LF-004

ユーザーの事前知識を統合するベイジアン・ネットワーク・ソフトの開発

Development of Bayesian network software which integrates user's prior knowledge

植野　真臣†
Maomi Ueno

1．まえがき

近年、ベイジアン・ネットワーク[1]がデータマイニングや Web Intelligence など実際の場面で広く用いられるようになってきた。ベイジアン・ネットワークはその予測効率の高さから様々な応用に対して注目されるようになってきたともいえる[2][3]。一般にベイジアン・ネットワークでは、(1)データからネットワーク構造の学習,(2)得られた構造と証拠データより未知の変数の値を推論、の二つのエンジンが必要であり、特に構造学習は変数数に対して NP 完全問題である。(1)の学習について Cooper and Herskovits (1992)の方法[4]や MDL を用いた学習法[5]-[8]が提案されている。いずれも一致性を持っており正当な手法とみなされるのであるが、実際のデータに当てはめたときのネットワークの推論効率を比較したところ、必ずしも一致性を持っている手法が推論効率が良いわけではないという結果が示されてきた[2][3]。むしろ相互情報量を用いた MWST 法が最適であるという結果は有名である[2][3]。このことはベイジアン・ネットワーク・モデルが複雑な現実データを表現するためにやや単純すぎること、言い換えると現実には真のモデルがベイジアン・ネットワーク・モデルに含まれていないことに原因すると考えられる。

本論では、従来、あるハイパーパラメータの値の下でのみ構造推定の一致性が証明されてきた[4]-[8]に対して、自由にハイパーパラメータを変化させても漸近的に一致性が成り立つことを証明し、事前知識を用いて一致性を持ったままベイジアン・ネットワークの推論効率を最適化する手法を提案する。具体的には、以下の提案を行う。

①様々な情報量基準に変化する予測分布の提案
最も一般的であるディレクレ-多項分布に基づくベイズ予測分布は、あるハイパーパラメータの値の下でのみ構造推定の一致性が証明されてきた[4]-[8]が本論で自由にハイパーパラメータを変化させても漸近的に一致性が成り立つことを証明し、その漸近予測分布を求める。この予測分布が、ハイパーパラメータを変化させることにより、様々な従来の情報量基準(MDL, BIC, AIC, ICOMP,CAIC)に収束することを示す。また、ユーザーの事前知識（アークのあるなし）をハイパーパラメータに埋め込むことも可能である。

②一致性と推論効率の両立
従来固定されていたハイパーパラメータの値を変化させながら、一致性を持ちさらに推論効率の高いベイジアン・ネットワーク構造学習法があることを証明する。これまで最も推論効率の良いとされていた一致性をもたないアドホック手法 MWST 法[2][3]よりも良い推論効率を得るハイパーパラメータがあることを実データより示した。

③ベイジアン・ネットワーク・ソフトウェア”Bayesian Discovery”の開発
事前分布のハイパーパラメータを自由に設定・最適化できるベイジアン・ネットワーク・ソフトウェア”Bayesian Discovery”に一致性と推論効率を両立させる機能を実装し、提案手法の有用性を高めた。

†電気通信大学大学院情報システム学研究科

2．ベイジアン・ネットワーク・ソフトウェア

これまでにも、多くの有用なベイジアン・ネットワーク・ソフトウェアが開発されてきたが、大別して、(1)確率推論のためのソフトウェア、(2)因果モデルのデータからの学習のための（因果発見）ソフトウェアの二つに分類できる。

(1)では、複雑なネットワーク構造を多重木に変換することで、確率伝播の計算量を減少させる Junction Tree アルゴリズムを搭載している Hugin (http://www.hugin. com), Junction Tree アルゴリズムを標準として、モデルを変えて推論結果を比較できる機能や Gibbs サンプリングの機能を持つ BayesianLab (http://www.bayesia.com)、通常のベイジアン･ネットワークのみでなく、ダイナミック･ベイジアン･ネットワークの推論をさまざまなアルゴリズムで行える Bayes Net Toolbox(BNT)(http://bnt. sourceforge. net/)などが有名であるし、(2)では、Cooper90[4]のアルゴリズムを搭載した商用ソフトの Bayesware Discover (http://www.bayesware.com) 、MML(Minimum Message Language)アプローチによる因果発見ソフトで、GA(遺伝アルゴリズム)、Metropolis search などの確率的探索手法を搭載したフリーソフト CaMML(Causal Discovery vis MML)(http://www.dataming.monash.edu.au/software/camml/)、PC アルゴリズムと GA(遺伝アルゴリズム)探索機能を持つ TETRAD (http://www.phil.cmu.edu/projects/tetrad/)、決定木の出力を同時に行えるフリーソフト WinMine (http://research.microsoft.com/~dmax/winmine/tooldoc.htm)、さらに、[Cooper90][4]、C4.5, AIC,MDL, ML を評価基準として持ち、ユーザーが新しい評価基準を定義でき、全探索、Greedy Search を探索機能として持つ国産初のソフト BayoNet[9] (http://www.msi.co.jp/BAYONET/ index.html) などが挙げられる。筆者も変数選択機能を持ち、様々なハイパーパラメータと情報量基準を持つソフトウェア「Bayesian Discovery」を開発してきた。今回、このソフトウェアにベイズ統計的基礎に忠実であり、矛盾無く、ユーザーの事前知識を因果学習に取り込むことができる機能を追加し、さらにこれを用いて一致性を持ち、さらに推論効率の高いモデル構築機能を追加したソフトを報告する。

3．ベイジアン・ネットワーク

3.1　モデル

今 $x=\{x_1,x_2,\cdots,x_N\}$ を離散 N 変数集合とし、各変数は r_i 個の状態集合 $\{1,\cdots,r_i\}$ の中からひとつの値をとるとする。ここで、変数 x_i が値 k をとるときに $x_i=k$ と書くことにし、バックグラウンド情報 ξ、$y=j$ を所与としたときの $x=k$ の条件付確率

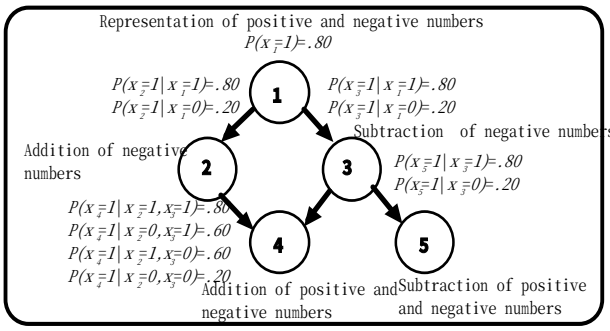

図１　ベイジアン・ネットワークの例

を $p(x=k \mid y=j,\xi)$ と書くことにする。Belief networks は、確率構造 S と条件付確率パラメータ集合Θによって(S, Θ)として表される。図１は、確率構造 S の一例である。例えば、$A \longrightarrow B$ という確率構造は、変数 B が変数 A に依存していることを示している。ここで、A は B の親ノードと呼ぶことにする。結果として、Belief networks の同時確率分布は、条件付パラメータによって、いわゆる“チェーンルール”を用いて以下の式（１）のように示される。

$$P(x_1, x_2, \cdots, x_N) = \prod_{i=1}^{N} p(x_i \mid x_1, x_2, \cdots, x_{i-1}) \quad (1)$$

ベイジアン・ネットワークでは、構造 S を所与として、同時確率分布を以下のように表すことができる。

$$P(x_1, x_2, \cdots, x_N \mid S) = \prod_{i=1}^{N} p(x_i \mid \Pi_i, S) \quad (2)$$

ここで、$\Pi_i \subseteq \{x_1, x_2, \cdots, x_{q_i}\}$ は、変数 i の親ノード集合を示している。

3.2　事後分布と予測分布

今、θ_{ijk} を親ノード変数集合Π_iが j 番目のパターンをとったときの $x_i = k$ となる条件付確率を示すパラメータとする。このとき、データ $\mathbf{X}$ を得たときの尤度は、以下のとおりである。

$$p(\mathbf{X} \mid \Theta_S, S) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\sum_{k=0}^{r_i-1} n_{ijk}!}{\prod_{k=0}^{r_i-1} n_{ijk}!} \prod_{k=0}^{r_i-1} \theta_{ijk}^{\,n_{ijk}} \quad (3)$$

共益自然事前分布である以下のディレクレ分布を事前分布に考えれば、

$$p(\Theta_S \mid S) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\sum_{k=0}^{r_i-1} n'_{ijk})}{\prod_{k=0}^{r_i-1} \Gamma(n'_{ijk})} \prod_{k=0}^{r_i-1} \theta_{ijk}^{\,n'_{ij}-1} \quad (4)$$

以下のような事後分布を得る。

$$p(\mathbf{X}, \Theta_S \mid S) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\sum_{k=0}^{r_i-1} (n'_{ijk}+n_{ijk}-1))}{\prod_{k=0}^{r_i-1} \Gamma(n'_{ijk}+n_{ijk}-1)} \prod_{k=0}^{r_i-1} \theta_{ijk}^{\,n'_{ijk}+n_{ijk}-1} \quad (5)$$

$$\propto \prod_{i=1}^{N} \prod_{j=1}^{q_i} \prod_{k=0}^{r_i-1} \theta_{ijk}^{\,n'_{ijk}+n_{ijk}-1}$$

ここで、MAP 推定量（maximum a posterior estimator）は、$\hat{\theta}_{ijk} = \frac{n'_{ijk}+n_{ijk}}{n'_{ij}+n_{ij}}$ となる。ただし、$n'_{ij} = \sum_{k=0}^{r_i-1} n'_{ijk}$ ，$n_{ij} = \sum_{k=0}^{r_i-1} n_{ijk}$ 。さらにこれらより、予測分布を求めると

$$p(\mathbf{X} \mid S) = \int_{\Theta_S} p(\mathbf{X}, \Theta_S \mid S) p(\Theta_S) d\Theta_S \quad (6)$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(n'_{ijk})}{\Gamma\left[\sum_{k=0}^{r_i-1} (n'_{ijk}+n_{ijk})\right]} \prod_{k=0}^{r_i-1} \frac{\Gamma(n'_{ijk}+n_{ijk})}{\Gamma(n'_{ijk})}$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(n'_{ij})}{\Gamma(n'_{ij}+n_{ij})} \prod_{k=0}^{r_i-1} \frac{\Gamma(n'_{ijk}+n_{ijk})}{\Gamma(n'_{ijk})}$$

4．予測分布と漸近展開

式(6)の予測分布の形状ではこの分布が一致性があるのかどうかが判断できない。そこで漸近展開を行う。これまでにも Suzuki[6][7]によって求められた漸近近似では、すべてのハイパーパラメータが 1/2 のとき一致性を持つ MDL 基準に収束することが証明されているし、Bouckaert[8]によって求められた漸近近似では、すべてのハイパーパラメータが 1 のとき一致性を持つ MDL 基準に収束することが証明されている。Suzuki[7]は、Bouckaert[8]の展開が誤りであることを指摘し[7]、すべてのハイパーパラメータが 1/2 が最適であることを指摘している。

本論で導く漸近分布は、Suzuki[7]、Bouckaert[8]の主張がどちらも正しいこと、さらにすべての条件付確率に自由にハイパーパラメータを設定しても一致性が成り立つことを示す。すなわち、式(6)の予測分布について以下の定理が証明される（証明は付録参照）。

[Theorem 1]

For $\forall n'_{ijk} > 0$,

$$\log p(\mathbf{X} \mid S) \le \log p(\mathbf{X}, \hat{\Theta}_S \mid S) - \left(\frac{\mathrm{K}}{2}\right) \log \frac{(n'+n)}{2\pi} + const + O(1/n) \quad (7)$$

一般に情報量基準についてペナルティ項 K にかかる係数 c が

$$\frac{\log\log n}{2} < c < \frac{n}{2}$$

のとき強一致性(Strong Consistency)を持つことが知られている。したがって、式(7)は強一致性を満たし、式(6)も満たしていることを示す。定理１の特徴は、すべての条件付確率パラメータに関するハイパーパラメータを自由に変化させても成り立つことであり、この定理に従えば、

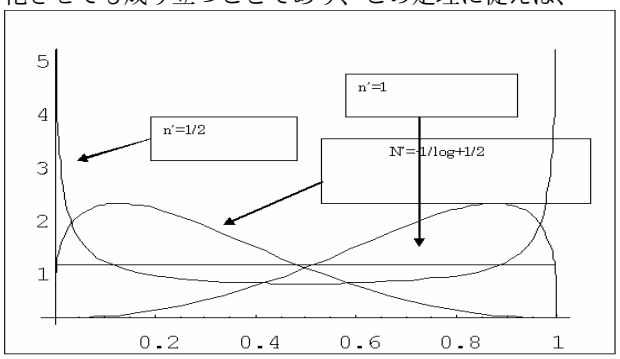

図２．ハイパーパラメータの値と周辺事前分布

Cooper らの K2[4]も一致性を持つし、Suzuki[6][7]の主張も Bouckert[8]の主張も正しいことになる。実際には、ハイパーパラメータ n'_{ijk} の値を変化させることにより、ディレクレ分布の周辺分布ベータ分布は図２のように変化する。すなわち、ハイパーパラメータ n'_{ijk} の値を大きくすればするほど、条件付確率パラメータの事前分布のモードは 0.5 付近で凸となり、推定値の縮約機能を高める。また、ハイパーパラメータ n'_{ijk} を大きくすればするほど対応するアークはつきにくくなるのである。

５．ハイパーパラメータの変化による様々な情報量基準の表現

また、ハイパーパラメータの値によって、予測分布の性質が異なってくる。ここでは、式(7)のハイパーパラメータｎ’を変化させて予測分布の性質を分析する。

(1) $n'=2\pi\exp(2)-n$ のとき

$$\log\ p(\mathbf{X}\mid S)\longrightarrow\log\ p(\mathbf{X},\hat{\Theta}_S\mid S)-K:AIC \qquad (8)$$

$n'>0$ より、この性質は、データ数 n が 46 以下まで成り立つ。すなわち、AIC は一般に一致性を持たないが、ここ

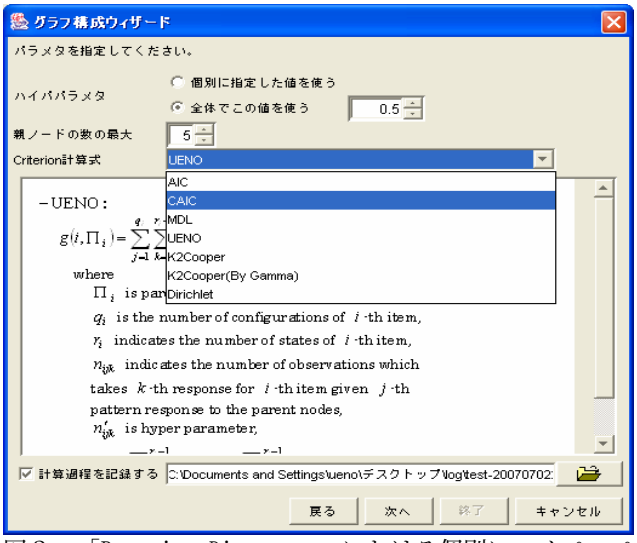

図３．「Bayesian Discovery」における個別にハイパーパラメータを設定できる機能

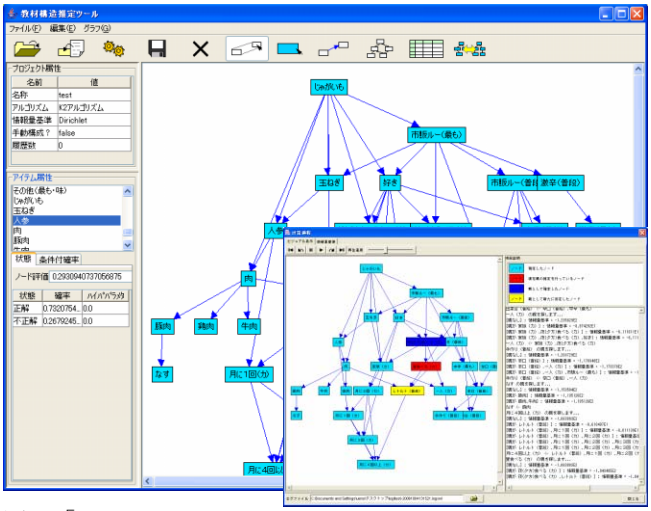

図４「Bayesian Discovery」のインターフェース

表１．ハイパーパラメータの変化による一致性の比較

n	AIC	MDL	K2	ICOMP
50	0	0	0	0
100	205	75	18	1
150	423	347	222	8
200	486	641	487	59
250	494	786	634	166
300	497	908	752	385
400	498	951	782	618
500	487	972	835	880
600	480	983	870	974
700	458	985	879	994
800	462	980	872	998
900	432	988	881	999
1000	424	988	897	999

でも少数データの時の予測分布に一致するだけである。

(2) $n'=(2\pi-1)n$ のとき

$$\log\ p(\mathbf{X}\mid S)\longrightarrow\log\ p(\mathbf{X},\hat{\Theta}_S\mid S)-\frac{1}{2}K\log n:MDL \qquad (9)$$

$n'=(2\pi-1)n$ のとき、MDL 基準に一致する。

(3) $n'=1$ のとき

$$\log\ p(\mathbf{X}\mid S)\longrightarrow\log\ p(\mathbf{X},\hat{\Theta}_S\mid S)-\frac{1}{2}K\log\frac{n+1}{2\pi}:ICOMP \qquad (10)$$

6．ソフト「Bayesian Discovery」

「Bayesian Discovery」は Java で開発されたベイジアン・ネットワーク・ソフトウェアである。このソフトでは、式(6)や式(7)の各条件付確率のハイパーパラメータを図３のように設定できることが特徴である。また、構造のサーチアルゴリズムは、 1.全数探索,2. CH Procedure[Cooper90],3.Greedy Algorithm(A),4. Greedy Algorithm(B),5.各アークのある・なしを遺伝子とみた GA(遺伝アルゴリズム)を用意している。すべての計算過程は LOG として保存され、途中経過もすべて可視化できるプログラムを備えている（図４）。

７．一致性と推論効率の評価

7.1 一致性評価

図１の構造よりデータを 50 から 1000 個発生させ、本ソフトを用いて５章で示した Cooper, MDL,ABIC, ICOMP に対応するハイパーパラメータを設定して構造推定を行った。このプロセスをそれぞれ 1000 回繰り返した結果、各手法について正しく構造を推定して 1000 回中モデルが的中した回数を表１に示した。表より、少数データでは AIC に対応するハイパーパラメータが良い値を示しているが、一致性を持っていないためによい漸近的な成績を出せず、さらにオーバーフィッティングとなっていることがわかる。一方、一致性を持つ MDL に対応したハイパーパラメータでは 1000 個のデータで 989 回の的中回数に収束している。式(7)のベイズ的アプローチでは MDL ほど成績はよくないが、897 回に収束している。本稿で示したようにハイパーパラメータが１の一様分布を仮定した ICOMP が最も成績が良い

ことがわかる。しかし、これらはあくまでもシミュレーション結果であり実データではない。文献[2][3]で示されるように必ずしも一致性を持っている手法が必ずしも確率推論の効率が良いわけではないという結果が示されてきた。本提案手法では、ハイパーパラメータを変化させることにより、一致性をもったままこの確率推論の効率を最適化できると考えている。

7.2 現実データでの推論効率評価

ここでは閲覧ユーザ数1000 万以上，サイト内全ページ数10 万以上の大規模Webサイト「ISIZE」における閲覧データを用いてランダムに選択されたサイト11個中、10個の証拠データを得たときに残りの1個を推論させたときの正答確率を、式(7)のハイパーパラメータを0.3，0.5，0.8,1.0と変化させたときとMDL、現在最も推論効率が良いとされるMWST法[2][3]を用いて構造学習を行ったときの100人のデータについて求め平均値を比較した。結果は表２のとおりである。

表2. ハイパーパラメータの変化による推論効率の比較

	n'=0.3	n'=0.5	n'=0.8	n'=1.0	MDL	MWST
的中率	.724	.781	.584	.482	.421	.722
標準偏差	0.169	0.158	0.176	0.217	0.223	0.161

表より、一致性の評価でよい成績を出していた MDL や n'=1.0 のときの予測分布が推論効率が非常によくないことがわかる。これは先行研究で示されたとおりである。また、文献[2][3]で示されたように伝達情報量を用いたアドホックな手法 MWST 法では高い推論効率を示せた。しかし、MWST 法には一致性がない。注目すべきは、式(7)のハイパーパラメータを変化させることにより、n'=0.5 のとき、最もよい推論効率を示すことができた。この方法では、一致性も満たしていることが重要であり、MWST 法よりも良い効率を得ることができたことが最も重要な知見である。
ハイパーパラメータが 1 以下になり小さくなると、対応するアークは真のモデルに対して冗長（因果が認められやすい）になりやすい。すなわち、ベイジアン・ネットワーク・モデルは一般的に多くの現実に対してやや単純すぎることを意味しており、そのことをハイパーパラメータが調整していることになるのである。

８．おわりに

本論では、新しいベイジアン・ネットワーク・ソフトウェア「Bayesian Discovery」を開発し報告した。その特徴は、条件付確率パラメータの事前分布に一般的な表現であるディレクレ-多項分布に基づくベイズアプローチにより、事前分布のハイパーパラメータの設定により本手法が様々な情報量基準を表現することができることである。さらにソフト上の個々の変数間についてのハイパーパラメータを設定することにより、ユーザーの事前知識を構造学習に組み込める。ハイパーパラメータの適切な設定により、一致性を持ち、推論効率の高いベイジアン・ネットワークが構成できることを示し、これまで最もよいとされてきた MWST 法よりも良い推論効率を示した。さらにハイパーパラメータの推測を最適化する経験的ベイズ手法を用いることにより、より推論効率の良いハイパーパラメータを探索できると考えられ、今後の課題にしたい。

参考文献

[1]Pearl, J: “Probabilistic reasoning in intelligent systems”, Morgan Kaufmann Publishers, California, 1988

[2]Cheng, J., Hatzis, C., Hayashi, H., Krogel, M.,Morishita, S., Page, D., and Sese, J.: KDDD cup 2001 report, *ACM SIGKDD Explorations*, Vol. 3, No. 2 ,2002.

[3]Cheng, J. and R.Greiner, : Comparing Bayesian Network Classi.ers, *proceedings of the fifteenth conferenceon uncertainty in artificial intelligence*,1999.

[4]Cooper,G.F. and Herskovits,E.: A Bayesian method for the induction of probabilistic networks from data, *Machine Learning* , **9**, pp.54-62, 1992

[5]Lam,W., and Bachus, F.:”Learning Bayesian belief networks:An approach based on the MDL principle. *Computational Intelligence* **10**, pp. 269-293, 1993

[6]Suzuki, J., “A construction of Bayesian networks from databases on an MDL principle”, Proc. of Ninth Conference on Uncertainty in Artificial Intelligence, Washington D.C., pp.266-273, 1993.

[7]Suzuki, J., “Learning Bayesian belief networks based on the MDL principle: An efficient algorithm using the branch and bound technique”, *IEICE Transactions on Information and Systems*, **Vol. E82-D**, No.2, 1999

[8]Bouckaert, R., “Properties of Bayesian network learning algorithm”, Proc of Tenth Conference on Uncertainty in Artificial Intelligence, California, pp.102-109, 1994

[9] 本村 陽一: ベイジアンネットソフトウェア BayoNet, 計測と制御, 42-8 , 693 -694, 2003

[10] 植野真臣：変数選択機能を持つベイジアン・ネットワーク・ソフトの開発、人工知能学会全国大会論文誌、2P1,2006

[付録: Theorem 1 の証明] 以下の Stirling series 展開を用いて以下が証明できる。

$$\log p(\mathbf{X}|S)=\sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\sum_{k=0}^{r_i-1}\log\Gamma(n'_{ijk}+n_{ijk})-\log\Gamma\left[\sum_{k=0}^{r_i-1}(n'_{ijk}+n_{ijk})\right]\right)+const$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\begin{array}{l}\sum_{k=0}^{r_i-1}(n'_{ijk}+n_{ijk})\log(n'_{ijk}+n_{ijk})-\sum_{k=0}^{r_i-1}(n'_{ijk}+n_{ijk})\log(n'_{ij}+n_{ij})+\frac{r_i-1}{2}\log(2\pi)\\ -\frac{1}{2}\sum_{k=0}^{r_i-1}\log(n'_{ijk}+n_{ijk})+\frac{1}{2}\log(n'_{ij}+n_{ij})\end{array}\right)$$

$$+const+O(1/n)$$

$$Since\ \ \log(n'_{ij}+n_{ij})\geq\log(n'_{ijk}+n_{ijk}),$$

$$\geq\sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\begin{array}{l}\sum_{k=0}^{r_i-1}(n'_{ijk}+n_{ijk})\log\frac{(n'_{ijk}+n_{ijk})}{(n'_{ij}+n_{ij})}+\frac{r_i-1}{2}\log(2\pi)\\ -\frac{1}{2}\sum_{k=0}^{r_i-1}\log(n'_{ij}+n_{ij})+\frac{1}{2}\log(n'_{ij}+n_{ij})\end{array}\right)+const+O(1/n)$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\begin{array}{l}\sum_{k=0}^{r_i-1}(n'_{ijk}+n_{ijk})\log\frac{(n'_{ijk}+n_{ijk})}{(n'_{ij}+n_{ij})}\\ -(\frac{r_i-1}{2})\log\frac{(n'_{ij}+n_{ij})}{2\pi}\end{array}\right)+const+O(1/n)$$

$$Since\ \ \log(n'+n)\geq\log(n'+n_{ij}),$$

$$\log p(\mathbf{X}|S)\leq\log p(\mathbf{X},\widehat{\Theta}_S|S)-\left(\frac{\mathrm{K}}{2}\right)\log\frac{(n'+n)}{2\pi}+const+O(1/n)$$

(証明終)