Performances de la régression linéaire dans le cas gaussien

Performance of linear regression in the gaussian case

P. COMON, THOMSON-SINTRA, Parc de Sophia Antipolis, BP 138, 06561 Valbonne Cedex.

Dans le numéro 3, volume 8 de la revue, un article [1] illustre l'utilisation de la régression linéaire pour la soustraction de bruit et en étudie les performances approchées. Cette note présente une alternative pour le calcul des performances des régressions linéaires en général. Nous essayons de montrer de manière convaincante que dans le cas gaussien, il peut parfois être plus simple (paradoxalement) de mener les calculs sans faire d'approximations qu'en en faisant. Nous avons repris des notations compatibles avec celles de [1].

Soient Y une variable aléatoire scalaire et R une variable aléatoire de dimension p. On suppose que Y et R sont conjointement gaussiennes centrées ; elles pourront soit être réelles, soit être complexes circulaires [2] dans la suite. On pose $B = E\{Y/R\}$. Alors B est la meilleure estimation bayesienne de Y et s'écrit :

$$B = \gamma_{YR}\,\gamma_R^{-1}\,R\,, \tag{1}$$

où γ_{YR} désigne le moment croisé $E\{YR^\dagger\}$ et γ_R la covariance de R. Posons $S = Y - B$, l'erreur de régression dans le cas idéal où ces moments sont connus. Les deux variables aléatoires S et B que nous venons de définir sont indépendantes par construction.

Supposons que l'on observe $M + 1$ réalisations indépendantes et identiquement distribuées du couple (Y, R), notées (Y_μ, R_μ), $0 \leq \mu \leq M$. On peut alors définir les moments estimés suivants :

$$Q_{YR} = \frac{1}{M}\sum_{\mu=1}^{M} Y_\mu R_\mu^\dagger\,, \quad \text{et} \quad Q_R = \frac{1}{M}\sum_{\mu=1}^{M} R_\mu R_\mu^\dagger\,. \tag{2}$$

On se propose d'utiliser ces moments pour calculer la régression de Y_0 sur R_0. En conséquence, une erreur supplémentaire est a priori commise et vaut

$$\varepsilon = (\gamma_{YR}\,\gamma_R^{-1} - Q_{YR}\,Q_R^{-1})\,R_0\,, \tag{3}$$

ce qui porte l'erreur totale de régression de S_0 à $S_0 + \varepsilon$. L'expression ci-dessus peut être simplifiée en remarquant que Q_{YR} s'écrit aussi, en utilisant (2) et (1) :

$$Q_{YR} = Q_{SR} + \gamma_{YR}\,\gamma_R^{-1}\,Q_R\,. \tag{4}$$

Il vient alors, avec des notations évidentes, que l'erreur supplémentaire vaut

$$\varepsilon = -\,Q_{SR}\,Q_R^{-1}\,R_0\,. \tag{5}$$

La variance de cette erreur, $E\{|\varepsilon|^2\}$, peut être calculée en prenant les espérances mathématiques sur R_0, S_μ et R_μ successivement, ce qui est autorisé puisque ces variables sont indépendantes :

$$\gamma_\varepsilon = E_{R_\mu}\left\{E_{S_\mu}\left\{E_{R_0}\{Q_{SR}\,Q_R^{-1}\,R_0\,R_0^\dagger\,Q_R^{-1}\,Q_{RS}\}\right\}\right\}.$$

D'où, puisque S est scalaire et que les réalisations μ sont indépendantes :

$$\gamma_\varepsilon = \frac{1}{M}\,E_{S_\mu}\{S_\mu S_\mu^\dagger\} \times E_{R_\mu}\left\{R_\mu^\dagger\,Q_R^{-1}\,E_{R_0}\{R_0 R_0^\dagger\}\,Q_R^{-1}\,R_\mu\right\},$$

de sorte que l'on obtient sans difficulté la relation :

$$\gamma_\varepsilon = \frac{\gamma_S}{M}\,\text{trace}\,\left\{E_{R_\mu}\{Q_R^{-1}\,\gamma_R\}\right\}. \tag{6}$$

Or, la variable MQ_R est, par définition, une variable de Wishart (réelle ou complexe) à M degrés de liberté et de moyenne γ_R ; à l'instar des variables suivant la loi du Chi-deux, la moyenne de l'inverse est par conséquent connue [3] [4], ce qui nous permet de conclure :

$$\gamma_\varepsilon = \theta_1\,\frac{p}{M}\,\gamma_S\,, \quad \text{avec} \tag{7}$$

$$\theta_1 = \frac{M}{M-p} \quad \text{dans le cas complexe circulaire}$$

$$\theta_1 = \frac{M}{M-p-1} \quad \text{dans le cas réel}$$

Remarques

1. On peut réécrire (7) uniquement en fonction des moments des observations en remarquant que

$$\gamma_S = \gamma_Y - \gamma_{YR}\,\gamma_R^{-1}\,\gamma_{RY}\,.$$

2. D'après l'expression (7), il est clair qu'il faut que le nombre de moyennes, M, soit supérieur à la dimension de la matrice Q_R, notée p. Cette condition est nécessaire pour que Q_R soit inversible.

3. Si un autre type de moyennage est utilisé, notamment un lissage spectral ou un moyennage exponentiel, il suffit de remplacer M dans (7) par le nombre de moyennes équivalent [1] [4].

4. Si les estimateurs Q_{YR} et Q_R sont indépendants, ce qui se produit par exemple lorsqu'ils sont construits sur deux jeux d'observations indépendants, alors la variance de l'erreur ε diminue sensiblement. Mais le calcul, dont on pourrait s'attendre à ce qu'il soit plus simple, est en réalité plus compliqué. En effet, (6) devient dans ce cas :

$$(8)\qquad \gamma_\varepsilon = \frac{\gamma_S}{M}\,\mathrm{trace}\,\left\{E\left\{Q_R^{-1}\,\gamma_R\,Q_R^{-1}\,\gamma_R\right\}\right\}.$$

Nous obtenons alors

$$(9)\quad \gamma_\varepsilon = \theta_2\,\frac{p}{M}\,\gamma_S\,,\quad \text{avec}$$

$$\theta_2 = \frac{M^3}{(M-p+1)(M-p)(M-p-1)}$$

dans le cas complexe circulaire

$$\theta_2 = \frac{M^2(M-1)}{(M-p)(M-p-1)(M-p-3)}$$

dans le cas réel.

Conclusion

Les expressions de l'erreur supplémentaire de régression ont été obtenues en (7) et (9) avec, à notre avis, peu de calculs. Pourtant, aucune approximation n'a été pour cela nécessaire, de sorte qu'il s'agit de résultats exacts. Le résultat (7) reprend de façon synthétique la démarche de [4] tout en la clarifiant ; le résultat (9) quant à lui est inédit, et complète le précédent dans le cas où le blanchisseur et le corrélateur sont indépendants.

Remerciements

Je remercie J. F. Guerre-Chaley pour sa lecture soignée de l'article [4], qui a permis d'identifier un certain nombre de coquilles compromettant malheureusement sa compréhension.

BIBLIOGRAPHIE

[1] J. F. Guerre-Chaley et D. Baudois, « Étude du Spectrofiltre Adaptatif Multivariable avec Facteur d'Oubli », *Traitement du Signal*, vol. 8, n° 3, 1991.

[2] D. R. Brillinger, *Time Series Data Analysis and Theory*, Holden Day, 1981.

[3] A. M. Kshirsagar, *Multivariate Analysis*, Dekker, 1972.

[4] P. Comon and D. T. Pham, « An Error Bound for a Noise Canceller », *IEEE Trans. ASSP*, oct. 1989, 1513-1517.