Bias reduction in the equicorrelated multivariate normal

Riduzione della distorsione nel modello normale multivariato equicorrelato

Elena Bortolato, Euloge Clovis Kenne Pagui

Abstract In the multivariate normal model, the maximum likelihood estimates can be highly inaccurate with small sample size, or in presence of many covariates. The variance and correlation may result in substantial bias and therefore compromise the inferential conclusions.The paper focuses on the equicorrelated normal model and uses the mean and median bias reduction methods to improve the accuracy of inference. The properties of the resulting estimators are assessed through extensive simulation studies and one application.

Abstract *Nel modello normale multivariato, le stime di massima verosimiglianza possono essere altamente imprecise nel caso in cui la numerosità campionaria non sia particolarmente elevata, o in presenza di molte covariate. Gli stimatori dei parametri di varianza e correlazione risultano distorti e possono compromettere l'attendibilità delle conclusioni inferenziali. Questo lavoro pone l'attenzione sul modello normale multivariato equicorrelato e applica i metodi di riduzione della distorsione in media e in mediana per migliorare l'accuratezza dell'inferenza. Le proprietà degli stimatori risultanti sono verificate mediante ampi studi di simulazione e si prende inoltre in considerazione un'applicazione ad un dataset reale.*

Key words: bias reduction, confidence intervals, likelihood, multivariate normal

1 Introduction

The equicorrelated multivariate model was intensively studied in the decades, both for theoretical properties of estimates (Basu, 1972; De and Mukhopadhyay, 2019), and for building flexible extentions and applications (Engle and Kelly, 2012). One of

Elena Bortolato
University of Padova, Department of Statistical Sciences, e-mail: elena.bortolato.1@phd.unipd.it

Euloge C. Kenne Pagui
University of Padova, Department of Statistical Sciences, e-mail: kenne@stat.unipd.it

the problematic aspects is related to the bias arising from the estimators of the variance and the correlation parameters. The standard approach to inference based on maximum likelihood (ML) might not be accurate when the sample size n is small, or in presence of many covariates. Eventhough the ML estimators of the regression parameters are unbiased, the variance and correlation parameters may result in substantial bias and therefore misleading the inferential conclusion. This affects not only the covariance and correlation parameters, but especially the standard errors of regression coefficients.
As a result, confidence intervals provided by Wald's construction, might be unreliable. In this paper, we show that applying adjustment to the score function according to the procedures derived by Firth (1993) and Kenne Pagui et al. (2017) aiming at mean and median bias reduction (BR) respectively, improves the accuracy of the inference. The performance of the ML, mean and median BR estimators are assessed through Monte Carlo simulations under different settings. An application to a real dataset is considered. Both mean and median bias reduction estimators show better coverages than that obtained with ML estimators.

2 Model specification

Consider n independent observations from a q-variate normal, $Y_i \sim N_q(\mu_i, V)$, $i = 1,...,n$, with $\mu_i = X_i\beta$, where X_i is a $q \times p$ design matrix and $\beta = (\beta_1,\ldots,\beta_p)$. Let $N = n \times q$, and $Y = (Y_1,\ldots,Y_n)^T$, then $Y \sim N_N(\mu, \mathscr{V})$, with $\mu = (\mu_1,\ldots,\mu_n)^T \in \mathbb{R}^N$. In the above, the $N \times N$ block diagonal matrix $\mathscr{V}$ has form

$$\mathscr{V} = \begin{pmatrix} V & 0 & \ldots & 0 \\ 0 & V & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & V \end{pmatrix}, \quad \text{with } V = \sigma^2 \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}$$

The model has a total of $p+2$ parameters. Denoting by Ω the inverse of $\mathscr{V}$, the log-likelihood is

$$\ell(\theta; y) = -\frac{n}{2}[(q-1)\log(1-\rho) + q\log\sigma^2 + \log(q\rho - \rho + 1)] - \frac{1}{2}(y - X\beta)^T\Omega(y - X\beta),$$

where $\theta = (\beta_1,\ldots,\beta_p,\sigma^2,\rho)^T$. The ML estimator $\hat{\theta} \dot{\sim} N_{p+2}(\theta, i^{-1}(\theta))$.

3 Bias reduction

Let $U(\theta) = \partial\ell(\theta)/\partial\theta$, $j(\theta) = -\partial^2\ell(\theta)/\partial\theta\partial\theta^T$ and $i(\theta) = E[i(\theta)]$ be the score vector, the observed information and the Fisher information.

The bias expansion of the ML estimator ($\hat{\theta}$) has form $E_\theta[\hat{\theta}-\theta] = b(\theta)+O(n^{-2})$, where $b(\theta) = i(\theta)^{-1}A^*(\theta)$ with $A^*(\theta)$ having components $A^*_r(\theta) = \frac{1}{2}\text{tr}\{i(\theta)^{-1}[P_r(\theta)+Q_r(\theta)]\}$. In the latter, $P_r(\theta)$ and $Q_r(\theta)$ are $p+2\times p+2$ matrices defined as $P_r(\theta) = E[U(\theta)U(\theta)^T U_r(\theta)]$, $\quad Q_r(\theta) = E[-j(\theta)U_r(\theta)]$, $r = 1,\ldots,p+2$. Firth (1993) proposed an adjusted score of form

$$U^*(\theta) = U(\theta)+A^*(\theta),$$

where the adjustment term $A^*(\theta)$ of order $O(1)$, is built in such a way that $b(\theta)$ is implicitly removed. The resulting estimator, θ^* (mean BR estimator) , solution of the $U^*(\theta) = 0$, has smaller bias than that of ML, that is $E_\theta[\theta^*] = \theta+O(n^{-2})$.
Kenne Pagui et al. (2017) in a similar way develop an adjusted score of form

$$\tilde{U}(\theta) = U(\theta)+\tilde{A}(\theta),$$

built in such a way that the resulting estimator, $\tilde{\theta}$ (median BR estimator), is componentwise third-order median unbiased, that is $Pr_\theta(\tilde{\theta}_r < \theta_r) = 1/2+O(n^{-3/2})$. The adjustment term is $\tilde{A}(\theta) = A^*(\theta)-i(\theta)F(\theta)$, where $F(\theta)$ is a vector of components $F_r = [i(\theta)^{-1}]^T_r\tilde{F}_r$, $r = 1,\ldots,p+2$. The vector $\tilde{F}_r$ has elements $\tilde{F}_{r,t} = \text{tr}\{h_r[(1/3)P_t+(1/2)Q_t]\}$, $t = 1,\ldots,p+2$ and the matrix h_r is defined as $h_r = \{[i(\theta)^{-1}]_r[i(\theta)^{-1}]^T_r\}/i^{rr}(\theta)$, where $[i(\theta)^{-1}]_r$ is the r-th column of $i(\theta)^{-1}$ and $i^{rr}(\theta)$ its r-th element. The estimators $\tilde{\theta}$ and θ^* have the same asymptotic distribution of the ML estimator and this can be used to construct confidence intervals.

4 Simulation studies

We present two simulation studies, in which we compare ML estimator with the mean and median BR estimators. The former focuses on independent and identical distribution case while the latter involves covariates. We draw 10000 samples from $Y \sim \mathcal{N}_N(\mu,\mathcal{V})$, with $n = 10$ and considering $q = 5,15$. The true parameter values are $\mu = 10, \sigma^2 = 5, \rho = 0.9$. The performance of the estimators are evaluated in terms of percentage of underestimation, PU= $R^{-1}\sum_{r=1}^R I_{\{\hat{\theta}_r\le\theta\}}$, with I denoting the indicator function, the relative bias, RB= $R^{-1}\sum_{r=1}^R(\hat{\theta}_r-\theta)/\theta$, empirical 95% Wald confidence interval (WALD) and influence of bias on mean square error, IBMSE = B^2/SD^2), which indicates the relative increase due to bias on the mean square error from its absolute minimum. Here, B and SD denote the bias and standard deviation, respectively. Then we repeat the experiment increasing the sample size to $n = 20$. Results are summarized in table 1. The ML estimator tends to underestimate the variance and correlation parameters and this is more evident for smaller n and larger q, the bias has also an hight impact on the IBMSE index. From the IBMSE, we note that the effect of the bias on the standard error of $\hat{\rho}$ is more pronounced. The mean and median BR estimators succeed in achieving their own desiderable goals, respectively, and the results are preferable than the ML estimator.

Bias reduction methods produce the empirical coverage of confidence intervals which is closer to the nominal 95% level compared to those obtained with the ordinary ML. To assess the properties of bias reduction methods in a regression frame-

			$q=5$			$q=15$		
			ML	mean BR	median BR	MLE	mean BR	median BR
$n=10$	PU	μ	50.00	50.00	50.00	50.01	50.01	50.01
		σ^2	64.88	55.90	49.78	65.92	57.20	50.72
		ρ	62.68	41.92	50.20	64.98	42.75	50.56
	RB	μ	0.03	0.03	0.03	0.02	0.02	0.02
		σ^2	-8.97	0.38	7.59	-9.66	-0.52	6.52
		ρ	-3.68	-0.57	-1.70	-3.64	-0.57	-1.56
	WALD	μ	90.00	91.85	92.60	90.42	92.03	92.58
		σ^2	80.60	85.74	88.00	80.11	85.15	87.54
		ρ	94.37	87.15	91.16	94.30	88.38	91.38
	IBMSE	μ	0.00	0.00	0.00	0.00	0.00	0.00
		σ^2	5.20	0.01	2.59	6.33	0.01	2.01
		ρ	21.71	0.72	5.82	25.84	0.87	6.02
$n=20$	PU	μ	49.08	49.08	49.08	50.25	50.25	50.25
		σ^2	59.66	53.72	49.25	60.35	54.51	49.74
		ρ	57.93	43.04	49.47	60.15	43.83	50.31
	RB	μ	0.09	0.09	0.09	-0.02	-0.02	-0.02
		σ^2	-4.22	0.44	3.75	-4.53	0.04	3.29
		ρ	-1.56	-0.08	-0.69	-1.61	-0.14	-0.69
	WALD	μ	92.63	93.42	93.78	92.86	93.54	93.89
		σ^2	87.13	89.80	91.26	86.78	89.66	91.06
		ρ	94.58	90.14	92.57	94.66	90.81	92.77
	IBMSE	μ	0.04	0.04	0.04	0.00	0.00	0.00
		σ^2	2.13	0.02	1.42	2.57	0.00	1.14
		ρ	12.20	0.04	2.72	15.09	0.15	3.16

Table 1: First simulation study under independent and identical distribution. Estimation of parameter θ=(μ,σ^2,ρ): $\mu=10$, $\sigma^2=5$, $\rho=0.9$.

work, we run a simulation study considering 10000 samples of size 20 from the model

$$\mu_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_4 x_{i4},$$

where x_{i1} is drawn from a Uniform in (-10,10); x_{i2} from an exponential distribution of rate $\frac{1}{2}$; x_{i3} is generated from a Bernoulli $B(1,0.5)$ and x_{i4} from a $B(1,0.2)$. The true values for the parameter is set to $\beta=(2,0.3,-1,3,-0.5)$, with, $\sigma^2=5$ and $\rho=0.9$. We first consider $q=2$. From table 2, the estimators of σ^2 and ρ obtained with the adjusted score fulfill the expected properties. Both mean and median BR estimators perform better than the ML one with respect to the four performance measures. In particular, with the mean BR the RB is significantly reduced with respect to ML while median BR has PU closer to 50% . Similar results are obtained with $q=5$. In this case, the estimator of β is unbiased and identical for the three methods. As a result, PU, RB and IBMSE are equal as shown in table 3. Under the

	$q=2$				$q=5$			
	PU	RB	WALD	IBMSE	PU	RB	WALD	IBMSE
$\hat{\sigma}^2$	83.60	-24.38	64.35	87.75	82.60	-23.10	64.48	83.39
$\hat{\rho}$	73.41	-5.32	95.52	45.47	79.89	-4.93	91.89	67.17
σ^{2*}	56.16	-0.71	86.40	0.04	55.17	0.07	87.02	0.00
ρ^*	44.87	-0.53	87.08	0.82	45.40	-0.38	88.50	0.71
$\tilde{\sigma}^2$	50.97	3.46	88.02	0.91	51.04	4.11	88.71	1.36
$\tilde{\rho}$	50.71	-1.38	90.07	4.96	50.42	-0.93	90.63	3.87

Table 2: Simulations with covariates: estimation of σ^2 and ρ.

two scenarions ($q=2$ and $q=5$), it is remarkable the good performance of the bias reduced estimators in terms of the coverages of confidence intervals.

		$q=2$					$q=5$				
		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
PU		50.19	50.01	48.90	49.16	50.82	49.83	50.67	50.44	50.24	48.96
RB		-0.27	-0.23	-0.44	0.45	6.15	0.04	-0.30	0.29	0.05	-2.90
WALD	ML	89.58	88.69	88.83	89.52	89.34	88.76	89.12	88.97	89.28	88.92
	mean BR	93.55	93.04	93.16	93.69	93.28	93.16	93.23	93.08	93.62	93.16
	median BR	93.88	93.59	93.68	94.02	93.70	93.63	93.69	93.59	93.94	93.48
IBMSE		0.00	0.01	0.02	0.02	0.04	0.00	0.01	0.01	0.00	0.01

Table 3: Simulations with covariates: estimation of regression coefficients.

5 Application

We consider the `Stroke` dataset (Dobson e Barnett, 2008), available in the R package MLGdata on CRAN. This was collected with the aim of study post-heart attack rehabilitation therapies. Patients were assigned to three experimental groups: A, treated with the innovative therapy; B, treated with traditional therapy in the same hospital as the patients of group A; C, treated with traditional therapy in a different hospital. For each of the 24 patients, 8 measures of functional ability were obtained in consecutive weeks. The study aimed to verify whether treatment A was more effective than the others. The model considered is

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5},$$

where $x_{i1}=1$ or $x_{i2}=1$ if the subject belongs to the B or C group therapy, x_{i3} refers to the week, while x_{i4}, x_{i5} represent the interaction terms between the group and the week. Results in table 4 show that the standard errors of the regression coefficients are different for the three approaches.

	ML	mean BR	median BR
β_0	29.82 (7.05)	29.82 (8.07)	29.82 (7.60)
β_1	3.35 (9.97)	3.35 (11.41)	3.35 (10.75)
β_2	-0.02 (9.97)	-0.02 (11.41)	-0.02 (10.75)
β_3	6.32 (0.46)	6.32 (0.45)	6.32 (0.46)
β_4	-1.99 (0.66)	-1.99 (0.63)	-1.99 (0.66)
β_5	-2.69 (0.65)	-2.69 (0.63)	-2.69 (0.66)
σ^2	425.57 (104.88)	547.69 (141.17)	490.86 (123.63)
ρ	0.83 (0.04)	0.88 (0.03)	0.85 (0.03)

Table 4: `Stroke` data: estimates and standard errors in parenthesis.

References

1. Basu, J. P. (1972). Statistical analysis of equicorrelated samples from multivariate population (Doctoral dissertation, Texas Tech University).
2. Firth, D. (1993). Bias reduction of maximum likelihood estimates. Biometrika, 80, 27 – 38.
3. Engle, R. and Kelly, B. (2012) Dynamic equicorrelation Journal of Business & Economic Statistics, 30(2), 212-228 Taylor & Francis
4. Kenne Pagui, E. C., Salvan, A. and Sartori, N. (2017). Median bias reduction of maximum likelihood estimates. Biometrika, 104, 923 – 938.
5. Kenne Pagui, E. C., Salvan, A. and Sartori, N. (2020). Efficient implementation of median bias reduction. arXiv preprint arXiv:2004.08630.
6. De, S. K., and Mukhopadhyay, N. (2019). Two-stage fixed-width and bounded-width confidence interval estimation methodologies for the common correlation in an equi-correlated multivariate normal distribution. Sequential Analysis, 38(2), 214-258.
7. Sartori, N., Salvan, A. and Pace, L. (2020). Package 'MLGdata'. https://CRAN.R-project.org/package=MLGdata