Mathematics > Statistics Theory

Consistent Model Selection of Discrete Bayesian Networks from Incomplete Data

Nikolay Balov

(Submitted on 23 May 2011 (v1), last revised 20 Feb 2012 (this version, v2))

A maximum likelihood based model selection of discrete Bayesian networks is considered. The model selection is performed through scoring function S, which, for a given network G and n-sample D_n, is defined to be the maximum log-likelihood l minus a penalization term $\lambda_n h$ proportional to network complexity $h(G)$, $$ S(G|D_n) = l(G|D_n) - \lambda_n h(G). $$ The data is allowed to have missing values at random that has prompted, to improve the efficiency of estimation, a replacement of the standard log-likelihood with the sum of sample average node log-likelihoods. The latter avoids the exclusion of most partially missing data records and allows the comparison of models fitted to different samples.
Provided that a discrete Bayesian network is identifiable for a given missing data distribution, we show that if the sequence λ_n converges to zero at a slower rate than $n^{-{1/2}}$ then the estimation is consistent. Moreover, we establish that BIC model selection ($\lambda_n=0.5\log(n)/n$) applied to the node-average log-likelihood is in general not consistent. This is in contrast to the complete data case where BIC is known to be consistent. The conclusions are confirmed by numerical examples.

Subjects:	**Statistics Theory (math.ST)**
MSC classes:	62F12
Cite as:	**arXiv:1105.4507 [math.ST]**
	(or **arXiv:1105.4507v2 [math.ST]** for this version)

Submission history

From: Nikolay Balov [view email]
[v1] Mon, 23 May 2011 14:08:55 GMT (57kb)
[v2] Mon, 20 Feb 2012 18:52:41 GMT (76kb)