Ann Inst Stat Math (2018) 70:421–438
https://doi.org/10.1007/s10463-016-0592-7

An information criterion for model selection with missing data via complete-data divergence

Hidetoshi Shimodaira[1,2] · **Haruyoshi Maeda**[1,3]

Received: 26 September 2015 / Revised: 4 November 2016 / Published online: 21 January 2017

Abstract We derive an information criterion to select a parametric model of complete-data distribution when only incomplete or partially observed data are available. Compared with AIC, our new criterion has an additional penalty term for missing data, which is expressed by the Fisher information matrices of complete data and incomplete data. We prove that our criterion is an asymptotically unbiased estimator of complete-data divergence, namely the expected Kullback–Leibler divergence between the true distribution and the estimated distribution for complete data, whereas AIC is that for the incomplete data. The additional penalty term of our criterion for missing data turns out to be only half the value of that in previously proposed information criteria PDIO and AICcd. The difference in the penalty term is attributed to the fact that our criterion is derived under a weaker assumption. A simulation study with the weaker assumption shows that our criterion is unbiased while the other two criteria are biased. In addition, we review the geometrical view of alternating minimizations of the EM algorithm. This geometrical view plays an important role in deriving our new criterion.

The research was supported in part by JSPS KAKENHI Grant (24300106, 16H02789).

✉ Hidetoshi Shimodaira
shimo@sigmath.es.osaka-u.ac.jp

[1] Division of Mathematical Science, Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama-cho, Toyonaka, Osaka 560-8531, Japan

[2] RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

[3] Present Address: Kawasaki Heavy Industries, Ltd., 1-1 Kawasaki-cho, Akashi, Hyogo 673-8666, Japan

Keywords Akaike information criterion · Alternating projections · Data manifold · EM algorithm · Fisher information matrix · Incomplete data · Kullback–Leibler divergence · Misspecification · Takeuchi information criterion