribution, i.e. part (3) of the theorem. It remains to show part (2) of the theorem; $P(\hat{\beta}_{\mathcal{A}^c} = 0) \to 1$.

Assume $\hat{\beta}_j \neq 0$ for $j \in \mathcal{A}^c$. From the first order conditions

$$2\Delta y_{-j}'(\Delta y - X_T(\hat{\rho}, \hat{\beta}')') + \lambda_T w_{2j}^{\gamma_2} \text{sign}(\hat{\beta}_j) = 0$$

or equivalently,

$$\frac{2\Delta y_{-j}' \left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T^{1/2}} + \frac{\lambda_T w_{2j}^{\gamma_2} \text{sign}(\hat{\beta}_j)}{T^{1/2}} = 0$$

First, consider the second term

$$\left| \frac{\lambda_T w_{2j}^{\gamma_2} \text{sign}(\hat{\beta}_j)}{T^{1/2}} \right| = \frac{\lambda_T w_{2j}^{\gamma_2}}{T^{1/2}} = \frac{\lambda_T}{T^{1/2-\gamma_2/2} \left| T^{1/2}\hat{\beta}_{I,j} \right|^{\gamma_2}} \to \infty$$

since $\sqrt{T}\hat{\beta}_{I,j}$ is tight. Regarding the first term,

$$\begin{aligned}\frac{2\Delta y_{-j}'\left(\Delta y - X_T(\hat{\rho},\hat{\beta}')'\right)}{T^{1/2}} &= \frac{2\Delta y_{-j}'\left(\epsilon - X_T[\hat{\rho}-\rho^*,\hat{\beta}'-\beta^{*\prime}]'\right)}{T^{1/2}}\\ &= \frac{2\Delta y_{-j}'\epsilon}{T^{1/2}} - \frac{2\Delta y_{-j}' X_T\sqrt{T}[\hat{\rho}-\rho^*,\hat{\beta}'-\beta^{*\prime}]'}{T}\end{aligned}$$

By (14), $\frac{\Delta y_{-j}'\epsilon}{T^{1/2}}\overset{\sim}{\rightarrow}N(0,\sigma^2 Q_{(j+1)})$ where in accordance with previous notation $Q_{(j+1)}$ is the $(j+1)$th diagonal element of Q. $\frac{\Delta y_{-j}'X_T}{T}\xrightarrow{p}(Q_{(j+1,1)},...,Q_{(j+1,p+1)})$ by (13). Hence, $\frac{\Delta y_{-j}'\epsilon}{T^{1/2}}$ and $\frac{\Delta y_{-j}'X_T}{T}$ are tight. The same is the case for $\sqrt{T}[\hat{\rho}-\rho^*,\hat{\beta}'-\beta^{*\prime}]$ since it converges weakly by (21)-(22). Hence,

$$P(\hat{\beta}_j\neq 0)\leq P\left(\frac{2\Delta y_{-j}'\left(\Delta y - X_T(\hat{\rho},\hat{\beta}')'\right)}{T^{1/2}} + \frac{\lambda_T w_{2j}^{\gamma_2}\text{sign}(\hat{\beta}_j)}{T^{1/2}} = 0\right)\to 0$$

□

Proof. Denote by $\hat{\eta}_\lambda = (\hat{\rho}_\lambda,\hat{\beta}_\lambda')'$ the adaptive Lasso estimator of $\eta^* = (\rho^*,\beta^{*\prime})'$ for the tuning parameter λ. Let $\hat{\epsilon}_\lambda = \Delta y - X_T\hat{\eta}_\lambda$ be the corresponding vector of error terms, and set $\hat{\mathcal{A}}_\lambda = \{j : \hat{\beta}_{\lambda,j}\neq 0\}$ and $\hat{\mathcal{B}}_\lambda = \{j : \hat{\eta}_{\lambda,j}\neq 0\}$. BIC_λ is the value of the information criterion for the adaptive Lasso with tuning parameter λ. For any $\mathcal{S}\subseteq\{1,...,p+1\}$, $X_{T,\mathcal{S}}$ denotes the matrix which has picked out all columns of X_T indexed by $\mathcal{S}$ [1]. Define $\hat{\epsilon}_{\mathcal{S},LS} = \Delta y - X_{T,\mathcal{S}}\hat{\beta}_{\mathcal{S},LS}$ to be the vector of error terms from a least squares regression only involving the columns of X_T indexed by $\mathcal{S}$. For any symmetric matrix A, let $\phi_{\min}(A)$ denote its smallest eigenvalue and let. Let $\{\lambda_T\}$ be a sequence satisfying the assumptions of Theorem 1.

a) Non-stationary case, $\rho^* = 0$. Thus, $\mathcal{B} = \mathcal{A}+1$.

Case 1: relevant variable left out, i.e. λ is such that $\hat{\mathcal{B}}_\lambda\not\supseteq\mathcal{B}$ (or, equivalently, as $\rho^*=0$, $\hat{\mathcal{A}}_\lambda\not\supseteq\mathcal{A}$). First, note that

$$\begin{aligned}\frac{\hat{\epsilon}_{\lambda_T}'\hat{\epsilon}_{\lambda_T}}{T} &= \frac{\left\|\Delta y - X_T\hat{\eta}_{\lambda_T}\right\|_{\ell_2}^2}{T}\\ &= \frac{\epsilon'\epsilon}{T} + \frac{1}{T}\left(\hat{\eta}_{\lambda_T}-\eta^*\right)' S_T S_T^{-1}X_T'X_T S_T^{-1}S_T\left(\hat{\eta}_{\lambda_T}-\eta^*\right) - 2\frac{1}{T}\epsilon' X_T S_T^{-1}S_T\left(\hat{\eta}_{\lambda_T}-\eta^*\right)\\ &= \sigma^2 + o_p(1) \qquad (23)\end{aligned}$$

since $S_T S_T^{-1}X_T'X_T S_T^{-1} = O_p(1)$ by (1) and $\epsilon' X_T S_T^{-1} = O_p(1)$ by (2). Furthermore, we used $S_T\left(\hat{\eta}_{\lambda_T}-\eta^*\right) = O_p(1)$ by Theorem 1. Therefore, because $|\hat{\mathcal{B}}_\lambda|\leq p+1$,

$$BIC_{\lambda_T} := \log\left(\frac{\hat{\epsilon}_{\lambda_T}'\hat{\epsilon}_{\lambda_T}}{T}\right) + |\hat{\mathcal{B}}_{\lambda_T}|\frac{\log(T)}{T} = \log(\sigma^2) + o_p(1) \qquad (24)$$

[1] This is not in conflict with the notation introduced in the main paper, as we have only indexed square matrices by sets so far.

Next, note that for any non-random set $\mathcal{S} \not\supseteq \mathcal{B}$

$$\hat{\eta}_{\mathcal{S},LS} = \left(X_{T,\mathcal{S}}' X_{T,\mathcal{S}}\right)^{-1} X_{T,\mathcal{S}}' \Delta y = \eta_{\mathcal{S}}^* + \left(X_{T,\mathcal{S}}' X_{T,\mathcal{S}}\right)^{-1} X_{T,\mathcal{S}}' X_{T,\mathcal{S}^c} \eta_{\mathcal{S}^c}^* + \left(X_{T,\mathcal{S}}' X_{T,\mathcal{S}}\right)^{-1} X_{T,\mathcal{S}}' \epsilon$$

such that by (1), (2), and $S_{T,\mathcal{S}^c}\eta_{\mathcal{S}^c}^*/\sqrt{T} = \eta_{\mathcal{S}^c}^*$ (since $\rho^* = 0$),

$$\begin{aligned}
\frac{S_{T,\mathcal{S}}\left(\hat{\eta}_{\mathcal{S},LS} - \eta_{\mathcal{S}}^*\right)}{\sqrt{T}} &= \left(S_{T,\mathcal{S}}^{-1} X_{T,\mathcal{S}}' X_{T,\mathcal{S}} S_{T,\mathcal{S}}^{-1}\right)^{-1} S_{T,\mathcal{S}}^{-1} X_{T,\mathcal{S}}' X_{T,\mathcal{S}^c} S_{T,\mathcal{S}^c}^{-1} \frac{S_{T,\mathcal{S}^c}\eta_{\mathcal{S}^c}^*}{\sqrt{T}} \\
&+ \left(S_{T,\mathcal{S}}^{-1} X_{T,\mathcal{S}}' X_{T,\mathcal{S}} S_{T,\mathcal{S}}^{-1}\right)^{-1} \frac{S_{T,\mathcal{S}}^{-1} X_{T,\mathcal{S}}' \epsilon}{\sqrt{T}} \\
&\overset{\sim}{\to} (A_{\mathcal{S}})^{-1} A_{\mathcal{S},\mathcal{S}^c} \eta_{\mathcal{S}^c}^*
\end{aligned}$$

As there are only finitely many $\mathcal{S} \not\supseteq \mathcal{B}$ the convergence is actually joint over these (for every $\mathcal{S}$ the converging matrix above is a continuous function of one and the same matrix $S_T^{-1} X_T' X_T S_T^{-1}$). Thus, for arbitrary $\mathcal{S} \not\supseteq \mathcal{B}$, letting $\hat{b}(\mathcal{S})$ be the $(p+1) \times 1$ vector with $\hat{\eta}_{\mathcal{S},LS}$ filled into all entries indexed by $\mathcal{S}$ and 0 in all entries indexed by $\mathcal{S}^c$, we get that $S_T(\hat{b}(\mathcal{S}) - \eta^*)/\sqrt{T} \overset{\sim}{\to} c(\mathcal{S})$ where $c(\mathcal{S})$ is a $(p+1) \times 1$ vector depending on $\mathcal{S}$ that has at least one entry different from zero (at least one entry will equal one of the $\beta_j^*,\ j \in \mathcal{A}$). Furthermore, $\hat{\epsilon}_{\mathcal{S},LS} = \Delta y - X_{T,\mathcal{S}}\hat{\eta}_{\mathcal{S},LS} = \epsilon - X_T(\hat{b}(\mathcal{S}) - \eta^*)$. This implies, using that a finite minimum (over $\mathcal{S} \not\supseteq \mathcal{B}$) is a continuous function and Σ is positive definite,

$$\begin{aligned}
\min_{\mathcal{S} \not\supseteq \mathcal{B}} \frac{\hat{\epsilon}_{\mathcal{S},LS}' \hat{\epsilon}_{\mathcal{S},LS}}{T} &\geq \frac{\epsilon'\epsilon}{T} + \min_{\mathcal{S} \not\supseteq \mathcal{B}} \frac{(\hat{b}(\mathcal{S}) - \eta^*)' S_T}{\sqrt{T}} S_T^{-1} X_T' X_T S_T^{-1} \frac{S_T(\hat{b}(\mathcal{S}) - \eta^*)}{\sqrt{T}} - 2 \max_{\mathcal{S} \not\supseteq \mathcal{B}} \epsilon' X_T S_T^{-1} \frac{S_T(\hat{b}(\mathcal{S}) - \eta^*)}{T} \\
&\overset{\sim}{\to} \sigma^2 + \min_{\mathcal{S} \not\supseteq \mathcal{B}} c(\mathcal{S})' A c(\mathcal{S}) \geq \sigma^2 + \phi_{\min}(A) \min_{\mathcal{S} \not\supseteq \mathcal{B}} c(\mathcal{S})' c(\mathcal{S}) \\
&\geq \sigma^2 + \phi_{\min}(\Sigma) \min_{\mathcal{S} \not\supseteq \mathcal{B}} c(\mathcal{S})' c(\mathcal{S}) \geq \sigma^2 + \underline{c}
\end{aligned}$$

for a $\underline{c} > 0$ since by assumption $c(\mathcal{S})$ has a non-zero entry of at magnitude at least $\min\{|\beta_j^*|, j \in \mathcal{A}\}$ which does not depend on $\mathcal{S}$. The above display also allows us to conclude that $F(t) = P\left(\sigma^2 + \min_{\mathcal{S} \not\supseteq \mathcal{B}} c(\mathcal{S})' A c(\mathcal{S}) \leq t\right) = 0$ for all $t < \sigma^2 + \underline{c}$. As such t are continuity points of F and so[2]

$$\limsup_{T \to \infty} P\left(\min_{\mathcal{S} \not\supseteq \mathcal{B}} \frac{\hat{\epsilon}_{\mathcal{S},LS}' \hat{\epsilon}_{\mathcal{S},LS}}{T} \leq \sigma^2 + \underline{c}/2\right) = 0 \tag{25}$$

Therefore, using that by construction $\hat{\epsilon}_\lambda' \hat{\epsilon}_\lambda \geq \hat{\epsilon}_{\hat{\mathcal{B}}_\lambda,LS}' \hat{\epsilon}_{\hat{\mathcal{B}}_\lambda,LS}$ (as least squares minimizes the sum of squared error terms), with probability tending to one

$$\begin{aligned}
BIC_\lambda = \log\left(\frac{\hat{\epsilon}_\lambda' \hat{\epsilon}_\lambda}{T}\right) + |\hat{\mathcal{B}}_\lambda| \frac{\log(T)}{T} &\geq \log\left(\frac{\hat{\epsilon}_{\hat{\mathcal{B}}_\lambda,LS}' \hat{\epsilon}_{\hat{\mathcal{B}}_\lambda,LS}}{T}\right) \geq \min_{\mathcal{S} \not\supseteq \mathcal{B}} \log\left(\frac{\hat{\epsilon}_{\mathcal{S},LS}' \hat{\epsilon}_{\mathcal{S},LS}}{T}\right) \\
&> \log(\sigma^2 + \underline{c}/2) > \log(\sigma^2)
\end{aligned} \tag{26}$$

[2] (25) also uses the following: Let U_T and V_T and be sequences of real random variables such that for all $T \geq 1$, $U_T \geq V_T$. If $V_T \overset{\sim}{\to} V$ and t is a continuity point of V, then $\limsup_{T\to\infty} P(U_T \leq t) \leq \limsup_{T\to\infty} P(V_T \leq t) = P(V \leq t)$. In our case $U_T = \min_{\mathcal{S} \not\supseteq \mathcal{B}} \frac{\hat{\epsilon}_{\mathcal{S},LS}' \hat{\epsilon}_{\mathcal{S},LS}}{T}$, $V_T = \frac{\epsilon'\epsilon}{T} + \min_{\mathcal{S} \not\supseteq \mathcal{B}} \frac{(\hat{b}(\mathcal{S}) - \eta^*)' S_T}{\sqrt{T}} S_T^{-1} X_T' X_T S_T^{-1} \frac{S_T(\hat{b}(\mathcal{S}) - \eta^*)}{\sqrt{T}} - 2 \max_{\mathcal{S} \not\supseteq \mathcal{B}} \epsilon' X_T S_T^{-1} \frac{S_T(\hat{b}(\mathcal{S}) - \eta^*)}{T}$, and $V = \sigma^2 + \min_{\mathcal{S} \not\supseteq \mathcal{B}} c(\mathcal{S})' A c(\mathcal{S})$.

for all $\lambda \geq 0 : \hat{B}_\lambda \not\supseteq \mathcal{B}$. In total, combining (24) and (26) and using that the latter is valid uniformly over $\lambda \geq 0 : \hat{B}_\lambda \not\supseteq \mathcal{B}$,

$$P\left(\inf_{\lambda \geq 0 : \hat{B}_\lambda \not\supseteq \mathcal{B}} BIC_\lambda > BIC_{\lambda_T}\right) \to 1$$

which implies that with probability tending to one BIC does not choose a λ for which the adaptive Lasso leaves out a relevant variable.
Case 2: Overfitted model, i.e. λ is such that $\mathcal{B} \subset \hat{\mathcal{B}}_\lambda$ ($\mathcal{B}$ is a proper subset of $\hat{\mathcal{B}}_\lambda$). Let $\mathcal{S}$ be any non-random set such that $\mathcal{B} \subset \mathcal{S}$. Then, by (1) and (2), and defining $\hat{b}(\mathcal{S})$ as previously

$$\begin{aligned}\hat\epsilon'_{\lambda_T}\hat\epsilon_{\lambda_T} - \hat\epsilon'_{\mathcal{S},LS}\hat\epsilon_{\mathcal{S},LS} &= \left\|\Delta y - X_T\hat\eta_{\lambda_T}\right\|^2_{\ell_2} - \left\|\Delta y - X_{T,\mathcal{S}}\hat\eta_{\mathcal{S},LS}\right\|^2_{\ell_2}\\ &= \left(\hat\eta_{\lambda_T} - \eta^*\right) S_T' S_T^{-1} X_T' X_T S_T^{-1} S_T \left(\hat\eta_{\lambda_T} - \eta^*\right) - 2\epsilon' X_T S_T^{-1} S_T \left(\hat\eta_{\lambda_T} - \eta^*\right) -\\ &\left(\hat b(\mathcal{S}) - \eta^*\right)' S_T S_T^{-1} X_T' X_T S_T^{-1} S_T \left(\hat b(\mathcal{S}) - \eta^*\right) + 2\epsilon' X_T S_T^{-1} S_T \left(\hat b(\mathcal{S}) - \eta^*\right)\\ &= O_{p,\mathcal{S}}(1)\end{aligned}$$

where $O_{p,\mathcal{S}}(1)$ indicates an $O_p(1)$ depending on $\mathcal{S}$. Furthermore, we used $S_T\left(\hat\eta_{\lambda_T} - \eta^*\right) = O_p(1)$ by Theorem 1, and $S_{T,\mathcal{S}}\left(\hat\eta_{\mathcal{S},LS} - \eta^*_{\mathcal{S}}\right) = O_p(1)$ by the properties of the least squares estimator in a model including all relevant variables. Therefore, as there are only finitely many sets $\mathcal{S}$ which contain $\mathcal{B}$, we conclude

$$\left|\hat\epsilon'_{\hat{\mathcal{B}}_\lambda,LS}\hat\epsilon_{\hat{\mathcal{B}}_\lambda,LS} - \hat\epsilon'_{\lambda_T}\hat\epsilon_{\lambda_T}\right| \leq \max_{\mathcal{S}:\mathcal{B}\subset\mathcal{S}}\left|\hat\epsilon'_{\lambda_T}\hat\epsilon_{\lambda_T} - \hat\epsilon'_{\mathcal{S},LS}\hat\epsilon_{\mathcal{S},LS}\right| = O_p(1) \tag{27}$$

which by (23) implies $\hat\epsilon'_{\hat{\mathcal{B}}_\lambda,LS}\hat\epsilon_{\hat{\mathcal{B}}_\lambda,LS}/T \xrightarrow{p} \sigma^2$. Thus, using $\hat\epsilon'_\lambda\hat\epsilon_\lambda \geq \hat\epsilon'_{\hat{\mathcal{B}}_\lambda,LS}\hat\epsilon_{\hat{\mathcal{B}}_\lambda,LS}$

$$\begin{aligned}T\left(BIC_\lambda - BIC_{\lambda_T}\right) &= T\left(\log(\hat\epsilon'_\lambda\hat\epsilon_\lambda) - \log(\hat\epsilon'_{\lambda_T}\hat\epsilon_{\lambda_T})\right) + \left(|\hat{\mathcal{B}}_\lambda| - |\hat{\mathcal{B}}_{\lambda_T}|\right)\log(T)\\ &\geq T\left(\log(\hat\epsilon'_{\hat{\mathcal{B}}_\lambda,LS}\hat\epsilon_{\hat{\mathcal{B}}_\lambda,LS}) - \log(\hat\epsilon'_{\lambda_T}\hat\epsilon_{\lambda_T})\right) + \left(|\hat{\mathcal{B}}_\lambda| - |\hat{\mathcal{B}}_{\lambda_T}|\right)\log(T)\end{aligned} \tag{28}$$

First, by the mean value theorem there exists a $\tilde c$ on the line segment joining $\hat\epsilon'_{\hat{\mathcal{B}}_\lambda,LS}\hat\epsilon_{\hat{\mathcal{B}}_\lambda,LS}$ and $\hat\epsilon'_{\lambda_T}\hat\epsilon_{\lambda_T}$ such that

$$T\left|\log(\hat\epsilon'_{\hat{\mathcal{B}}_\lambda,LS}\hat\epsilon_{\hat{\mathcal{B}}_\lambda,LS}) - \log(\hat\epsilon'_{\lambda_T}\hat\epsilon_{\lambda_T})\right| = T\frac{\left|\hat\epsilon'_{\hat{\mathcal{B}}_\lambda,LS}\hat\epsilon_{\hat{\mathcal{B}}_\lambda,LS} - \hat\epsilon'_{\lambda_T}\hat\epsilon_{\lambda_T}\right|}{\tilde c} \leq \frac{\left|\hat\epsilon'_{\hat{\mathcal{B}}_\lambda,LS}\hat\epsilon_{\hat{\mathcal{B}}_\lambda,LS} - \hat\epsilon'_{\lambda_T}\hat\epsilon_{\lambda_T}\right|}{\hat\epsilon'_{\hat{\mathcal{B}}_\lambda,LS}\hat\epsilon_{\hat{\mathcal{B}}_\lambda,LS}/T \wedge \hat\epsilon'_{\lambda_T}\hat\epsilon_{\lambda_T}/T} = O_p(1)$$

by (27) and convergence in probability of the denominator to $\sigma^2 > 0$. Finally, $\left(|\hat{\mathcal{B}}_\lambda| - |\hat{\mathcal{B}}_{\lambda_T}|\right)\log(T)$ tends to infinity in probability as $|\hat{\mathcal{B}}_{\lambda_T}| = |\mathcal{B}|$ with probability tending to one and $|\hat{\mathcal{B}}_\lambda| > |\mathcal{B}|$. Therefore, as the above arguments are valid uniformly in $\lambda \geq 0 : \mathcal{B} \subset \hat{B}_\lambda$, we conclude

$$P\left(\inf_{\lambda \geq 0 : \mathcal{B} \subset \hat{B}_\lambda}(BIC_\lambda - BIC_{\lambda_T}) > 0\right) = P\left(\inf_{\lambda \geq 0 : \mathcal{B} \subset \hat{B}_\lambda} T(BIC_\lambda - BIC_{\lambda_T}) > 0\right) \to 1$$

which completes the proof in the non-stationary setting.
b) Next we consider the stationary setting where $\rho^* \neq 0$. Thus, the non-zero entries of η^* have indices $\mathcal{B} = \{1\} \cup (\mathcal{A} + 1)$ the true active subset of $\{1, ..., p+1\}$.

Case 1: relevant variable left out, i.e. λ is such that $\hat{\mathcal{B}}_\lambda \not\supseteq \mathcal{B}$. First, note that

$$\begin{aligned}\frac{\hat{\epsilon}'_{\lambda_T}\hat{\epsilon}_{\lambda_T}}{T} &= \frac{\left\|\Delta y - X_T\hat{\eta}_{\lambda_T}\right\|^2_{\ell_2}}{T} \\ &= \frac{\epsilon'\epsilon}{T} + \left(\hat{\eta}_{\lambda_T} - \eta^*\right)' \frac{X_T'X_T}{T}\left(\hat{\eta}_{\lambda_T} - \eta^*\right) - 2\frac{1}{T}\frac{\epsilon' X_T}{\sqrt{T}}\sqrt{T}\left(\hat{\eta}_{\lambda_T} - \eta^*\right) \\ &= \sigma^2_\epsilon + o_p(1) \end{aligned} \tag{29}$$

since $X_T'X_T/T = O_p(1)$ by (13) and $\epsilon' X_T/\sqrt{T} = O_p(1)$ by (14). Furthermore, we used $\sqrt{T}\left(\hat{\eta}_{\lambda_T} - \eta^*\right) = O_p(1)$ by Theorem 1. Therefore, because $|\hat{\mathcal{B}}_\lambda| \leq p+1$,

$$BIC_{\lambda_T} := \log\left(\frac{\hat{\epsilon}'_{\lambda_T}\hat{\epsilon}_{\lambda_T}}{T}\right) + |\hat{\mathcal{B}}_{\lambda_T}|\frac{\log(T)}{T} = \log(\sigma^2_\epsilon) + o_p(1) \tag{30}$$

Next, note that for any non-random set $\mathcal{S} \not\supseteq \mathcal{B}$

$$\hat{\eta}_{\mathcal{S},LS} = \left(X'_{T,\mathcal{S}}X_{T,\mathcal{S}}\right)^{-1}X'_{T,\mathcal{S}}\Delta y = \eta^*_{\mathcal{S}} + \left(X'_{T,\mathcal{S}}X_{T,\mathcal{S}}\right)^{-1}X'_{T,\mathcal{S}}X_{T,\mathcal{S}^c}\eta^*_{\mathcal{S}^c} + \left(X'_{T,\mathcal{S}}X_{T,\mathcal{S}}\right)^{-1}X'_{T,\mathcal{S}}\epsilon$$

such that by (13) and (14)

$$\left(\hat{\eta}_{\mathcal{S},LS} - \eta^*_{\mathcal{S}}\right) = \left(\frac{X'_{T,\mathcal{S}}X_{T,\mathcal{S}}}{T}\right)^{-1}\frac{X'_{T,\mathcal{S}}X_{T,\mathcal{S}^c}}{T}\eta^*_{\mathcal{S}^c} \quad + \left(\frac{X'_{T,\mathcal{S}}X_{T,\mathcal{S}}}{T}\right)^{-1}\frac{X'_{T,\mathcal{S}}\epsilon}{T} \xrightarrow{p} (Q_{\mathcal{S}})^{-1}Q_{\mathcal{S},\mathcal{S}^c}\eta^*_{\mathcal{S}^c}$$

Thus, for arbitrary $\mathcal{S} \not\supseteq \mathcal{B}$, letting $\hat{b}(\mathcal{S})$ be the $(p+1)\times 1$ vector with $\hat{\eta}_{\mathcal{S},LS}$ filled into all entries indexed by $\mathcal{S}$ and 0 in all entries indexed by $\mathcal{S}^c$, we get that $(\hat{b}(\mathcal{S}) - \eta^*) \xrightarrow{p} c(\mathcal{S})$ where $c(\mathcal{S})$ is a $(p+1)\times 1$ vector depending on $\mathcal{S}$ that has at least one entry different from zero (at least one entry equals one of the $\eta^*_j,\ j \in \mathcal{B}$). Furthermore, $\hat{\epsilon}_{\mathcal{S},LS} = \Delta y - X_{T,\mathcal{S}}\hat{\eta}_{\mathcal{S},LS} = \epsilon - X_T(\hat{b}(\mathcal{S}) - \eta^*)$. This implies, using that a finite minimum (over $\mathcal{S} \not\supseteq \mathcal{B}$) is a continuous function and Q is positive definite,

$$\begin{aligned}\min_{\mathcal{S}\not\supseteq\mathcal{B}}\frac{\hat{\epsilon}'_{\mathcal{S},LS}\hat{\epsilon}_{\mathcal{S},LS}}{T} &\geq \frac{\epsilon'\epsilon}{T} + \min_{\mathcal{S}\not\supseteq\mathcal{B}}(\hat{b}(\mathcal{S}) - \eta^*)'\frac{X_T'X_T}{T}(\hat{b}(\mathcal{S}) - \eta^*) - 2\max_{\mathcal{S}\not\supseteq\mathcal{B}}\frac{\epsilon' X_T}{T}(\hat{b}(\mathcal{S}) - \eta^*) \\ &\xrightarrow{p} \sigma^2 + \min_{\mathcal{S}\not\supseteq\mathcal{B}} c(\mathcal{S})'Qc(\mathcal{S}) \geq \sigma^2 + \phi_{\min}(Q)\min_{\mathcal{S}\not\supseteq\mathcal{B}} c(\mathcal{S})'c(\mathcal{S}) \geq \sigma^2 + \underline{c}\end{aligned}$$

for a $\underline{c} > 0$ since by assumption $c(\mathcal{S})$ has a non-zero entry of at least $\min\{|\beta^*_j|, j \in \mathcal{A}\} \wedge |\rho^*|$ which does not depend on $\mathcal{S}$. Therefore, using that by construction $\hat{\epsilon}'_\lambda\hat{\epsilon}_\lambda \geq \hat{\epsilon}'_{\hat{\mathcal{B}}_\lambda,LS}\hat{\epsilon}_{\hat{\mathcal{B}}_\lambda,LS}$ (as least squares minimizes the sum of squared error terms), with probability tending to one

$$\begin{aligned}BIC_\lambda = \log\left(\frac{\hat{\epsilon}'_\lambda\hat{\epsilon}_\lambda}{T}\right) + |\hat{\mathcal{B}}_\lambda|\frac{\log(T)}{T} &\geq \log\left(\frac{\hat{\epsilon}'_{\hat{\mathcal{B}}_\lambda,LS}\hat{\epsilon}_{\hat{\mathcal{B}}_\lambda,LS}}{T}\right) \geq \min_{\mathcal{S}\not\supseteq\mathcal{B}}\log\left(\frac{\hat{\epsilon}'_{\mathcal{S},LS}\hat{\epsilon}_{\mathcal{S},LS}}{T}\right) \\ &> \log(\sigma^2 + c/2) > \log(\sigma^2)\end{aligned} \tag{31}$$

for all $\lambda \geq 0 : \hat{B}_\lambda \not\supseteq \mathcal{B}$. In total, combining (30) and (31), and using that the latter is valid uniformly over $\lambda \geq 0 : \hat{B}_\lambda \not\supseteq \mathcal{B}$,

$$P\left(\inf_{\lambda\geq 0:\hat{\mathcal{B}}_\lambda\not\supseteq\mathcal{B}} BIC_\lambda > BIC_{\lambda_T}\right)\to 1$$

which implies that with probability tending to one BIC does not choose a λ for which the adaptive Lasso leaves out a relevant variable.
Case 2: Overfitted model, i.e. λ is such that $\mathcal{B}\subset\hat{\mathcal{B}}_\lambda$ ($\mathcal{B}$ is a proper subset of $\hat{\mathcal{B}}_\lambda$), or equivalently, $\mathcal{A}\subset\hat{\mathcal{A}}_\lambda$. Let $\mathcal{S}$ be any non-random set such that $\mathcal{B}\subset\mathcal{S}$. Then, by (13) and (14), and defining $\hat{b}(\mathcal{S})$ as previously

$$\begin{aligned}
\hat{\epsilon}'_{\lambda_T}\hat{\epsilon}_{\lambda_T}-\hat{\epsilon}'_{\mathcal{S},LS}\hat{\epsilon}_{\mathcal{S},LS}&=\left\|\Delta y-X_T\hat{\eta}_{\lambda_T}\right\|^2_{\ell_2}-\left\|\Delta y-X_{T,\mathcal{S}}\hat{\eta}_{\mathcal{S},LS}\right\|^2_{\ell_2}\\
&=\left(\hat{\eta}_{\lambda_T}-\eta^*\right)\sqrt{T}'\frac{X_T'X_T}{T}\sqrt{T}\left(\hat{\eta}_{\lambda_T}-\eta^*\right)-2\frac{\epsilon'X_T}{\sqrt{T}}\sqrt{T}\left(\hat{\eta}_{\lambda_T}-\eta^*\right)-\\
&\left(\hat{b}(\mathcal{S})-\eta^*\right)'\sqrt{T}\frac{X_T'X_T}{T}\sqrt{T}\left(\hat{b}(\mathcal{S})-\eta^*\right)+2\frac{\epsilon'X_T}{\sqrt{T}}\sqrt{T}\left(\hat{b}(\mathcal{S})-\eta^*\right)\\
&=O_{p,\mathcal{S}}(1)
\end{aligned}$$

where $O_{p,\mathcal{S}}(1)$ indicates an $O_p(1)$ depending on $\mathcal{S}$. Furthermore, we used $\sqrt{T}\left(\hat{\eta}_{\lambda_T}-\eta^*\right)=O_p(1)$ by Theorem 1 and $\sqrt{T}\left(\hat{\eta}_{\mathcal{S},LS}-\eta^*_{\mathcal{S}}\right)=O_p(1)$ by the properties of the least squares estimator in a model including all relevant variables. Therefore, as there are only finitely many sets $\mathcal{S}$ which contain $\mathcal{B}$, we conclude

$$\left|\hat{\epsilon}'_{\hat{\mathcal{B}}_\lambda,LS}\hat{\epsilon}_{\hat{\mathcal{B}}_\lambda,LS}-\hat{\epsilon}'_{\lambda_T}\hat{\epsilon}_{\lambda_T}\right|\leq\max_{\mathcal{S}:\mathcal{B}\subset\mathcal{S}}\left|\hat{\epsilon}'_{\lambda_T}\hat{\epsilon}_{\lambda_T}-\hat{\epsilon}'_{\mathcal{S},LS}\hat{\epsilon}_{\mathcal{S},LS}\right|=O_p(1)\tag{32}$$

which by (29) implies $\hat{\epsilon}'_{\hat{\mathcal{B}}_\lambda,LS}\hat{\epsilon}_{\hat{\mathcal{B}}_\lambda,LS}/T\xrightarrow{p}\sigma^2$. Thus, using that by construction $\hat{\epsilon}'_\lambda\hat{\epsilon}_\lambda\geq\hat{\epsilon}'_{\hat{\mathcal{B}}_\lambda,LS}\hat{\epsilon}_{\hat{\mathcal{B}}_\lambda,LS}$,

$$\begin{aligned}
T\left(BIC_\lambda-BIC_{\lambda_T}\right)&=T\left(\log(\hat{\epsilon}'_\lambda\hat{\epsilon}_\lambda)-\log(\hat{\epsilon}'_{\lambda_T}\hat{\epsilon}_{\lambda_T})\right)+\left(|\hat{\mathcal{B}}_\lambda|-|\hat{\mathcal{B}}_{\lambda_T}|\right)\log(T)\\
&\geq T\left(\log(\hat{\epsilon}'_{\hat{\mathcal{B}}_\lambda,LS}\hat{\epsilon}_{\hat{\mathcal{B}}_\lambda,LS})-\log(\hat{\epsilon}'_{\lambda_T}\hat{\epsilon}_{\lambda_T})\right)+\left(|\hat{\mathcal{B}}_\lambda|-|\hat{\mathcal{B}}_{\lambda_T}|\right)\log(T)
\end{aligned}\tag{33}$$

First, by the mean value theorem there exists a $\tilde{c}$ on the line segment joining $\hat{\epsilon}'_{\hat{\mathcal{B}}_\lambda,LS}\hat{\epsilon}_{\hat{\mathcal{B}}_\lambda,LS}$ and $\hat{\epsilon}'_{\lambda_T}\hat{\epsilon}_{\lambda_T}$ such that

$$T\left|\log(\hat{\epsilon}'_{\hat{\mathcal{B}}_\lambda,LS}\hat{\epsilon}_{\hat{\mathcal{B}}_\lambda,LS})-\log(\hat{\epsilon}'_{\lambda_T}\hat{\epsilon}_{\lambda_T})\right|=T\frac{\left|\hat{\epsilon}'_{\hat{\mathcal{B}}_\lambda,LS}\hat{\epsilon}_{\hat{\mathcal{B}}_\lambda,LS}-\hat{\epsilon}'_{\lambda_T}\hat{\epsilon}_{\lambda_T}\right|}{\tilde{c}}\leq\frac{\left|\hat{\epsilon}'_{\hat{\mathcal{B}}_\lambda,LS}\hat{\epsilon}_{\hat{\mathcal{B}}_\lambda,LS}-\hat{\epsilon}'_{\lambda_T}\hat{\epsilon}_{\lambda_T}\right|}{\hat{\epsilon}'_{\hat{\mathcal{B}}_\lambda,LS}\hat{\epsilon}_{\hat{\mathcal{B}}_\lambda,LS}/T\wedge\hat{\epsilon}'_{\lambda_T}\hat{\epsilon}_{\lambda_T}/T}=O_p(1)$$

by (32) and convergence in probability of the denominator to $\sigma^2_\epsilon>0$. Finally, $\left(|\hat{\mathcal{B}}_\lambda|-|\hat{\mathcal{B}}_{\lambda_T}|\right)\log(T)$ tends to infinity in probability as $|\hat{\mathcal{B}}_{\lambda_T}|=|\mathcal{B}|$ with probability tending to one and $|\hat{\mathcal{B}}_\lambda|>|\mathcal{B}|$. Therefore, as the above arguments are valid uniformly in $\lambda\geq 0:\mathcal{B}\subset\hat{B}_\lambda$, we conclude

$$P\left(\inf_{\lambda\geq 0:\mathcal{B}\subset\hat{B}_\lambda}(BIC_\lambda-BIC_{\lambda_T})>0\right)=P\left(\inf_{\lambda\geq 0:\mathcal{B}\subset\hat{B}_\lambda}T(BIC_\lambda-BIC_{\lambda_T})>0\right)\to 1$$

which completes the proof in the stationary setting. □

Proof of Theorem 6. We begin with part a). The setting is the same as in the proof of Theorem 1a). Follow the proof of that theorem, with identical notation, until (3) with $\gamma_1 = \gamma_2 = 1$. Next, notice that

$$\lambda_T w_1 \left|\frac{u_1}{T}\right| = \lambda_T \frac{1}{|\hat{\rho}_I|}\left|\frac{u_1}{T}\right| = |u_1|\, \lambda_T \frac{1}{|T\hat{\rho}_I|} \overset{\sim}{\to} \lambda \frac{|u_1|}{|C_1|} \tag{34}$$

by (1) and (2) (and the form of the initial least squares estimator $\hat{\rho}_I$) since C_1 has no mass at 0. Furthermore, if $\beta_j^* \neq 0$

$$\begin{aligned}
\lambda_T w_{2j}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) &= \lambda_T \left|\frac{1}{\hat{\beta}_{I,j}}\right| \frac{u_{2j}}{\sqrt{T}} \left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) / \left(\frac{u_{2j}}{\sqrt{T}}\right) \\
&= \frac{\lambda_T}{T^{1/2}} \left|\frac{1}{\hat{\beta}_{I,j}}\right| u_{2j} \left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) / \left(\frac{u_{2j}}{\sqrt{T}}\right) \\
&\to 0 \text{ in probability}
\end{aligned} \tag{35}$$

since (i): $\lambda_T / T^{1/2} \to 0$, (ii): $\left|1/\hat{\beta}_{I,j}\right| \to \left|1/\beta_j^*\right| < \infty$ in probability and (iii): $u_{2j}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) / \left(\frac{u_{2j}}{\sqrt{T}}\right) \to u_{2j}\text{sign}(\beta_j^*)$. Finally, if $\beta_j^* = 0$,

$$\lambda_T w_{2j}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) = \frac{\lambda_T}{T^{1/2}}\left|\frac{1}{\hat{\beta}_{I,j}}\right| |u_{2j}| = \lambda_T \left|\frac{1}{\sqrt{T}\hat{\beta}_{I,j}}\right| |u_{2j}| \overset{\sim}{\to} \lambda \frac{|u_{2j}|}{|C_{2j}|} \tag{36}$$

by (1) and (2) (and the form of the initial least squares estimator $\hat{\beta}_{I,j}$) since (i): $\lambda_T \to \lambda$ and (ii): C_{2j} is 0 with probability 0 such that $x \mapsto \left|1/x\right|$ is continuous almost everywhere with respect to the limiting measure. Putting together (3) and (34)-(36) one concludes

$$V_T(u) \overset{\sim}{\to} u'Au - 2u'B + \lambda \frac{|u_1|}{|C_1|} + \lambda \sum_{j=1}^{p} \frac{|u_{2j}|}{|C_{2j}|} \mathbf{1}_{\{\beta_j^* = 0\}} := \Psi(u)$$

Hence, since $V_T(u)$ is convex and $\Psi(u)$ has a unique minimum it follows from Knight (1999) that $\arg\min V_T(u) \overset{\sim}{\to} \arg\min \Psi(u)$

We now turn to proving part b). The setting is the same as in the proof of Theorem 1b). Follow the proof of that theorem, with identical notation, until (17) (as we now assume $\lambda_T \to \lambda \in [0, \infty)$ we clearly have $\lambda_T/\sqrt{T} \to 0$ as required in that theorem) with $\gamma_1 = \gamma_2 = 1$. For the case of $\beta_j^* = 0$ one has

$$\lambda_T w_{2j}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) = \frac{\lambda_T}{T^{1/2}}\left|\frac{1}{\hat{\beta}_{I,j}}\right| |u_{2j}| = \lambda_T \left|\frac{1}{\sqrt{T}\hat{\beta}_{I,j}}\right| |u_{2j}| \overset{\sim}{\to} \lambda \frac{|u_{2j}|}{|\tilde{C}_{2j}|} \tag{37}$$

by (13) and (14) (and the form of the initial least squares estimator $\hat{\beta}_{I,j}$) since (i): $\lambda_T \to \lambda$, (ii): $\tilde{C}_{2j}$ is 0 with probability 0 such that $x \mapsto \left|1/x\right|$ is continuous almost everywhere with respect to the limiting measure. Putting together (17)-(17) and (37) one concludes

$$\tilde{V_T}(u) \overset{\sim}{\to} u'Qu - 2u'\tilde{B} + \lambda \sum_{j=1}^{p} \frac{|u_{2j}|}{|\tilde{C}_{2j}|} \mathbf{1}_{\{\beta_j^* = 0\}} := \tilde{\Psi}(u)$$

Hence, since $\tilde{V}_T(u)$ is convex and $\tilde{\Psi}(u)$ has a unique minimum it follows from Knight (1999) that $\arg\min \tilde{V}_T(u) \tilde{\to} \arg\min \tilde{\Psi}(u)$. □

Proof of Theorem 7. a) First, consider the non-stationary setting. Just as in the proof of part a) of Theorem 6 above we can follow the proof of Theorem 1a) and make the necessary changes. In particular, one only has to omit w_1 and w_{2j} from (4)-(6), respectively and use that $\lambda_T/T \to \lambda$ and $\mu_T/\sqrt{T} \to \mu$ to conclude part a).

b) Just as in the proof of Theorem 6b) above we can follow the proof of Theorem 1b) and make the necessary changes. In particular, one only has to omit w_1 and w_{2j} from (16)-(18), respectively, use that $\rho^* \in (-2, 0)$ (by stationarity of y_t), $\lambda_T/T \to \lambda$ and $\mu_T/\sqrt{T} \to \mu$ to conclude part b).

□

References

Hamilton, J. D. (1994). *Time Series Analysis.* Cambridge University Press, Cambridge.

Knight, K. (1999). Epi-convergence in distribution and stochastic equi-semicontinuity. *Unpublished manuscript*.

Wang, H. and C. Leng (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis 52*(12), 5277–5286.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.