A Small-Sample Choice of the Tuning Parameter in Ridge Regression

Philip S. Boonstra, Bhramar Mukherjee, and Jeremy M. G. Taylor

Department of Biostatistics, University of Michigan, Ann Arbor 48109

Supplementary Material

S1 Generalized maximum profile marginal likelihood (GMPML)

In estimating variance components, Harville (1977) suggests to maximize the *restricted* log-likelihood which offsets the log-likelihood to account for bias introduced from estimating "fixed" effects. Casting ridge regression in the mixed model framework, $\boldsymbol{\beta}$ is treated as random and so does not contribute bias to estimation. However, $\boldsymbol{y}$ is centered and $\boldsymbol{x}$ is standardized, which together implicitly estimate β_0 with $\hat{\beta}_0 = 0$. Thus, there is one unknown parameter hidden in the mean of the distribution $\boldsymbol{y}|\lambda, \sigma^2$, and the restricted marginal log-likelihood, denoted as $m_R(\lambda, \sigma^2)$, is as follows (Section 4.3, Harville, 1977):

$$\begin{aligned} m_R(\lambda, \sigma^2) &= m(\lambda, \sigma^2) - \frac{1}{2} \ln \left| \mathbf{1}_n^\top (\boldsymbol{I}_n - \boldsymbol{P}_\lambda) \mathbf{1}_n / \sigma^2 \right| \\ &= -\frac{n-1}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{P}_\lambda) \boldsymbol{y} + \frac{1}{2} \ln \left| \boldsymbol{I}_n - \boldsymbol{P}_\lambda \right| - \frac{1}{2} \ln \mathbf{1}_n^\top (\boldsymbol{I}_n - \boldsymbol{P}_\lambda) \mathbf{1}_n \end{aligned}$$

By standardization of $\boldsymbol{x}$, it can be shown that the last term simplifies to a constant: $-(1/2)\ln(n)$. Replacing each instance of σ^2 with the restricted estimate $\hat{\sigma}^2_\lambda = \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{P}_\lambda)\boldsymbol{y}/(n-1)$, the maximization in (3.3) follows.

S2 Maximum adjusted profile h-likelihood (MAPHL)

The h-loglihood (Lee and Nelder, 1996) is given by

$$\ell_H(\boldsymbol{\beta}, \lambda, \sigma^2) = \ell(\boldsymbol{\beta}, \sigma^2) - p_\lambda(\boldsymbol{\beta}, \sigma^2).$$

When the dispersion and variance components, respectively σ^2 and λ, are unknown, Lee and Nelder propose maximization of the adjusted h-loglihood (Section 4.3, Lee and

Nelder, 1996), to simultaneously estimate $\boldsymbol{\beta}$, λ, and σ^2. This, too, is a restricted log-likelihood. In contrast to $m_R(\lambda, \sigma^2)$ above, the h-loglihood must be adjusted for both β_0 and $\boldsymbol{\beta}$, because there is no marginalization. This adjusted h-loglihood is defined as

$$\begin{aligned}\ell_{HA}(\boldsymbol{\beta}, \lambda, \sigma^2) &= \ell_H(\boldsymbol{\beta}, \lambda, \sigma^2) + \frac{1}{2}\ln(n\sigma^2) + \frac{1}{2}\ln\left|\sigma^2(\boldsymbol{x}^\top\boldsymbol{x} + \lambda)^{-1}\right| \\ &= -\frac{n-1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{x\beta})^\top(\boldsymbol{y} - \boldsymbol{x\beta}) - \frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta} + \frac{1}{2}\ln\left|\lambda(\boldsymbol{x}^\top\boldsymbol{x} + \lambda)^{-1}\right| + \frac{1}{2}\ln(n) \\ &= -\frac{n-1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{x\beta})^\top(\boldsymbol{y} - \boldsymbol{x\beta}) - \frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta} + \frac{1}{2}\ln\left|\boldsymbol{I}_n - \boldsymbol{P}_\lambda\right| + \frac{1}{2}\ln(n).\end{aligned}$$

Sequentially maximizing $\ell_{HA}(\boldsymbol{\beta}, \lambda, \sigma^2)$ with respect to each of $\boldsymbol{\beta}$, λ and σ^2 yields expressions (3.4)–(3.6).

S3 Hyperpenalty based on Gamma distribution

We have

$$\begin{aligned}-p_\lambda(\boldsymbol{\beta}, \sigma^2) - h(\lambda) &= -\frac{p}{2}\ln(\sigma^2) + \frac{p}{2}\ln(\lambda) - \frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta} + (a-1)\ln(\lambda) - b\lambda \\ &= -\frac{p}{2}\ln(\sigma^2) + \frac{p+2a-2}{2}\ln(\lambda) - \lambda\left(\frac{\boldsymbol{\beta}^\top\boldsymbol{\beta}}{2\sigma^2} + b\right)\end{aligned}$$

From the conjugacy of the gamma hyperpenalty, the update for λ follows immediately:

$$\arg\max_{\lambda|\boldsymbol{\beta},\sigma^2}\{-p_\lambda(\boldsymbol{\beta}, \sigma^2) - h(\lambda)\} = \frac{p+2a-2}{\boldsymbol{\beta}^\top\boldsymbol{\beta}/\sigma^2 + 2b}$$

S4 Optimal choice of λ when $\boldsymbol{x}^\top\boldsymbol{x} \neq n\boldsymbol{I}_p$

In the independent-covariates setting, ie when $\boldsymbol{x}^\top\boldsymbol{x} = n\boldsymbol{I}_p$, the choice of λ which minimizes prediction error is $\lambda^* \equiv \arg\min_\lambda \mathrm{E}[(\boldsymbol{x\beta} - \boldsymbol{x\beta}_\lambda)^\top(\boldsymbol{x\beta} - \boldsymbol{x\beta}_\lambda)] = p\sigma^2/\boldsymbol{\beta}^\top\boldsymbol{\beta}$ (Hoerl et al., 1975; Hoerl and Kennard, 1970). Here we give an approximation of λ^* in the general $\boldsymbol{x}^\top\boldsymbol{x}$ setting, assuming n is large. This is based on the following expansion for $\boldsymbol{P}_\lambda = \boldsymbol{x}(\boldsymbol{x}^\top\boldsymbol{x} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{x}^\top$, shown through recursive applications of the Woodbury matrix identity:

$$\begin{aligned}\boldsymbol{P}_\lambda &= \boldsymbol{x}(\boldsymbol{x}^\top\boldsymbol{x})^{-1}\left(\sum_{j=0}^{\infty}(-\lambda\boldsymbol{x}^\top\boldsymbol{x}^{-1})^j\right)\boldsymbol{x}^\top \\ &= \boldsymbol{x}(\boldsymbol{x}^\top\boldsymbol{x})^{-1}\boldsymbol{x}^\top - \lambda\boldsymbol{x}(\boldsymbol{x}^\top\boldsymbol{x})^{-2}\boldsymbol{x}^\top + \lambda^2\boldsymbol{x}(\boldsymbol{x}^\top\boldsymbol{x})^{-3}\boldsymbol{x}^\top + \boldsymbol{R},\end{aligned}$$

with $\boldsymbol{R} = \boldsymbol{x}(\boldsymbol{x}^\top\boldsymbol{x})^{-1}\left(\sum_{j=3}^{\infty}(-\lambda\boldsymbol{x}^\top\boldsymbol{x}^{-1})^j\right)\boldsymbol{x}^\top$. Up to a constant not depending on λ,

$$\begin{aligned}
&\mathrm{E}[(\boldsymbol{x\beta}-\boldsymbol{x\beta}_\lambda)^\top(\boldsymbol{x\beta}-\boldsymbol{x\beta}_\lambda)] \\
&\quad= -2\boldsymbol{\beta}^\top\boldsymbol{x}^\top\boldsymbol{P}_\lambda\boldsymbol{x\beta}+\boldsymbol{\beta}^\top\boldsymbol{x}^\top\boldsymbol{P}_\lambda^2\boldsymbol{x\beta}+\sigma^2\mathrm{Tr}\;[\boldsymbol{x}^\top\boldsymbol{P}_\lambda^2\boldsymbol{x}] \\
&\quad= \lambda^2\boldsymbol{\beta}^\top(\boldsymbol{x}^\top\boldsymbol{x})^{-1}\boldsymbol{\beta}-2\lambda^3\boldsymbol{\beta}^\top(\boldsymbol{x}^\top\boldsymbol{x})^{-2}\boldsymbol{\beta}+\lambda^4\boldsymbol{\beta}^\top(\boldsymbol{x}^\top\boldsymbol{x})^{-3}\boldsymbol{\beta}+\boldsymbol{\beta}^\top f(\boldsymbol{R})\boldsymbol{\beta} \\
&\qquad+\sigma^2\mathrm{Tr}\;[-2\lambda(\boldsymbol{x}^\top\boldsymbol{x})^{-1}+3\lambda^2(\boldsymbol{x}^\top\boldsymbol{x})^{-2}-2\lambda^3(\boldsymbol{x}^\top\boldsymbol{x})^{-3}+\lambda^4(\boldsymbol{x}^\top\boldsymbol{x})^{-4}+g(\boldsymbol{R})].
\end{aligned}$$

We ignore the functions of the remainder term, $f(\boldsymbol{R})$ and $g(\boldsymbol{R})$, as well as the terms containing λ^3 and λ^4, and differentiate with respect to λ to find its minimum:

$$\begin{aligned}
\frac{d}{d\lambda}\mathrm{E}[(\boldsymbol{x\beta}-\boldsymbol{x\beta}_\lambda)^\top(\boldsymbol{x\beta}-\boldsymbol{x\beta}_\lambda)] &\approx 2\lambda\boldsymbol{\beta}^\top(\boldsymbol{x}^\top\boldsymbol{x})^{-1}\boldsymbol{\beta}-2\sigma^2\mathrm{Tr}\;[(\boldsymbol{x}^\top\boldsymbol{x})^{-1}]+6\lambda\mathrm{Tr}\;[(\boldsymbol{x}^\top\boldsymbol{x})^{-2}] \\
&\overset{\text{set}}{=} 0 \\
\Rightarrow \lambda^* &\approx \frac{\sigma^2\mathrm{Tr}\;[(\boldsymbol{x}^\top\boldsymbol{x})^{-1}]}{\boldsymbol{\beta}^\top(\boldsymbol{x}^\top\boldsymbol{x})^{-1}\boldsymbol{\beta}+3\mathrm{Tr}\;[(\boldsymbol{x}^\top\boldsymbol{x})^{-2}]}
\end{aligned}$$

The expression $3\mathrm{Tr}\;[(\boldsymbol{x}^\top\boldsymbol{x})^{-2}]$ is of a smaller order of n than the remaining expressions. Thus, for sufficiently large n, we have $\lambda^* \approx \sigma^2\mathrm{Tr}\;[(\boldsymbol{x}^\top\boldsymbol{x})^{-1}]/\boldsymbol{\beta}^\top(\boldsymbol{x}^\top\boldsymbol{x})^{-1}\boldsymbol{\beta}$, and for n approaching ∞, $n(\boldsymbol{x}^\top\boldsymbol{x})^{-1} \overset{\mathrm{Pr}}{\to} \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$ and therefore $\lambda^* \overset{\mathrm{Pr}}{\to} \sigma^2\mathrm{Tr}\;[\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}]/\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\boldsymbol{\beta}$.

S5 Algorithm for generating $\boldsymbol{\Sigma}_{\boldsymbol{X}}$

Let the matrix $\boldsymbol{\Omega}(\rho)$ be a $p\times p$ block-wise compound symmetric matrix of 10 blocks such that the within-block correlation is ρ. Hardin et al. (2013) propose a strategy to add noise to the off-diagonal elements of $\boldsymbol{\Omega}(\rho)$, both within and between blocks, simultaneously maintaining positive-definiteness, so as to make the resulting correlation matrix more realistic. The algorithm is as follows. First, choose a positive integer m, a constant ψ, and a constant $\delta \in [0, 1-\rho]$. Then, for $i = 1,\ldots,p$, sample $\boldsymbol{u}_i = \{u_{i1},\ldots,u_{im}\}^\top \overset{iid}{\sim} N_m\{\psi\boldsymbol{1}_m, \boldsymbol{I}_m\}$. Construct the $m\times p$ matrix $\boldsymbol{U}$ given by

$$\boldsymbol{U} = \left(\frac{\boldsymbol{u}_1}{\sqrt{\boldsymbol{u}_1^\top\boldsymbol{u}_1}}, \frac{\boldsymbol{u}_2}{\sqrt{\boldsymbol{u}_2^\top\boldsymbol{u}_2}}, \;\cdots\;, \frac{\boldsymbol{u}_p}{\sqrt{\boldsymbol{u}_p^\top\boldsymbol{u}_p}}\right).$$

Then, the covariance matrix of $\boldsymbol{x}$ is given by

$$\boldsymbol{\Sigma}_{\boldsymbol{X}} = \boldsymbol{\Omega}(\rho)+\delta(\boldsymbol{U}^\top\boldsymbol{U}-\boldsymbol{I}_p), \tag{S5.1}$$

where $\boldsymbol{U}$ is re-generated at each iteration. The element in the ith row and the jth column of $\boldsymbol{U}^\top\boldsymbol{U}$ is given by

$$\frac{\sum_{k=1}^{m}u_{ik}u_{jk}}{\sqrt{\boldsymbol{u}_i^\top\boldsymbol{u}_i}\sqrt{\boldsymbol{u}_j^\top\boldsymbol{u}_j}}.$$

From this, the diagonal elements of $\boldsymbol{U}^\top\boldsymbol{U}$ are unit-valued, and therefore only the off-diagonal elements of $\boldsymbol{\Omega}(\rho)$ are affected by adding $\delta(\boldsymbol{U}^\top\boldsymbol{U}-\boldsymbol{I}_p)$. The off-diagonal elements of $\delta(\boldsymbol{U}^\top\boldsymbol{U}-\boldsymbol{I}_p)$ are in $[-\delta,\delta]$. When $m=1$, it can be shown that the off-diagonal elements equal δ with probability $\Pr(u_{i1}u_{j1}>0)$ and $-\delta$ otherwise. As $m\rightarrow\infty$, each off-diagonal element converges in probability to $\delta\psi^2/(\psi^2+1)$. Based on these results, ψ affects the location shift of the perturbation, m affects the variance, and δ affects both the location and the variance. The underlying structure of $\boldsymbol{\Omega}(\rho)$ will be better preserved in the subsequent $\boldsymbol{\Sigma_X}$ from Equation (S5.1) when (i) m is large and $\psi^2/(\psi^2+1)$ is close to zero or (ii) δ is close to zero. Conversely, the structure of $\boldsymbol{\Omega}(\rho)$ will be masked in $\boldsymbol{\Sigma_X}$ when (i) m is small or the magnitude of ψ is large and (ii) δ is close to $1-\rho$.

In the simulation study, we use two configurations of ρ and $\{m,\psi,\delta\}$. For the "approximately uncorrelated" scenario, we use $\rho=0$ and perturb $\boldsymbol{\Omega}(0)$ using $\{m,\psi,\delta\}=\{25,0.25,0.5\}$. Figure S1 gives levelplots of $\boldsymbol{\Omega}(0)$ for $p=100$ and a sample realization of the resulting $\boldsymbol{\Sigma_X}$ under this $\{m,\psi,\delta\}$ configuration. The average of the off-diagonal elements of the plotted $\boldsymbol{\Sigma_X}$ is 0.023, which is nearly equal to $\delta\psi^2/(\psi^2+1)$, since all off-diagonal elements of the original matrix $\boldsymbol{\Omega}(0)$ are zero. In addition, most of the correlations are modest: the minimum, first quartile, median, third quartile, and maximum are, respectively, -0.316, -0.045, 0.025, 0.093, and 0.325. For the "correlated scenario", we use $\rho=0.4$ and and perturb $\boldsymbol{\Omega}(0.4)$ using $\{m,\psi,\delta\}=\{2,1,0.4\}$. Figure S2 gives levelplots of $\boldsymbol{\Omega}(0.4)$ for $p=100$ and a sample realization of the resulting $\boldsymbol{\Sigma_X}$ under this $\{m,\psi,\delta\}$ configuration. The block-structure of $\boldsymbol{\Omega}(0.4)$ is evident in $\boldsymbol{\Sigma_X}$ but muted. The average of the off-diagonal elements of the plotted $\boldsymbol{\Sigma_X}$ is 0.247, and there is greater variation: the minimum, first quartile, median, third quartile, and maximum are, respectively, -0.400, 0.137, 0.316, 0.387, and 0.800. This additional variation is due to both the underlying block structure of $\boldsymbol{\Omega}(0.4)$ and the $\{m,\psi,\delta\}$ configuration.

S6 Additional Simulation Results

Table 5.2 in the main text presents rMSPE corresponding to the subset of simulation settings for which $\pi=0.3$ and $R^2\in\{0.2,0.4,0.8\}$, where π is the first-order autoregressive parameter in expression (5.2) to create $\boldsymbol{\beta}$. Tables S1–S3 are analogous versions of Table 5.2 from the main text for the remaining simulation settings, i.e. when $\pi=0$ and/or $R^2\in\{0.05,0.6,0.95\}$. Finally, Figure S3 graphically presents the empirical mean squared error of the 632-estimate of R^2.

Bibliography

Hardin, J., Garcia, S. R., Golan, D., et al. (2013). A method for generating realistic correlation matrices. *Ann. Appl. Statist.* **7,** 1733–1762.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* **72,** 320–338.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12,** 55–67.

Hoerl, A. E., Kennard, R. W., and Baldwin, K. F. (1975). Ridge regression: some simulations. *Communications in Statistics-Theory and Methods* **4,** 105–123.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *J. Roy. Statist. Soc. Ser. B* **58,** 619–678.

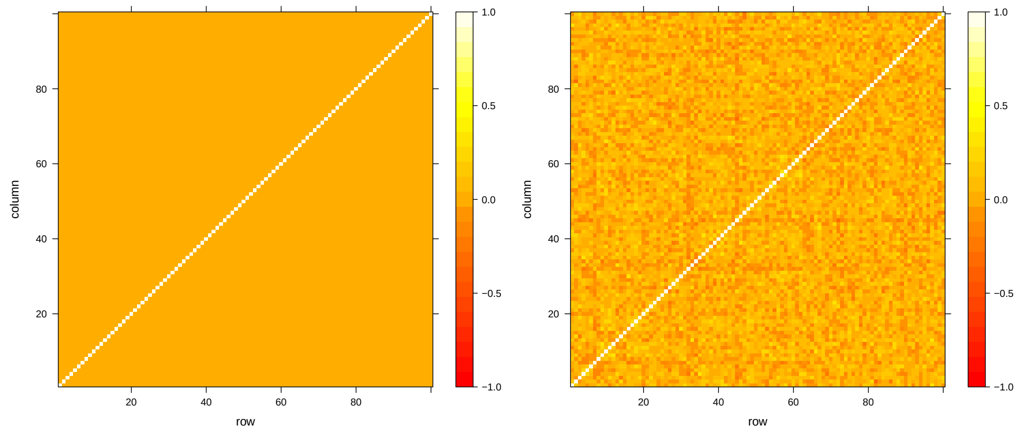

Figure S1: **Approximately Uncorrelated** Levelplots of $\mathbf{\Omega}(0)$ (left panel) and a sample realization of the resulting perturbed matrix $\mathbf{\Sigma}_{\boldsymbol{X}}$ (right panel) using $\{m, \psi, \delta\} = \{25, 0.25, 0.5\}$ (see Supplement S5).

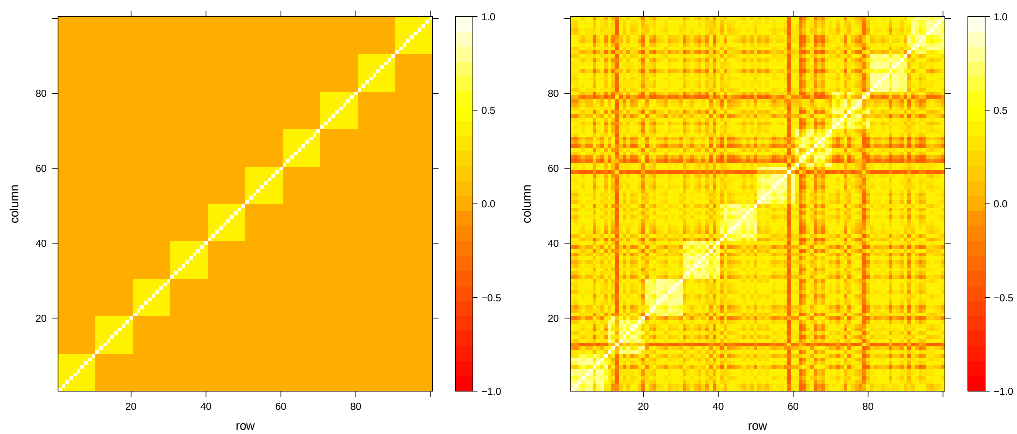

Figure S2: **Correlated** Levelplots of $\mathbf{\Omega}(0.4)$ (left panel) and a sample realization of the resulting perturbed matrix $\mathbf{\Sigma}_{\boldsymbol{X}}$ (right panel) using $\{m, \psi, \delta\} = \{2, 0, 0.4\}$ (see Supplement S5)

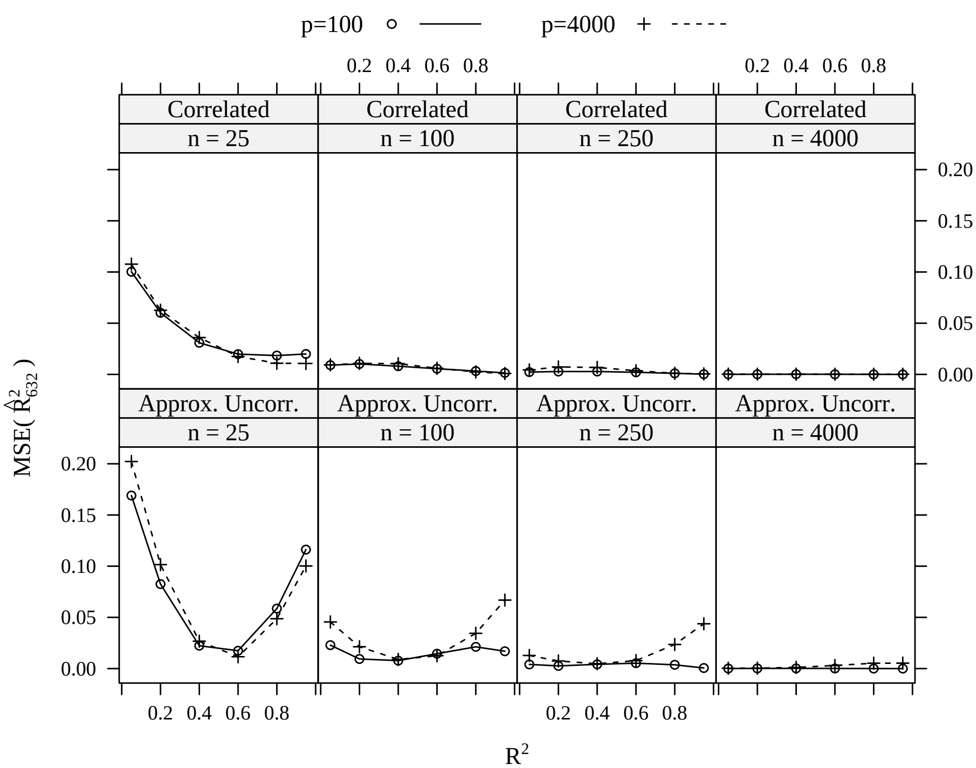

Figure S3: Empirical mean squared error (MSE), $E[\hat{R}^2_{632} - R^2]^2$, where $\hat{R}^2_{632}$ is the 632-estimate of R^2 proposed in the main text.

Table S1: **Additional results when $\pi = 0$:** Average rMSPE, defined in (5.3), for the 11 methods in Table 5.1 and HYP(R^2), which is the hyperpenalty approach using the true value of R^2. Values in **bold** are the column-wise minima, excluding HYP(R^2), and those with an '*' are less than twice the column-wise minimum. The '⇒' indicates a new method.

$p = 100$, Approximately Uncorrelated

Method/$\{n/R^2\}$	25/0.2	100/0.2	250/0.2	4000/0.2	25/0.4	100/0.4	250/0.4	4000/0.4	25/0.8	100/0.8	250/0.8	4000/0.8
5-CV	64.1	23.4	*8.9	***0.8**	72.1	*25.9	*8.8	***0.8**	*73.8	*34.1	*10.4	***0.7**
BIC	369.7	$> 10^4$	103.9	31.6	249.4	$> 10^4$	253.5	16.5	*62.8	$> 10^4$	264.0	3.4
AIC$_C$	*22.0	18.5	*8.5	***0.8**	69.8	40.7	14.6	***0.8**	281.6	214.4	33.1	***0.7**
GCV	79.3	242.8	*7.8	***0.8**	82.5	267.2	*8.2	***0.8**	*63.4	434.8	*9.1	***0.7**
⇒GCV$_C$	*23.4	28.4	*7.8	***0.8**	37.2	*27.5	*8.2	***0.8**	*84.2	49.1	*9.1	***0.7**
RGCV$_{0.3}$	***19.1**	51.1	46.4	13.7	63.7	139.1	115.5	*1.0	276.5	579.5	109.5	***0.7**
MPML	311.9	18.5	***6.8**	***0.8**	234.3	*19.5	***6.7**	***0.8**	*63.7	*25.5	***7.6**	***0.7**
GMPML	57.4	18.0	***6.8**	***0.8**	68.5	***18.9**	***6.7**	***0.8**	*65.4	***24.0**	***7.6**	***0.7**
MAPHL	343.2	21.4	***6.8**	***0.8**	230.6	*22.8	*6.8	***0.8**	***56.2**	*33.0	***7.6**	***0.7**
LR	180.4	17.4	19.2	11.4	164.1	40.7	46.8	16.0	*73.9	176.3	182.0	21.6
⇒HYP($\hat{R}^2_{632}$)	*23.7	***8.2**	*8.0	***0.8**	***16.5**	*19.3	*12.0	***0.8**	*59.2	*46.6	*8.3	***0.7**
HYP(R^2)	*8.2*	*7.1*	*5.3*	*0.7*	*11.1*	*7.7*	*4.3*	*0.8*	*12.8*	*15.2*	*7.5*	*0.7*

$p = 100$, Positively Correlated

Method/$\{n/R^2\}$	25/0.2	100/0.2	250/0.2	4000/0.2	25/0.4	100/0.4	250/0.4	4000/0.4	25/0.8	100/0.8	250/0.8	4000/0.8
5-CV	96.9	*24.6	*7.3	***0.8**	103.7	*22.1	*7.2	*0.9	*95.1	*22.5	*7.8	***0.8**
BIC	600.5	$> 10^4$	62.5	15.0	447.8	$> 10^4$	73.4	24.4	170.9	$> 10^4$	138.1	17.7
AIC$_C$	*43.9	*15.1	*5.4	***0.8**	96.8	*18.2	*5.7	***0.8**	355.9	47.0	*13.6	***0.8**
GCV	121.7	109.8	*6.9	***0.8**	119.7	97.1	*6.6	***0.8**	*87.0	132.0	*7.4	***0.8**
⇒GCV$_C$	*57.7	27.9	*6.8	***0.8**	*63.5	*20.7	*6.5	***0.8**	***70.8**	*25.0	***7.3**	***0.8**
RGCV$_{0.3}$	*47.8	49.9	29.2	5.0	122.9	74.0	29.3	11.5	377.4	88.8	66.4	1.9
MPML	248.8	*15.4	*5.1	*1.0	268.8	*13.9	*6.6	*1.0	166.8	*27.4	*11.3	***0.8**
GMPML	*67.6	*15.3	***5.0**	*1.0	*79.5	***13.4**	*6.4	*1.0	*79.9	*25.3	*11.0	***0.8**
MAPHL	258.1	*16.5	*5.5	*1.1	218.6	*16.1	*7.4	*1.0	*118.2	*34.2	*12.1	***0.8**
LR	*52.0	*25.2	16.8	4.3	*82.2	40.6	24.5	7.6	171.7	81.6	58.6	15.5
⇒HYP($\hat{R}^2_{632}$)	***35.8**	***13.9**	*6.9	*1.4	***41.1**	*16.6	***5.5**	1.8	*87.1	***20.1**	18.1	*1.2
HYP(R^2)	*17.2*	*10.9*	*5.1*	*1.4*	*26.0*	*10.3*	*3.0*	*1.8*	*31.8*	*15.7*	*20.2*	*1.2*

$p = 4000$, Approximately Uncorrelated

Method/$\{n/R^2\}$	25/0.2	100/0.2	250/0.2	4000/0.2	25/0.4	100/0.4	250/0.4	4000/0.4	25/0.8	100/0.8	250/0.8	4000/0.8
5-CV	42.8	20.0	*6.9	***1.0**	47.5	*20.3	*7.5	***1.1**	40.5	*13.0	*6.2	*1.2
BIC	101.1	116.1	93.2	41.1	50.7	68.8	61.6	72.4	***6.9**	*13.3	15.3	370.5
AIC$_C$	34.4	21.4	10.5	2.1	102.1	50.6	22.5	8.0	323.1	163.5	78.0	86.1
GCV	43.3	18.4	*6.5	***1.0**	46.8	*18.7	*7.7	***1.1**	36.6	***12.4**	*6.5	*1.2
⇒GCV$_C$	*17.9	13.9	*5.5	***1.0**	42.4	***14.4**	***5.1**	***1.1**	110.3	*17.0	***5.4**	*1.2
RGCV$_{0.3}$	29.2	36.6	35.2	5.5	91.9	95.9	72.3	19.8	307.6	279.8	137.4	211.6
MPML	100.9	60.5	*5.9	***1.0**	50.7	62.4	11.3	*1.2	***6.9**	*13.3	15.3	***1.1**
GMPML	40.2	17.0	*5.7	***1.0**	47.0	*18.3	*5.6	*1.2	33.8	***12.4**	*7.0	***1.1**
MAPHL	100.8	115.9	93.0	6.7	50.5	68.6	61.4	7.3	***6.9**	*13.2	15.2	6.4
LR	100.7	13.7	12.3	6.3	50.7	*27.3	21.5	14.0	***6.9**	55.9	48.4	73.2
⇒HYP($\hat{R}^2_{632}$)	***9.4**	***6.5**	***4.9**	*1.2	***16.1**	*15.3	*6.7	2.4	74.5	34.6	11.1	7.1
HYP(R^2)	*16.7*	*12.9*	*6.9*	*1.0*	*29.6*	*12.8*	*4.8*	*1.5*	*20.9*	*6.2*	*2.3*	*12.3*

$p = 4000$, Positively Correlated

Method/$\{n/R^2\}$	25/0.2	100/0.2	250/0.2	4000/0.2	25/0.4	100/0.4	250/0.4	4000/0.4	25/0.8	100/0.8	250/0.8	4000/0.8
5-CV	74.5	*17.5	*7.8	*0.8	79.3	*19.2	9.6	***0.8**	*71.0	*14.4	*7.4	***1.0**
BIC	333.4	98.4	85.7	10.8	253.1	74.4	73.8	15.7	*96.6	29.3	52.3	42.1
AIC$_C$	*46.0	*12.8	***4.8**	***0.6**	*74.6	*15.1	*6.3	***0.8**	235.8	49.8	16.9	5.3
GCV	81.8	*19.0	*8.6	*0.8	*74.9	*19.9	9.6	***0.8**	*71.2	*14.2	*7.4	***1.0**
⇒GCV$_C$	*49.7	*14.9	*6.1	*0.8	*48.4	*13.8	*6.4	***0.8**	*61.3	*10.9	***4.3**	***1.0**
RGCV$_{0.3}$	*52.7	33.0	18.5	*0.9	100.0	47.7	25.9	*1.4	248.4	77.6	24.3	12.7
MPML	248.2	*14.9	*5.3	*0.9	237.9	*19.2	*5.0	*1.4	*96.6	28.9	22.2	2.2
GMPML	*65.7	*14.3	*5.2	*0.9	*67.3	*14.4	***4.7**	*1.4	*71.7	*10.6	*6.8	2.2
MAPHL	332.5	98.1	85.0	7.3	252.3	74.2	73.1	8.1	*96.1	29.1	51.7	9.0
LR	82.7	*18.3	10.9	2.3	90.7	*26.1	17.4	3.1	126.3	64.0	31.2	8.6
⇒HYP($\hat{R}^2_{632}$)	***33.8**	***11.5**	*5.1	*0.7	***38.0**	***13.3**	*5.0	2.1	***56.8**	***10.3**	*4.9	32.5
HYP(R^2)	*17.0*	*5.8*	*3.3*	*0.7*	*14.5*	*7.2*	*3.6*	*2.0*	*31.0*	*7.7*	*3.9*	*35.4*

Table S2: **Additional results when $\pi = 0.3$ and $R^2 \in \{0.05, 0.60, 0.95\}$**: Average rMSPE, defined in (5.3), for the 11 methods in Table 5.1 and HYP(R^2), which is the hyperpenalty approach using the true value of R^2. Values in **bold** are the column-wise minima, excluding HYP(R^2), and those with an '*' are less than twice the column-wise minimum. The '⇒' indicates a new method.

$p = 100$, Approximately Uncorrelated

Method/$\{n/R^2\}$	25/0.05	100/0.05	250/0.05	4000/0.05	25/0.6	100/0.6	250/0.6	4000/0.6	25/0.95	100/0.95	250/0.95	4000/0.95
5-CV	63.6	16.3	7.6	*0.9	75.7	*29.5	*9.6	*$\mathbf{0.7}$	74.3	*63.6	*11.7	*0.6
BIC	479.7	$> 10^4$	10.7	29.3	145.5	$> 10^4$	362.6	7.5	*21.2	8804.0	34.3	*1.1
AIC$_C$	*3.6	*4.8	*4.7	*0.9	137.0	90.0	24.6	*$\mathbf{0.7}$	414.4	843.5	25.2	*$\mathbf{0.5}$
GCV	79.7	242.9	7.0	*0.9	77.9	304.2	*8.6	*$\mathbf{0.7}$	51.0	503.7	*9.9	*$\mathbf{0.5}$
⇒GCV$_C$	19.0	17.4	6.9	*0.9	56.3	39.2	*8.6	*$\mathbf{0.7}$	128.1	*86.5	*9.9	*$\mathbf{0.5}$
RGCV$_{0.3}$	*$\mathbf{3.2}$	*7.6	6.1	9.3	128.8	290.5	173.2	*$\mathbf{0.7}$	410.1	510.7	*11.2	*$\mathbf{0.5}$
MPML	367.5	12.2	6.2	*$\mathbf{0.8}$	144.6	*20.3	*$\mathbf{6.9}$	*$\mathbf{0.7}$	*21.5	*49.2	*9.2	*$\mathbf{0.5}$
GMPML	52.9	11.4	6.1	*$\mathbf{0.8}$	72.1	*$\mathbf{19.5}$	*$\mathbf{6.9}$	*$\mathbf{0.7}$	60.1	*$\mathbf{43.5}$	*9.2	*$\mathbf{0.5}$
MAPHL	444.4	18.8	7.0	*$\mathbf{0.8}$	133.3	*24.8	*$\mathbf{6.9}$	*$\mathbf{0.7}$	*$\mathbf{19.0}$	*55.5	*9.2	*$\mathbf{0.5}$
LR	195.4	*$\mathbf{4.5}$	*4.2	4.8	127.8	84.9	97.1	19.4	*31.7	553.3	383.8	24.0
⇒HYP($\hat{R}^2_{632}$)	44.1	9.2	*$\mathbf{2.7}$	*1.3	*$\mathbf{26.1}$	*35.1	*11.9	*$\mathbf{0.7}$	110.8	96.7	*$\mathbf{9.1}$	*$\mathbf{0.5}$
HYP(R^2)	*2.7*	*2.3*	*2.5*	*0.9*	*10.6*	*6.5*	*4.0*	*0.7*	*15.1*	*101.7*	*9.6*	*0.5*

$p = 100$, Positively Correlated

Method/$\{n/R^2\}$	25/0.05	100/0.05	250/0.05	4000/0.05	25/0.6	100/0.6	250/0.6	4000/0.6	25/0.95	100/0.95	250/0.95	4000/0.95
5-CV	79.4	20.1	8.4	*0.9	109.2	*25.2	*9.3	*$\mathbf{0.8}$	*87.0	*45.0	*11.0	*$\mathbf{0.6}$
BIC	738.2	$> 10^4$	20.3	10.6	324.3	$> 10^4$	137.7	19.9	*60.0	8070.9	149.8	2.2
AIC$_C$	*12.2	*9.8	*5.9	*0.9	149.8	33.5	13.3	*$\mathbf{0.8}$	551.9	358.4	33.4	*$\mathbf{0.6}$
GCV	105.5	89.5	8.1	*0.9	118.8	142.8	*8.3	*$\mathbf{0.8}$	*67.6	267.3	*9.6	*$\mathbf{0.6}$
⇒GCV$_C$	42.4	22.9	8.0	*0.9	*67.0	28.8	*8.3	*$\mathbf{0.8}$	108.6	63.8	*9.6	*$\mathbf{0.6}$
RGCV$_{0.3}$	*$\mathbf{9.7}$	*11.6	10.4	3.0	196.2	109.1	57.8	2.1	487.6	443.3	48.1	*$\mathbf{0.6}$
MPML	197.8	*12.4	*6.0	*$\mathbf{0.8}$	259.0	*14.8	*$\mathbf{5.8}$	*$\mathbf{0.8}$	*61.7	*30.5	*$\mathbf{8.7}$	*$\mathbf{0.6}$
GMPML	47.3	*12.1	*6.0	*$\mathbf{0.8}$	*85.9	*$\mathbf{14.3}$	*$\mathbf{5.8}$	*$\mathbf{0.8}$	*72.8	*$\mathbf{28.4}$	*$\mathbf{8.7}$	*$\mathbf{0.6}$
MAPHL	273.0	14.5	*6.1	*$\mathbf{0.8}$	190.7	*17.2	*6.0	*$\mathbf{0.8}$	*$\mathbf{49.8}$	*34.9	*$\mathbf{8.7}$	*$\mathbf{0.6}$
LR	29.7	*8.1	*6.7	3.1	113.1	69.4	59.8	15.2	181.2	357.5	260.5	22.6
⇒HYP($\hat{R}^2_{632}$)	45.8	*$\mathbf{6.9}$	*$\mathbf{3.9}$	*1.3	*$\mathbf{53.6}$	*22.5	13.4	*$\mathbf{0.8}$	133.3	63.7	*9.4	*$\mathbf{0.6}$
HYP(R^2)	*7.0*	*4.2*	*3.0*	*1.1*	*24.7*	*10.9*	*7.9*	*0.8*	*33.2*	*34.4*	*10.2*	*0.6*

$p = 4000$, Approximately Uncorrelated

Method/$\{n/R^2\}$	25/0.05	100/0.05	250/0.05	4000/0.05	25/0.6	100/0.6	250/0.6	4000/0.6	25/0.95	100/0.95	250/0.95	4000/0.95
5-CV	44.9	14.9	7.2	*0.9	*43.2	*17.5	*7.9	*1.1	40.1	*10.4	*4.4	*1.9
BIC	164.8	175.2	130.3	22.4	*31.0	35.9	34.5	187.0	*$\mathbf{4.8}$	*$\mathbf{6.0}$	*$\mathbf{2.7}$	1735.8
AIC$_C$	*3.7	*3.7	*4.2	*$\mathbf{0.8}$	133.4	83.8	42.4	36.7	318.0	210.3	118.2	506.2
GCV	46.3	14.3	6.8	*0.9	*39.4	*17.6	*8.1	*$\mathbf{1.0}$	28.8	*9.0	*4.7	*2.0
⇒GCV$_C$	10.0	11.0	*6.0	*0.9	51.7	*$\mathbf{13.7}$	*$\mathbf{5.1}$	*$\mathbf{1.0}$	110.2	24.3	10.0	*2.0
RGCV$_{0.3}$	*$\mathbf{3.3}$	*$\mathbf{3.6}$	*5.7	5.5	120.2	150.9	109.5	91.0	300.3	343.0	176.7	1055.8
MPML	164.4	33.5	*6.0	*$\mathbf{0.8}$	*31.0	36.5	21.3	*$\mathbf{1.0}$	*$\mathbf{4.8}$	*$\mathbf{6.0}$	*$\mathbf{2.7}$	*$\mathbf{1.6}$
GMPML	41.7	12.9	*5.9	*$\mathbf{0.8}$	*39.0	*15.4	*5.4	*$\mathbf{1.0}$	29.7	*7.7	*3.1	*$\mathbf{1.6}$
MAPHL	164.5	175.0	130.0	5.0	*30.9	35.8	34.3	5.7	*$\mathbf{4.8}$	*$\mathbf{6.0}$	*$\mathbf{2.7}$	13.8
LR	163.9	*4.7	*4.2	2.5	*31.0	40.6	33.7	37.2	*$\mathbf{4.8}$	51.4	66.3	326.5
⇒HYP($\hat{R}^2_{632}$)	24.8	9.2	*$\mathbf{3.1}$	*1.0	*$\mathbf{23.9}$	*25.8	11.0	11.0	81.9	55.6	24.7	29.6
HYP(R^2)	*3.0*	*2.7*	*3.0*	*1.0*	*15.2*	*7.7*	*3.3*	*2.9*	*7.0*	*5.5*	*1.6*	*61.0*

$p = 4000$, Positively Correlated

Method/$\{n/R^2\}$	25/0.05	100/0.05	250/0.05	4000/0.05	25/0.6	100/0.6	250/0.6	4000/0.6	25/0.95	100/0.95	250/0.95	4000/0.95
5-CV	71.9	17.3	8.0	*0.9	*76.0	*16.5	10.1	*$\mathbf{1.0}$	*64.1	*10.7	*5.9	*1.2
BIC	410.7	124.1	101.8	6.8	188.1	50.1	57.8	34.1	*46.8	*9.5	12.9	469.7
AIC$_C$	*15.3	*9.5	*4.7	*$\mathbf{0.7}$	111.3	*22.7	10.1	4.7	395.5	92.1	42.2	122.2
GCV	79.0	16.6	*6.9	*0.9	*73.7	*16.6	9.1	*$\mathbf{1.0}$	*52.5	*9.9	*5.4	*1.1
⇒GCV$_C$	37.6	*12.4	*5.6	*0.9	*49.0	*$\mathbf{11.7}$	*6.0	*$\mathbf{1.0}$	*80.6	*16.1	*4.9	*1.1
RGCV$_{0.3}$	*$\mathbf{13.5}$	*11.7	8.8	2.1	147.1	63.6	26.9	11.3	319.2	100.1	51.3	300.7
MPML	231.7	*12.3	*5.3	*$\mathbf{0.7}$	191.2	24.5	*7.0	*$\mathbf{1.0}$	*46.8	*9.6	12.7	*$\mathbf{1.0}$
GMPML	46.2	*11.8	*5.3	*$\mathbf{0.7}$	*69.7	*12.4	*$\mathbf{4.4}$	*$\mathbf{1.0}$	*51.7	*$\mathbf{8.8}$	*5.3	*$\mathbf{1.0}$
MAPHL	409.7	123.9	101.1	6.2	187.4	49.8	57.2	7.1	*$\mathbf{46.6}$	*9.5	12.7	6.1
LR	51.0	*8.9	*5.4	1.7	115.1	37.9	25.4	9.2	*68.2	86.5	49.5	93.5
⇒HYP($\hat{R}^2_{632}$)	42.0	*$\mathbf{7.6}$	*$\mathbf{3.8}$	*$\mathbf{0.7}$	*$\mathbf{42.9}$	*13.0	*4.8	3.7	*65.7	*13.5	*$\mathbf{4.7}$	36.4
HYP(R^2)	*10.6*	*4.7*	*2.0*	*0.7*	*17.4*	*8.7*	*3.6*	*3.5*	*35.7*	*6.8*	*3.6*	*70.5*

Table S3: **Additional results when $\pi = 0$ and $R^2 \in \{0.05, 0.60, 0.95\}$**: Average rMSPE, defined in (5.3), for the 11 methods in Table 5.1 and HYP(R^2), which is the hyperpenalty approach using the true value of R^2. Values in **bold** are the column-wise minima, excluding HYP(R^2), and those with an '*' are less than twice the column-wise minimum. The '⇒' indicates a new method.

***p* = 100, Approximately Uncorrelated**

Method/$\{n/R^2\}$	25/0.05	100/0.05	250/0.05	4000/0.05	25/0.6	100/0.6	250/0.6	4000/0.6	25/0.95	100/0.95	250/0.95	4000/0.95
5-CV	62.6	16.6	7.4	*0.9	76.8	*29.5	*9.5	***0.7**	70.1	*60.0	*11.8	***0.5**
BIC	476.5	$> 10^4$	11.7	27.3	147.9	$> 10^4$	336.3	8.0	*17.3	9120.5	37.6	1.1
AIC$_C$	*4.0	*5.0	*4.7	***0.8**	150.2	85.8	23.2	***0.7**	454.4	798.1	26.5	***0.5**
GCV	77.0	190.0	6.7	***0.8**	77.4	307.3	*8.4	***0.7**	47.9	502.8	*9.9	***0.5**
⇒GCV$_C$	18.3	18.4	6.5	***0.8**	57.9	*38.3	*8.4	***0.7**	131.0	93.9	*10.0	***0.5**
RGCV$_{0.3}$	***3.5**	19.4	6.5	8.8	142.8	287.3	165.4	***0.7**	453.5	513.2	*11.7	***0.5**
MPML	371.9	12.6	6.1	***0.8**	146.8	*21.0	***6.8**	***0.7**	*17.5	*50.3	***9.3**	***0.5**
GMPML	50.5	11.7	6.0	***0.8**	72.5	***19.9**	***6.8**	***0.7**	54.8	***44.2**	***9.3**	***0.5**
MAPHL	442.9	19.1	6.7	***0.8**	135.1	*26.3	*6.9	***0.7**	***15.5**	*59.8	***9.3**	***0.5**
LR	194.0	***4.6**	*4.4	4.5	128.1	80.9	91.2	18.9	*29.5	524.1	376.8	23.9
⇒HYP($\hat{R}^2_{632}$)	43.0	*8.9	***2.6**	*1.1	***27.4**	*31.8	*10.5	***0.7**	114.0	*84.7	*9.4	***0.5**
HYP(R^2)	*3.1*	*2.6*	*2.7*	*0.8*	*11.2*	*6.1*	*4.2*	*0.7*	*12.5*	*112.7*	*10.2*	*0.5*

***p* = 100, Positively Correlated**

Method/$\{n/R^2\}$	25/0.05	100/0.05	250/0.05	4000/0.05	25/0.6	100/0.6	250/0.6	4000/0.6	25/0.95	100/0.95	250/0.95	4000/0.95
5-CV	80.9	19.8	9.1	***0.7**	*108.3	*21.6	*6.9	***0.8**	*57.9	***31.5**	*9.7	***0.7**
BIC	745.1	$> 10^4$	22.6	8.8	312.1	$> 10^4$	90.3	28.5	*38.8	$> 10^4$	237.0	4.6
AIC$_C$	*12.7	*9.9	*6.3	***0.7**	183.8	*25.3	*7.4	***0.8**	869.7	171.4	29.6	***0.7**
GCV	105.4	88.5	*8.5	***0.7**	*111.4	102.3	*6.5	***0.8**	*47.6	195.1	***8.7**	***0.7**
⇒GCV$_C$	42.0	19.9	*8.3	***0.7**	*69.1	*27.2	***6.4**	***0.8**	129.8	*47.0	***8.7**	***0.7**
RGCV$_{0.3}$	***10.5**	*11.0	11.5	2.1	237.1	87.7	31.7	13.3	445.4	244.8	140.3	***0.7**
MPML	189.8	*12.3	*6.4	*0.8	259.6	*17.2	*9.0	*0.9	*38.8	*46.4	*11.2	***0.7**
GMPML	46.5	*12.1	*6.4	*0.8	*88.8	*16.1	*8.8	*0.9	*42.7	*41.4	*11.0	***0.7**
MAPHL	276.4	14.2	*6.5	*0.8	180.4	*21.4	*9.9	*0.9	***30.9**	*58.5	*11.5	***0.7**
LR	28.7	*8.6	*7.3	2.1	129.1	56.1	34.1	11.1	133.4	178.3	158.0	20.7
⇒HYP($\hat{R}^2_{632}$)	45.0	***6.7**	***4.4**	***0.7**	***61.4**	***14.9**	*7.7	*1.5	149.5	*59.1	29.1	*0.8
HYP(R^2)	*7.7*	*4.9*	*3.8*	*0.6*	*32.8*	*6.7*	*6.1*	*1.5*	*26.8*	*111.6*	*37.1*	*0.8*

***p* = 4000, Approximately Uncorrelated**

Method/$\{n/R^2\}$	25/0.05	100/0.05	250/0.05	4000/0.05	25/0.6	100/0.6	250/0.6	4000/0.6	25/0.95	100/0.95	250/0.95	4000/0.95
5-CV	48.3	14.4	6.9	*0.8	*39.3	*18.4	*7.4	*1.2	32.9	8.7	*3.7	*1.6
BIC	157.3	168.5	131.4	22.0	*21.1	36.9	37.3	149.6	***2.3**	***3.1**	*2.9	1438.9
AIC$_C$	*5.5	***5.2**	*4.1	***0.7**	194.4	88.4	43.1	27.2	472.5	249.7	123.2	409.1
GCV	47.1	13.4	6.2	*0.8	*36.0	*17.5	*7.8	***1.1**	24.6	7.9	*3.9	*1.6
⇒GCV$_C$	9.5	*10.3	5.5	*0.8	65.8	***13.5**	***5.2**	***1.1**	143.9	29.3	8.8	*1.6
RGCV$_{0.3}$	***4.6**	*5.9	6.8	5.0	179.6	166.3	106.3	66.7	460.3	374.5	170.9	897.8
MPML	157.1	28.6	5.7	*0.8	*21.1	38.0	26.8	***1.1**	***2.3**	*3.4	*2.9	***1.3**
GMPML	40.3	11.8	5.6	*0.8	*35.6	*18.3	*6.8	***1.1**	22.1	*5.0	*2.9	***1.3**
MAPHL	157.0	168.2	131.1	5.0	***21.0**	36.8	37.1	6.3	*2.4	***3.1**	***2.8**	11.7
LR	157.1	***5.2**	*4.2	2.3	*21.1	40.2	33.1	30.1	***2.3**	44.6	61.5	272.7
⇒HYP($\hat{R}^2_{632}$)	23.3	*7.5	***2.8**	*0.8	*37.5	*22.6	*8.6	5.1	110.9	49.3	15.6	14.2
HYP(R^2)	*4.7*	*5.1*	*4.4*	*0.9*	*29.4*	*8.7*	*3.0*	*2.6*	*9.1*	*4.0*	*1.6*	*91.5*

***p* = 4000, Positively Correlated**

Method/$\{n/R^2\}$	25/0.05	100/0.05	250/0.05	4000/0.05	25/0.6	100/0.6	250/0.6	4000/0.6	25/0.95	100/0.95	250/0.95	4000/0.95
5-CV	76.6	15.4	*7.7	*0.8	*75.1	*16.5	*7.9	***0.9**	48.2	*8.8	*6.5	***1.0**
BIC	406.1	126.0	99.7	6.8	168.7	48.9	64.8	22.6	***20.1**	*10.1	24.0	181.6
AIC$_C$	*16.0	*9.3	*5.0	***0.6**	124.0	26.2	9.2	*1.6	625.0	110.2	44.4	40.7
GCV	80.2	16.6	*7.9	*0.8	*73.1	*16.6	*7.5	***0.9**	*32.5	*8.9	*6.4	***1.0**
⇒GCV$_C$	41.4	12.9	*6.3	*0.8	*49.9	*12.3	*5.0	***0.9**	118.2	*13.5	***4.3**	***1.0**
RGCV$_{0.3}$	***13.6**	*11.9	9.9	2.1	161.6	69.7	23.7	3.4	392.4	109.7	55.6	101.8
MPML	234.2	*11.8	*5.4	***0.6**	170.7	24.3	*6.2	2.0	***20.1**	*10.1	24.0	*1.6
GMPML	52.3	*11.6	*5.3	***0.6**	*70.3	***11.6**	*4.4	2.0	*28.0	*9.5	15.5	*1.6
MAPHL	405.2	125.7	99.1	5.9	168.1	48.7	64.2	8.8	***20.1**	*10.0	23.6	7.4
LR	62.4	*8.7	*6.2	1.7	121.8	42.9	24.0	4.4	*30.8	78.4	43.0	36.3
⇒HYP($\hat{R}^2_{632}$)	43.5	***6.2**	***4.1**	***0.6**	***46.7**	*11.8	***4.2**	7.7	83.1	***7.4**	*7.5	142.4
HYP(R^2)	*11.2*	*4.9*	*2.3*	*0.6*	*22.1*	*8.7*	*3.3*	*7.9*	*27.8*	*6.0*	*8.6*	*182.0*