Quaderni di Dipartimento

Moment Priors for
Bayesian Model Choice with Applications
to Directed Acyclic Graphs

Guido Consonni
(Università di Pavia)

Luca La Rocca
(Università di Modena e Reggio Emilia)

# 115 (06-10)

Dipartimento di economia politica
e metodi quantitativi
Università degli studi di Pavia
Via San Felice, 5
I-27100 Pavia

Giugno 2010

Moment Priors for Bayesian Model Choice with Applications to Directed Acyclic Graphs

Guido Consonni & Luca La Rocca

Università di Pavia, Italy *Università di Modena e Reggio Emilia, Italy*

guido.consonni@unipv.it luca.larocca@unimore.it

Summary

We propose a new method for the objective comparison of two nested models based on *non-local* priors. More specifically, starting with a default prior under each of the two models, we construct a *moment prior* under the larger model, and then use the fractional Bayes factor for a comparison. Non-local priors have been recently introduced to obtain a better separation between nested models, thus accelerating the learning behaviour, relative to currently used *local* priors, when the smaller model holds. Although the argument showing the superior performance of non-local priors is asymptotic, the improvement they produce is already apparent for small to moderate samples sizes, which makes them a useful and practical tool. As a by-product, it turns out that routinely used objective methods, such as ordinary fractional Bayes factors, are alarmingly slow in learning that the smaller model holds. On the downside, when the larger model holds, non-local priors exhibit a weaker discriminatory power against sampling distributions close to the smaller model. However, this drawback becomes rapidly negligible as the sample size grows, because the learning rate of the Bayes factor under the larger model is exponentially fast, whether one uses local or non-local priors. We apply our methodology to directed acyclic graph models having a Gaussian distribution. Because of the recursive nature of the joint density, and the assumption of global parameter independence embodied in our prior, calculations need only be performed for individual vertices admitting a distinct parent structure under the two graphs; additionally we obtain closed-form expressions as in the ordinary conjugate case. We provide illustrations of our method for a simple three-variable case, as well as for a more elaborate seven-variable situation. Although we concentrate on pairwise comparisons of nested models, our procedure can be implemented to carry-out a search over the space of all models.

June 2010. This work was partially supported by PRIN grant 2007XECZ7L_001 (MIUR-Italy). The authors would like to thank Alberto Roverato, Giovanni Marchetti and Piero Veronese for useful discussions; thanks also to Davide Altomare for checking some numerical results.

Keywords and Phrases:

Fractional Bayes factor; Gaussian graphical model; Non-local prior; Objective Bayes.

1. INTRODUCTION

Bayesian model choice is an important and fascinating area. In particular, the choice of suitable parameter priors is still a challenge, especially if an objective analysis is pursued; the latter being almost inevitable when the number of models is large, because subjective elicitations are not a viable option. Unfortunately, standard default priors for estimation or prediction, which are known to perform very well within the standard single-model paradigm, are not appropriate for Bayesian model comparison, if the marginal likelihood is used as a measure of the support for the model, as with Bayes factors. Standard default priors are obviously unsuitable when they are improper, because the marginal likelihood would be defined only up to an arbitrary constant. Interestingly, however, they are also inappropriate when they are proper (as it may happen for discrete data models). The reason is best understood when comparing two nested models and can be succinctly put as follows: the prior on the larger model tends to be too diffuse for typical data sets, thus unduly favouring the smaller model. This feature is closely related to the Jeffreys-Lindley's paradox; see Robert (2001, sect. 5.2.5).

Several attempts have been made to produce objective Bayesian model comparisons. The notions of partial Bayes factor, intrinsic Bayes factor and fractional Bayes factor stand out as major contributions; see Pericchi (2005) for a comprehensive review. More specific contributions have appeared in specialized areas, notably variable selection in linear models; see Liang *et al.* (2008) and references therein.

A recent area of research concerns the rate of learning of Bayesian model selection procedures, and this has important implications on the choice of priors. Consider for simplicity two nested models. Most currently used parameter priors, whether subjective or objective, share a common structural feature: they are *local*, i.e., the prior under the larger model does not vanish on the parameter subspace characterising the smaller model. This aspect is epitomized in testing a sharp null hypothesis on the mean of a normal model with known variance. Typical conjugate priors on the mean parameter under the alternative hypothesis have a mode on the null, and this is also true for intrinsic priors. While there are good reasons to follow this practice (basically to mitigate the inherent larger diffuseness of the prior under the alternative), the implications on the ability of the Bayes factor to learn the true model are disturbing. Essentially, the asymptotic learning rate is exponential when the larger model holds, while it behaves only as a power of the sample size when the smaller model is assumed to be true. To countervail this phenomenon, Johnson and Rossell (2010) recently suggested that priors for nested model comparison should be *non-local* (thus vanishing on the null) and showed that such priors can be effectively constructed (in particular as *moment priors*). The main advantages of non-local priors can be summarized under two headings: from a descriptive viewpoint, they embody a notion of *separation* between the larger and the smaller model; from an inferential perspective, they produce an accelerated learning behaviour when the smaller model holds.

We believe that the rationale underpinning non-local priors is sound and attractive. On the other hand, we are convinced of the need to produce Bayesian model choice procedures applicable in contexts where prior information is very limited or cannot be elicited in a reasonable amount of time. In this spirit, the paper Consonni, Forster and La Rocca (2010) combines non-local and intrinsic priors to obtain an enhanced Bayesian test for the equality of two proportions. In the same spirit, we here merge the idea of non-local priors with the methodology based on fractional Bayes factors, and apply our method to the comparison of Gaussian graphical models, focussing on directed acyclic graphs; see Cowell *et al.* (1999).

The structure of this paper is as follows. Section 2 presents some background material on non-local priors, fractional Bayes factors and directed acyclic graph models. Section 3 presents our new method, namely fractional Bayes factors based on moment priors, and presents our main result for the comparison of two nested Gaussian directed acyclic graph models (Theorem 1); some asymptotic considerations are also developed about the rate of learning of our procedure. Finally, Section 4 illustrates the performance of our method with two examples. The Appendix contains a lemma for the expression of some raw moments of the multivariate normal distribution, as well as the proof of Theorem 1.

2. BACKGROUND

2.1. *Non-Local and Moment Priors*

For data y, consider two models $M_0 : f_0(y|\theta_0)$ and $M_1 : f_1(y|\theta_1)$ with M_0 *nested* in M_1, so that each distribution in M_0 coincides with some $f_1(y|\theta_1)$ in M_1. Let $p_1(\theta_1)$ denote the parameter prior under M_1, and similarly for $p_0(\theta_0)$ under M_0. We assume that model comparison takes place through the Bayes factor (BF) and write $BF_{10}(y) = m_1(y)/m_0(y)$ for the BF of M_1 against M_0 (or simply in favour of M_1), where $m_k(y)$ is the marginal likelihood of M_k, i.e., $m_k(y) = \int f_k(y|\theta_k)p_k(\theta_k)\, d\theta_k$. Usually $p_1(\theta_1)$ is a local prior, i.e., assuming continuity, it is strictly positive over the subspace Θ_0 characterising the smaller model M_0.

Assume that the data $y^{(n)} = (y_1, \ldots, y_n)$ arise under i.i.d. sampling from some (unknown) distribution q. We say that the smaller model holds if q belongs to M_0, while we say that the larger model holds if q belongs to M_1 but not to M_0. If M_0 holds, then $BF_{10}(y^{(n)}) = O_p(n^{-(d_1-d_0)/2})$, as $n \to \infty$, where d_k is the dimension of M_k, $k = 0, 1$, and $d_1 > d_0$; if M_1 holds, then $BF_{01}(y^{(n)}) = e^{-Kn+O_p(\sqrt{n})}$, as $n \to \infty$, for some $K > 0$ (Kullback-Leibler divergence of M_0 from q). For a proof of this result, which shows an imbalance in the learning rate of the Bayes factor, see Dawid (1999). It is clear from Dawid's proof that, by forcing the prior density under M_1 to vanish on Θ_0, one can speed up the decrease of $BF_{10}(y^{(n)})$ when M_0 holds. This is indeed the approach taken by Johnson and Rossell (2010) when defining non-local priors. We focus here on a specific family of non-local priors. Let $g(\theta_1)$ be a continuous function vanishing on Θ_0. For a given local prior $p_1(\theta_1)$, define a new non-local prior as

$$p_1^M(\theta_1) \propto g(\theta_1)p_1(\theta_1),$$

which we name a *generalized moment prior*. For instance, if θ_1 is a scalar parameter in $\mathbb{R}$ and θ_0 a fixed value, we may take $g(\theta_1) = (\theta_1 - \theta_0)^{2h}$, where h is a positive integer ($h = 0$ returns the starting local prior); this is precisely the *moment prior* introduced by Johnson and Rossell (2010) for testing a sharp hypothesis on a scalar parameter. It can be proved that in this case $BF_{10}(y^{(n)}) = O_p(n^{-h-1/2})$ when M_0 holds, while $BF_{01}(y^{(n)}) = e^{-Kn+O_p(\sqrt{n})}$ when M_1 holds. In the former case the extra power h means that, for instance, if $h = 1$ the rate changes from sublinear to superlinear. While the above argument is asymptotic, we shall see that it is clearly reflected in finite sample size results. However, for small samples, a price is paid in terms of discriminatory power when the sampling distribution is in the low prior density region around Θ_0. We shall see that this price is affordable, and worth paying, at least if $h = 1$. The idea of moment priors outlined above can be suitably extended to the multivariate case; we shall give an example in Section 3.

2.2. *Fractional Bayes Factors*

Objective priors are often improper and thus they cannot be naively used to compute Bayes factors, even when the marginal likelihoods $m_k(y)$ are positive and finite for all y, because of the presence of arbitrary constants which do not cancel out when taking their ratios. A basic tool to overcome this difficulty is represented by the partial Bayes factor, which however depends on the specific choice of a training data set. Two ways to overcome this difficulty are the intrinsic Bayes factor by Berger and Pericchi (1996) and the fractional Bayes factor (FBF) by O'Hagan (1995). Here we focus on the latter. Let $0 < b < 1$ be a quantity depending on the sample size n, and define

$$w_k(y; b) = \frac{\int f_k(y \mid \theta_k)p(\theta_k)d\theta_k}{\int f_k^b(y \mid \theta_k)p(\theta_k)d\theta_k},$$

where $f_k^b(y \,|\, \theta_k)$ is the sampling density raised to the b-th power, $p_k(\theta_k)$ is the prior, and the integrals are assumed to be finite and nonzero. Informally, we refer to $w_k(y;b)$ as the *fractional marginal likelihood* for the k-th model.

The FBF (in favour of M_1) is then given by $FBF_{10}(y;b) = w_1(y;b)/w_0(y;b)$. It is easy to see that the FBF is an ordinary BF computed from the "likelihood" $f_k^{(1-b)}(y|\theta_k)$ and a data-dependent prior proportional to $p_k(\theta_k)f_k^b(y|\theta_k)$, i.e., a posterior based on a fraction b of the likelihood; usually b will be small, so that the dependence on the data of the prior will be weak. Consistency of the FBF is achieved as long as $b \to 0$ for $n \to \infty$. O'Hagan (1995, sect. 6) suggests three possible choices for b: i) $b = n_0/n$, where n_0 is the minimal (integer) training sample size for which the fractional marginal likelihood is well defined; ii) $b = \max\{n_0, \sqrt{n}\}/n$; iii) $b = \max\{n_0, \log n\}/n$. Choice i) is suggested as the standard option, when robustness issues are of little concern, while ii) is recommended when robustness is a serious concern, with iii) representing an intermediate option. One of the attractive properties of the FBF is its simplicity of implementation: with exponential families and conjugate priors its expression is typically available in closed-form.

2.3. *Directed Acyclic Graph Models*

Graphical models represent a powerful statistical tool in multivariate analysis, yielding dependence models that can be easily visualized and communicated; see Lauritzen (1996). Here, we are concerned with comparing graphical models in order to learn the dependence structure of a set of variables $\{U_1, \ldots, U_q\}$, using a Bayesian approach. This entails assigning a prior distribution on the space of models, together with a parameter prior within each model; we discuss the latter issue only, because our focus is on parameter priors.

There are several classes of graphs of direct use in statistics, among which undirected graphs, directed acyclic graphs (DAGs) and chain graphs are well-known. In this paper we concentrate on DAG models, assuming that there exists *a priori* a total ordering of the variables involved (e.g., temporal). Furthermore, we take the distribution of the random variables to be jointly normal.

Let $D = (V, E)$ be a DAG, where $V = \{1, \ldots, q\}$ is a set of vertices and $E \subseteq V \times V$ is a set of directed edges. We assume that the total ordering of the variables forms a well-numbering of the vertices according to D, so that, if there is a directed path from vertex i to vertex j in D, then $i < j$. For $W \subseteq V$, denote by U_W the set of all variables U_j with $j \in W$. The Gaussian graphical model corresponding to D is the family of all q-variate normal distributions such that, if there is no edge $i \to j$ in D, then U_j is conditionally independent of U_i given all variables $U_{\{1,\ldots,j\}\setminus\{i,j\}}$. We denote this DAG model as M_D.

Notice that the joint density of $(U_1, \ldots, U_q)$ can then be written as

$$f(u_1, \ldots, u_q \mid \beta, \gamma) = \prod_{j=1}^{q} f(u_j \mid u_{\mathrm{pa}(j)}; \beta_j, \gamma_j), \tag{1}$$

where $\mathrm{pa}(j)$ denotes the parents of j in D, i.e., all vertices preceding j such that each of them is joined by a directed edge to j. Since each conditional distribution in (1) is a univariate normal, the vector parameter β_j represents the regression coefficients in the conditional expectation of U_j given $U_{\mathrm{pa}(j)}$, namely $(\mathbf{1}, u'_{\mathrm{pa}(j)})\beta_j$, while γ_j is the corresponding conditional precision (inverse of variance). By convention, the first element of the vector β_j is the intercept β_{j0}, while the remaining elements are written as β_{jk} with $k \in \mathrm{pa}(j)$. If $\mathbb{E}(U_j) = 0$ for all j, then $\beta_{j0} = 0$, $j = 1, \ldots, q$, and the intercept can be dropped, so that β_j has dimension $|\mathrm{pa}(j)|$.

3. FRACTIONAL BAYES FACTOR BASED ON MOMENT PRIORS

We present in this section our proposal for a new Bayesian testing procedure, based on combining the advantages of the FBF with those of the moment prior, in order to obtain an objective method with enhanced learning behaviour. We now detail our procedure for the problem of comparing Gaussian DAG models.

Because of the recursive structure of the likelihood (1), it is natural to assume that $p(\beta, \gamma)$ satisfies the assumption of global parameter independence: $p(\beta, \gamma) = \prod_j p(\beta_j, \gamma_j)$; see Geiger and Heckerman (2002). A natural default prior is then $p^D(\beta_j, \gamma_j) \propto \gamma_j^{-1}$. Now consider two Gaussian DAG models $M_0 = M_{D_0}$ and $M_1 = M_{D_1}$ with the same vertex ordering and with M_0 nested in M_1. For each vertex j, let L_j be the subset of the parents pointing to j in D_1 but not in D_0. We define the moment prior (of order h) for vertex j, under M_1, as

$$p_1^M(\beta_j, \gamma_j) \propto \gamma_j^{-1} \prod_{l \in L_j} \beta_{jl}^{2h}, \tag{2}$$

where h is a positive integer. Notice that $h = 0$ gives back the starting default prior. The overall moment prior will be obtained by multiplying together the priors (2):

$$p_1^M(\beta, \gamma) = \prod_{j=1}^{q} p_1^M(\beta_j, \gamma_j) \propto \prod_{j=1}^{q} \left\{ \gamma_j^{-1} \prod_{l \in L_j} \beta_{jl}^{2h} \right\}. \tag{3}$$

To compute the FBF based on the moment prior (3), we need the expression for the fractional marginal likelihood pertaining to vertex j both under model

M_0 and under model M_1. The former is standard, because it is based on the default prior, while the latter is provided in the theorem below (whose proof is deferred to the Appendix to ease the flow of ideas). Notice that, to simplify notation, we omit in the statement the subscript j; thus we use y instead of y_j, while β and γ stand for β_j and γ_j.

Theorem 1 *For a DAG model M_1, consider a vertex likelihood $f(y \mid y_{pa}; \beta, \gamma)$, which is an n-variate normal distribution with expectation $X\beta$ and variance matrix $\gamma^{-1} I_n$, where X is an $n \times p$ matrix whose columns contain the observations on the parent variables (adding as first column the vector $\mathbf{1}_n$ whenever appropriate). For the comparison of M_1 with respect to a nested DAG model M_0, assume a vertex moment prior $p_1^M(\beta, \gamma) \propto \gamma^{-1} \prod_{l \in L} \beta_l^{2h}$, where $L \subseteq pa$ is the subset of the parents pointing to the vertex in D_1 but not in D_0. Then, the vertex fractional marginal likelihood based on the moment prior is*

$$w_1(y \mid X, b) = \left(\pi b S^2\right)^{-\frac{n(1-b)}{2}} \frac{\sum_{i=0}^{h|L|} 4^{-i} H_i^{(h)}(\hat{\beta}, (X'X)^{-1}) \Gamma(\frac{n-p-2i}{2})(S^2)^i}{\sum_{i=0}^{h|L|} 4^{-i} H_i^{(h)}(\hat{\beta}, (X'X)^{-1}) \Gamma(\frac{nb-p-2i}{2})(S^2)^i}, \quad (4)$$

where $0 < b < 1$ is the sample size dependent fraction satisfying $nb > p+2h|L|$, and $H_i^{(h)}(\mu, \Sigma)$ is defined in formula (6) of the Appendix. From a purely formal viewpoint, the expression of $\hat{\beta}$ is that of the usual OLS estimate, while that of S^2 corresponds to the residual sum of squares; the analogy is merely formal because the matrix X contains observations on stochastic variables, namely those associated to the parents of the vertex under consideration.

Using Theorem 1, we can conclude that the fractional marginal likelihood based on the moment prior is $w_1(y; b) = \prod_{j=1}^q w_1(y_j \mid X_j, b)$, where each individual factor $w_1(y_j \mid X_j, b)$ is as in (4). It is important to realise that the quantity $w_1(y; b)$ is contingent upon the choice of the specific nested DAG model M_0 used for the comparison: this determines the nature of the sets $L_j \subseteq \mathrm{pa}_j$ used in constructing the moment prior. The FBF of M_1 against M_0 is now given by the ratio of the two fractional marginal likelihoods:

$$FBF_{10}(y; b) = \frac{w_1(y; b)}{w_0(y; b)} = \prod_{j=1}^q \frac{w_1(y_j \mid X_{1j}, b)}{w_0(y_j \mid X_{0j}, b)} = \prod_{j=1}^q FBF_{10}^{(j)}(y_j; X_{1j}, b), \quad (5)$$

where each individual $w_1(y_j \mid X_{1j}, b)$ is computed using formula (4), while each individual $w_0(y_j \mid X_{0j}, b)$ is directly available using standard calculations for the FBF in the normal linear model (O'Hagan and Forster, 2004; sect. 11.40) and in principle it can also be deduced from (4) upon setting $h = 0$ throughout.

Notice that $FBF_{10}(y; b)$ is a product of FBFs pertaining to single vertices: $FBF_{10}^{(j)}(y_j; X_{1j}, b)$. In addition, it is well-known, and immediate to realise, that in order to compute the quantity $FBF_{10}^{M}(y;b)$ one requires only those FBFs referring to vertices with different parent structures under the two DAGs D_1 and D_0; otherwise $FBF_{10}^{(j)}(y_j;X_{1j},b)$ is identically one.

3.1. *Asymptotics*

The proof in Dawid (1999) suggests that $FBF_{10}(y^{(n)};b) = O_p(n^{-(h+1)\sum_j |L_j|/2})$, if M_0 holds, while $BF_{01}(y^{(n)}) = e^{-Kn+O_p(\sqrt{n})}$, for some $K > 0$, if M_1 holds. However, Dawid's argument is not directly applicable, because the FBF uses a data dependent prior. Nevertheless, the intuition is correct, and the same result can be obtained directly (at least when nb is held constant) from (4) and (5). Assuming that M_0 holds, one first writes $S_1^2/S_0^2 = \exp\{(S_1^2 - S_0^2)/S_0^2 + o_p((S_1^2 - S_0^2)/S_0^2)\}$ in $w_1(y \mid X_1, b)/w_0(y \mid X_0, b)$, focussing on a single vertex, and acknowledges that $(S_1^2/S_0^2)^n$ converges in law to the exponential transform of an F distribution. Then, the Gamma function is approximated by Stirling's formula, and one notices that $n\hat{\beta}_l^2$ converges in law to a χ^2 distribution, for all $l \in L$. Working out the details, and considering all vertices together, the desired result is achieved. On the other hand, if M_1 holds, the factor S_0^2/S_1^2 converges in probability to a value lesser than one and the exponential behaviour is obtained, as the remaining factor is dealt with by means of Stirling's formula.

4. EXAMPLES

We illustrate our method by means of two examples. The first one relates to a three-vertex DAG: we show the learning behaviour of our FBF based on moment priors, and its discriminatory power, as a function of a simply interpretable parameter; we also apply our results to a real data set. The second example concerns a seven-variable real data set on the issue of publishing productivity, which has been previously analysed in the literature and thus allows some comparison with alternative methods.

4.1. *Three-Variable DAG Models*

Let (X, Z, Y) be three random variables jointly distributed according to a normal distribution. We can think of Y as a response variable, while X and Z are potential explanatory variables. Assume that X precedes Z, so that the total ordering of the three variables is X, Z, Y. In the sequel, we shall provide a concrete example, where X is *Age*, Z is *Weight* and Y is *Systolic blood pressure*. A typical hypothesis of interest is $Y \perp\!\!\!\perp X \mid Z$, so that the effect of X on Y vanishes when we condition on Z; this is represented by the DAG D_0 in Figure 1, whereas the DAG D_1 in the same figure represents the full model with no conditional independencies.

Assume now, for simplicity, that each variable has expectation zero and

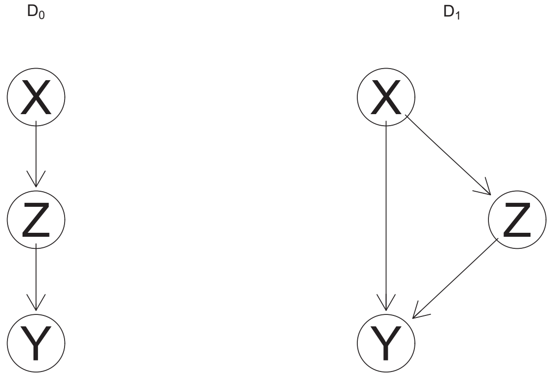

Figure 1: *Full and reduced DAG for the three-variable example.*

variance one, and let the correlation matrix be given by

$$\begin{array}{cc} & \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{array}{c} X \\ Y \\ Z \end{array} & \begin{bmatrix} 1 & r & a \\ r & 1 & s \\ a & s & 1 \end{bmatrix} \end{array}$$

with r, s and a constrained by positive definiteness. Then, the partial correlation between X and Y given Z is given by

$$\rho_{XY\,|\,Z} = \frac{a - rs}{\sqrt{1-r^2}\sqrt{1-s^2}}.$$

To fix ideas let $r = s = 0.5$, so that the only free parameter is a, and the condition of positive definiteness on the correlation matrix leads to $-0.5 < a < 1$. Notice that $\rho_{XY\,|\,Z} = (4a-1)/3$, which is free to vary over the interval $(-1, 1)$.

If $a = rs = 0.25$, then $\rho_{XY\,|\,Z} = 0$ and thus $Y \perp\!\!\!\perp X \mid Z$ (and conversely) so that there is no edge between X and Y; this provides the reduced, or null, model M_0 corresponding to the DAG D_0. On the other hand, if $a \neq rs$, then the full model M_1 corresponding to the DAG D_1 holds. Clearly, the only local likelihood that matters for the comparison of M_1 and M_0 is the conditional distribution of Y given (X, Z), which is normal with conditional expectation $(a-rs)/(1-r^2)X+(s-ar)/(1-r^2)Z = (4a/3)X+[2(1-a)/3]Z$. As a typical value of a such that M_1 holds, we consider $a = 5/8 = 0.625$, corresponding to $\rho_{XY\,|\,Z} = 0.5$ (an intermediate situation).

Figure 2 reports the posterior probability of M_0 as a function of the sample size n; here and in the following we assume equal prior probabilities for the two models under comparison. Results are available for each combination of $a = 0.25$ and $a = 0.625$, and for three choices of FBF: the standard one (corresponding to $h = 0$) and two FBFs based on moment priors (with $h = 1$ and $h = 2$). It is assumed that the data produce, for each n, the same correlation matrix R_a, say, as in the population (after having fixed $r = s = 0.5$). In this way, we are able to capture more neatly the effect of the sample size n. Recall from Theorem 1 that the fraction b must satisfy the condition $nb > p + 2h|L|$. Here, since the variables have expectation zero, $p = 2$ is the number of parents of Y in the larger model, and $|L| = 1$, because we only consider dropping one edge, namely $X \to Y$; hence, the condition is $nb > 2 + 2h$, which we round to the next integer, thus taking $nb = 3 + 2h$.

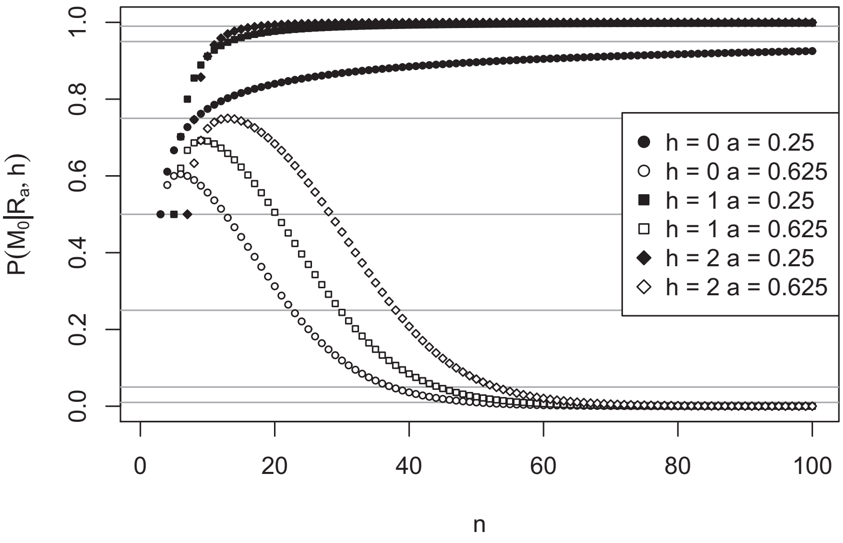

Figure 2: *Learning behaviour. The horizontal grey lines refer to possible decision thresholds at levels 1%, 5%, 25%, 50%, 75%, 95%, 99%.*

One can see from Figure 2 that, when M_0 holds (solid symbols), learning is much faster under the FBF based on moment priors than it is under the usual FBF. In fact, under the latter, the rate of growth of the learning curve is so slow that even after 100 observations the 95% threshold is not attained; this should be compared with $n = 14$ under the moment prior (with $h = 1$). On the other hand, when M_1 holds (hollow symbols), the ordinary FBF performs better; yet the decline of the curve under the two moment priors is rapid enough to reach a conclusion (e.g., by hitting the 5% threshold at $n = 45$ when $h = 1$). We regard Figure 2 as an important piece of evidence in favour of our method, and one that suggests $h = 1$ as a better compromise between learning under M_0 and learning under M_1.

In Figure 3 we study the ability to discriminate between the two models for three FBFs: the standard one ($h = 0$) and two based on actual moment priors ($h = 1$ and $h = 2$). Assuming $n = 50$, this is done by plotting the posterior probability of M_1 as a function of the free parameter a over its whole range of variability ($-0.5 < a < 1$). Recall that $a = 0.25$ corresponds to conditional independence (model M_0) and that the farther away a is from this threshold the farther away is the sampling distribution from M_0. It is apparent that the ordinary FBF (solid line) is not able to provide enough evidence against M_1, when M_0 is true, because the minimum value for the posterior probability of M_1 is about 10%, while it is less than 1% for the other two curves. Clearly, the better performance of the moment priors at $a = 0.25$ produces a lower value for the posterior probability of M_1 also for $a \neq 0.25$, and thus technically belonging to M_1. However, by the time $a \leq -0.11$ or $a \geq 0.61$, when $h = 1$, this posterior probability has attained the 95% threshold. Only substantive knowledge in the area can tell whether this type of discrimination is strong enough for the given sample size. We do believe, however, that plots like Figure 2 and Figure 3 represent a valid tool for assessing the appropriateness of the testing procedure under consideration, and one which may lead to a further refinement on the value of the fraction b to meet other subject-matter requirements.

We conclude this subsection by analysing a data set also discussed in Wermuth (1993). The data refer to $n = 98$ healthy male adults. For each individual, the variables (Age, Weight, Systolic blood pressure) were recorded. Table 1 reports some summary statistics. It is apparent that the partial correlation between Age and Systolic blood pressure given Weight is very small (-0.007), thus suggesting a model of conditional independence. This model is indeed confirmed by each of the FBFs we consider. Specifically, we obtain that the posterior probability of M_0 is 0.9244 (h=0), 0.9983 ($h = 1$) and 0.9999 ($h = 2$) in the three cases. Assuming prior odds equal to one, we could convert the Jeffreys' scale for the Bayes factor (Robert, 2001, p. 228) into the corresponding one for the posterior probability of M_1. It would then appear that, under the moment prior, there is decisive evidence against M_1, whereas this evidence is only strong under the local prior.

4.2. *Publishing Productivity*

This data set is part of a larger study aimed at investigating the interrelationship among variables potentially related to publishing productivity among academics. The data were discussed in Spirtes, Glymour and Scheines (2000,

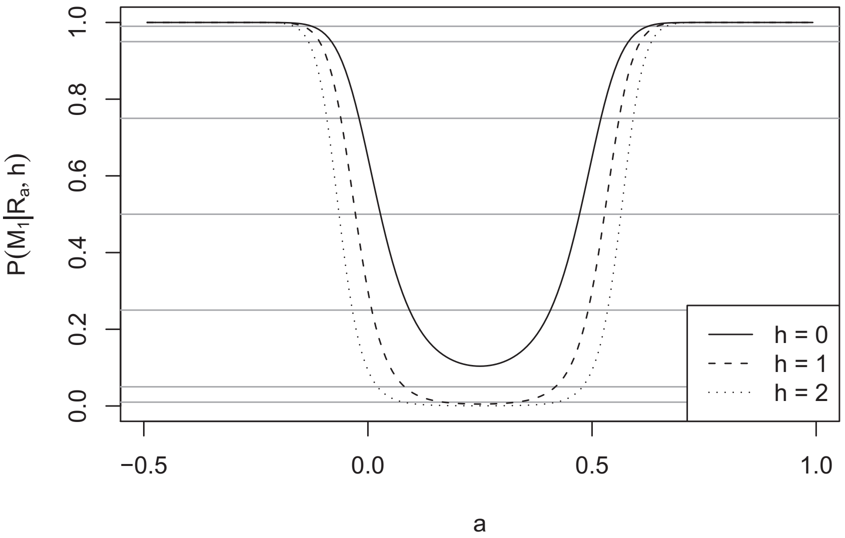

Figure 3: *Discriminatory power. The horizontal grey lines refer to possible decision thresholds at levels 1%, 5%, 25%, 50%, 75%, 95%, 99%.*

Table 1: *Observed marginal correlations (lower half) and partial correlations (upper half) for $n = 98$ healthy male adults.*

Variable	X (Age)	Z (Weight)	Y (Systolic blood pressure)
X (Age)	1.000	0.369	−0.007
Z (Weight)	0.390	1.000	0.348
Y (Systolic blood pressure)	0.139	0.371	1.000

Example 5.8.1) and also analysed in Drton and Perlman (2008), using a frequentist simultaneous testing procedure named SIN. The sample comprises $n = 162$ subjects and seven variables, which we write in the order considered by Drton and Perlman (2008): 1. subject's sex (Sex); 2. score of the subject's ability (Ability); 3. measure of the quality of the graduate program attended (GPQ); 4. preliminary measure of productivity (PreProd); 5. quality of the first job (QFJ); 6. publication rate (Pubs); 7. citation rate (Cites). Table 2 reports some summary statistics.

The SIN method (at simultaneous level 0.05) selected the DAG D_0 in Figure 4, whereas stepwise backward selection (at individual level 0.05) using the MIM software package (Edwards, 2000) yielded the super-graph on D_1 in the same figure (which includes the additional edges Ability → Pubs and

Table 2: *Observed marginal correlations (lower half) and pairwise partial correlations given the rest of the variables that do not follow the pair in the given ordering (upper half) for $n = 162$ academics.*

	Sex	Ability	GPQ	PreProd	QFJ	Pubs	Cites
Sex	1.00	−0.10	0.08	0.06	0.10	0.45	−0.09
Ability	−0.10	1.00	0.62	0.25	−0.02	0.17	0.07
GPQ	0.00	0.62	1.00	−0.09	0.23	−0.07	0.07
PreProd	0.03	0.25	0.09	1.00	0.05	0.14	0.26
QFJ	0.10	0.16	0.28	0.07	1.00	0.39	0.16
Pubs	0.43	0.18	0.15	0.19	0.41	1.00	0.43
Cites	0.13	0.29	0.25	0.34	0.37	0.55	1.00

QFJ → Cites). We decided to compare the two models using the three FBFs with $h = 0, 1, 2$, and obtained the following values for the posterior probability of M_0: 0.2907 ($h = 0$), 0.9814 ($h = 1$) and 0.9999 ($h = 2$). Hence, FBFs based on non-local priors support the simplification operated by SIN, with respect to MIM, while the ordinary FBF gives a different result.

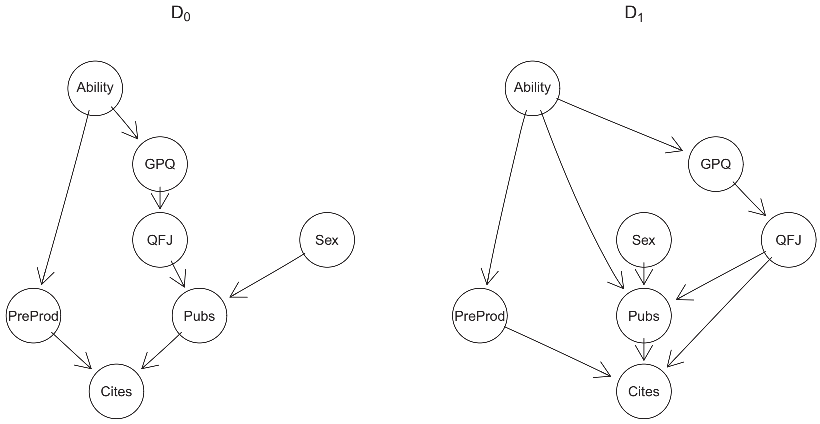

Figure 4: *Full and reduced DAG for the publishing productivity example.*

We also compared the SIN model with all simpler models obtained by removing individual edges from it. The results are reported in Table 3. It is apparent that the FBFs based on moment priors with $h = 1$ (as well as the ordinary ones) do not suggest any further simplification, whereas letting $h = 2$ suggests removing Ability → PreProd (and GPQ → QFJ). This provides us with some evidence that moment priors with $h = 1$ do not favour overly simple models.

Table 3: *Posterior probabilities of models obtained by removing individual edges from D_0.*

Edge Removed	h = 0	h = 1	h = 2
Ability → GPQ	2.21E-16	4.63E-16	1.40E-15
Ability → PreProd	8.33E-02	4.15E-01	8.62E-01
GPQ → QFJ	2.26E-02	1.30E-01	5.32E-01
Sex → Pubs	1.55E-06	5.02E-06	2.20E-05
QFJ → Pubs	7.84E-06	2.73E-05	1.28E-04
PreProd → Cites	1.66E-02	9.43E-02	4.34E-01
Pubs → Cites	1.07E-10	2.64E-10	9.13E-10

5. DISCUSSION

In this paper we have presented a novel approach for the comparison of Gaussian DAG models within an objective Bayes framework. For a given total ordering of the variables, we write the joint density under an assumed DAG model as a product of recursive conditional normal distributions; in this way the absence of an edge from a potential parent of a vertex in the DAG is mirrored in the value zero taken on by the corresponding regression coefficient. For each DAG-model we assume global parameter independence for the parameter prior, and assign a standard default improper prior on each vertex regression coefficient and conditional variance. In order to compare two nested models, we turn the default prior under the larger model into a moment prior, and then apply the fractional Bayes factor methodology. We demonstrate that the learning behaviour of our method outperforms the traditional fractional Bayes factor when the smaller model holds; moreover, when the larger model holds, the learning behaviour is only marginally worse, for small samples, but rapidly becomes comparable as the sample size grows.

A further, important, area of application, which was not explicitly touched on in this paper, is that of model search. We believe that our methodology can be successfully applied in this context, with the help of a suitable search algorithm over the space of all models. Since our approach is based on a pairwise comparison of nested models, some form of *encompassing* is required, if an MCMC strategy is adopted; see, in the context of variable selection, Liang *et al.* (2008) using mixtures of g-priors, or Casella and Moreno (2006) using intrinsic priors. An alternative option is to use a Feature-Inclusion Stochastic Search, as implemented in Scott and Carvalho (2006) for undirected decomposable graphical models. The underlying parameter priors for this search algorithm will be path-based pairwise moment priors; see Berger and Molina (2005) in the context of variable selection using g-priors. Preliminary results indicate that our method compares favourably with lasso and adaptive lasso techniques to identify DAG-models having a fixed ordering of the variables as reported recently in Shojaie and Michailidis (2010).

REFERENCES

Berger, J. O. and Pericchi, L. R.(1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91**, 109–122.

Berger, J. O. and Molina, G. (2005). Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica* **59**, 3–15

Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection. *J. Amer. Statist. Assoc.* **101**, 157–167.

Consonni, G., Forster, J. J. and La Rocca, L. (2010). Enhanced objective Bayesian testing for the equality of two proportions. Submitted.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. New York: Springer.

Dawid, A. P. (1999). The trouble with Bayes factors. Research Report 202, University College London, Department of Statistical Science.

Drton, M. and Perlman, M. D. (2008). A SINful approach to Gaussian graphical model selection *J. Statist. Planning and Inference* **138**, 1179–1200.

Edwards, D. M. (2000). *Introduction to Graphical Modelling*. New York: Springer.

Geiger, D. and Heckerman, D.(2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* **30**, 1412-1440.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *J. Roy. Statist. Soc. B* **72**, 143–170.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford: University Press.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103**, 410-423.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *J. Roy. Statist. Soc. B* **57**, 99-138 (with discussion).

O'Hagan, A. and Forster J. J. (2004). *Kendall's Advanced Theory of Statistics. Vol 2B. Bayesian Inferences* (2nd edition). Arnold: London.

Pericchi, L. R. (2005). Model selection and hypothesis testing based on objective probabilities and Bayes factors. *Bayesian Thinking: Modeling and Computation, Handbook of Statistics* **25** (Dey, D. K. and Rao, C. R., eds). Amsterdam: Elsevier, 115–149.

Robert, C. P. (2001). *The Bayesian Choice* (2nd edition). New York: Springer

Scott, J. G. and Carvalho, C. M. (2006). Feature-inclusion stochastic search for Gaussian graphical models. *J. Comp. Graphical Statist.* **17**, 780–808.

Shojaie, A. and Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika* . To appear.

Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, Prediction and Search* (2nd edition). Cambridge, MA: The MIT Press.

Wermuth, N. (1993). Association structures with few variables: characteristics and examples. *Population Health Research* (Dean, K., ed.). Sage: London, 181–203.

APPENDIX

Lemma. Let $U = (U_1, \ldots, U_p) \sim N_p(\mu, \Sigma)$, where $\mu' = (\mu_1, \ldots, \mu_p)$ and $\Sigma = [\sigma_{lm}]$; $l, m = 1, \ldots, p$. Fix $d \le p$ and a positive integer h; then

$$\mathbb{E}\left[\prod_{l=1}^{d} U_l^{2h}\right] = \sum_{i=0}^{hd} \frac{1}{2^i} H_i^{(h)}(\mu, \Sigma),$$

where

$$H_i^{(h)}(\mu, \Sigma) = \sum_{j \in J_h(i)} \prod_{l=1}^{d} (2h)! \prod_{m=1}^{d} \frac{\sigma_{lm}^{j_{lm}}}{j_{lm}!} \prod_{l=1}^{d} \frac{\mu_l^{j_l^\star}}{j_l^\star!}, \tag{6}$$

having defined

$$j_l^\star = 2h - \sum_{m=1}^{d} j_{lm} - \sum_{m=1}^{d} j_{ml}$$

and

$$J_h(i) = \left\{ j : \sum_{l=1}^{d} \sum_{m=1}^{d} j_{lm} = i \;\&\; \forall l : j_l^\star \ge 0 \right\}.$$

Remark. In formula (6) we have used the convention $0^0 = 1$. Notice that

$$H_i^{(h)}(\mu, a\Sigma) = a^i H_i^{(h)}(\mu, \Sigma),$$

i.e., $H_i^{(h)}(\mu, \cdot)$ is homogeneous of order i. Although, for simplicity, we state the result for the first d components of U, it clearly holds for any d components of U.

Proof. The *moment generating function* of $U^{(d)} = (U_1, \ldots, U_d)$ is given by

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{t'U^{(d)}\right\}\right] &= \exp\left\{\sum_{\ell=1}^{d} t_\ell \mu_\ell + \frac{1}{2}\sum_{\ell=1}^{d}\sum_{m=1}^{d} t_\ell \Sigma_{\ell m} t_m\right\} \\
&= \sum_{n=0}^{\infty} \frac{1}{n!}\left(\sum_{\ell=1}^{d} t_\ell \mu_\ell + \frac{1}{2}\sum_{\ell=1}^{d}\sum_{m=1}^{d} t_\ell \Sigma_{\ell m} t_m\right)^n \\
&= \sum_{n=0}^{\infty}\sum_{i=0}^{n} \frac{(t'\mu)^{n-i}(t'\Sigma t)^i}{2^i i!(n-i)!} \\
&= \sum_{n=0}^{\infty}\sum_{i=0}^{n} \frac{1}{2^i}\left[\sum_{k_\ell}\prod_{\ell=1}^{d} \frac{(t_\ell \mu_\ell)^{k_\ell}}{k_\ell!}\right]\left[\sum_{j_{\ell m}}\prod_{\ell=1}^{d}\prod_{m=1}^{d} \frac{(t_\ell \Sigma_{\ell m} t_m)^{j_{\ell m}}}{j_{\ell m}!}\right],
\end{aligned}$$

where the summations over $j_{\ell m}$ and k_ℓ are restricted to $j_{11}+j_{12}+\cdots+j_{d(d-1)}+j_{dd}=i$ and $k_1+\cdots+k_d=n-i$, respectively. The desired raw moment is the coefficient of $t_1^{2h}\cdots t_d^{2h}$ in the above expansion multiplied by $((2h)!)^d$. This gives $\mu_1^{2h}\cdots\mu_d^{2h}$ plus all terms obtained by replacing one or more factors $\mu_\ell\mu_m$ with $\Sigma_{\ell m}$, that is, by letting $k_\ell=j^\star_\ell$. The function $H_i^{(h)}(\mu,\Sigma)$ groups the terms with a given amount i of the overall exponent $2hd$ assigned to elements of Σ, and is thus homogenous of order i; the index set $J_h(i)$ identifies the possible values of j for given i.

Proof of Theorem 1. The generic vertex sampling density is

$$f(y|X;\beta,\gamma)=\left(\frac{\gamma}{\pi}\right)^{n/2}\exp\left\{-\frac{\gamma}{2}||y-X\beta||_2^2\right\}.$$

The corresponding moment prior is

$$p_1^M(\beta,\gamma)\propto\gamma^{-1}\prod_{l\in L}\beta_l^{2h}.$$

Let

$$I(y;X,b)=\int\int\left(\frac{\gamma}{\pi}\right)^{nb/2}\exp\left\{-\frac{b\gamma}{2}||y-X\beta||_2^2\right\}\gamma^{-1}\prod_{l\in L}\beta_l^{2h}\,d\beta\,d\gamma.$$

Then, the fractional marginal likelihood is $w_1(y;b)=I(y;X,1)/I(y;X,b)$.

Consider now $I(y;X,b)$. This can be written as

$$I(y;X,b)=(2\pi)^{nb/2}\int\gamma^{\frac{nb}{2}-1}\exp\left\{-b\frac{\gamma}{2}S^2\right\}J(y;X,b)\,d\gamma,$$

with

$$S^2=||y-X\hat\beta||_2^2,\quad \hat\beta=(X'X)^{-1}X'y,$$

and

$$\begin{aligned}J(y;X,b)&=\int\exp\left\{-\frac{b\gamma}{2}||X(\beta-\hat\beta)||_2^2\right\}\prod_{l\in L}\beta_l^{2h}\,d\beta\\&=\left(\frac{2\pi}{\gamma b}\right)^{p/2}|X'X|^{-1/2}\mathbb{E}\left[\prod_{l\in L}\beta_l^{2h}\right],\end{aligned}$$

where p is one plus the cardinality of pa (in the general case where the expected value μ of the q-variate normal population is different from zero) or just the cardinality of pa (if $\mu=0$ so that all local intercepts are zero); the expectation is taken with respect to the p-variate normal $N_p(\hat\beta,\gamma^{-1}b^{-1}(X'X)^{-1})$.

Letting $|L|$ be the cardinality of the set L, and using the result of the Lemma, we obtain

$$J(y; X, b) = \left(\frac{2\pi}{\gamma b}\right)^{p/2} |X'X|^{-1/2} \sum_{i=0}^{h|L|} \left(\frac{1}{2b\gamma}\right)^{i} H_i^{(h)}(\hat{\beta}, (X'X)^{-1}),$$

where $H_i^{(h)}(\mu, \Sigma)$ is defined in (6). As a consequence we can write

$$\begin{aligned} I(y; X, b) \quad = \quad & (2\pi)^{\frac{p-nb}{2}} b^{-\frac{p}{2}} |X'X|^{-1/2} \cdot \\ & \cdot \textstyle\sum_{i=0}^{h|L|} \left(\frac{1}{2b}\right)^{i} H_i^{(h)}(\hat{\beta}, (X'X)^{-1}) \int \gamma^{\frac{nb-p}{2}-i-1} \exp\{-\frac{b\gamma}{2} S^2\}\, d\gamma. \end{aligned} \qquad (7)$$

The integral in (7) exists provided $nb > p + 2h|L|$ and in that case we obtain

$$\begin{aligned} I(y; X, b) \quad = \quad & (2\pi)^{\frac{p-nb}{2}} b^{-\frac{p}{2}} |X'X|^{-1/2} \cdot \\ & \cdot \textstyle\sum_{i=0}^{h|L|} H_i^{(h)}(\hat{\beta}, (X'X)^{-1}) \Gamma(\frac{nb-p}{2} - i) \left(\frac{bS^2}{2}\right)^{\frac{p-nb}{2}} \left(\frac{S^2}{4}\right)^{i} \\ = \quad & \pi^{\frac{p-nb}{2}} b^{-\frac{nb}{2}} |X'X|^{-1/2} (S^2)^{\frac{p-nb}{2}} \cdot \\ & \cdot \textstyle\sum_{i=0}^{h|L|} \frac{1}{4^i} H_i^{(h)}(\hat{\beta}, (X'X)^{-1}) \Gamma(\frac{nb-p-2i}{2}) (S^2)^{i}. \end{aligned}$$

Finally we obtain

$$\begin{aligned} w_1(y; b) \quad & = \quad \frac{I(y; X, 1)}{I(y; X, b)} \\ & = \quad \left(\pi b S^2\right)^{-\frac{n(1-b)}{2}} \frac{\sum_{i=0}^{h|L|} 4^{-i} H_i^{(h)}(\hat{\beta}, (X'X)^{-1}) \Gamma(\frac{n-p-2i}{2}) (S^2)^{i}}{\sum_{i=0}^{h|L|} 4^{-i} H_i^{(h)}(\hat{\beta}, (X'X)^{-1}) \Gamma(\frac{nb-p-2i}{2}) (S^2)^{i}} \end{aligned} \qquad (8)$$