Supplementary Material to "Large-scale model selection in misspecified generalized linear models"

Emre Demirkaya, Yang Feng, Pallavi Basu and Jinchi Lv

This Supplementary Material contains additional numerical studies, examples to compute HGBIC$_p$, all the proofs of main results, and additional technical details.

A. Examples of HGBIC$_p$

We illustrate how HGBIC$_p$ can be calculated in linear regression and logistic regression. In both cases, we assume that $\widehat{\beta}$ is the maximum likelihood estimator. Then, the estimated natural parameter vector is $\widehat{\theta} = X\widehat{\beta}$. In order to estimate the two forms of Fisher Information matrices given in (7) and (8), we use the plug-in estimator $\widehat{H}_n$ as defined in Section 3.3. To proceed, we need to make use of the link function $b(\theta)$, and its first and second derivatives.

In linear regression, the link function takes the from $b(\theta) = \theta^2/2$. So, $b'(\theta) = \theta$ and $b''(\theta) = 1$. Then, $\widehat{A}_n = X^T X$, and $\widehat{B}_n = X^T \mathrm{diag}\{(y - \widehat{\theta}) \circ (y - \widehat{\theta})\} X$ where $\mathrm{diag}\{(y - \widehat{\theta}) \circ (y - \widehat{\theta})\}$ is the diagonal matrix whose i-th entry is $(y_i - \widehat{\theta}_i)^2$.

In logistic regression, the link function takes the from $b(\theta) = \log(1 + e^\theta)$. Then, the derivatives are $b'(\theta) = e^\theta/(1 + e^\theta)$ and $b''(\theta) = e^\theta/(1 + e^\theta)^2$. First, we can form diagonal matrix $\Sigma(X\widehat{\beta}_n)$ whose i-th diagonal entry is $\exp(\widehat{\theta}_i)/\{1 + \exp(\widehat{\theta}_i)\}^2$, and calculate $\widehat{A}_n = A_n(\widehat{\beta}_n) = X^T \Sigma(X\widehat{\beta}_n) X$. Next, we need another diagonal matrix, namely $\mathrm{diag}[\{y - \mu(X\widehat{\beta}_n)\} \circ \{y - \mu(X\widehat{\beta}_n)\}]$ with the i-th diagonal entry being $[y_i - \exp(\widehat{\theta}_i)/\{1 + \exp(\widehat{\theta}_i)\}]^2$. Then, we calculate $\widehat{B}_n = X^T \mathrm{diag}[\{y - \mu(X\widehat{\beta}_n)\} \circ \{y - \mu(X\widehat{\beta}_n)\}] X$.

In either case, we obtain the plug-in estimates of A_n and B_n following Section 3.3. Finally, we get $\widehat{H}_n = \widehat{A}_n^{-1} \widehat{B}_n$. We use log determinant and the trace of $\widehat{H}_n$ to calculate HGBIC$_p$.

B. Numerical studies on logistic regression with interaction

Our second simulation example is the high-dimensional logistic regression with interaction. We simulated 100 data sets from the logistic regression model with interaction and an n-dimensional parameter vector

$$\theta = Z\beta + 2x_{p+1} + 2x_{p+2}, \tag{A.1}$$

where $Z = (x_1, \ldots, x_p)$ is an $n \times p$ design matrix, $x_{p+1} = x_1 \circ x_2$ and $x_{p+2} = x_3 \circ x_4$ are two interaction terms, and the rest of the setting is the same as in the simulation example in Section 4.2. For each data set, the n-dimensional response vector y was sampled from the Bernoulli distribution with success probability vector $\{e^{\theta_1}/(1 + e^{\theta_1}), \ldots, e^{\theta_n}/(1 + e^{\theta_n})\}^T$ with $\theta = (\theta_1, \ldots, \theta_n)^T$ given in (A.1). As in Section 4.2, we consider the case where all covariates are independent of each other. We chose $\beta_0 = (2.5, -1.9, 2.8, -2.2, 3, 0, \ldots, 0)^T$ and set sample size $n = 300$. Although the data was generated from the logistic regression model with parameter vector (A.1), we fit the logistic regression model without the two interaction terms. This provides another example of misspecified models. As argued in Section 4.2, the oracle working model is $\mathrm{supp}(\beta_0) = \{1, \ldots, 5\}$ which corresponds to the logistic regression model with the first five covariates. To build a sequence of sparse models, we applied Lasso followed by maximum-likelihood refitting based on the support of the estimated model.

Since the goal in logistic regression is usually classification, we replace the prediction error with the classification error rate. Tables 3 and 4 show similar conclusions to those in Section 4.2. Again HGBIC$_p$ outperformed all other model selection criteria with greater advantage as the dimensionality increases (e.g., $p \geq 800$). As in Example 4.2, we also present the trend of the false discovery proportion and the true positive rate as ζ varies in Figure 2. From the figure, we observe that it is a more difficult setting than the multiple index model to reach model selection consistency. The proposed HGBIC$_p$ criterion with

Table 3. *Average results over 100 repetitions for Example B with all entries multiplied by 100.*

Consistent selection probability with sure screening probability in parentheses									
p	AIC	BIC	EBIC	GIC	GAIC	GBIC	GBIC$_p$	HGBIC$_p$	Oracle
100	0(100)	30(100)	71(100)	67(100)	3(100)	32(100)	60(100)	98(98)	100(100)
200	0(100)	27(100)	72(100)	73(100)	1(100)	29(100)	50(100)	94(96)	100(100)
400	0(100)	12(100)	80(100)	85(100)	0(100)	16(100)	44(100)	94(94)	100(100)
800	0(100)	2(99)	65(98)	75(98)	0(99)	4(99)	31(99)	92(93)	100(100)
1600	0(100)	0(100)	55(96)	74(95)	0(99)	1(100)	14(98)	83(84)	100(100)
3200	0(100)	0(100)	64(99)	79(98)	0(100)	1(100)	13(100)	78(81)	100(100)
Mean classification error (in percentage) with standard error in parentheses									
100	25.3(0.3)	16.1(0.2)	15.5(0.2)	15.5(0.2)	17.1(0.2)	16.1(0.2)	15.6(0.2)	15.2(0.2)	15.2(0.2)
200	24.9(0.3)	17.2(0.3)	15.9(0.2)	15.8(0.2)	17.9(0.2)	16.9(0.3)	16.1(0.2)	15.5(0.2)	15.4(0.2)
400	25.0(0.3)	19.7(0.4)	15.5(0.3)	15.4(0.2)	18.7(0.3)	17.8(0.3)	16.3(0.2)	15.3(0.2)	15.2(0.2)
800	24.7(0.3)	21.9(0.4)	16.2(0.2)	15.9(0.2)	18.9(0.2)	18.8(0.3)	16.8(0.3)	15.8(0.2)	15.5(0.2)
1600	26.0(0.4)	24.3(0.4)	16.2(0.2)	15.8(0.2)	19.4(0.3)	20.2(0.3)	17.2(0.2)	15.9(0.3)	15.4(0.2)
3200	25.7(0.3)	24.4(0.4)	16.0(0.2)	15.7(0.2)	19.3(0.2)	20.7(0.3)	17.9(0.3)	16.0(0.2)	15.3(0.2)

the choice of $\zeta = 1$ appears to strike a good balance between the false discovery proportion and the true positive rate.

To evaluate how much the fitted misspecified model deviates from the true model, we have now calculated the average correlation (AC) between the fitted mean vector and the true mean vector, and the consistent selection probability (CSP) over 100 repetitions, for the multiple index model and logistic regression model in Table 5. It shows that the proposed information criterion is in general robust to different levels of model misspecification.

Table 4. *Average false positives over 100 repetitions for Example B.*

p	AIC	BIC	EBIC	GIC	GAIC	GBIC	GBIC$_p$	HGBIC$_p$
100	55.71	1.57	0.37	0.43	4.51	1.48	0.54	0.00
200	40.83	3.24	0.46	0.37	5.10	2.14	0.80	0.02
400	35.25	11.74	0.28	0.16	6.43	4.27	1.16	0.00
800	31.78	18.22	0.55	0.26	5.83	6.00	1.59	0.01
1600	30.25	22.65	0.65	0.26	5.87	8.02	2.06	0.01
3200	28.41	22.31	0.50	0.25	5.26	8.61	2.74	0.03

Table 5. *Average correlation (AC) between the fitted mean vector and the true mean vector, and the consistent selection probability (CSP), for the multiple index model and logistic regression model*

	Multiple index model		Logistic regression model	
p	AC	CSP	AC	CSP
100	0.97	1.00	0.89	0.98
200	0.97	0.99	0.89	0.94
400	0.97	0.99	0.89	0.94
800	0.97	0.98	0.89	0.92
1600	0.97	0.98	0.88	0.83
3200	0.97	0.95	0.88	0.78

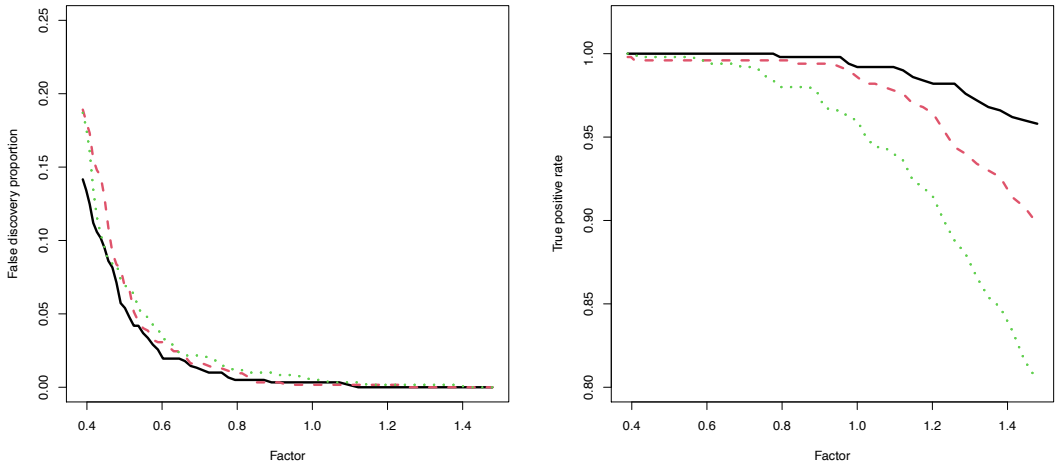

Fig. 2. The average false discovery proportion (left panel) and the true positive rate (right panel) as the factor ζ varies for Example B when $p = 200$ (black solid), $p = 800$ (red dashes), and $p = 3200$ (green dot-dash).

C. Numerical studies on Poisson regression with interaction

Our third simulation example is the high-dimensional Poisson regression with interaction. We simulated 100 data sets from the Poisson regression model with interaction and an n-dimensional parameter vector

$$\theta = 1.5 + Z\beta + x_{p+1} + x_{p+2}, \tag{A.1}$$

where $Z = (x_1, \ldots, x_p)$ is an $n \times p$ design matrix, and $x_{p+1} = x_1 \circ x_2$ and $x_{p+2} = x_3 \circ x_4$ are two interaction terms. Here the rows of the $n \times p$ design matrix Z are sampled as independent copies via the following process. First, we generate $w_{p\times 1} \sim t_5(0, \Sigma)$, a multi-variate t-distribution with 5 degrees of freedom and scale matrix Σ, where $\Sigma_{ij} = \rho^{|i-j|}$ for all i and j. We consider two correlation structures, namely $\rho = 0$ and $\rho = 0.6$. Then, we define $x_{p\times 1} = (x_1, \ldots, x_p)^T$ as a scaled version of w to the unit interval, where $x_j = (w_j - \min_k w_k)/(\max_k w_k - \min_k w_k)$, for $j = 1, \ldots, p$. Finally, for each data set, the n-dimensional response vector y was sampled from the Poisson distribution with mean vector $(e^{\theta_1}, \ldots, e^{\theta_n})^T$ with $\theta = (\theta_1, \ldots, \theta_n)^T$ given in (A.1). Here, we choose $\beta_0 = (1.25, 1, 0.75, 1.25, 1, 0, \ldots, 0)^T$ and set sample size $n = 200$. Although the data was generated from the Poisson regression model with parameter vector (A.1), we fit the Poisson regression model without the two interaction terms. From the data generation process, it is clear that Y is independent of $(x_6, \ldots, x_p)$ conditional on $(x_1, \ldots, x_5)$. Thus the oracle working model is $\mathrm{supp}(\beta_0) = \{1, \ldots, 5\}$, which corresponds to the Poisson regression model with the first five covariates; it is sometimes referred to as the Markov blanket for Y (Candès et al., 2018). To build a sequence of sparse models, we applied Lasso followed by maximum-likelihood refitting based on the support of the estimated model.

Tables 6, 7, 8, and 9 show similar conclusions to those in Sections 4.2 and B. Again HGBIC$_p$ outperformed all other model selection criteria in terms of higher consistent selection probability and smaller mean squared prediction error, with a more prominent improvement when the covariates are dependent ($\rho = 0.6$).

To test the robustness of the proposed method, we consider a fourth simulation example, which is similar to the third example with a change in the covariate generating process. In particular, the rows of the $n \times p$ design matrix Z are sampled as independent copies from a mixture of two t-distributions via the fol-

Table 6. *Average results over 100 repetitions for Example C with all entries in the top panel multiplied by 100 when* $\rho = 0$.

Consistent selection probability with sure screening probability in parentheses									
p	AIC	BIC	EBIC	GIC	GAIC	GBIC	GBIC$_p$	HGBIC$_p$	Oracle
100	0(100)	12(100)	39(100)	36(100)	25(100)	12(100)	44(100)	97(100)	100(100)
200	0(100)	12(100)	57(100)	60(100)	23(100)	18(100)	54(100)	98(100)	100(100)
400	0(100)	15(100)	71(100)	73(100)	15(100)	21(100)	54(100)	100(100)	100(100)
800	0(100)	14(100)	73(100)	81(100)	13(100)	15(100)	59(100)	100(100)	100(100)
1600	0(100)	12(100)	81(100)	85(100)	8(100)	18(100)	52(100)	99(100)	100(100)
3200	0(100)	12(100)	82(100)	92(100)	5(100)	13(100)	41(100)	99(100)	100(100)
Mean prediction error with standard error in parentheses									
100	577(39)	424(32)	381(25)	381(25)	396(28)	423(32)	367(25)	334(21)	333(21)
200	451(16)	289(11)	254(8)	250(8)	279(10)	284(11)	250(8)	234(7)	233(7)
400	378(16)	232(9)	201(6)	201(6)	240(9)	223(7)	205(6)	192(5)	192(5)
800	337(12)	205(7)	173(4)	171(4)	211(8)	198(6)	176(4)	166(3)	166(3)
1600	313(11)	191(7)	159(4)	158(4)	206(7)	182(5)	169(5)	154(4)	154(4)
3200	284(9)	174(6)	145(3)	143(3)	189(7)	171(5)	153(4)	142(3)	142(3)

Table 7. *Average false positives over 100 repetitions for Example C when* $\rho = 0$.

p	AIC	BIC	EBIC	GIC	GAIC	GBIC	GBIC$_p$	HGBIC$_p$
100	25.14	3.42	1.60	1.65	3.05	3.27	1.14	0.03
200	34.21	3.71	1.03	0.79	3.92	3.14	0.90	0.02
400	35.84	3.72	0.63	0.57	5.33	2.86	0.98	0.00
800	40.47	4.02	0.40	0.27	5.07	2.96	0.80	0.00
1600	38.80	4.28	0.24	0.20	7.83	2.87	1.07	0.01
3200	38.05	4.07	0.23	0.08	6.83	3.34	1.05	0.01

Table 8. *Average results over 100 repetitions for Example C with all entries in the top panel multiplied by 100 when* $\rho = 0.6$.

Consistent selection probability with sure screening probability in parentheses									
p	AIC	BIC	EBIC	GIC	GAIC	GBIC	GBIC$_p$	HGBIC$_p$	Oracle
100	0(100)	1(100)	13(100)	12(100)	42(100)	1(100)	39(100)	89(100)	100(100)
200	0(100)	1(100)	8(100)	8(100)	46(100)	1(100)	42(100)	92(100)	100(100)
400	0(100)	1(100)	20(100)	24(100)	34(100)	1(100)	38(100)	92(100)	100(100)
800	0(100)	0(100)	29(100)	31(100)	29(100)	1(100)	32(100)	94(100)	100(100)
1600	0(100)	1(100)	35(100)	48(100)	19(100)	2(100)	29(100)	95(100)	100(100)
3200	0(100)	0(100)	45(100)	59(100)	9(100)	2(100)	24(100)	97(100)	100(100)
Mean prediction error with standard error in parentheses									
100	1943(114)	1245(91)	1099(80)	1086(79)	955(73)	1204(84)	923(62)	796(55)	779(55)
200	2102(209)	1202(118)	1013(89)	998(88)	854(78)	1165(106)	864(80)	725(60)	715(60)
400	1301(88)	859(65)	652(50)	636(49)	659(49)	819(64)	593(44)	525(41)	511(39)
800	1160(84)	716(64)	536(47)	521(45)	575(58)	673(62)	537(48)	425(37)	416(32)
1600	1039(81)	624(48)	463(34)	442(33)	539(46)	589(46)	475(36)	377(29)	374(28)
3200	754(63)	439(29)	332(24)	318(23)	435(34)	426(30)	361(28)	275(16)	275(16)

lowing process. First, we generate $w^{(1)}_{p\times 1} \sim t_5(0, \Sigma)$, a multi-variate t-distribution with 5 degrees of freedom and scale matrix Σ, where $\Sigma_{ij} = 0.5^{|i-j|}$ for all i and j. Then, we define $x^{(1)}_{p\times 1} = (x^{(1)}_1, \ldots, x^{(1)}_p)^T$ as a scaled version of $w^{(1)}$ to the unit interval, where $x^{(1)}_j = (w^{(1)}_j - \min_k w^{(1)}_k)/(\max_k w^{(1)}_k - \min_k w^{(1)}_k)$, for $j = 1, \ldots, p$. Following a similar process, we generate an independent $w^{(2)}_{p\times 1} \sim t_4(0, I_p)$, a multi-
625 variate t-distribution with 4 degrees of freedom and scale matrix I_p. And $x^{(2)}_{p\times 1}$ is its corresponding

Table 9. *Average false positives over 100 repetitions for Example C when* $\rho = 0.6$.

p	AIC	BIC	EBIC	GIC	GAIC	GBIC	GBIC$_p$	HGBIC$_p$
100	30.37	5.83	3.18	2.95	1.53	5.22	1.29	0.13
200	36.36	5.70	2.89	2.51	1.27	5.04	1.09	0.11
400	29.86	5.93	1.81	1.59	2.03	4.64	1.01	0.09
800	34.47	6.14	1.58	1.33	3.04	4.86	1.40	0.06
1600	34.24	5.68	1.43	0.91	3.82	4.44	1.47	0.05
3200	34.01	5.76	0.93	0.67	5.62	4.46	1.79	0.03

Table 10. *Average results over 100 repetitions for Example C with the mixture distribution and with all entries in the top panel multiplied by 100.*

Consistent selection probability with sure screening probability in parentheses									
p	AIC	BIC	EBIC	GIC	GAIC	GBIC	GBIC$_p$	HGBIC$_p$	Oracle
100	0(100)	4(100)	30(100)	28(100)	30(100)	5(100)	44(100)	92(100)	100(100)
200	0(100)	10(100)	40(100)	43(100)	27(100)	13(100)	54(100)	92(100)	100(100)
400	0(100)	10(100)	40(100)	46(100)	16(100)	10(100)	37(100)	99(100)	100(100)
800	0(100)	8(100)	59(100)	67(100)	15(100)	12(100)	46(100)	97(100)	100(100)
1600	0(100)	8(100)	59(100)	70(100)	11(100)	12(100)	39(100)	98(100)	100(100)
3200	0(100)	8(100)	65(100)	77(100)	13(100)	13(100)	51(100)	99(100)	100(100)
Mean prediction error with standard error in parentheses									
100	1565(216)	1194(187)	1077(174)	1079(174)	1069(176)	1175(186)	1016(162)	958(158)	945(158)
200	932(98)	625(76)	539(71)	533(71)	616(87)	609(76)	533(72)	480(63)	474(63)
400	1025(114)	623(70)	537(69)	527(68)	649(74)	589(66)	543(70)	468(56)	468(56)
800	625(65)	397(40)	314(29)	304(27)	404(49)	370(32)	328(30)	284(26)	283(26)
1600	635(67)	361(40)	271(28)	268(28)	355(41)	331(35)	290(35)	255(27)	253(27)
3200	569(52)	310(23)	230(14)	225(14)	319(29)	295(22)	242(16)	218(14)	216(13)

scaled version. Subsequently, we generate $x_{p\times 1}$ according to the mixture distribution $x_{p\times 1} = \iota x^{(1)}_{p\times 1} + (1-\iota)x^{(2)}_{p\times 1}$, where ι is an independent Bernoulli random variable with distribution Ber(0.5). The remaining data generation process and fitting procedure are the same as in the third simulation example.

Tables 10 and 11 show similar conclusions as the tables for the third simulation example. Again HGBIC$_p$ outperformed all other model selection criteria in terms of higher consistent selection probability, smaller mean squared prediction error, and much smaller false positives. This demonstrates the robustness of the proposed model selection criteria.

Table 11. *Average false positives over 100 repetitions for Example C with the mixture distribution.*

p	AIC	BIC	EBIC	GIC	GAIC	GBIC	GBIC$_p$	HGBIC$_p$
100	26.80	4.95	2.24	2.16	2.35	4.25	1.15	0.08
200	32.39	5.07	1.75	1.49	3.06	4.04	0.84	0.08
400	35.19	4.53	1.35	1.02	5.25	3.43	1.06	0.01
800	34.15	4.46	0.84	0.64	4.89	3.44	1.07	0.03
1600	38.36	4.36	0.54	0.36	4.92	3.01	1.06	0.02
3200	35.84	4.87	0.52	0.28	5.19	3.55	0.97	0.01

D. DISCUSSION ON ASSUMPTION 5

Assumption 5 is about the continuity of the matrix-valued functions $V_n(\beta)$ and $\widetilde{V}_n(\beta_1, \ldots, \beta_d)$ at $\beta_{n,0}$. To obtain some further insights, let us expand $V_n(\beta)$ in terms of the design matrix X, the covariance matrix of response vector $\mathrm{cov}(Y)$, and the link function $b(\theta)$ as

$$V_n(\beta) = \{X^T \mathrm{cov}(Y) X\}^{-1/2} X^T \Sigma(X\beta) X \{X^T \mathrm{cov}(Y) X\}^{-1/2},$$

where $\Sigma(\theta) = \mathrm{diag}\{b''(\theta_1), \ldots, b''(\theta_n)\}$. Assumption 5 puts some restrictions on the minimum and maximum eigenvalues of $V_n(\beta) - V_n$ in a shrinking neighborhood of $\beta_{n,0}$. In particular, $V_n(\beta) - V_n$ has the same spectrum as the corresponding matrix $\{X^T \mathrm{cov}(Y) X\}^{1/2} \{V_n(\beta) - V_n\} \{X^T \mathrm{cov}(Y) X\}^{-1/2}$, which can be simplified as

$$\begin{aligned}
&\{X^T \mathrm{cov}(Y) X\}^{1/2} \{V_n(\beta) - V_n\} \{X^T \mathrm{cov}(Y) X\}^{-1/2} \\
&= \{X^T \Sigma(X\beta) X - X^T \Sigma(X\beta_{n,0}) X\} \{X^T \mathrm{cov}(Y) X\}^{-1} \\
&= X^T \{\Sigma(X\beta) - \Sigma(X\beta_{n,0})\} X \{X^T \mathrm{cov}(Y) X\}^{-1}.
\end{aligned}$$

Hence, Assumption 5 can be regarded as the continuity condition of the matrix-valued function $X^T \{\Sigma(X\beta) - \Sigma(X\beta_{n,0})\} X \{X^T \mathrm{cov}(Y) X\}^{-1}$ at $\beta_{n,0}$. Moreover, the only factor that depends on β in this expression is the difference $\Sigma(X\beta) - \Sigma(X\beta_{n,0})$, which is a diagonal matrix. The diagonal entries are given by $b''(\cdot)$, which is assumed to be a continuous function. Thus the difference $\Sigma(X\beta) - \Sigma(X\beta_{n,0})$ can be kept small in an appropriately scaled neighborhood of $\beta_{n,0}$. A similar analysis can be conducted on $\widetilde{V}_n(\beta_1, \ldots, \beta_d)$.

Furthermore, it is easy to see that Assumption 5 is satisfied under the linear model. Indeed, under linear model, the second derivative of the link function is constant, so are both functions $V_n(\beta)$ and $\widetilde{V}_n(\beta_1, \ldots, \beta_d)$. Hence, the differences in the assumption are zero matrices, and Assumption 5 holds naturally. For the logistic regression model and Poisson regression model, the second derivative of the link function is a continuous function $b''(\theta) = \exp(\theta)/\{1 + \exp(\theta)\}^2$ and $b''(\theta) = \exp(\theta)$, respectively.

E. PROOFS OF MAIN RESULTS

We provide the proofs of Theorems 1–3 in this section. We aim to establish the asymptotic consistency of the maximum likelihood estimator uniformly over all models $\mathfrak{M}_m$ such that $|\mathfrak{M}_m| \leq K$ where $K = o(n)$. For this purpose, we extend our notation. $\beta_{n,0}(\mathfrak{M}_m)$ denotes the parameter vector for the working model and is defined as the minimizer of the Kullback–Leibler divergence whose support is $\mathfrak{M}_m$: $\beta_{n,0}(\mathfrak{M}_m) = \arg\min_{\beta \in \mathcal{B}(\mathfrak{M}_m)} I\{g_n; f_n(\cdot; \beta, \tau)\}$. $\beta_{n,0}(\mathfrak{M}_m)$ is estimated by the maximum likelihood estimator $\widehat{\beta}(\mathfrak{M}_m)$ which is defined as $\widehat{\beta}(\mathfrak{M}_m) = \arg\max_{\beta \in \mathcal{B}(\mathfrak{M}_m)} \ell_n(\beta)$.

E.1. *Proof of Theorem* 1

We consider the decomposition of $S(y, \mathfrak{M}_m; F_n)$ in (13) and deal with terms $\log[E_{\mu_{\mathfrak{M}_m}}\{U_n(\beta)^n\}]$ and $\log \alpha_{\mathfrak{M}_m}$ separately by invoking Taylor's expansion. In fact, $\log[E_{\mu_{\mathfrak{M}_m}}\{U_n(\beta)^n\}]$ is based on $\ell_n^*(y, \beta)$, the deviation of the quasi-log-likelihood from its maximum, while $\log \alpha_{\mathfrak{M}_m}$ is the log-prior probability which depends on $D_m = E[I\{g_n; f_n(\cdot, \widehat{\beta}_{n,m})\} - I\{g_n; f_n(\cdot, \beta_{n,m,0})\}]$, expected difference in the Kullback–Leibler divergences. In light of consistency of the estimator $\widehat{\beta}_n$ as shown in Lemma 1, we focus only on the neighborhood of $\beta_{n,0}$.

First, we make a few remarks on the technical details of the proof. Throughout the proof, we condition on the event $\widetilde{Q}_n = \{\widehat{\beta}_n \in N_n(\delta_n)\}$, where $N_n(\delta_n) = \{\beta \in \mathbb{R}^d : \|(n^{-1}B_n)^{1/2}(\beta - \beta_{n,0})\|_2 \leq (n/d)^{-1/2}\delta_n\}$, $B_n = X^T \mathrm{cov}(Y) X$, $\delta_n = O\{L_n(\log p)^{1/2}\}$ and $\widehat{\beta}_n$ is the unrestricted MLE. The eigenvalues of $n^{-1}A_n(\beta)$ and $n^{-1}B_n$ are bounded away from 0 and ∞ by Assumptions 1 and 3. This follows from the fact that eigenvalues of $M^T N M$ lie between $\lambda_{\min}(N)\lambda_{\min}(M^T M)$ and $\lambda_{\max}(N)\lambda_{\max}(M^T M)$ for any matrix M and positive semidefinite symmetric matrix N. Therefore, from Lemma 1 we have that $pr(\widetilde{Q}_n) \to 1$ as $n \to \infty$.

To establish this theorem we require a possibly dimension dependent bound on the quantity $\|n^{-1/2}X\widehat{\beta}_n\|_2$. This can be achieved by putting some restriction on the parameter space. Let $M_n(\alpha) = \{\beta \in \mathbb{R}^d : \|X\beta\|_\infty \leq \alpha \log n\}$ be a neighborhood, where α is some positive constant. One way of bounding the quantity $\|n^{-1/2}X\widehat{\beta}_n\|_2$ is to restrict the maximum likelihood estimator $\widehat{\beta}_n$ on the set $M_n(\alpha)$. Here, the constant α can be chosen as large as desired to make $M_n(\alpha)$ large enough, whereas the neighborhood $N_n(\delta_n)$ is asymptotically shrinking. Then, we have $N_n(\delta_n) \subset M_n(\alpha)$ for all sufficiently large n, which implies that conditional on $\widetilde{Q}_n$, the restricted MLE coincides with its unrestricted version. Hereafter in this proof, $\widehat{\beta}_n$ will be referred to as the restricted ML unless specified otherwise.

Part I: expansion of the term $\log[E_{\mu_{\mathfrak{M}_m}}\{U_n(\beta)^n\}]$. Recall that $U_n(\beta) = \exp\{n^{-1}\ell_n^*(y,\beta\}$ and $\ell_n^*(y,\beta) = \ell_n(y,\beta) - \ell_n(y,\widehat{\beta}_n)$. First, we observe that the maximum value of the function $\ell_n^*(y,\beta)$ is attained at $\beta = \widehat{\beta}_n$. Moreover, we have $\partial^2\ell_n^*(y,\beta)/\partial\beta^2 = -A_n(\beta)$ from (8) where $A_n(\beta) = X^T\Sigma(X\beta)X$. Then, we consider Taylor's expansion of the log-likelihood function $\ell_n(y,\cdot)$ around $\widehat{\beta}_n$ in a new neighborhood $\widetilde{N}_n(\delta_n) = \{\beta \in \mathbb{R}^d : \|(n^{-1}B_n)^{1/2}(\beta - \widehat{\beta}_n)\|_2 \leq (n/d)^{-1/2}\delta_n\}$. We get

$$\begin{aligned}\ell_n^*(y,\beta) &= \frac{1}{2}(\beta - \widehat{\beta}_n)^T\left\{\partial^2\ell_n^*(y,\beta_*)/\partial\beta^2\right\}(\beta - \widehat{\beta}_n) \qquad \text{(A.1)}\\ &= -\frac{n}{2}\delta^T V_n(\beta_*)\delta,\end{aligned}$$

where β_* lies on the line segment joining β and $\widehat{\beta}_n$, $\delta = n^{-1/2}B_n^{1/2}(\beta - \widehat{\beta}_n)$, and $V_n(\beta) = B_n^{-1/2}A_n(\beta)B_n^{-1/2}$. Since $\widehat{\beta}_n \in \widetilde{N}_n(\delta_n)$, by the convexity of the neighborhood $\widetilde{N}_n(\delta_n)$ we have $\beta_* \in \widetilde{N}_n(\delta_n)$. We also note that conditional on the event $\widetilde{Q}_n$, it holds that $\widetilde{N}_n(\delta_n) \subset N_n(2\delta_n)$.

Now, we will bound $U_n(\beta)^n$ over the region $\widetilde{N}_n(\delta_n)$ using Taylor's expansion in (A.1). By Assumption 5, we get

$$q_1(\beta)1_{\widetilde{N}_n(\delta_n)}(\beta) \leq -n^{-1}\ell_n^*(y,\beta)1_{\widetilde{N}_n(\delta_n)}(\beta) \leq q_2(\beta)1_{\widetilde{N}_n(\delta_n)}(\beta), \qquad \text{(A.2)}$$

where $q_1(\beta) = \frac{1}{2}\delta^T\{V_n - \rho_n(\delta_n)I_d\}\delta$ and $q_2(\beta) = \frac{1}{2}\delta^T\{V_n + \rho_n(\delta_n)I_d\}\delta$. Then, we consider the linear transformation $h(\beta) = (n^{-1}B_n)^{1/2}\beta$. For sufficiently large n, we obtain

$$E_{\mu_{\mathfrak{M}_m}}\{e^{-nq_2(\beta)}1_{\widetilde{N}_n(\delta_n)}(\beta)\} \leq E_{\mu_{\mathfrak{M}_m}}\{U_n(\beta)^n 1_{\widetilde{N}_n(\delta_n)}(\beta)\} \leq E_{\mu_{\mathfrak{M}_m}}\{e^{-nq_1(\beta)}1_{\widetilde{N}_n(\delta_n)}(\beta)\}, \qquad \text{(A.3)}$$

where $\mu_{\mathfrak{M}_m}$ denotes the prior distribution on $h(\beta) \in \mathbb{R}^d$ for the model $\mathfrak{M}_m$.

The final expansion of $\log[E_{\mu_{\mathfrak{M}_m}}\{U_n(\beta)^n\}]$ results from combination of Lemmas 7–10. The expressions $E_{\mu_{\mathfrak{M}_m}}\{U_n(\beta)^n 1_{\widetilde{N}_n^c(\delta_n)}\}$ and $\int_{\delta\in\mathbb{R}^d} e^{-nq_j}1_{\widetilde{N}_n^c(\delta_n)}d\mu_0$ for $j = 1, 2$ in Lemmas 8 and 10 converge to zero faster than any polynomial rate in n since $\widetilde{\kappa}_n = \lambda_{\min}(V_n)/2$ is bounded away from 0. Moreover, Lemmas 7 and 9 yield

$$\log[E_{\mu_{\mathfrak{M}_m}}\{U_n(\beta)^n\}] = \log\left\{\left(\frac{2\pi}{n}\right)^{d/2}|V_n \pm \rho_n(\delta_n)I_d|^{-1/2}\right\} + \log c_4,$$

where $c_4 \in [c_3, c_3^{-1}]$. Finally, we observe that

$$\begin{aligned}|V_n \pm \rho_n(\delta_n)I_d|^{-1/2} &= |V_n|^{-1/2}|I_d \pm \rho_n(\delta_n)V_n^{-1}|^{-1/2} = |V_n|^{-1/2}[1 + O\{\rho_n(\delta_n)\mathrm{tr}(V_n^{-1})\}]^{-1/2}\\ &= |V_n|^{-1/2}[1 + O\{\rho_n(\delta_n)d\lambda_{\min}^{-1}(V_n)\}]^{-1/2} = |V_n|^{-1/2}\{1 + o(1)\},\end{aligned}$$

where we use Assumption 5. So, we obtain

$$\begin{aligned}\log[E_{\mu_{\mathfrak{M}_m}}\{U_n(\beta)^n\}] &= \log\left[\left(\frac{2\pi}{n}\right)^{d/2}|V_n|^{-1/2}\{1 + o(1)\}\right] + \log c_4\\ &= -\frac{\log n}{2}d + \frac{1}{2}\log|A_n^{-1}B_n| + \frac{\log(2\pi)}{2}d + \log c_4 + o(1). \qquad \text{(A.4)}\end{aligned}$$

This completes the expansion of $\log[E_{\mu_{\mathfrak{M}_m}}\{U_n(\beta)^n\}]$.

Part II: expansion of the prior term $\log \alpha_{\mathfrak{M}_m}$. Now, we consider the prior term $\log \alpha_{\mathfrak{M}_m}$ which depends on $\widehat{\beta}_n$ through D_m. Simple calculation shows that

$$\log \alpha_{\mathfrak{M}_m} = -D_m + \log C - d \log p. \tag{A.5}$$

We aim to provide a decomposition of D_m in terms of H_n. Observe that $-D_m = E\{\eta_n(\widehat{\beta}_n)\} - \eta_n(\beta_{n,0})$ where $\eta_n(\beta) = E\{\ell_n(\widetilde{y}, \beta)\}$, and $\widetilde{y}$ is an independent copy of y. We expand $E\{\eta_n(\widehat{\beta}_n)\}$ around $\eta_n(\beta_{n,0})$. In the generalized linear models setup, we observe that $\ell_n(\widetilde{y}, \beta) = \widetilde{y}^T X\beta - 1^T b(X\beta)$ and $\eta_n(\beta) = \{E(\widetilde{y}^T)\}X\beta - 1^T b(X\beta)$. Then, we split $E\{\eta_n(\widehat{\beta}_n)\}$ in the region $\widetilde{Q}_n$ and its complement, that is,

$$\begin{aligned} E\{\eta_n(\widehat{\beta}_n)\} &= E\{\eta_n(\widehat{\beta}_n)1_{\widetilde{Q}_n}\} + E\{\eta_n(\widehat{\beta}_n)1_{\widetilde{Q}_n^c}\} \\ &= E\{\eta_n(\widehat{\beta}_n)1_{\widetilde{Q}_n}\} + E[\{(E\widetilde{y})^T(X\widehat{\beta}_n) - 1^T b(X\widehat{\beta}_n)\}1_{\widetilde{Q}_n^c}]. \end{aligned} \tag{A.6}$$

First, we aim to show that the second term on the right-hand side of (A.6) is $o(1)$. Performing componentwise Taylor's expansion of $b(\cdot)$ around 0 and evaluating at $X\widehat{\beta}_n$, we obtain $b(X\widehat{\beta}_n) = b(0) + b'(0)X\widehat{\beta}_n + r$, where $r = (r_1, \ldots, r_n)^T$ with $r_i = 2^{-1}b''\{(X\beta_i^*)_i\}(X\widehat{\beta}_n)_i^2$ and $\beta_1^*, \ldots, \beta_n^*$ lying on the line segment joining $\widehat{\beta}_n$ and 0. Thus, we get

$$E[|\{E(\widetilde{y})\}^T X\widehat{\beta}_n - 1^T b(X\widehat{\beta}_n) \mid 1_{\widetilde{Q}_n^c}] \leq O\{n\log n + n + n(\log n)^2\}pr(\widetilde{Q}_n^c) = o(1) \tag{A.7}$$

for sufficiently large n. The last inequality follows from the fact that $pr(\widetilde{Q}_n^c)$ converges to zero faster than any polynomial rate. To verify the orders, we recall that $\widehat{\beta}_n$ is the constrained MLE and $b''(\cdot)$ is bounded away from 0 and ∞. Thus, we obtain following bounds for the four terms $|\{E(\widetilde{y})\}^T X\widehat{\beta}_n| = O(n\log n)$, $|1^T b(0)| = O(n)$, $|b'(0)1^T X\widehat{\beta}_n| = O(n\log n)$, and $|1^T r| = O\{n(\log n)^2\}$.

Now, we consider the first term on the right-hand side of (A.6). We begin by expanding $\eta_n(\beta)$ around $\beta_{n,0}$ conditioned on the event $\widetilde{Q_n}$. By the definition of $\beta_{n,0}$, $\eta_n(\beta)$ attains its maximum at $\beta_{n,0}$. By evaluating Taylor's expansion of $\eta_n(\cdot)$ around $\beta_{n,0}$ at $\widehat{\beta}_n$, we derive

$$\begin{aligned} \eta_n(\widehat{\beta}_n) &= \eta_n(\beta_{n,0}) - \frac{1}{2}(\widehat{\beta}_n - \beta_{n,0})^T A_n(\beta^*)(\widehat{\beta}_n - \beta_{n,0}) \\ &= \eta_n(\beta_{n,0}) - \frac{1}{2}(\widehat{\beta}_n - \beta_{n,0})^T A_n(\widehat{\beta}_n - \beta_{n,0}) - \frac{s_n}{2}, \end{aligned}$$

where $A_n(\cdot) = -\partial^2\ell_n(y,\cdot)/\partial\beta^2$, $A_n = A_n(\beta_{n,0})$, and β^* is on the line segment joining $\beta_{n,0}$ and $\widehat{\beta}_n$. The second equality is obtained by taking $s_n = (\widehat{\beta}_n - \beta_{n,0})^T\{A_n(\beta^*) - A_n\}(\widehat{\beta}_n - \beta_{n,0})$. Furthermore, setting $C_n = B_n^{-1/2}A_n$ and $v_n = C_n(\widehat{\beta}_n - \beta_{n,0})$ simplifies the above expression to

$$\eta_n(\widehat{\beta}_n) = \eta_n(\beta_{n,0}) - \frac{1}{2}v_n^T\{(C_n^{-1})^T A_n C_n^{-1}\}v_n - \frac{s_n}{2}. \tag{A.8}$$

In (A.8), we first handle the term s_n. On the event $\widetilde{Q_n}$, by the convexity of the neighborhood $N_n(\delta_n)$ we have $\beta^* \in N_n(\delta_n)$. Then, Assumption 5 implies

$$\begin{aligned} \left|s_n 1_{\widetilde{Q_n}}\right| &= \left|(\widehat{\beta}_n - \beta_{n,0})^T\{A_n(\beta^*) - A_n\}(\widehat{\beta}_n - \beta_{n,0})\right| 1_{\widetilde{Q_n}} \\ &= \left|\{B_n^{1/2}(\widehat{\beta}_n - \beta_{n,0})\}^T\{V_n(\beta^*) - V_n\}\{B_n^{1/2}(\widehat{\beta}_n - \beta_{n,0})\}\right| 1_{\widetilde{Q_n}} \\ &\leq \rho_n(\delta_n)\delta_n^2 d 1_{\widetilde{Q_n}}, \end{aligned} \tag{A.9}$$

where $V_n(\cdot) = B^{-1/2}A_n(\cdot)B_n^{-1/2}$ and $V_n = V(\beta_{n,0})$. We then deduce that $E(s_n 1_{\widetilde{Q_n}}) = o(1)$, since $\rho_n(\delta_n)\delta_n^2 d 1_{\widetilde{Q_n}} = o(1)$ by Assumption 5. Therefore, (A.8) becomes

$$E\{\eta_n(\widehat{\beta}_n)1_{\widetilde{Q_n}}\} = E[\eta_n(\beta_{n,0}) - \frac{1}{2}v_n^T\{(C_n^{-1})^T A_n C_n^{-1}\}v_n 1_{\widetilde{Q_n}}] + o(1). \tag{A.10}$$

We provide a decomposition of v_n to handle the term $v_n^T\{(C_n^{-1})^T A_n C_n^{-1}\}v_n$ in (A.10). Define $\Psi(\beta_n) = X^T\{y - \mu(X\beta_n)\}$. From the score equation we have $\Psi(\widehat{\beta}_n) = 0$. From (6), it holds that $X^T\{E(y) - \mu(X\beta_{n,0})\} = 0$. For any $\beta_1, \ldots, \beta_d \in \mathbb{R}^d$, denote by $\widetilde{A}_n(\beta_1, \ldots, \beta_d)$ a $d \times d$ matrix with jth row the corresponding row of $A_n(\beta_j)$ for each $j = 1, \ldots, d$. Then, we define matrix-valued function $\widetilde{V}_n(\beta_1, \ldots, \beta_d) = B_n^{-1/2}\widetilde{A}_n(\beta_1, \ldots, \beta_d)B_n^{-1/2}$. Assuming the differentiability of $\Psi(\cdot)$ and applying the mean-value theorem componentwise around $\beta_{n,0}$, we obtain

$$\begin{aligned} 0 = \Psi_n(\widehat{\beta}_n) &= \Psi_n(\beta_{n,0}) - \widetilde{A}_n(\beta_1, \ldots, \beta_d)(\widehat{\beta}_n - \beta_{n,0}) \\ &= X^T\{y - E(y)\} - \widetilde{A}_n(\beta_1, \ldots, \beta_d)(\widehat{\beta}_n - \beta_{n,0}), \end{aligned}$$

where each of $\beta_1, \ldots, \beta_d$ lies on the line segment joining $\widehat{\beta}_n$ and $\beta_{n,0}$. Therefore, we have the decomposition

$$v_n = C_n(\widehat{\beta}_n - \beta_{n,0}) = u_n + w_n, \tag{A.11}$$

where $u_n = B_n^{-1/2}X^T\{y - E(y)\}$ and $w_n = -\{\widetilde{V}_n(\beta_1, \ldots, \beta_d) - V_n\}\{B_n^{1/2}(\widehat{\beta}_n - \beta_{n,0})\}$.

We handle the quadratic term $v_n^T\{(C_n^{-1})^T A_n C_n^{-1}\}v_n$ in (A.10) by using the decomposition of v_n. For simplicity of notation, denote by $R_n = (C_n^{-1})^T A_n C_n^{-1}$. Recall that $C_n = B_n^{-1/2}A_n$. With some calculations we obtain

$$\begin{aligned} E(u_n^T R_n u_n) &= E[\{y - E(y)\}^T X A_n^{-1} X^T\{y - E(y)\}] \\ &= E(\mathrm{tr}[A_n^{-1}X^T\{y - E(y)\}\{y - E(y)\}^T X]) = \mathrm{tr}(A_n^{-1}B_n). \end{aligned}$$

We get $E(u_n^T R_n u_n 1_{\widetilde{Q_n}}) = E(u_n^T R_n u_n) - E(u_n^T R_n u_n 1_{\widetilde{Q_n}^c})$. From Lemma 1, we have $pr(\widetilde{Q_n}^c) \to 0$ as $n \to \infty$. We set $\widetilde{\mu}_n = max\{\mathrm{tr}(A_n^{-1}B_n), 1\}$, hereby μ_n is bounded away from zero. We apply Vitali's convergence theorem to show that $E(u_n^T R_n u_n 1_{\widetilde{Q_n}^c}) = o(\widetilde{\mu}_n)$. To establish uniform integrability, we use Lemma 6 which states that $\sup_n E\{|(u_n^T R_n u_n)/\widetilde{\mu}_n|^{1+\gamma}\} < \infty$ for some constant $\gamma > 0$. This leads to $E(u_n^T R_n u_n 1_{\widetilde{Q_n}^c}) = o(\widetilde{\mu}_n)$. Hence we have

$$\frac{1}{2}E(u_n^T R_n u_n 1_{\widetilde{Q_n}}) = \frac{1}{2}\mathrm{tr}(A_n^{-1}B_n) + o(\widetilde{\mu}_n). \tag{A.12}$$

Now, it remains to show that

$$E\{(w_n^T R_n w_n + 2w_n^T R_n u_n)1_{\widetilde{Q_n}}\} = o(\widetilde{\mu}_n). \tag{A.13}$$

Using the definition of R_n and w_n, we can bound $w_n^T R_n w_n$:

$$w_n^T R_n w_n = \|R_n^{1/2}w_n\|_2^2 \le \|\widetilde{V}_n(\beta_1, \ldots, \beta_d) - V_n\|_2^2 \delta_n^2 d\,\mathrm{tr}(A_n^{-1}B_n).$$

So, on the event $\widetilde{Q_n}$, it holds that $E(w_n^T R_n w_n 1_{\widetilde{Q_n}}) = o(\widetilde{\mu}_n)$ by Assumption 5. For the cross term $w_n^T R_n u_n$, applying the Cauchy–Schwarz inequality yields

$$\begin{aligned} |E(w_n^T R_n u_n 1_{\widetilde{Q_n}})| &\le E(\|R_n^{1/2}w_n\|_2^2 1_{\widetilde{Q_n}})^{1/2} E(\|u_n^T R_n^{1/2}\|_2^2)^{1/2} \\ &\le E\{\|\widetilde{V}_n(\beta_1, \ldots, \beta_d) - V_n\|_2 1_{\widetilde{Q_n}} \delta_n d^{1/2}\mathrm{tr}(A_n^{-1}B_n)\}. \end{aligned}$$

Thus, we obtain that $E(w_n^T R_n u_n 1_{\widetilde{Q_n}}) = o(\widetilde{\mu}_n)$. $E\{|\eta_n(\beta_{n,0})|1_{\widetilde{Q_n}^c}\}$ is of order $o(1)$ by similar calculations as in (A.7). Then, combining (A.6), (A.10), (A.12), and (A.13) yields

$$E\{\eta_n(\widehat{\beta}_n)\} = \eta_n(\beta_{n,0}) - \frac{1}{2}\mathrm{tr}(A_n^{-1}B_n) + o(\widetilde{\mu}_n). \tag{A.14}$$

From (A.5) and (A.14), we can obtain the expansion

$$\log \alpha_{\mathfrak{M}_m} = -\frac{1}{2}\mathrm{tr}(A_n^{-1}B_n) + \log C - d\log p + o(\widetilde{\mu}_n).$$

Therefore, combining Parts I and II completes the proof of Theorem 1.

E.2. *Proof of Theorem* 2

At the beginning of the proof, we demonstrate that the theorem follows from the consistency of $\widehat{A}_n$ and $\widehat{B}_n$. Next, we establish the consistency of $\widehat{A}_n$ and $\widehat{B}_n$. The consistency of $\widehat{A}_n$ follows directly from the Lipschitz assumption; however, the consistency of $\widehat{B}_n$ is harder to prove. To accomplish this, we break down $\widehat{B}_n$ and invoke Bernstein-type tail inequalities and concentration theorems to handle challenging pieces.

We first introduce some notation to simplify the presentation of the proof. $\lambda_k(\cdot)$ denotes the eigenvalues arranged in increasing order. Denote the spectral radius of $d \times d$ square matrix M by $\rho(M) = \max_{1\le k\le d}\{|\lambda_k(M)|\}$. $\|\cdot\|_2$ denotes the matrix operator norm. $o_P(\cdot)$ denotes the convergence in probability of the matrix operator norm.

We want to show that $\log|\widehat{H}_n| = \log|H_n| + o_P(1)$ and $\mathrm{tr}(\widehat{H}_n) = \mathrm{tr}(H_n) + o_P(1)$. To establish both equalities, it is enough to show that $\widehat{H}_n = H_n + o_P(1/d)$. Indeed, assume that $\widehat{H}_n = H_n + o_P(1/d)$ is established. In that case, we observe that

$$|\mathrm{tr}(\widehat{H}_n) - \mathrm{tr}(H_n)| = |\mathrm{tr}(\widehat{H}_n - H_n)| \le d\rho(\widehat{H}_n - H_n) = d\|\widehat{H}_n - H_n\|_2 = o_P(1),$$

where the equality of the spectral radius and the operator norm follows from the symmetry of the matrix $\widehat{H}_n - H_n$. Moreover, we have

$$\begin{aligned}
|\log|\widehat{H}_n| - \log|H_n|| &\le d \max_{1\le k\le d} |\log\lambda_k(\widehat{H}_n) - \log\lambda_k(H_n)| \\
&= d \max_{1\le k\le d} \log\left[\max\left\{\frac{\lambda_k(\widehat{H}_n)}{\lambda_k(H_n)}, \frac{\lambda_k(H_n)}{\lambda_k(\widehat{H}_n)}\right\}\right] \\
&\le d \max_{1\le k\le d} \left[\max\left\{\frac{\lambda_k(\widehat{H}_n)}{\lambda_k(H_n)}, \frac{\lambda_k(H_n)}{\lambda_k(\widehat{H}_n)}\right\} - 1\right] \\
&\le d \max_{1\le k\le d} \frac{|\lambda_k(\widehat{H}_n) - \lambda_k(H_n)|}{\min\{\lambda_k(\widehat{H}_n), \lambda_k(H_n)\}}.
\end{aligned} \tag{A.15}$$

Recall that the smallest and largest eigenvalues of both $n^{-1}B_n$ and $n^{-1}A_n$ are bounded away from 0 and ∞. (See the note in the beginning of the proof of Theorem 1.) So, we get $\lambda_k(H_n) = O(1)$ and $\lambda_k^{-1}(H_n) = O(1)$ uniformly for all $1 \le k \le d$. An application of Weyl's theorem shows that $|\lambda_k(\widehat{H}_n) - \lambda_k(H_n)| \le \rho(\widehat{H}_n - H_n)$ for each k. We have $\rho(\widehat{H}_n - H_n) = \|\widehat{H}_n - H_n\|_2 = o_P(1/d)$. Hence, the right-hand side of (A.15) is $o_P(1)$.

Now, we proceed to show that $\widehat{H}_n = H_n + o_P(1/d)$. It suffices to prove that $n^{-1}\widehat{A}_n = n^{-1}A_n + o_P(1/d)$ and $n^{-1}\widehat{B}_n = n^{-1}B_n + o_P(1/d)$. To see the sufficiency,

$$\begin{aligned}
\widehat{H}_n - H_n &= (n^{-1}\widehat{A}_n)^{-1}(n^{-1}d\widehat{B}_n) - (n^{-1}A_n)^{-1}(n^{-1}dB_n) \\
&= (n^{-1}\widehat{A}_n)^{-1}(n^{-1}d\widehat{B}_n) - (n^{-1}\widehat{A}_n)^{-1}(n^{-1}dB_n) \\
&\quad + (n^{-1}\widehat{A}_n)^{-1}(n^{-1}dB_n) - (n^{-1}A_n)^{-1}(n^{-1}dB_n).
\end{aligned}$$

Then, $\widehat{H}_n = H_n + o_P(1/d)$ can be obtained by repeated application of the following properties of the operator norm: $\|(I_d - M)^{-1}\|_2 \leq 1/(1 - \|M\|_2)$ if $\|M\|_2 < 1$, $\|MN\|_2 \leq \|M\|_2\|N\|_2$, and $\|M + N\|_2 \leq \|M\|_2 + \|N\|_2$, where M and N are $d \times d$ matrices Horn & Johnson (1985).

Part 1: prove $n^{-1}\widehat{A}_n = n^{-1}A_n + o_P(1/d)$. From Lemma 1 we have, $\|\widehat{\beta}_n - \beta_{n,0}\|_2 = O_P\{(n/d)^{-1/2}\delta_n\}$, which entails $\widehat{\beta}_n = \beta_{n,0} + O_P\{(n/d)^{-1/2}\delta_n\}$. Then it follows from the Lipschitz assumption for $n^{-1}A_n(\beta)$ in the neighborhood $N_n(\delta_n)$ that $n^{-1}\widehat{A}_n = n^{-1}A_n + o_P(1/d)$.

Part 2: prove $n^{-1}\widehat{B}_n = n^{-1}B_n + o_P(1/d)$. We need to control the term $y - \mu(X\widehat{\beta}_n)$. In correctly specified models, $\mu(X\beta_{n,0})$ and $E(y)$ are the same. So, it is enough to introduce the mean $E(y)$ which is close to both y and $\mu(X\widehat{\beta}_n)$. However, it is harder to control the term $y - \mu(X\widehat{\beta}_n)$ in misspecified models since we need to deal with both $\mu(X\beta_{n,0})$ and $E(y)$.

First, we use the fact that $\mu(X\beta_{n,0})$ and $\mu(X\widehat{\beta}_n)$ are close. To accomplish this, we add and subtract $\mu(X\beta_{n,0})$ to get the following decomposition:

$$\begin{aligned} n^{-1}\widehat{B}_n &= n^{-1}X^T \mathrm{diag}\left[\left\{y - \mu(X\widehat{\beta}_n)\right\} \circ \left\{y - \mu(X\widehat{\beta}_n)\right\}\right] X \\ &= G_1 + G_2 + G_3, \end{aligned}$$

where

$$\begin{aligned} G_1 &= n^{-1}X^T \mathrm{diag}[\{y - \mu(X\beta_{n,0})\} \circ \{y - \mu(X\beta_{n,0})\}]X, \\ G_2 &= 2n^{-1}X^T \mathrm{diag}[\{y - \mu(X\beta_{n,0})\} \circ \{\mu(X\beta_{n,0}) - \mu(X\widehat{\beta}_n)\}]X, \\ G_3 &= n^{-1}X^T \mathrm{diag}[\{\mu(X\widehat{\beta}_n) - \mu(X\beta_{n,0})\} \circ \{\mu(X\widehat{\beta}_n) - \mu(X\beta_{n,0})\}]X. \end{aligned}$$

Next, we introduce $E(y)$ to obtain terms $y - E(y)$ and $E(y) - \mu(X\beta_{n,0})$ both of which can be kept small. We split G_1 as $G_1 = G_{11} + G_{12} + G_{13}$ and G_2 as $G_2 = G_{21} + G_{22}$, where

$$\begin{aligned} G_{11} &= n^{-1}X^T \mathrm{diag}[\{y - E(y)\} \circ \{y - E(y)\}]X, \\ G_{12} &= 2n^{-1}X^T \mathrm{diag}[\{y - E(y)\} \circ \{E(y) - \mu(X\beta_{n,0})\}]X, \\ G_{13} &= n^{-1}X^T \mathrm{diag}[\{E(y) - \mu(X\beta_{n,0})\} \circ \{E(y) - \mu(X\beta_{n,0})\}]X, \\ G_{21} &= 2n^{-1}X^T \mathrm{diag}[\{y - E(y)\} \circ \{\mu(X\beta_{n,0}) - \mu(X\widehat{\beta}_n)\}]X, \\ G_{22} &= 2n^{-1}X^T \mathrm{diag}[\{E(y) - \mu(X\beta_{n,0})\} \circ \{\mu(X\beta_{n,0}) - \mu(X\widehat{\beta}_n)\}]X. \end{aligned}$$

Now, we will control each of the above terms separately. Before we begin, we observe that for any matrices M and N, we have

$$\begin{aligned} pr(d\|M - N\|_2 \geq t) &\leq pr(d\|M - N\|_F \geq t) \\ &\leq d^2 \max_{1 \leq j,k \leq d} pr(|M^{jk} - N^{jk}| \geq t/d^2), \end{aligned} \tag{A.16}$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm and M^{jk} denotes the (j,k)th entry of M. Therefore, it is enough to bound $pr(|M^{jk} - N^{jk}| \geq t/d^2)$ by $o(1/d^2)$ to show that $M = N + o_p(1/d)$.

Part 2a) prove $G_{11} = n^{-1}B_n + o_P(1/d)$. We will use Bernstein-type tail inequality. First, $E(G_{11}) = n^{-1}B_n$ and $G_{11}^{jk} = n^{-1}\sum_{i=1}^n [x_{ij}x_{ik}\{y_i - E(y_i)\}^2] = \sum_{i=1}^n a_i^{jk} q_i^2$, where $a_i^{jk} = n^{-1}x_{ij}x_{ik}\mathrm{var}(y_i)$ and $q_i = \{\mathrm{var}(y_i)\}^{-1/2}\{y_i - E(y_i)\}$. Let $a\ a^{jk} = (a_1^{jk}, \ldots, a_n^{jk})^T$. Then we have $\|a^{jk}\|_2^2 = O(n^{4u_3 - 1})$ since $\|X\|_\infty = O(n^{u_3})$ from Assumption 3. It may be noted that q_i's are 1-sub-exponential random variables from Assumption 1 and so q_i^2's are 2-sub-exponential random variables. Furthermore, $\sup_{1 \leq i \leq n} \mathrm{var}(q_i^2) = O(1)$. To see this, we note

$$\mathrm{var}(q_i^2) \leq E(q_i^4) \leq 4^4[4^{-1}\{E(q_i^4)\}^{1/4}]^4 \leq 4^4 \left(\sup_{m \geq 1}\left[m^{-1}\{E(|q_i|^m)\}^{1/m}\right]\right)^4 = O(1),$$

where we use Lemma 5. Then combining (A.16) with Lemma 12 for a choice of $\alpha = 2$, we deduce

$$pr\{d\|G_{11} - E(G_{11})\|_2 \geq t\} \leq d^2 \max_{1\leq j,k\leq d} pr\{|G_{11}^{jk} - E(G_{11}^{jk})| \geq t/d^2\}$$
$$\leq Cd^2 \exp(-Ct^{1/2}n^{1/4-u_3}/d)$$

for some constant C. Since $d = O(n^{\kappa_1})$ and $u < 1/4 - u_3$, the right-hand side of above equation tends to zero. Thus, we obtain $G_{11} = E(G_{11}) + o_P(1/d) = n^{-1}B_n + o_P(1/d)$.

Part 2b) prove $G_{12} = o_P(1/d)$. Similar to the previous part, we invoke Bernstein-type tail inequality. Observe that $G_{12}^{jk} = n^{-1}\sum_{i=1}^n 2[x_{ij}x_{ik}\{E(y) - \mu(X\beta_{n,0})\}_i\{y_i - E(y_i)\}] = \sum_{i=1}^n \tilde{a}_i^{jk} q_i$, where $\tilde{a}_i^{jk} = 2n^{-1}\mathrm{var}(y_i)^{1/2}x_{ij}x_{ik}\{E(y) - \mu(X\beta_{n,0})\}_i$ and $q_i = \{\mathrm{var}(y_i)\}^{-1/2}\{y_i - E(y_i)\}$. Then, we get $\|\tilde{a}^{jk}\|_2^2 = O(n^{4u_3+u_2/2-3/2})$ by Assumptions 2 and 3.

By Lemma 11, we have

$$pr(d\|G_{12}\|_2 \geq t) \leq d^2 \max_{1\leq j,k\leq d} pr(|G_{12}^{jk}| \geq t/d^2)$$
$$\leq Cd^2 \exp\{-Ctn^{3/4-2u_3-u_2/4}/d^2\}$$

for some constant C. Since $d = O(n^{\kappa_1})$ and $3/4 - 2u_3 - u_2/4 - 2\kappa_1 > 0$, the right-hand side of above equation tends to zero. Hence, we have $G_{12} = o_P(1/d)$.

Part 2c) prove $G_{13} = o(1/d)$. We derive

$$\|G_{13}\|_2^2 \leq \|n^{-1}\sum_{i=1}^n (x_i x_i^T[E(y_i) - \{\mu(X\beta_{n,0})\}_i]^2)\|_F^2$$
$$= \Sigma_{1\leq j,k\leq d}\left(\sum_{i=1}^n a_i^{jk}[E(y_i) - \{\mu(X\beta_{n,0})\}_i]^2/\mathrm{var}(y_i)\right)^2$$
$$\leq \sum_{i=1}^n ([E(y_i) - \{\mu(X\beta_{n,0})\}_i]^2/\mathrm{var}(y_i))^2 \Sigma_{1\leq j,k\leq d}\|a^{jk}\|_2^2,$$

where the last step follows from the componentwise Cauchy–Schwarz inequality. From Assumptions 2 and 3, we get $\|G_{13}\|_2^2 = O(n^{u_2}d^2n^{4u_3-1})$. Therefore, $G_{13} = o(1/d)$ since $d = O(n^{\kappa_1})$ and $u_2 + 4\kappa_1 + 4u_3 - 1 < 0$.

Part 2d) prove $G_{21} = o(1/d^2)$. Bounding G_{21} is the trickiest part. The use of classical Bernstein-type inequalities are prohibited since the summation includes two random quantities y and $\widehat{\beta}$. Instead, we will apply concentration inequalities.

We start by truncating the random variable y by conditioning on the set $\Omega_n = \{\|W\|_\infty \leq C_1 \log n\}$ which is defined in Lemma 2. Since $\widehat{\beta}_n$ belongs to the neighborhood $N_n(\delta_n)$ by Lemma 1, we get

$$|G_{21}^{jk}| = |2n^{-1}\sum_{i=1}^n x_{ij}x_{ik}\{y_i - E(y_i)\}\{\mu(X\beta_{n,0}) - \mu(X\widehat{\beta}_n)\}_i|$$
$$\leq \sup_{\beta_n \in N_n(\delta_n)} 2n^{-1}|\sum_{i=1}^n x_{ij}x_{ik}\{y_i - E(y_i)\}\{\mu(X\beta_{n,0}) - \mu(X\beta_n)\}_i|.$$

Then, we can separate the right-hand side by conditioning on Ω_n. So, we have $|G_{21}^{jk}| \leq G_{211}^{jk} + G_{212}^{jk}$ where

$$G_{211}^{jk} = \sup_{\beta_n \in N_n(\delta_n)} 2n^{-1}|\sum_{i=1}^n x_{ij}x_{ik}\{y_i - E(y_i)\}\{\mu(X\beta_{n,0}) - \mu(X\beta_n)\}_i 1_{\Omega_n}|,$$

$$G_{212}^{jk} = \sup_{\beta_n \in N_n(\delta_n)} 2n^{-1}|\sum_{i=1}^n x_{ij}x_{ik}\{y_i - E(y_i)\}\{\mu(X\beta_{n,0}) - \mu(X\beta_n)\}_i(1 - 1_{\Omega_n})|.$$

First, we bound $E(G_{211}^{jk})$. We take a Rademacher sequence $\{\epsilon_i\}_{i=1}^n$ independent of y. Then, we apply symmetrization and contraction inequalities in Bühlmann & van de Geer (2011) as follows.

$$\begin{aligned}
E(G_{211}^{jk}) =& E\left[\sup_{\beta_n \in N_n(\delta_n)} 2n^{-1}\left|\sum_{i=1}^n x_{ij}x_{ik}\{y_i - E(y_i)\}\{\mu(X\beta_{n,0}) - \mu(X\beta_n)\}_i 1_{\Omega_n}\right|\right] \\
\leq& 4n^{-1}E\left[\sup_{\beta_n \in N_n(\delta_n)}\left|\sum_{i=1}^n \epsilon_i x_{ij}x_{ik}y_i\{\mu(X\beta_{n,0}) - \mu(X\beta_n)\}_i 1_{\Omega_n}\right|\right] \\
\leq& 4n^{-1}c_0 E\left\{\sup_{\beta_n \in N_n(\delta_n)}\left|\sum_{i=1}^n \epsilon_i x_{ij}x_{ik}y_i(X\beta_{n,0} - X\beta_n)_i 1_{\Omega_n}\right|\right\} \\
\leq& 4n^{-1}c_0 \sup_{\beta_n \in N_n(\delta_n)} \|\beta_{n,0} - \beta_n\|_2 E\left(\left\|\sum_{i=1}^n \epsilon_i x_{ij}x_{ik}y_i 1_{\Omega_n}x_i\right\|\right),
\end{aligned}$$

where the last step follows from the Cauchy–Schwarz inequality. We observe that $\sup_{\beta_n \in N_n(\delta_n)} \|\beta_{n,0} - \beta_n\|_2 \leq n^{-1/2}d^{1/2}\delta_n$ and $E(\|\sum_{i=1}^n \epsilon_i x_{ij}x_{ik}y_i 1_{\Omega_n}x_i\|_2) \leq \{\sum_{i=1}^n x_{ij}^2 x_{ik}^2 E(y_i^2 1_{\Omega_n})\|x_i\|_2^2\}^{1/2}$. So, we can bound $E(G_{211}^{jk})$ by $4c_0 n^{-3/2}d^{1/2}\delta_n\{\sum_{i=1}^n x_{ij}^2 x_{ik}^2 E(y_i^2 1_{\Omega_n})\|x_i\|_2^2\}^{1/2}$. Using Assumptions 2 and 3, we obtain $E(G_{211}^{jk}) = O(n^{-1+2u_3}d\delta_n\widetilde{m}_n)$. Since $d = O(n^{\kappa_1})$ and $-1 + 2u_3 + 3\kappa_1 + 2u_1 + \kappa_2/2 < 0$, we deduce $E(G_{211}^{jk}) = o(1/d^2)$.

Furthermore, we need to bound $2|x_{ij}x_{ik}y_i\{\mu(X\beta_{n,0}) - \mu(X\beta_n)\}_i 1_{\Omega_n}|$ for any $\beta_n \in N_n(\delta_n)$ in order to use the concentration theorem in Bühlmann & van de Geer (2011). We use Lemma 2 to bound y_i:

$$\begin{aligned}
&2|x_{ij}x_{ik}y_i\{\mu(X\beta_{n,0}) - \mu(X\beta_n)\}_i 1_{\Omega_n}| \\
&\leq 2|x_{ij}||x_{ik}||\{y_i - E(y_i) + E(y_i)\}|1_{\Omega_n}|\{\mu(X\beta_{n,0}) - \mu(X\beta_n)\}_i| \\
&\leq 2|x_{ij}||x_{ik}|\{|E(y_i)| + C_1\log(n)\}|\{\mu(X\beta_{n,0}) - \mu(X\beta_n)\}_i|.
\end{aligned}$$

Since $b''(X\beta) \leq c_0^{-1}$ for any β joining the line segment $\beta_{n,0}$ and β_n, we have $|\{\mu(X\beta_{n,0}) - \mu(X\beta_n)\}_i| \leq c_0^{-1}\|x_i\|_2\|\beta_{n,0} - \beta_n\|_2$ for any $\beta_n \in N_n(\delta_n)$. When we put last two inequalities together with Assumptions 2 and 3, we get $2|x_{ij}x_{ik}y_i\{\mu(X\beta_{n,0}) - \mu(X\beta_n)\}_i 1_{\Omega_n}| \leq c_{i,\beta_n}$ where $c_{i,\beta_n} = O(n^{2u_3}\widetilde{m}_n)\|x_i\|_2\|\beta_{n,0} - \beta_n\|_2$. Moreover, we have

$$\begin{aligned}
\sup_{\beta_n \in N_n(\delta_n)} n^{-1}\sum_{i=1}^n c_{i,\beta_n}^2 &\leq O(n^{-1+4u_3}\widetilde{m}_n^2) \sup_{\beta_n \in N_n(\delta_n)} \|\beta_{n,0} - \beta_n\|_2^2 \sum_{i=1}^n \|x_i\|_2^2 \\
&\leq O(n^{-1+4u_3}\widetilde{m}_n^2 d^2\delta_n^2)
\end{aligned}$$

where we use the fact that $\|\beta_{n,0} - \beta_n\|_2^2 = O(n^{-1}d\delta_n^2)$ for any $\beta_n \in N_n(\delta_n)$. Thus, we can use the concentration inequality in Bühlmann & van de Geer (2011) which yields

$$pr\{G_{211}^{jk} \geq E(G_{211}^{jk}) + t\} \leq C\exp\left(-C\frac{nt^2}{n^{-1+4u_3}\widetilde{m}_n^2 d^2\delta_n^2}\right), \tag{A.17}$$

for some constant C.

Now, take any $\tilde{t} > 0$. We know that $E(G_{211}^{jk}) < \tilde{t}/(2d^2)$ for large enough n. Then by taking $t = \tilde{t}/(2d^2)$ in equation (A.17), we obtain

$$pr(G_{211}^{jk} \geq \tilde{t}/d^2) \leq C\exp\left(-C\frac{\tilde{t}^2}{n^{-2+4u_3}\widetilde{m}_n^2 d^6\delta_n^2}\right).$$

Since $-2 + 4u_3 + 6\kappa_1 + 4u_1 + \kappa_2 < 0$, we have $pr(G_{211}^{jk} \geq \tilde{t}/d^2) = o(1/d^2)$.

Lastly, $G_{212}^{jk} = 0$ on the event Ω_n which holds with probability at least $1 - O(n^{-\delta})$ by Lemma 2. Therefore, we obtain $G_{21} = o(1/d^2)$ by using (A.16).

Part 2e) prove $G_{22} = o(1/d)$. First, we apply the Cauchy–Schwarz inequality to obtain

$$|G_{22}^{jk}|^2 = \left(2 \sum_{i=1}^{n} \left[n^{-1} \mathrm{var}(y_i)^{1/2} x_{ij} x_{ik} \{ \mu(X\beta_{n,0}) - \mu(X\widehat{\beta}_n) \}_i \right] \left[\frac{\{E(y) - \mu(X\beta_{n,0})\}_i}{\mathrm{var}(y_i)^{1/2}} \right] \right)^2$$
$$\leq 4 \sum_{i=1}^{n} n^{-2} \mathrm{var}(y_i) x_{ij}^2 x_{ik}^2 \{ \mu(X\beta_{n,0}) - \mu(X\widehat{\beta}_n) \}_i^2 \sum_{i=1}^{n} \frac{\{E(y) - \mu(X\beta_{n,0})\}_i^2}{\mathrm{var}(y_i)}$$

Since $\widehat{\beta}_n$ lies in the region $N_n(\delta_n)$ with high probability and $b''(\cdot)$ is bounded, $\{\mu(X\beta_{n,0}) - \mu(X\widehat{\beta}_n)\}_i^2$ can be bounded by $\|x_i\|_2^2 O(n^{-1} d \delta_n^2)$. Assumption 2 and the Cauchy–Schwarz inequality yield $\sum_{i=1}^{n} \{\mathrm{var}(y_i)\}^{-1} \{E(y) - \mu(X\beta_{n,0})\}_i^2 \leq O(n^{1/2+u_2/2})$. We further use Assumptions 1 and 3 to obtain $|G_{22}^{jk}|^2 = O(n^{-3/2+4u_3+u_2/2} d^2 \delta_n^2)$. Since $d = O(n^{\kappa_1})$ and $-3/2 + 4u_3 + u_2/2 + 6\kappa_1 + 2u_1 + \kappa_2 < 0$, we get $|G_{22}^{jk}|^2 = o(1/d^4)$. Thus, we obtain $G_{22} = o_p(1/d)$.

Part 2f) prove $G_3 = o(1/d)$. We decompose (i, j)th entry of G_3 as follows

$$|G_3^{jk}| = n^{-1} \left| \sum_{i=1}^{n} x_{ij} x_{ik} \{ \mu(X\beta_{n,0}) - \mu(X\widehat{\beta}_n) \}_i^2 \right|$$
$$\leq n^{-1} \sum_{i=1}^{n} |x_{ij}| |x_{ik}| \{ \mu(X\beta_{n,0}) - \mu(X\widehat{\beta}_n) \}_i^2$$
$$= O(n^{-1+2u_3} d^2 \delta_n^2),$$

where the last line is similar to Part 2e. So, $|G_3^{jk}| = o(1/d^2)$ since $-1 + 2u_3 + 4\kappa_1 + 2u_1 + \kappa_2 < 0$. Therefore, we get $G_3 = o(1/d)$.

We have finished the proof of Part 2. This concludes the proof of Theorem 2 with the desired probability bound $1 - O(n^{-\delta} + p^{1-8c_2\gamma_n^2})$.

E.3. *Proof of Theorem* 3

Theorem 3 is a direct consequence of Theorem 2, Lemma 1, and assumption (17). To see this, observe that the difference in the sample version HGBIC_p can be written as the sum of the population version HGBIC_p^* and the terms consisting of differences of likelihood, $\mathrm{tr}(H_n)$ and $\log(\det(H_n))$ between the sample and population versions. That is,

$$\mathrm{HGBIC}_p(\mathfrak{M}_m) - \mathrm{HGBIC}_p(\mathfrak{M}_1) = \mathrm{HGBIC}_p^*(\mathfrak{M}_m) - \mathrm{HGBIC}_p^*(\mathfrak{M}_1)$$
$$- 2\{\ell_n(y, \widehat{\beta}_{n,m}) - \ell_n(y, \beta_{n,m,0})\} + 2\{\ell_n(y, \widehat{\beta}_{n,1}) - \ell_n(y, \beta_{n,1,0})\}$$
$$+ \{\mathrm{tr}(\widehat{H}_{n,m}) - \mathrm{tr}(H_{n,m})\} - \{\mathrm{tr}(\widehat{H}_{n,1}) - \mathrm{tr}(H_{n,1})\}$$
$$- (\log|\widehat{H}_{n,m}| - \log|H_{n,m}|) + (\log|\widehat{H}_{n,1}| - \log|H_{n,1}|).$$

The equation (17) suggests that the first line is bounded below by Δ_n for any $m > 1$. Then we focus on the remaining terms. Let $m = 2, \ldots, M$ be fixed. The consistency of the maximum likelihood estimator in Lemma 1 implies that $-2\{\ell_n(y, \widehat{\beta}_{n,m}) - \ell_n(y, \beta_{n,m,0})\} + 2\{\ell_n(y, \widehat{\beta}_{n,1}) - \ell_n(y, \beta_{n,1,0})\}$ converges to zero with probability at least $1 - O(n^{-\delta})$ for some constant $\delta > 0$. Moreover, Theorem 2 proves that the last two lines are also of order $o(\Delta_n)$ with probability at least $1 - O(n^{-\delta})$ provided that Δ_n is converging slowly enough. Therefore, $\{\mathrm{HGBIC}_p(\mathfrak{M}_m) - \mathrm{HGBIC}_p(\mathfrak{M}_1)\} > \Delta_n/2$ for large enough n with probability $1 - O(n^{-\delta})$ for any fixed $m > 1$. Applying the union bound over all $M = o(n^{\delta})$ competing models completes the proof of Theorem 3.

F. TECHNICAL LEMMAS

F.1. *Lemma* 1 *and its proof*

LEMMA 1 (UNIFORM CONSISTENCY OF THE MAXIMUM LIKELIHOOD ESTIMATOR). *Assume Assumptions* 1–3 *hold. If* $L_n\{Kn^{-1}\log(p)\}^{1/2} \to 0$*, then*

$$\sup_{|\mathfrak{M}|\leq K,\, \mathfrak{M}\subset\{1,\ldots,p\}} (|\mathfrak{M}|)^{-1/2}\|\widehat{\beta}(\mathfrak{M}) - \beta_{n,0}(\mathfrak{M})\|_2 = O_p\left\{L_n(n^{-1}\log p)^{1/2}\right\},$$

where $L_n = 2\widetilde{m}_n + C_1\log n$*.* $\widetilde{m}_n$ *is a diverging sequence which appears in Assumption* 2 *and* C_1 *is the positive constant from Lemma* 2*.*

Proof. First, we construct the auxiliary parameter vector $\widehat{\beta}_u(\mathfrak{M})$ as follows. For any sequence N_n, we take $u = \{1 + \|\widehat{\beta}(\mathfrak{M}) - \beta_{n,0}(\mathfrak{M})\|_2/N_n\}^{-1}$ and define $\widehat{\beta}_u(\mathfrak{M}) = u\widehat{\beta}(\mathfrak{M}) + (1-u)\beta_{n,0}(\mathfrak{M})$. We have $\|\widehat{\beta}_u(\mathfrak{M}) - \beta_{n,0}(\mathfrak{M})\|_2 = u\|\widehat{\beta}(\mathfrak{M}) - \beta_{n,0}(\mathfrak{M})\|_2 \leq N_n$ by the definition of u. So, $\widehat{\beta}_u(\mathfrak{M})$ belongs to the neighborhood $\mathcal{B}_{\mathfrak{M}}(N_n) = \{\beta \in \mathbb{R}^d,\, \mathrm{supp}(\beta) = \mathfrak{M} : \|\beta - \beta_{n,0}(\mathfrak{M})\|_2 \leq N_n\}$. Moreover, we observe that $\|\widehat{\beta}_u(\mathfrak{M}) - \beta_{n,0}(\mathfrak{M})\|_2 \leq N_n/2$ implies $\|\widehat{\beta}(\mathfrak{M}) - \beta_{n,0}(\mathfrak{M})\|_2 \leq N_n$. Thus, it is enough to bound $\|\widehat{\beta}_u(\mathfrak{M}) - \beta_{n,0}(\mathfrak{M})\|_2$ to prove the theorem.

Now, we consider $\|\widehat{\beta}_u(\mathfrak{M}) - \beta_{n,0}(\mathfrak{M})\|_2$. First, the concavity of ℓ_n and the definition of $\widehat{\beta}(\mathfrak{M})$ yield

$$\begin{aligned}\ell_n\{\widehat{\beta}_u(\mathfrak{M})\} &\geq u\ell_n\{\widehat{\beta}(\mathfrak{M})\} + (1-u)\ell_n\{\beta_{n,0}(\mathfrak{M})\}\\ &\geq u\ell_n\{\widehat{\beta}_u(\mathfrak{M})\} + (1-u)\ell_n\{\beta_{n,0}(\mathfrak{M})\}.\end{aligned}$$

So, by rearranging terms, we get

$$-\ell_n\{\beta_{n,0}(\mathfrak{M})\} + \ell_n\{\widehat{\beta}_u(\mathfrak{M})\} \geq 0. \tag{A.1}$$

Besides, for any $\beta \in \mathcal{B}_{\mathfrak{M}}(N_n)$, we have

$$E[\ell_n\{\beta_{n,0}(\mathfrak{M})\} - \ell_n(\beta)] = I\{g_n; f_n(\cdot;\beta,\tau)\} - I\{g_n; f_n(\cdot;\beta_{n,0}(\mathfrak{M}),\tau)\} \geq 0, \tag{A.2}$$

by the optimality of $\beta_{n,0}(\mathfrak{M})$. Combining (A.1) and (A.2) gives

$$\begin{aligned}0 &\leq E[\ell_n\{\beta_{n,0}(\mathfrak{M})\} - \ell_n\{\widehat{\beta}_u(\mathfrak{M})\}]\\ &\leq -\ell_n\{\beta_{n,0}(\mathfrak{M})\} + \ell_n\{\widehat{\beta}_u(\mathfrak{M})\} + E[\ell_n\{\beta_{n,0}(\mathfrak{M})\} - \ell_n\{\widehat{\beta}_u(\mathfrak{M})\}]\\ &\leq \sup_{\beta\in\mathcal{B}_{\mathfrak{M}}(N_n)} \left|\ell(\beta) - E\{\ell_n(\beta)\} - (\ell_n\{\beta_{n,0}(\mathfrak{M})\} - E[\ell_n\{\beta_{n,0}(\mathfrak{M})\}])\right|\\ &= nT_{\mathfrak{M}}(N_n),\end{aligned} \tag{A.3}$$

since $\widehat{\beta}_u(\mathfrak{M}) \in \mathcal{B}_{\mathfrak{M}}(N_n)$.

On the other hand, for any $\beta \in \mathcal{B}_{\mathfrak{M}}(N_n)$,

$$\begin{aligned}E[\ell_n\{\beta_{n,0}(\mathfrak{M})\} - \ell_n(\beta)] &= E(Y^T)Z_{\mathfrak{M}}\{\beta_{n,0}(\mathfrak{M}) - \beta\} - 1^T[b\{Z_{\mathfrak{M}}\beta_{n,0}(\mathfrak{M})\} - b(Z_{\mathfrak{M}}\beta)]\\ &= \mu\{Z_{\mathfrak{M}}\beta_{n,0}(\mathfrak{M})\}Z_{\mathfrak{M}}\{\beta_{n,0}(\mathfrak{M}) - \beta\} - 1^T[b\{Z_{\mathfrak{M}}\beta_{n,0}(\mathfrak{M})\} - b(Z_{\mathfrak{M}}\beta)],\end{aligned}$$

since $\beta_{n,0}(\mathfrak{M})$ satisfies the score equation: $Z_{\mathfrak{M}}^T\{E(Y) - \mu(Z_{\mathfrak{M}}\beta)\} = 0$. Furthermore, applying the second order Taylor expansion yields

$$E[\ell_n\{\beta_{n,0}(\mathfrak{M})\} - \ell_n(\beta)] = \frac{1}{2}\left\{\beta_{n,0}(\mathfrak{M}) - \beta\right\}^T Z_{\mathfrak{M}}^T\Sigma(Z_{\mathfrak{M}}\bar{\beta})Z_{\mathfrak{M}}\left\{\beta_{n,0}(\mathfrak{M}) - \beta\right\},$$

where $\bar{\beta}$ lies on the line segment connecting $\beta_{n,0}(\mathfrak{M})$ and β. Then, we use Assumption 3 and the assumption that $c_0 \leq b''(Z\beta) \leq c_0^{-1}$ for any $\beta \in \mathcal{B}$. So, we get $E[\ell_n\{\beta_{n,0}(\mathfrak{M})\} - \ell_n(\beta)] \geq \frac{1}{2}nc_0c_2\|\beta_{n,0}(\mathfrak{M}) - \widehat{\beta}_u(\mathfrak{M})\|_2^2$. Therefore, for any $\beta \in \mathcal{B}_{\mathfrak{M}}(N_n)$,

$$\|\beta_{n,0}(\mathfrak{M}) - \beta\|_2^2 \leq 2(c_0c_2)^{-1}n^{-1}E[\ell_n\{\beta_{n,0}(\mathfrak{M})\} - \ell_n(\beta)]. \tag{A.4}$$

Finally, we take a slowly diverging sequence γ_n such that $\gamma_n L_n\{K\log(p)/n\}^{1/2} \to 0$. Then, we choose $N_n = \gamma_n L_n\{|\mathfrak{M}|n^{-1}\log(p)\}^{1/2}$. Since $\widehat{\beta}_u(\mathfrak{M}) \in \mathcal{B}_{\mathfrak{M}}(N_n)$, we combine equations (A.3) and (A.4) to obtain

$$\sup_{|\mathfrak{M}|\le K}(|\mathfrak{M}|)^{-1/2}\|\beta_{n,0}(\mathfrak{M}) - \widehat{\beta}_u(\mathfrak{M})\|_2 \le \sup_{|\mathfrak{M}|\le K}\left\{\frac{T_{\mathfrak{M}}(N_n)}{|\mathfrak{M}|}\right\}^{1/2}\{2(c_0c_2)^{-1}n^{-1}\}^{1/2}$$
$$= O_p[L_n\{n^{-1}\log(p)\}^{1/2}],$$

where the last step follows from Lemma 4. This completes the proof of Lemma 1.

F.2. *Lemma* 2 *and its proof*

LEMMA 2. *Assume that $Y_1,\ldots,Y_n$ are independent and satisfy Assumption* 1. *Then, for any constant $\delta > 0$, there exist large enough positive constants C_1 and C_2 such that*

$$\|W\|_\infty \le C_1\log n, \tag{A.5}$$

with probability at least $1 - O(n^{-\delta})$ and,

$$\|n^{-1/2}E(W \mid \Omega_n)\|_2 = O\{(\log n)n^{-C_2}\}, \tag{A.6}$$

where $\Omega_n = \{\|W\|_\infty \le C_1\log n\}$.

Proof. We take $t = C_1\log n$ in Assumption 1. So we get

$$pr(\|W\|_\infty \le C_1\log n) \ge 1 - n\max_{i\le n} pr(|W_i| > C_1\log n) \ge 1 - c_1 n^{1-c_1^{-1}C_1}.$$

å We choose C_1 large enough so that $1 - c_1^{-1}C_1 \le 0$. Thus, we have $pr(\|W\|_\infty \le C_1\log n) = 1 - O(n^{-\delta})$ where we pick $\delta = c_1^{-1}C_1 - 1 > 0$. This proves the first part of the lemma.

Now, we proceed the proof of the second part of the lemma. We will bound each term $E(W_i|\Omega_n)$ for $i = 1,\ldots,n$. Since W_i's for $i = 1,\ldots,n$ are independent, the conditional expectation $E(W_i|\Omega_n)$ can be written as follows

$$E(W_i \mid \Omega_n) = E(W_i \mid |W_i| \le C_1\log n) = \frac{E\{W_i 1(|W_i| \le C_1\log n)\}}{pr(|W_i| \le C_1\log n)}.$$

Since $E(W) = 0$ by definition, we get $E\{W_i 1(|W_i| \le C_1\log n)\} = -E\{W_i 1(|W_i| > C_1\log n)\}$. Last two equalities result in

$$|E(W_i \mid \Omega_n)| \le \frac{E\{|W_i| 1(|W_i| > C_1\log n)\}}{pr(|W_i| \le C_1\log n)}.$$

We already showed that the denominator $pr(|W_i| \le C_1\log n)$ can be bounded below by $1 - O(n^{-\delta})$ uniformly in i. Thus, it suffices to bound the numerator $E\{|W_i|1(|W_i| > C_1\log n)\}$. Indeed, we have

$$E\{|W_i|1(|W_i| > C_1\log n)\} = \int_0^\infty pr\{|W_i|1(|W_i| > C_1\log n) \ge t\}dt$$
$$= \int_0^{C_1\log n} pr\{|W_i|1(|W_i| > C_1\log n) \ge t\}dt$$
$$+ \int_{C_1\log n}^\infty pr\{|W_i|1(|W_i| > C_1\log n) \ge t\}dt$$
$$= \int_0^{C_1\log n} pr(|W_i| \ge C_1\log n)dt + \int_{C_1\log n}^\infty pr(|W_i| \ge t)dt$$
$$\le C_1\log(n)pr(|W_i| \ge C_1\log n) + \int_{C_1\log n}^\infty c_1\exp(-c_1^{-1}t)dt$$
$$\le C_1\log(n)c_1\exp(-c_1^{-1}C_1\log n) + c_1^2\exp(-c_1^{-1}C_1\log n),$$

where we use Assumption 1 in the last two steps. This concludes the proof of Lemma 2 by choosing $C_2 = c_1^{-1}C_1$.

F.3. *Lemma* 3 *and its proof*

LEMMA 3. *Under Assumption* 2*, the function* ρ *defined by* $\rho(x_i^T\beta, Y_i) = Y_i x_i^T\beta - b(x_i^T\beta)$ *is Lipschitz continuous with the Lipschitz constant* $L_n = 2\widetilde{m}_n + C_1 \log n$ *conditioned on the set* $\Omega_n = \{\|W\|_\infty \leq C_1 \log n\}$ *given in Lemma* 2.

Proof. We consider the difference $\rho(x_i^T\beta_1, Y_i) - \rho(x_i^T\beta_2, Y_i)$ for any β_1 and β_2 in $\mathbb{R}^p$. We observe that

$$|\rho(x_i^T\beta_1, Y_i) - \rho(x_i^T\beta_2, Y_i)| \leq |Y_i||x_i^T(\beta_1 - \beta_2)| + |b(x_i^T\beta_1) - b(x_i^T\beta_2)|.$$

We can bound $|Y_i|$ on Ω_n using Assumption 2 as $|Y_i| \leq \|Y\|_\infty \leq \|EY\|_\infty + \|W\|_\infty \leq \widetilde{m}_n + C_1\log(n)$. Then we apply the mean-value theorem to obtain $|b(x_i^T\beta_1) - b(x_i^T\beta_2)| \leq |b'(\tilde{\beta})||x_i^T(\beta_1 - \beta_2)|$ where $\tilde{\beta}$ lies on the line segment connecting β_1 and β_2. Thus, we get $|b(x_i^T\beta_1) - b(x_i^T\beta_2)| \leq \widetilde{m}_n|x_i^T(\beta_1 - \beta_2)|$ by Assumption 2. Hereby, we showed that $|\rho(x_i^T\beta_1, Y_i) - \rho(x_i^T\beta_2, Y_i)| \leq (2\widetilde{m}_n + C_1\log n)|x_i^T\beta_1 - x_i^T\beta_2|$ conditioned on Ω_n. Thus, $\rho(\cdot, Y_i)$ is Lipschitz continuous with the Lipschitz constant $L_n = 2\widetilde{m}_n + C_1\log n$ conditioned on the set Ω_n. This completes the proof of Lemma 3.

F.4. *Lemma* 4 *and its proof*

LEMMA 4. *Assume that Assumptions* 1–3 *hold. Define the neighborhood* $\mathcal{B}_{\mathfrak{M}}(N) = \{\beta \in \mathbb{R}^d, \operatorname{supp}(\beta) = \mathfrak{M} : \|\beta - \beta_{n,0}(\mathfrak{M})\|_2 \leq N\}$ *and*

$$T_{\mathfrak{M}}(N) = \sup_{\beta \in \mathcal{B}_{\mathfrak{M}}(N)} n^{-1}\left|\ell_n(\beta) - \ell_n\{\beta_{n,0}(\mathfrak{M})\} - E[\ell_n(\beta) - \ell_n\{\beta_{n,0}(\mathfrak{M})\}]\right|.$$

If γ_n *is a slowly diverging sequence such that* $\gamma_n L_n\{Kn^{-1}\log(p)\}^{1/2} \to 0$*, then*

$$\sup_{|\mathfrak{M}|\leq K} (|\mathfrak{M}|)^{-1/2} T_{\mathfrak{M}}\left[\gamma_n L_n\{|\mathfrak{M}|n^{-1}\log(p)\}^{1/2}\right] = O(L_n^2 n^{-1}\log p)$$

with probability at least $(1 - e^2 p^{1-8c_2\gamma_n^2})\{1 - O(n^{-\delta})\}$*, where* $L_n = 2\widetilde{m}_n + C_1\log n$*.*

Proof. To prove the lemma, we condition on the set $\Omega_n = \{\|Y - EY\|_\infty \leq C_1\log n\}$. We observe that

$$\begin{aligned}
&\left|\ell_n(\beta) - \ell_n\{\beta_{n,0}(\mathfrak{M})\} - E[\ell_n(\beta) - \ell_n\{\beta_{n,0}(\mathfrak{M})\}]\right| \\
&\quad\leq \left|\ell_n(\beta) - \ell_n\{\beta_{n,0}(\mathfrak{M})\} - E[\ell_n(\beta) - \ell_n\{\beta_{n,0}(\mathfrak{M})\} \mid \Omega_n]\right| \\
&\qquad + |E[\ell_n(\beta) - \ell_n(\beta_{n,0}(\mathfrak{M}))] - E[\ell_n(\beta) - \ell_n\{\beta_{n,0}(\mathfrak{M})\} \mid \Omega_n]|,
\end{aligned}$$

by the triangle inequality. Thus, $T_{\mathfrak{M}}(N_n)$ can be bounded by the sum of the following two terms:

$$\begin{aligned}
\tilde{T}_{\mathfrak{M}}(N_n) &= \sup_{\beta \in \mathcal{B}_{\mathfrak{M}}(N_n)} n^{-1}\left|\ell_n(\beta) - \ell_n\{\beta_{n,0}(\mathfrak{M})\} - E[\ell_n(\beta) - \ell_n\{\beta_{n,0}(\mathfrak{M})\} \mid \Omega_n]\right|, \text{ and} \\
R_{\mathfrak{M}}(N_n) &= \sup_{\beta \in \mathcal{B}_{\mathfrak{M}}(N_n)} n^{-1}(E[\ell_n(\beta) - \ell_n\{\beta_{n,0}(\mathfrak{M})\}] - E[\ell_n(\beta) - \ell_n\{\beta_{n,0}(\mathfrak{M})\} \mid \Omega_n])
\end{aligned}$$

That is,

$$T_{\mathfrak{M}}(N_n) \leq \tilde{T}_{\mathfrak{M}}(N_n) + R_{\mathfrak{M}}(N_n). \tag{A.7}$$

In the rest of the proof, we will show the following bounds

$$R_{\mathfrak{M}}(N_n) = o\left(L_n^2 \frac{\log p}{n}\right), \tag{A.8}$$

and

$$\tilde{T}_{\mathfrak{M}}(N_n) = O_p\left(L_n^2 \frac{\log p}{n}\right). \tag{A.9}$$

First, we consider $R_{\mathfrak{M}}(N_n)$. We split $R_{\mathfrak{M}}(N_n)$ by the Cauchy–Schwarz inequality so that

$$\begin{aligned} R_{\mathfrak{M}}(N_n) &= \sup_{\beta \in \mathcal{B}_{\mathfrak{M}}(N_n)} n^{-1} |\{E(Y) - E(Y \mid \Omega_n)\}^T X \{\beta - \beta_{n,0}(\mathfrak{M})\}| \\ &\leq \|n^{-1/2}\{E(Y) - E(Y \mid \Omega_n)\}\|_2 \sup_{\beta \in \mathcal{B}_{\mathfrak{M}}(N_n)} \|n^{-1/2} X \{\beta - \beta_{n,0}(\mathfrak{M})\}\|_2. \end{aligned}$$

We have

$$\|n^{-1/2}\{E(Y) - E(Y \mid \Omega_n)\}\|_2 = \|n^{-1/2}\{E(W \mid \Omega_n)\}\|_2 = O(n^{-C_2} \log n)$$

by Lemma 2. We also have

$$\|n^{-1/2} X \{\beta - \beta_{n,0}(\mathfrak{M})\}\|_2 \leq \{\lambda_{\max}(n^{-1} X_{\mathfrak{M}}^T X_{\mathfrak{M}})\}^{1/2} \|\beta - \beta_{n,0}(\mathfrak{M})\|_2 \leq c_2^{-1/2} N_n,$$

for any $\beta \in \mathcal{B}_{\mathfrak{M}}(N_n)$.

Therefore, $R_{\mathfrak{M}}(\beta) = O(N_n n^{-C_2} \log n)$. So, (A.8) follows by taking C_2 large enough.

Next, we deal with the term $\tilde{T}_{\mathfrak{M}}(N_n)$ by showing (A.9). We observe that the difference $\ell_n(\beta) - \ell_n\{\beta_{n,0}(\mathfrak{M})\}$ can be written as

$$\begin{aligned} \ell_n(\beta) - \ell_n\{\beta_{n,0}(\mathfrak{M})\} &= \sum_{i=1}^n \left(Y_i \{x_i^T \beta - x_i^T \beta_{n,0}(\mathfrak{M})\} - [b(x_i^T \beta) - b\{x_i^T \beta_{n,0}(\mathfrak{M})\}] \right) \\ &= \sum_{i=1}^n \left[\rho(x_i^T \beta, Y_i) - \rho\{x_i^T \beta_{n,0}(\mathfrak{M}), Y_i\} \right]. \end{aligned}$$

In Lemma 3, we showed that $\rho(x_i^T \beta, Y_i) = Y_i x_i^T \beta - b(x_i^T \beta)$ is Lipschitz continuous with the Lipschitz constant L_n conditioned on the set Ω_n.

Next, we choose a Rademacher sequence $\{\epsilon_i\}_{i=1}^n$. Then, we apply symmetrization and concentration inequalities in Bühlmann & van de Geer (2011) as follows:

$$\begin{aligned} & E\{\tilde{T}_{\mathfrak{M}}(N_n) \mid \Omega_n\} \\ & \leq 2E \left(\sup_{\beta \in \mathcal{B}_{\mathfrak{M}}(N_n)} n^{-1} \left| \sum_{i=1}^n \epsilon_i \left[\rho(x_i^T \beta, Y_i) - \rho\{x_i^T \beta_{n,0}(\mathfrak{M}), Y_i\} \right] \right| \mid \Omega_n \right) \\ & \leq 4 L_n E \left[\sup_{\beta \in \mathcal{B}_{\mathfrak{M}}(N_n)} n^{-1} \left| \sum_{i=1}^n \epsilon_i \{x_i^T \beta - x_i^T \beta_{n,0}(\mathfrak{M})\} \right| \mid \Omega_n \right]. \end{aligned}$$

Furthermore, we have

$$\begin{aligned} & E \left[\sup_{\beta \in \mathcal{B}_{\mathfrak{M}}(N_n)} n^{-1} \left| \sum_{i=1}^n \epsilon_i \{x_i^T \beta - x_i^T \beta_{n,0}(\mathfrak{M})\} \right| \mid \Omega_n \right] \\ & \leq E \left\{ n^{-1} \sup_{\beta \in \mathcal{B}_{\mathfrak{M}}(N_n)} \|\beta - \beta_{n,0}(\mathfrak{M})\|_2 \| \sum_{i=1}^n \epsilon_i (x_i)_{\mathfrak{M}} \|_2 \mid \Omega_n \right\} \\ & \leq E \left\{ n^{-1} N_n \| \sum_{i=1}^n \epsilon_i (x_i)_{\mathfrak{M}} \|_2 \mid \Omega_n \right\} = n^{-1} N_n E \left[\left\{ \sum_{j \in \mathfrak{M}} \left(\sum_{i=1}^n \epsilon_i x_{ij} \right)^2 \right\}^{1/2} \right] \\ & \leq n^{-1} N_n \left[\sum_{j \in \mathfrak{M}} E \left\{ \left(\sum_{i=1}^n \epsilon_i x_{ij} \right)^2 \right\} \right]^{1/2} = N_n n^{-1/2} |\mathfrak{M}|^{1/2}, \end{aligned}$$

where we use the Cauchy–Schwarz inequality and the assumption $\sum_{i=1}^n x_{ij}^2 = n$. Therefore, we obtain the bound

$$E\{\tilde{T}_{\mathfrak{M}}(N_n) \mid \Omega_n\} \leq 4L_n N_n n^{-1/2} |\mathfrak{M}|^{1/2}. \tag{A.10}$$

For any $\beta \in \mathcal{B}_{\mathfrak{M}}(N_n)$, we have

$$\begin{aligned} & n^{-1} \sum_{i=1}^n |\rho\{x_i^T \beta_{n,0}(\mathfrak{M}), Y_i\} - \rho(x_i^T \beta, Y_i)|^2 \\ & \leq n^{-1} L_n^2 \sum_{i=1}^n |x_i^T \beta_{n,0}(\mathfrak{M}) - x_i^T \beta|^2 \\ & = n^{-1} L_n^2 \{\beta_{n,0}(\mathfrak{M}) - \beta\}^T X_{\mathfrak{M}}^T X_{\mathfrak{M}} \{\beta_{n,0}(\mathfrak{M}) - \beta\} \\ & \leq \ L_n^2 c_2^{-1} N_n^2. \end{aligned}$$

Then we apply Theorem 14.2 in Bühlmann & van de Geer (2011) to obtain

$$pr\left[\tilde{T}_{\mathfrak{M}}(N_n) \geq E\{\tilde{T}_{\mathfrak{M}}(N_n) \mid \Omega_n\} + t \mid \Omega_n\right] \leq \exp\left(\frac{-nc_2 t^2}{8L_n^2 N_n^2}\right).$$

Now, we take $t = 4L_n N_n n^{-1/2} |\mathfrak{M}|^{1/2} u$ for some positive u that will be chosen later. So, we get $pr\{\tilde{T}_{\mathfrak{M}}(N_n) \geq 4L_n N_n n^{-1/2} |\mathfrak{M}|^{1/2}(1+u) \mid \Omega_n\} \leq \exp(-2c_2 u^2 |\mathfrak{M}|)$ by using (A.10).

We choose $N_n = L_n n^{-1/2} |\mathfrak{M}|^{1/2}(1+u)$. So, it follows that

$$pr\left\{\frac{\tilde{T}_{\mathfrak{M}}(N_n)}{|\mathfrak{M}|} \geq 4L_n^2 n^{-1}(1+u)^2 \mid \Omega_n\right\} \leq \exp(-8c_2 u^2 |\mathfrak{M}|).$$

Thus, we have

$$\begin{aligned} & pr\left\{\sup_{|\mathfrak{M}| \leq K} \frac{\tilde{T}_{\mathfrak{M}}(N_n)}{|\mathfrak{M}|} \geq 4L_n^2 n^{-1}(1+u)^2 \mid \Omega_n\right\} \leq \sum_{|\mathfrak{M}| \leq K} pr\left\{\frac{\tilde{T}_{\mathfrak{M}}(N_n)}{|\mathfrak{M}|} \geq 4L_n^2 n^{-1}(1+u)^2 \mid \Omega_n\right\} \\ & \leq \sum_{k \leq K} \binom{p}{k} \exp(-8c_2 u^2 k) \leq \sum_{k \leq K} \left(\frac{pe}{k}\right)^k \exp(-8c_2 u^2 k). \end{aligned}$$

Now, we choose $u = \gamma_n (\log p)^{1/2}$. So, for n large enough, we get

$$\begin{aligned} \sum_{k \leq K} \left(\frac{pe}{k}\right)^k \exp(-8c_2 u^2 k) &= \sum_{k \leq K} \left(\frac{pe}{k}\right)^k p^{-8c_2 \gamma_n^2 k} = \sum_{k \leq K} \frac{\{ep^{(1-8c_2\gamma_n^2)}\}^k}{k^k} \\ &\leq \sum_{k \leq K} \frac{ep^{(1-8c_2\gamma_n^2)}}{k!} \leq e^2 p^{1-8c_2\gamma_n^2}. \end{aligned}$$

So far, the probability of the event $\tilde{T}_{\mathfrak{M}}(N_n) = O(L_n^2 \log p/n)$, which we call A, is bounded below conditional on Ω_n. Simple calculation yields $pr(A) \geq pr(A \cap \Omega_n) = pr(\Omega_n) pr(A \mid \Omega_n)$. Thus, $pr(A) \geq (1 - e^2 p^{1-8c_2\gamma_n^2})(1 - O(n^{-\delta}))$. So, (A.9) follows.

We have shown (A.8) and (A.9), which control the terms $\tilde{T}_{\mathfrak{M}}(N_n)$ and $R_{\mathfrak{M}}(N_n)$, respectively. Thus, (A.7) concludes the proof of Lemma 4.

F.5. *Lemma 5 and its proof*

LEMMA 5. *Let q_i's be n independent, but not necessarily identically distributed, scaled and centered random variables with uniform sub-exponential decay, that is,*

$$pr(|q_i| > t) \leq C \exp(-C^{-1} t)$$

for some positive constant C. Let $\|q_i\|_{\psi_1}$ denote the sub-exponential norm defined by

$$\|q_i\|_{\psi_1} := \sup_{m\geq 1}\left[m^{-1}\{E(|q_i|^m)\}^{1/m}\right].$$

Then, we have $\|q_i\|_{\psi_1} \leq e^{1/e}C\max(C,1)$ for all i.

Proof. From the condition on sub-exponential tails, we derive

$$\begin{aligned}E(|q_i|^m) &= m\int_0^\infty x^{m-1}pr(|q_i|\geq x)dx \leq Cm\int_0^\infty x^{m-1}\exp(-C^{-1}x)dx\\ &= CmC^m\int_0^\infty u^{m-1}\exp(-u)du = CmC^m\Gamma(m) \leq CmC^m m^m,\end{aligned}$$

where the last line follows from the definition of the Gamma function. Taking the mth root, we have

$$\{E(|q_i|^m)\}^{1/m} \leq (Cm)^{1/m}Cm.$$

Rewriting above equation, we obtain

$$m^{-1}\{E(|q_i|^m)\}^{1/m} \leq m^{1/m}C^{1/m}C \leq e^{1/e}\max(C,1)C,$$

for all $m\geq 1$. Since the bound is independent of m, it holds that $\|q_i\|_{\psi_1} \leq e^{1/e}C\max(C,1)$ for all i. This completes the proof of Lemma 5.

F.6. *Lemma* 6 *and its proof*

LEMMA 6. *Under Assumption* 1*, for some constant $\gamma > 0$, we have*

$$\sup_n E\{|(u_n^T R_n u_n)/\widetilde{\mu}_n|^{1+\gamma}\} < \infty,$$

where $u_n = B_n^{-1/2}X^T\{Y - E(Y)\}$, $R_n = B_n^{1/2}A_n^{-1}B_n^{1/2}$, and $\widetilde{\mu}_n = \max\{\mathrm{tr}(A_n^{-1}B_n), 1\}$.

Proof. From the expression of $u_n^T R_n u_n$, we have

$$\begin{aligned}u_n^T R_n u_n =& \{Y - E(Y)\}^T X A_n^{-1}X^T\{Y - E(Y)\}\\ =& [\{Y - E(Y)\}^T\mathrm{cov}(Y)^{-1/2}][\mathrm{cov}(Y)^{1/2}XA_n^{-1}X^T\mathrm{cov}(Y)^{1/2}][\mathrm{cov}(Y)^{-1/2}\{Y - E(Y)\}].\end{aligned}$$

Denote $S_n = \mathrm{cov}(Y)^{1/2}XA_n^{-1}X^T\mathrm{cov}(Y)^{1/2}$ and $q = \mathrm{cov}(Y)^{-1/2}\{Y - E(Y)\}$. We decompose $u_n^T R_n u_n$ into two terms, the summations of the diagonal entries and the off-diagonal entries, respectively,

$$u_n^T R_n u_n = q^T S_n q = \sum_{i=1}^n s_{ii}q_i^2 \; + \sum_{1\leq i\neq j\leq n} s_{ij}q_i q_j,$$

where s_{ij} and q_i denote the (i,j)th entry of S_n and ith entry of q. Then, we have

$$\begin{aligned}E\{(u_n^T R_n u_n)^2\} \;=& \sum_{i=1}^n s_{ii}^2 E(q_i^4) \; + \sum_{1\leq i\neq j\leq n} s_{ii}s_{jj}E(q_i^2)E(q_j^2)\\ &+\; 2\sum_{1\leq i\neq j\leq n} s_{ij}^2 E(q_i^2)E(q_j^2).\end{aligned}$$

Using Assumption 1 and the sub-Gaussian norm bound in Lemma 5, both quantities $E(q_i^4)$ and $E(q_i^2)E(q_j^2)$ can be uniformly bounded by a common constant. Hence

$$E\{(u_n^T R_n u_n)^2\} \leq O(1)\cdot[\{\mathrm{tr}(S_n)\}^2 + \mathrm{tr}(S_n^2)].$$

Since S_n is positive semidefinite it holds that $\mathrm{tr}(S_n^2) \leq \{\mathrm{tr}(S_n)\}^2$. Finally noting that $\mathrm{tr}(S_n) = \mathrm{tr}(A_n^{-1}B_n) \leq \widetilde{\mu}_n$, we see that $\sup_n E\{|(u_n^T R_n u_n)/\widetilde{\mu}_n|^{1+\gamma}\} < \infty$ for $\gamma = 1$, which concludes the proof of Lemma 6.

G. ADDITIONAL TECHNICAL DETAILS

Lemmas 7–10 below are similar to those in Lv & Liu (2014). Their proofs can be found in Lv & Liu (2014) or with minor modifications.

LEMMA 7. *Under Assumption* 4*, for* $j = 1, 2$*, we have*

$$c_5 \int_{\delta\in\mathbb{R}^d} e^{-nq_j} 1_{\widetilde{N}_n(\delta_n)} d\mu_0 \le E_{\mu_{\mathfrak{M}}}\left\{e^{-nq_j} 1_{\widetilde{N}_n(\delta_n)}\right\} \le c_6 \int_{\delta\in\mathbb{R}^d} e^{-nq_j} 1_{\widetilde{N}_n(\delta_n)} d\mu_0. \tag{A.1}$$

LEMMA 8. *Conditional on the event* $\widetilde{Q}_n$*, for sufficiently large* n *we have*

$$\begin{aligned} E_{\mu_{\mathfrak{M}}}\left\{U_n(\beta)^n 1_{\widetilde{N}_n^c(\delta_n)}\right\} &\le \exp[-\{\widetilde{\kappa}_n - \rho_n(\delta_n)/2\} d\delta_n^2] \\ &\le \exp\{-(\widetilde{\kappa}_n/2) d\delta_n^2\}, \end{aligned} \tag{A.2}$$

where $\widetilde{\kappa}_n = \lambda_{\min}(V_n)/2$.

LEMMA 9. *It holds that*

$$\int_{\delta\in\mathbb{R}^d} e^{-nq_1} d\mu_0 = \left(\frac{2\pi}{n}\right)^{d/2} |V_n - \rho_n(\delta_n) I_d|^{-1/2} \tag{A.3}$$

and

$$\int_{\delta\in\mathbb{R}^d} e^{-nq_2} d\mu_0 = \left(\frac{2\pi}{n}\right)^{d/2} |V_n + \rho_n(\delta_n) I_d|^{-1/2}. \tag{A.4}$$

LEMMA 10. *For* $j = 1, 2$*, it holds that*

$$\int_{\delta\in\mathbb{R}^d} e^{-nq_j} 1_{\widetilde{N}_n^c(\delta_n)} d\mu_0 \le \left(\frac{2\pi}{n\widetilde{\kappa}_n}\right)^{d/2} \exp\left[-\{(\widetilde{\kappa}_n d\delta_n^2)^{1/2} - (d)^{1/2}\}^2/2\right]. \tag{A.5}$$

LEMMA 11 (VERSHYNIN (2012)). *For independent sub-exponential random variables* $\{y_i\}_{i=1}^n$*, we have that the sub-exponential norm of* $q_i = \{\mathrm{var}(y_i)\}^{-1/2}\{y_i - E(y_i)\}$ *is bounded by some positive constant* C_3*. Moreover, the following Bernstein-type tail probability bound holds*

$$pr\left(|\sum_{i=1}^n a_i q_i| \ge t\right) \le 2\exp\left\{-C_3 \min\left(\frac{t^2}{C_3^2\|a\|_2^2}, \frac{t}{C_3\|a\|_\infty}\right)\right\}$$

for $a \in \mathbb{R}^n$*,* $t \ge 0$.

Lemma 11 rephrases Proposition 5.16 of Vershynin (2012) for the case where $\|q_i\|_{\Psi_1} \le C_3$. Further, for our proof we need to characterize the concentration of the square of a sub-exponential random variable. In this regard, we define a general α-sub-exponential random variable ξ_α which satisfies

$$pr(|\xi_\alpha| > t^\alpha) \le H\exp(-t/H)$$

for $H, t > 0$. The usual sub-exponential q_i's are 1-sub-exponential random variables. It may be useful to note that $\alpha = 1/2$ corresponds to sub-Gaussian random variables.

LEMMA 12 (ERDŐS ET AL. (2012)). *For independent* α*-sub-exponential random variables* q_i^2*, the following Bernstein-type tail probability bound holds*

$$pr\left\{\left|\sum_{i=1}^n a_i q_i^2 - E\left(\sum_{i=1}^n a_i q_i^2\right)\right| \ge t\right\} \le C_4 \exp\left[-C_4\left\{\frac{t}{\sup_i \mathrm{var}^{1/2}(q_i^2)\|a\|_2}\right\}^{2/(2+\alpha)}\right]$$

for $a \in \mathbb{R}^n$*,* $t \ge \sup_i \mathrm{var}^{1/2}(q_i^2)\|a\|_2$*, and* $C_4 > 0$ *depending on the choice of* α, H.

The proof of Lemma 12 follows from that of Lemma 8.2 in Erdős et al. (2012).

[*Received on* ?? ???? 20??. *Editorial decision on* ?? ???? 20??]