

Further Results on Bayesian Method of Moments Analysis of the Multiple Regression Model

Justin Tobias and Arnold Zellner*

November 11, 1998

Abstract

The Bayesian Method of Moments (BMOM) was introduced in 1994 to permit investigators to make inverse probability statements regarding parameters' possible values given the data when the form of the likelihood function is unknown. BMOM has been applied in analyses of several statistical and econometric models including location, multiple and multivariate regression, and simultaneous equation models. In Zellner (1996, 1997a) and Zellner and Sacks (1996) some previous BMOM analyses of the multiple regression model have appeared that permit derivation of post-data densities for parameters and future observations to be calculated without use of a likelihood function, prior density, or Bayes' Theorem. In the present paper, we extend previous analyses by showing how information about a variance parameter and its relation to regression coefficients affects post-data densities. We also discuss estimation of functions of parameters and model selection techniques using BMOM and traditional Bayesian approaches. In addition, following Good (1950) and Kullback (1959), we discuss the link between cross-entropy and the average log odds. We then use the average log odds criterion in an experiment to compare results obtained from BMOM and traditional Bayes approaches using data generated from known models.

1 Introduction

The Bayesian method of moments (BMOM) was introduced to permit investigators to compute post-data densities for parameters and future observations when not enough information is available to

*Research financed in part by the National Science Foundation and by income from the H.G.B. Alexander Endowment Trust Fund, Graduate School of Business, University of Chicago. We would also like to thank the Editor, three anonymous referees, and participants of the Student Econometrics Workshop at the University of Chicago for helpful comments and suggestions. This paper was presented at the Econometric Society meeting, Caltech, June 1997.

formulate a satisfactory likelihood function; see Green and Strawderman (1996) for a study in which the authors did not have enough information to formulate a likelihood function and use was made of the BMOM. The BMOM approach provides a solution to the famous inverse problem posed by Bayes (1763) and hence the name Bayesian method of moments. BMOM has been studied and applied to various models in van der Merve and Viljoen (1998), Zellner (1995,1996,1997b), Green and Strawderman (1996), Zellner and Sacks (1996), Currie (1996), and Zellner Min, Dallaire and Currie (1994). In the BMOM approach, two basic assumptions are made that permit evaluation of post-data moments of parameters from a given set of data. Then, from among various methods of determining densities from given moments, the maximum entropy (maxent) approach is used to choose a proper density with the given moments that maximizes entropy. For discussion and applications of maxent, see *e.g.* Jaynes (1982,1988), Shore and Johnson (1980), Cover and Thomas (1991), and Zellner and Highfield (1988). Also, see Zellner (1997) for coherent procedures for updating BMOM maxent post-data densities for parameters and future observations.

The methods developed in this paper augment previous BMOM analyses of location, multiple regression, multivariate regression and simultaneous equations models. In the current paper we review and extend the existing theory of BMOM in analysis of the standard multiple regression model. In particular, we derive post-data densities for parameters and future, as yet unobserved observations using various moment side conditions and show how use of alternative moment side conditions affects the shapes and properties of maxent densities. For certain moment side conditions, post-data densities are very similar to those derived in a traditional Bayesian approach based on improper diffuse prior densities for parameters and a normal likelihood function. Further, as in Chaloner and Brant (1988), Zellner and Moulton (1985), and Zellner (1975), the BMOM approach yields moments and densities for realized error terms and functions of them that are very useful for diagnostic checking purposes. Last, it is shown how posterior odds can be computed to compare models produced by BMOM and those produced with a traditional Bayesian approach.

After presenting the above and pointing out key relations between BMOM results and traditional Bayesian results, we analyze generated data to illustrate BMOM empirical results. Also, we compute various measures, including traditional Bayes' factors to compare models produced by approaches based on different assumptions. As discussed in Min and Zellner (1993) and Palm and Zellner (1992), posterior odds can be utilized to compare and/or combine alternative predictive models. On this capability, Barnard (1997) has commented favorably on the value of being able to compare and select between or among BMOM and TB models.

The plan of the paper is as follows. In Section 2, we review the BMOM assumptions relating to the multiple regression model and demonstrate how various moment conditions are derived. Then these moments are used as side conditions in deriving proper maxent post-data densities for parameters and future observations. Included is a demonstration of the moment side conditions for the variance parameter that lead to a maxent density in the inverted gamma form, a form that is encountered in a traditional Bayes approach based on an improper diffuse prior and iid normal error terms. Also, the dependence of the variance of the variance parameter on the sample size is investigated. Finally, some comments on sampling properties of BMOM and traditional Bayes estimates are provided. Section 4 is devoted to presenting the results of generated data. Similarities and differences of results produced by various BMOM and traditional Bayes approaches are discussed. In Section 4, various measures including Bayes factors are employed to compare alternative models. The paper concludes with a summary in Section 5.

2 Review and Extension of BMOM

2.1 Post-Data Moments for Regression Parameters

Let y , an $n \times 1$ vector of given observations be assumed to be related to X , a given $n \times k$ matrix of rank k , as follows

$$(1) \quad y = X\beta + u,$$

where β is a $k \times 1$ vector of regression coefficients with fixed unknown values, and u is an $n \times 1$ vector of realized error terms. As in earlier work on the analysis of realized error terms (see Chaloner and Brant (1988), Zellner (1975) and Zellner and Moulton (1985)), we regard β and u to be subjectively random. We shall introduce assumptions that will enable us to obtain the moments of the elements of β and u and then use the principle of maximum entropy to obtain proper post-data densities.

From equation (1) we have

$$(2) \quad \hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'u.$$

On taking the post-data expectation of both sides of (2) given the data $D = (y, X)$,

$$(3) \quad \hat{\beta} = E(\beta | D) + (X'X)^{-1}X'E(u | D),$$

where E denotes the subjective, post-data expectation operator. We now introduce the following assumption,

Assumption 1 $X'E(u | D) = 0$

namely that the columns of X are orthogonal to the vector $E(u | D)$. This assumption would not be satisfied if relevant variables correlated with X are omitted from (1), if the included independent variables are measured with error, or if other errors are made in formulating the form of (1). Given that assumption 1 is satisfied, we have from (3),

$$(4) \quad E(\beta | D) = \hat{\beta} = (X'X)^{-1}X'y.$$

That is, the post-data mean of β is equal to the least squares estimate. Also, the mean of β given in (4) is an optimal point estimate relative to a quadratic loss function, $L(\beta, \tilde{\beta}) = (\beta - \tilde{\beta})'Q(\beta - \tilde{\beta})$, where $\tilde{\beta}$ is some estimate and Q is a given positive definite, symmetric matrix. The value of $\tilde{\beta}$ that minimizes expected loss is the mean given in (4).

Further, assumption 1 yields the following post-data mean for u ,

$$(5) \quad E(u | D) = y - XE(\beta | D) = y - X\hat{\beta} = \hat{u}$$

where \hat{u} is the least squares residual vector that satisfies $X'\hat{u} = 0$. Note that from (5) we can write

$$\begin{aligned} u - \hat{u} &= y - X\beta - (y - X\hat{\beta}) \\ &= X(X'X)^{-1}X'u \\ &= X(X'X)^{-1}X'(u - \hat{u}) \end{aligned}$$

where the last step follows from the orthogonality condition mentioned above, $X'\hat{u} = 0$. We can thus write

$$\text{Var}(u | D) = E[(u - \hat{u})(u - \hat{u})' | D] = X(X'X)^{-1}X'E[(u - \hat{u})(u - \hat{u})' | D]X(X'X)^{-1}X',$$

which defines a functional equation that the post-data covariance matrix for u , $V(u | D)$ must satisfy. Since there are only k free elements of u in the equations in (1), $V(u | D)$ must be of rank k . In view of these considerations, our second assumption is ¹

¹Note that substituting assumption 2 into the formula for $\text{Var}(u | \sigma^2, D)$ solves the fixed point problem.

Assumption 2 $\text{Var}(u \mid \sigma^2, D) = \sigma^2 X(X'X)^{-1}X'$,

where σ^2 is a variance parameter to be defined below. We use assumption 2 to evaluate the post-data covariance matrix of β as follows

$$\begin{aligned}
 \text{Var}(\beta \mid \sigma^2, D) &= E \left[(\beta - \hat{\beta}) (\beta - \hat{\beta})' \mid D \right] \\
 &= (X'X)^{-1} X' E \left[(u - \hat{u}) (u - \hat{u})' \mid D \right] X (X'X)^{-1} \\
 (6) \qquad \qquad \qquad &= \sigma^2 (X'X)^{-1}.
 \end{aligned}$$

Assumption 2 will also enable us to evaluate the post-data moments of σ^2 . In the arguments of this section, we assume that an intercept appears in the regression matrix. From assumption 1, observe that the presence of an intercept implies that the post-data mean of $\bar{u} = \sum_{i=1}^n u_i/n$ is zero, which simplifies the derivations to follow. We define the expectation of σ^2 given the data as follows

$$(7) \qquad E[\sigma^2 \mid D] \equiv E \left[\sum_{i=1}^n \frac{(u_i - E(\bar{u} \mid D))^2}{n} \mid D \right] = E \left(\frac{u'u}{n} \mid D \right).$$

We also note that

$$\begin{aligned}
 E(u'u \mid D) &= nE(\sigma^2 \mid D) = (y - X\hat{\beta})' (y - X\hat{\beta}) + E \left[(\beta - \hat{\beta})' X'X (\beta - \hat{\beta}) \mid D \right] \\
 &= \hat{u}'\hat{u} + \text{tr} \left[X'XE \left\{ (\beta - \hat{\beta}) (\beta - \hat{\beta})' \mid D \right\} \right] \\
 (8) \qquad \qquad &= \hat{u}'\hat{u} + kE(\sigma^2 \mid D),
 \end{aligned}$$

where the first line follows from (7). Solving (8) for $E(\sigma^2 \mid D)$, we obtain the result

$$(9) \qquad E(\sigma^2 \mid D) = \frac{\hat{u}'\hat{u}}{n-k} \equiv s^2.$$

Thus relative to quadratic loss, $L(\sigma^2, \tilde{\sigma}^2) = (\sigma^2 - \tilde{\sigma}^2)^2$, the estimate $\tilde{\sigma}^2$ which minimizes post-data expected loss is $E(\sigma^2 \mid D) = s^2$, with s^2 defined in (9). Note that this post-data expectation differs from the post-data mean of σ^2 in a diffuse prior - normal likelihood traditional Bayesian approach, namely $E_{TB}(\sigma^2 \mid D) = vs^2/(v-2)$, with $v = n - k > 2$. For small values of v , the last expression is much larger than s^2 .

The results in equations (6) and (9) yield

$$(10) \qquad \text{Var}(\beta \mid D) = s^2 (X'X)^{-1}.$$

Note that the results in (4) and (10) are what one might use for moments of parameters in a large sample, approximate traditional Bayesian approach when there are difficulties in formulating a likelihood function and/or prior density. Here the results in (4) and (10) are exact results and are not large sample approximations.

We will now explain how maximum entropy is used to compute a density function from given moment conditions. We will show that using the first and second moments in (4) and (10) as constraints, the proper density that maximizes entropy is the following normal density for β given the data D^2

$$\beta \sim N(\hat{\beta}, s^2(X'X)^{-1}).$$

The entropy, or negative information, of a density function relative to uniform measure is defined as

$$(11) \quad H(f) = - \int f(x) \log f(x) dx.$$

Thus, it is seen that entropy measures the average log height of a density function. Viewed as an objective function, maximizing the entropy of a given density is to minimize the amount of information in that density. We can then consider the problem

$$(12) \quad \begin{aligned} \max_f & - \int f(x) \log f(x) dx \\ \text{subject to} & \int x^i f(x) dx = \mu_i, \quad i = 0, 1, \dots, n \end{aligned}$$

with $\mu_0 = 1$. That is, we seek to find the most conservative, or least informative density that incorporates the information in the moment side conditions. Using the calculus of variations, we find that the maxent density solving (12) is of the form

$$(13) \quad f^*(x) = \exp(-(\lambda_0 + \lambda_1 x + \dots + \lambda_n x^n)),$$

where λ_i is the Lagrange multiplier associated with the moment side condition $E[x^i]$ in (12). Substituting this maxent density into the $n + 1$ moment conditions in (12) defines $n + 1$ nonlinear integral equations in $n + 1$ unknowns. Zellner and Highfield (1988) describe an iterative solution to this problem by taking a first-order expansion of the system of moment equations about initial values for the λ 's. They show existence and uniqueness of a solution and provide applications of

²Note that if β and σ^2 are assumed a priori independent given the data, D , then the entropy of their joint density, $p(\beta, \sigma^2 | D) = f(\beta | D)g(\sigma^2 | D)$ is $H(p) = H(fg) = H(f) + H(g)$. Thus, f in (13) maximizes the first term of the entropy of the joint density and a proper g that maximizes the second term, $H(g)$, subject to moment constraints can be derived. See below for examples.

the procedures. We apply this algorithm in the application of section 4 and find that for all models used, convergence is achieved to a tolerance of 10^{-4} in under 50 iterations.

It is also interesting to note that if the moments $E(\log(x))$ and $E(1/x)$ are employed as side conditions, the proper maxent density is

$$f^*(x) = \exp\left(-(\lambda_0 + \lambda_1 \log(x) + \lambda_2 \frac{1}{x})\right) \propto x^{-\lambda_1} \exp\left(-\frac{\lambda_2}{x}\right),$$

which has the form of an inverted gamma density with parameters λ_1 and λ_2 . The inverted gamma is also the form of the traditional Bayesian posterior density for σ^2 based on a diffuse prior and iid normal likelihood function. We shall return to this case when discussing post-data densities for the scale parameter.

Adding more moment conditions imposes more constraints on the problem, and hence, the entropy of the resulting maxent density will be reduced (unless the constraints are redundant). Given the post-data moment conditions we have described thus far, we can impose these conditions as constraints and find the maxent density which is as “flat” or uninformative as possible given these conditions. By minimizing the amount of information in the density, we are letting the shape of the post-data densities be determined only by the data via derived moment conditions. A natural question, then is to ask why not add as many moment conditions as possible? Some additional moment conditions may be redundant, and it may be that we are overfitting the model by adding too many constraints. We will address this issue in discussing model selection techniques and the computation of posterior odds in sections 3 and 4.

Given the above maxent results, the proper maxent density for β given (4) and (10) is multivariate normal,

$$(14) \quad f(\beta | D) \sim N(\hat{\beta}, s^2(X'X)^{-1}).$$

This exact finite sample density can be employed to compute optimal point estimates, marginal densities, intervals, *etc.* Further, inequality restrictions on the elements of β can easily be imposed in a traditional Bayesian approach using the methods described in Geweke (1986). That is, by making draws from the normal density in (14) and accepting only those draws that satisfy the given inequality constraints, post-data densities can be obtained that incorporate information regarding the restricted values of the elements of the coefficient vector. We can also impose these inequality constraints in the BMOM approach by restricting the region of integration so that the inequality

constraints are satisfied. We can then maximize the entropy of our density over the appropriate region subject to given moment side conditions. Last, as shown below in computed examples, for moderate sample sizes, the BMOM post-data density in (14) is very similar in form to traditional Bayesian posterior densities derived using a normal likelihood and a diffuse prior.

Next, using moment conditions in (4) and (6), the following is the proper maxent density for β given σ^2 and D ,

$$(15) \quad f(\beta \mid \sigma^2, D) \sim N(\hat{\beta}, \sigma^2(X'X)^{-1})$$

This density can be employed when the value of σ^2 is known. When its value is unknown, it is interesting to note that use of (15) in conjunction with the first expression in (8) permits us to evaluate higher-order moments of σ^2 . That is, from the definition of σ^2 , we have

$$(16) \quad \sigma^2 = \frac{u'u}{n} = \frac{1}{n} \left(vs^2 + (\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta}) \right) = \frac{1}{n}(vs^2 + \sigma^2 Q),$$

where $v = n - k$ and $Q \equiv (\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta})/\sigma^2$, which has a chi-square density with k degrees of freedom provided that β has the normal density in (15). This fact can be employed to evaluate the moments of σ^2 as illustrated below.

First, note that the powers of σ^2 are defined from (16) as

$$(17) \quad \sigma^{2j} = \frac{1}{n^j} \left(vs^2 + \sigma^2 Q \right)^j, \quad j = 1, 2, \dots$$

The right-hand side of this equation can be simplified using the binomial expansion. Taking expectations through the expanded equation produces an equation to be solved for $E(\sigma^{2j} \mid D)$. Using the moments of the chi-square variable and solving the resulting expression, we obtain the following recursive formula for obtaining the desired moments for all $j > 1$,

$$(18) \quad E(\sigma^{2j} \mid D) = \frac{(vs^2)^j + \sum_{i=1}^{j-1} \binom{j}{i} (vs^2)^{j-i} E(\sigma^{2i} \mid D) \left[k(k+2) \cdots (k+2(i-1)) \right]}{n^j - \left[k(k+2) \cdots (k+2(j-1)) \right]}.$$

From this expression we find that the first two moments are given as

$$(19) \quad E(\sigma^2 \mid D) = s^2 \quad \text{and} \quad E(\sigma^4 \mid D) = \frac{s^4(v^2 + 2vk)}{n^2 - k(k+2)}.$$

Hence, the posterior variance of σ^2 is given by

$$(20) \quad Var(\sigma^2 \mid D) = s^4 \left(\frac{2k}{n^2 - k(k+2)} \right)$$

It is seen that for given s^4 and k , the post-data variance of σ^2 declines with rate n^2 given the assumptions above. Further, the variance of the scale parameter is increasing with k , the dimension of the regression coefficient vector. To compare this to the TB result, we note that the traditional Bayes posterior variance for σ^2 when a normal likelihood function and diffuse prior are employed is $2s^4v^2/(v-2)^2(v-4)$. If σ^2 has an exponential density (a case which will be described in the following sections), then σ^2 will have a post-data variance of s^4 . Thus we have a range of alternative densities which allow the sample size to have different effects on the post-data variance of σ^2 . We shall describe in sections 3 and 4 how to compute posterior odds for these alternative models and thus select the model which is most supported by the data.

Higher-order moments of σ^2 can be obtained recursively by evaluating the expression in (18). Further, moments of functions of σ^2 , say $g(\sigma^2)$ can be evaluated numerically by making draws from the chi-square density with k degrees of freedom and noting from (16) that $g(\sigma^2) = g(vs^2/(n-Q))$. Hence, for each draw from the chi-square distribution, we can compute $g(\cdot)$. Repeating this process and averaging over the resulting values will give an approximation to the mean of $g(\sigma^2)$. These can be used in connection with imposing $E(\log(\sigma^2) | D)$, and $E(1/\sigma^2 | D)$ as side conditions and deriving the proper maxent post-data density. We can evaluate the values of these moments numerically and then proceed to solve for the Lagrange multipliers as discussed previously. In Table 1 below, we consider the assumptions used and the resulting post-data densities for σ^2 in selected BMOM and TB models. These models will then be used in the application of section 4.

Table 1³
Restrictions and Post-Data Densities for σ^2

Model	Restrictions	Density Function
BMOM(1)	$E(\sigma^2 D) = s^2$	$\frac{1}{s^2} \exp\left(-\frac{\sigma^2}{s^2}\right)$
BMOM(2)	$E(\sigma^2 D) = s^2, \quad E(\sigma^4 D) = \frac{s^4(v^2+2vk)}{n^2-k(k+2)}$	$\exp(-(\lambda_0 + \lambda_1\sigma^2 + \lambda_2\sigma^4))$
BMOM(IG)	$E(\log \sigma^2 D) = \mu_1, \quad E\left(\frac{1}{\sigma^2} D\right) = \mu_2$	$\exp\left(-(\lambda_0 + \lambda_1 \log(\sigma^2) + \lambda_2 \frac{1}{\sigma^2})\right)$
TB(IG)	$p(\beta, \sigma^2 D) \propto p_N(y \beta, \sigma^2)p(\beta, \sigma^2)$	$\propto \frac{1}{\sigma^{v+2}} \exp\left(-\frac{vs^2}{2\sigma^2}\right)$

As seen from Table 1, when just the first moment of σ^2 is employed, the proper maxent density for σ^2 is in the exponential form, and when two moments of σ^2 are employed, the maxent post-data density is in a truncated normal form. Finally, when means of $\log \sigma^2$ and of $1/\sigma^2$ are used

as side conditions, the maxent post-data density for σ^2 is in the inverted gamma form, just as in the case in a diffuse prior, normal likelihood traditional Bayesian analysis. From the table, it is seen that the values of the moments of σ^2 are different in small samples. For example, the mean of σ^2 under BMOM(1) and BMOM(2) is s^2 , whereas in the diffuse prior, normal likelihood function approach, the posterior mean of σ^2 is $vs^2/(v-2)$ which is larger than s^2 . Under BMOM(IG) the mean will be functions of the Lagrange multipliers λ_1 and λ_2 . Although the densities are both of the inverted gamma form, the Lagrange multipliers may depart from the corresponding parameters of the TB(IG) density, producing different values for moments of σ^2 (see, for example Table 4 and Figure 1 in the appendix). Of course, for large n , the difference between the values of the means is negligible.

Since use of alternative assumptions leads to different densities for σ^2 , a question as to which assumptions to employ naturally arises. In some applications, higher order moments may be redundant and in others their use may lead to a better model. In sections 3 and 4 we describe and implement model selection techniques using predictive densities and standard Bayesian procedures. In this way, we can choose the model which incorporates the appropriate amount of information regarding the variance parameter σ^2 .

Shown below in Table 2 are alternative marginal post-data densities for an element, β_i , of the coefficient vector β . For BMOM(N) and TB, we know that the vector β is distributed multivariate normal and multivariate Student-t, respectively.⁴

⁴See Zellner (1971) for a discussion of TB results.

Table 2⁵
Restrictions and Post-Data Densities for β_i

Model	Restrictions	Density Function
BMOM(N)	$E(\beta) = \hat{\beta}, \quad \text{Var}(\beta) = s^2(X'X)^{-1}$	$N(\hat{\beta}_i, s_i^2)$
BMOM(1)	$\int_0^\infty p_N(\beta \sigma^2, D)p_{EXP}(\sigma^2 D)d\sigma^2$	$\propto \frac{1}{s_i} \exp\left(-\frac{\sqrt{2} \beta_i - \hat{\beta}_i }{s_i}\right)$
BMOM(2)	$\int_0^\infty p_N(\beta \sigma^2, D)p_{TN}(\sigma^2 D)d\sigma^2$	Evaluate Numerically
TB	$p(\beta, \sigma^2 D) \propto p_N(y \beta, \sigma^2)p(\beta, \sigma^2)$	Student-t $(\hat{\beta}_i, \frac{v}{v-2}s_i^2, v)$

If we do not impose an independence assumption with respect to β and σ^2 , we can write their joint post-data density as $p(\beta, \sigma^2 | D) = f(\beta | \sigma^2, D)g(\sigma^2 | D)$ and maximize the entropy of the joint density with respect to the choice of f and g subject to their being proper and to moment side conditions such as considered above. When this is done, the maxent density for β given σ^2 , the data and the first two conditional moment restrictions in (4) and (6) is $f(\beta | \sigma^2, D) \sim N(\hat{\beta}, \sigma^2(X'X)^{-1})$. Also, the maxent post-data density for σ^2 using just the first moment condition, $E(\sigma^2 | D) = s^2$ is in the gamma form. Other moment conditions can be imposed to provide a range of possible forms for the marginal post-data density for σ^2 . As mentioned earlier, a draw can be made from the post-data density for σ^2 and inserted into the conditional normal post-data density for β . Repeating this process and taking a draw from the conditional normal density for each σ^2 draw enables numerical estimation of the density function, moments, intervals, and other statistics of interest for the coefficient vector β . Also, by first integrating over all elements of β but one, say β_i , in the conditional normal density for β given σ^2 and D , the joint post-data density for β_i and σ^2 is obtained that can be analyzed by bivariate integration techniques as an alternative to the Monte Carlo method mentioned above. In some cases, these integrations can be performed analytically.

Finally, we note that the post-data densities for σ^2 can be employed to compute post-data densities for the realized error terms and functions of the realized errors that are often useful for diagnostic purposes as has been recognized in traditional Bayesian analyses; see *e.g.* Chaloner and Brant

⁵In the table, BMOM(i) $i = 1, 2$ denotes that i moments were employed in deriving the marginal post-data density for σ^2 . BMOM(N) is the normal maxent density using the post-data moments in (4) and (10). In the table, we present the forms of density functions for an element of the β vector. Further, we let s_i^2 be the (i, i) element of the matrix $s^2(X'X)^{-1}$. BMOM(IG) is not carried along in this table, but the results will be similar in functional form to TB, with the Lagrange multipliers entering as parameters of the density function. For TB, the arguments of the density are the mean, variance, and degrees of freedom, respectively.

(1988), Zellner and Moulton (1985) and Zellner (1975). Since $u_i = y_i - x_i'\beta$, given y_i, x_i , and the post-data density for β , it is possible to compute numerically or perhaps analytically the density function, moments and intervals for the u_i 's or functions of them. These can be employed to analyze outlier problems or to obtain the distributions of interesting and useful functions of the u_i 's, say $\rho_1 = \sum_{i=2}^n u_i u_{i-1} / \sum_{i=1}^n u_i^2$.⁶ That similar analyses can be carried forward when the form of the likelihood function is unknown is noteworthy.

Having derived a range of post-data densities for parameters and indicating how BMOM realized error term analysis can be performed, we now turn to derive post-data predictive densities for future observations.

2.2 BMOM Predictive Densities

Here we assume that a $q \times 1$ vector of as yet unobserved future values of the dependent variable, denoted y_f , satisfies the following q equations,

$$(21) \quad y_f = X_f \beta + u_f,$$

where X_f is a $q \times k$ matrix with elements having known values, β is the $k \times 1$ vector of regression coefficients considered in sections 2.1 and 2.2, and u_f is a $q \times 1$ vector of as yet unrealized error terms. We shall make the same assumptions regarding the properties of u_f as made in previous BMOM work and from these assumptions deduce the moments of and maxent densities for y_f given the past data, (y, X) , X_f , and assumptions. First we assume that $E(u_f | D') = 0$ which expresses the belief that there is no systematic element in the future error vector. Second, we shall assume that the future, as yet unobserved error terms each have the same variance σ^2 and are mutually uncorrelated and also uncorrelated with the elements of β .⁷ With these assumptions, the first two moments of y_f given $D' = (y, X, X_f)$ are

$$(22) \quad E(y_f | D') = X_f \hat{\beta}$$

and

$$(23) \quad \text{Var}(y_f | \sigma^2, D') = M s^2,$$

⁶See Zellner and Hong (1989) and Hong (1989) for analysis using a traditional Bayesian approach.

⁷If the future errors were assumed to have non-zero means and were correlated, we could incorporate this information into our analysis.

where $M \equiv I_q + X_f(X'X)^{-1}X'_f$. The proper predictive maxent post-data density for y_f subject to (22) and (23) is

$$(24) \quad f(y_f | D') \sim N(X_f \hat{\beta}, Ms^2),$$

which can be used to compute marginal densities, predictive intervals, and other quantities of interest.

In addition to the result in (24) that parallels that for estimation in (14) we can also derive the following maxent conditional predictive density for y_f given D' that incorporates the conditional moments $E(y_f | \sigma^2, D') = X_f \hat{\beta}$, and $\text{Var}(y_f | \sigma^2, D') = M\sigma^2$:

$$(25) \quad f(y_f | \sigma^2, D') \sim N(X_f \hat{\beta}, M\sigma^2).$$

This conditional normal density can be multiplied by any of the marginal post-data maxent densities for σ^2 that are shown in Table 1. The marginal predictive density of y_f can be computed numerically by drawing a value of σ^2 from its marginal density, inserting this drawn value into (25), and drawing a vector y_f from the conditional normal density. Repeating this procedure will provide draws from the marginal density and thus enable calculation of predictive intervals, moments, etc. of y_f . Also, we can compute traditional Bayesian predictive densities based on, say, normal sampling assumptions and diffuse priors. As will be shown in the next section, posterior odds can be employed to evaluate alternative models and to provide a means for combining alternative models and their predictions as discussed and applied in a traditional Bayesian framework by Min and Zellner (1993). The BMOM predictive densities that we shall consider are shown in Table 3. We present density functions (when available) for the case in which y_f is a scalar. As discussed in Table 2, BMOM(N) and TB results generalize to the multivariate case.

Table 3 ⁸Restrictions and Post-Data Predictive Densities for a scalar y_f

Model	Restrictions	Density Function
BMOM(N)	$E(y_f D) = \hat{y}_f, \quad \text{Var}(y_f D) = s_e^2$	$N(\hat{y}_f, s_e^2)$
BMOM(1)	$\int_0^\infty p_N(y_f \sigma^2, D) p_{EXP}(\sigma^2 D) d\sigma^2$	$\propto \frac{1}{s_e} \exp\left(-\frac{\sqrt{2} y_f - \hat{y}_f }{s_e}\right)$
BMOM(2)	$\int_0^\infty p_N(y_f \sigma^2, D) p_{TN}(\sigma^2 D) d\sigma^2$	Evaluate Numerically
TB	$p(y_f D') = \int p_N(y_f \beta, \sigma^2, D') p(\beta, \sigma^2 D) d\beta d\sigma^2$	Student-t $(\hat{y}_f, \frac{v}{v-2} s_e^2, v)$

Given the alternative BMOM models for a scalar y_f in Table 3, a vector y_f , or predictive densities based on other moment side conditions, there is a need for model comparison and selection methods, a problem which we take up in the following section.

3 Model Comparison and Selection Techniques

If we have observed the values of the elements of y_f , the predictive post-data densities discussed above can be utilized to compute Bayes factors. That is, for two alternative predictive densities $f_1(y_f | D')$ and $f_2(y_f | D')$, the TB posterior odds, K_{12} , is given by

$$(26) \quad K_{12} = \frac{\pi_1 f_1(y_f | D')}{\pi_2 f_2(y_f | D')},$$

namely the product of the prior odds π_1/π_2 times the Bayes factor. On inserting $y_f = y_f^0$, the observed value of y_f , and a value for the prior odds, a numerical value for K_{12} is obtained; see section 4 for computed examples. Further as Good (1950) and Kullback (1959) have noted, on taking logs of both sides of (26), and averaging with respect to $f_1(y_f | D')$, the following result is obtained

$$(27) \quad W_{12} = \int \log K_{12} f_1(y_f | D') dy_f - \log \frac{\pi_1}{\pi_2} = \int f_1(y_f | D') \log \frac{f_1(y_f | D')}{f_2(y_f | D')} dy_f.$$

⁸Here, as in Table 2, BMOM(i), $i = 1, 2$ denotes the use of a maxent density for σ^2 with the use of just a first moment constraint and first and second moment constraints, respectively. BMOM(N) is the normal maxent post-data density using conditions (22) and (23). TB utilizes a normal likelihood function, diffuse prior and Bayes rule to obtain the joint posterior, $p(\beta, \sigma^2 | D)$. Multiplying this density by the conditional normal for y_f given β and σ^2 and integrating out the nuisance parameters σ^2 and β leads to a marginal predictive density for y_f in the univariate Student-t form. The BMOM(IG) density will also be of the Student-t form with the Lagrange multipliers entering as parameters of the density. We have also defined $s_e^2 = s^2(1 + x_f(X'X)^{-1}x_f')$.

Thus, the averaged log posterior odds minus the log prior odds, or “the weight of the evidence”, is equal to the cross entropy of f_1 with respect to f_2 , and is denoted by $CE(f_1, f_2)$, a non-negative measure of the distance between f_1 and f_2 . Further, from (27) we have

$$(28) \quad W = W_{12} + W_{21} = CE(f_1, f_2) + CE(f_2, f_1).$$

The quantity W in (28) above is symmetric with respect to f_1 and f_2 and is well-known as the Jeffreys-Kullback-Leibler distance measure. This derivation illustrates the connection between TB model comparison methods and the concept of entropy and cross entropy.

From (26), $K_{12} = P_1/P_2$, where P_i is the posterior probability associated with model $f_i(y_f | D')$, $i = 1, 2$. Since these two models are not exhaustive, $P_1 + P_2 < 1$. Even so, in the standard two-action, two-state model selection problem, (see *e.g.* DeGroot (1970)), if the loss structure is symmetric and $K_{12} = P_1/P_2 > 1$, then it is optimal in terms of minimizing expected loss to choose $f_1(y_f | D')$. If $K_{12} = P_1/P_2 < 1$, $f_2(y_f | D')$ is the optimal choice. See Palm and Zellner (1992) and Min and Zellner (1993) for application of these techniques in choosing between or combining alternative forecasting models. Of course, loss functions such as those described in Dehling et al (1996) can also be employed in evaluating expected losses associated with choices of alternative densities. Other model selection criteria can also be employed; see Judge *et al.* (1985) for a discussion of alternative criteria.

We now turn to presenting applications of the techniques described in this and the previous sections to illustrate how BMOM can be implemented in practice.

4 Computed Examples

In this section we compare results obtained from alternative BMOM and traditional Bayesian (TB) models in analyses of data from known models. We use the average log odds criterion, discussed in section 3, to compare the performance of the alternative models. Recall from equation (21) that with equal prior odds, we can equate the average log odds of model 1 relative to model 2 with the cross entropy of f_1 with respect to f_2 as follows:

$$\int \log K_{12} f_1(y_f | D') dy_f = \int \log \frac{f_1(y_f | D')}{f_2(y_f | D')} f_1(y_f | D') dy_f,$$

where K_{12} represents the traditional Bayes posterior odds based on equal prior odds. In this section, we let f_1 denote the “true” model, or the actual density from which the generated data are drawn.

The remaining density in the expression above, f_2 , represents a BMOM or TB predictive density which has been derived from the given generated sample of observations. By computing the integral above, we can determine which of the alternative BMOM or TB models is “closest” to the model from which the data were actually generated. Further, if we have the predictive densities from two competing models, say f_2 and f_3 , we can compute the average log odds of f_3 relative to f_2 by noting that

$$\begin{aligned} \int \log K_{12} f_1(y_f | D') dy_f - \int \log K_{13} f_1(y_f | D') dy_f &= \int \log \frac{f_3(y_f | D')}{f_2(y_f | D')} f_1(y_f | D') dy_f \\ &= \int \log K_{32} f_1(y_f | D') dy_f. \end{aligned}$$

Thus, the difference between the average log odds of the true model, f_1 , relative to f_2 and f_3 is equal to the average log odds of f_3 relative to f_2 , where the averaging is done with respect to the “true” predictive density. To illustrate, suppose that model 2 is “closer” to the true model than model 3 in terms of our average log odds criterion. In this case, the expression above will be negative,⁹ implying that the average log odds of model 3 with respect to model 2 is also negative. This result indicates that on average, model 2 will be preferred to model 3, which is a reasonable result, given that model 2 was assumed to be “closer” to the true model.

We generate both normal and non-normal data and compare the results obtained from BMOM and TB models for both types of data. For each type of data and each estimation procedure, we employ a sample with 200 observations. From this sample, the moments necessary to compute BMOM predictive densities and statistics needed to calculate the TB predictive density are obtained. The model that we use is a simple regression model given as follows:

$$y_i = \alpha_0 + \alpha_1 x_i + e_i,$$

with $\alpha_0 = .1$, and $\alpha_1 = 4$. The x_i 's are drawn independently from the uniform distribution on $[0,1]$. In one experiment, we draw the e 's from an iid normal distribution with mean zero and variance $\sigma^2 = 4$. In the other experiment, we draw the e 's independently from a student- t distribution with mean zero, $v = 4$ degrees of freedom, and variance $v\sigma^2/(v - 2) = 4$.

In these experiments, we consider three BMOM models: BMOM(N), BMOM(2), and BMOM(IG). The BMOM(N) predictive density is based on equation (24), which is the proper maxent density that results from imposing (22) and (23) as moment side conditions. The BMOM(2) and BMOM(IG) densities are obtained by averaging the conditional normal predictive density in (25) over the

⁹Note that the average log odds, or cross entropy, must be non-negative.

truncated normal and inverted gamma marginal posteriors for σ^2 , respectively. These marginal posteriors for the scalar parameter were discussed in section 2 and are presented in Table 1. For the TB case, we consider the traditional Bayesian model based on a diffuse prior, $p(\alpha_0, \alpha_1, \sigma) \propto 1/\sigma$ and an iid normal likelihood function. As is well known, the predictive density for this model has a multivariate Student- t distribution, as shown in Table 3. We draw a sample of 220 observations from the “true” model, f_1 , use the first 200 observations to estimate the BMOM and TB models, and hold out a 20×1 vector of future observations to evaluate ordinates of the predictive densities, $f_1(y_f | D')$, and $f_2(y_f | D')$. This process is repeated, and the integral is estimated by taking a sample average of the values of the computed log odds.

Presented in Tables 4-5 of the appendix are estimates of the first and second raw moments of $\log K_{12}$ and K_{12} , and percentiles of the sampling distributions of $\log K_{12}$ and K_{12} obtained from the generated normal data and the generated student- t data. Moments are computed by taking sample averages of the values of $\log K_{12}$ and K_{12} , which are computed for each draw from f_1 . Table 4 presents the results based on the generated normal data, and Table 5 presents the results using the Student- t data. We also use the set of values for $\log K_{12}$ and K_{12} to plot estimates of the sampling densities of the log odds and odds in Figures 1-2. Figures 1 and 1a plot the sampling densities of the log odds and odds, respectively, for the generated normal data, and Figures 2 and 2a present the same densities for the generated Student- t data. These densities are obtained by smoothing the values of $\log K_{12}$ and K_{12} using a kernel, and nonparametric density estimates¹⁰ are presented in the Figures. As suggested in the following tables, results obtained from BMOM(2) and BMOM(IG) are nearly identical to results using BMOM(N), and thus we only present the estimated sampling distributions for the BMOM(N) and TB models.

From the graphs in Figures 1 and 2, we see that results obtained from the BMOM(N) and TB models are quite similar for both the normal and Student- t generated data. In the normal model, the sampling density of the log odds appears to be centered at zero and approximately symmetric, suggesting that the BMOM(N) and TB models perform nearly as well as the true model. From Table 4, we see that the average log odds, or cross entropy, is slightly positive for all models, indicating that the true model is only slightly preferred to the alternative TB and BMOM models. The results for the generated Student- t data contrast with the results obtained from the normal case. The sampling densities of the log odds and odds for the generated Student- t data have a pronounced

¹⁰We chose a Gaussian kernel and fixed bandwidth. The bandwidth was chosen so that a reasonable degree of smoothness was achieved. Other density estimation methods, such as maxent with the moments in Tables 4-5, or higher-order moments, can also be employed.

right-skew. This indicates that there is a significant positive probability that the true model will be strongly favored over the TB or BMOM(N) models when the underlying data distribution is generated according to a Student- t distribution with a small degrees of freedom parameter. Table 5 shows that the expected log odds for the true model relative to TB is smaller than the expected log odds of the true model relative to the BMOM(N) model. Using the expression derived earlier, this suggests that the TB model will be favored on average relative to the BMOM(N) model, although the performance of the TB and BMOM(N) models is quite similar.

We also analyzed generated data with larger values for the degrees of freedom parameter v , and find that the resulting sampling densities of $\log K_{12}$ approach the densities presented in the normal case in Figure 1a as the degrees of freedom parameter increases. For a small value for the degrees of freedom parameter, the normal likelihood function in the TB case departs significantly from the true data generating density. The heavy right tail of the approximate posterior distribution of $\log K_{12}$ indicates that there is a positive probability that both BMOM(N) and TB will perform poorly relative to the true model in the case of the generated student- t data.

We also note from Tables 4 and 5 that the results obtained from the BMOM(N), BMOM(2), and BMOM(IG) models are very similar. In this example, the the inverted gamma and truncated normal densities for the scale parameter, σ^2 are very informative and spiked about the value s^2 . The BMOM(N) model is obtained by taking expectations over σ^2 , and behaves as if the posterior distribution of σ^2 is degenerate about s^2 . Since the inverted gamma and truncated normal densities for the scale parameter are quite informative in this example, it is not surprising that results obtained from the alternate BMOM models are quite similar. To introduce a possible departure in the results obtained from our BMOM models, we repeated this experiment for a significantly smaller sample size. We generated 40 observations in an identical manner, using the last 10 observations as a hold-out sample, and the first 30 observations to estimate the TB and BMOM models. In this case, our BMOM models produced different results, and the BMOM models were occasionally favored over the TB model in terms of average log odds.

The results in this section suggest that BMOM results, based on fewer assumptions than those required for TB analysis, are quite competitive with results obtained from a Traditional Bayesian analysis. BMOM analysis does not require the researcher to specify a likelihood function or prior density, and offers a valuable alternative to TB procedures when there is difficulty in specifying these quantities. The use of additional moment side conditions, such as those described in Zellner,

Tobias and Ryu (1998), can also be employed to give possibly improved predictive performance and may be robust to departures from normality. We also linked the average posterior odds with cross-entropy, and demonstrated how this criterion can be used to compare and select among competing models. The average log odds criterion used in this section is quite operational, and can easily be applied in other experiments to evaluate the performance of alternative models.

5 Conclusion

In this paper we have indicated how to apply the Bayesian Method of Moments in analysis of the standard multiple regression model when information is not available to formulate a likelihood function. On introducing simple assumptions relating to the moments of the realized error terms and the future, as yet unobserved error terms, we derived post-data moments of parameters and future values of the dependent variable. Using these moments as side conditions, proper maxent densities for parameters and future values of the dependent variable were derived that can easily be computed from the data. Further, the methods developed in this paper can be used to extend previous BMOM analyses of models with autocorrelated or heteroscedastic error terms, see e.g Currie (1996), Zellner and Sacks (1996), and Zellner (1997b), where use is made of the Gibbs sampler to compute post-data densities. Last, it was shown how alternative BMOM and TB predictive densities can be compared by use of posterior odds. As shown in computed examples, some BMOM maxent densities are very similar to TB densities, while others are not. Thus it is fortunate that it is possible to use posterior odds to ascertain the extent to which alternative assumptions and models are supported by information in the data.

With respect to future research, we, along with H. Ryu, are applying BMOM techniques to various semi-parametric models such as considered in Ryu (1993). Further attention is being given to understanding how alternative assumptions regarding the form of the dependence between regression coefficients and the variance parameter affects results statistically and economically. It has been recognized that the form of the mean-variance dependency is a critical factor in explaining decision-making under uncertainty in financial economics and other areas as well as in formulating likelihood functions. Last, work relating to multivariate regression similar to that contained in the present paper will extend the BMOM multivariate results in Zellner (1995).

6 Appendix

Estimates of the First Two Sampling Moments and Sampling Percentiles of $\log K_{1i}$
and K_{1i} Using Generated Normal Data.

Table 4a: $\log K_{1i}$

Moment	Model Used			
	TB	BMOM(N)	BMOM(2)	BMOM(IG)
$E[\log K_{12}]$.1595	.1613	.1612	.1613
$E[(\log K_{12})^2]$.3223	.3422	.3419	.3418
Percentiles	TB	BMOM(N)	BMOM(2)	BMOM(IG)
10 th	-.5014	-.5015	-.5010	-.5007
25 th	-.1486	-.1444	-.1444	-.1437
50 th	.1656	.1453	.1457	.1459
75 th	.4460	.4379	.4382	.4380
90 th	.8063	.8417	.8417	.8418

Table 4b: K_{1i}

Moment	Model Used			
	TB	BMOM(N)	BMOM(2)	BMOM(IG)
$E[K_{12}]$	1.3634	1.3891	1.3887	1.3887
$E[(K_{12})^2]$	2.6077	2.9406	2.9353	2.9354
Percentiles	TB	BMOM(N)	BMOM(2)	BMOM(IG)
10 th	.6057	.6056	.6060	.6061
25 th	.8619	.8655	.8655	.8661
50 th	1.1801	1.1564	1.1569	1.1571
75 th	1.5621	1.5494	1.5499	1.5496
90 th	2.2395	2.3203	2.3202	2.3205

Estimates of the First Two Sampling Moments and Sampling Percentiles of $\log K_{1i}$
and K_{1i} Using Generated Student- t Data.

Table 5a: $\log K_{1i}$

Moment	Model Used			
	TB	BMOM(N)	BMOM(2)	BMOM(IG)
$E[\log K_{12}]$.1747	.3533	.3502	.3501
$E[(\log K_{12})^2]$	7.4587	20.1079	19.7937	19.7699
Percentiles	TB	BMOM(N)	BMOM(2)	BMOM(IG)
10 th	-1.0441	-1.0791	-1.0786	-1.0788
25 th	-.8787	-.9126	-.9122	-.9119
50 th	-.4589	-.4562	-.4562	-.4561
75 th	.4871	.5468	.5457	.5464
90 th	1.7771	1.9340	1.9318	1.9324

Table 5b: K_{1i}

Moment	Model Used			
	TB	BMOM(N)	BMOM(2)	BMOM(IG)
$E[K_{12}]$	2.2806	3.0076	2.9946	2.9942
$E[(K_{12})^2]$	44.0355	87.5010	86.4856	86.4001
Percentiles	TB	BMOM(N)	BMOM(2)	BMOM(IG)
10 th	.3505	.3367	.3369	.3369
25 th	.4138	.3990	.3993	.3991
50 th	.6188	.6163	.6162	.6168
75 th	1.4468	1.6145	1.6124	1.6132
90 th	4.5837	5.4890	5.4781	5.4836

Nonparametric Estimates of the Sampling Density of $\log K_{12}$ Using BMOM(N) and TB models. Generated Normal Data with $\sigma^2 = 4$. 200 Observations used for estimation, 20 in Hold-out Sample. 2,500 Iterations.¹¹

Figure 1a: Sampling Density of Log Odds

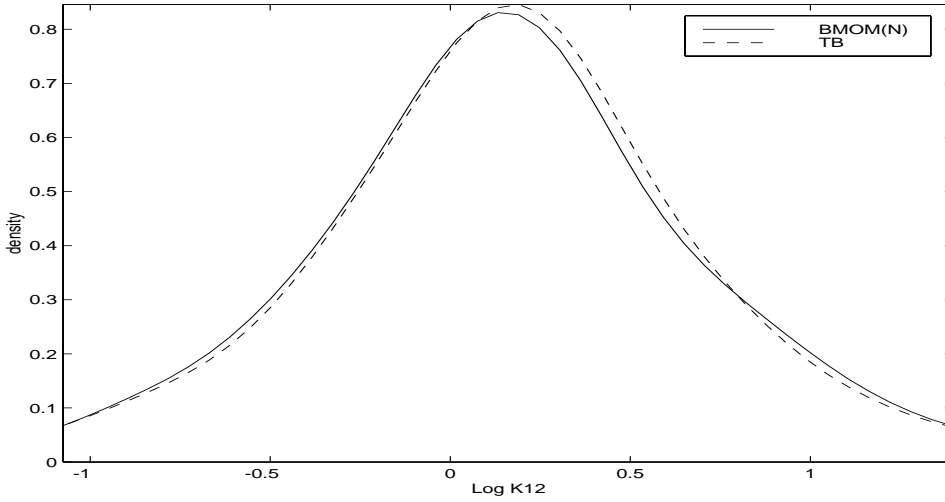
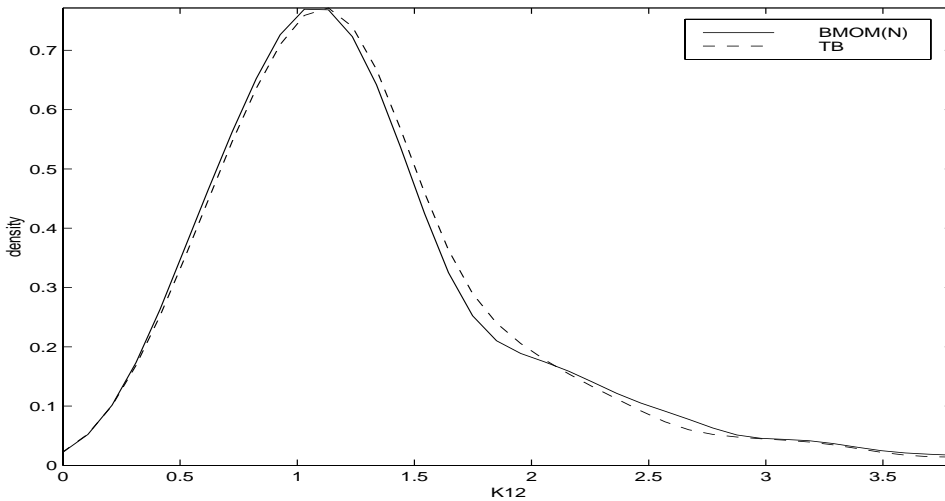


Figure 1b: Sampling Density of Odds



¹¹For each draw from the true model, f_1 we compute the corresponding value for $\log K_{12}$. Given 2,500 of these draws, we then smooth them using a Gaussian kernel and fixed bandwidth. Note that results obtained from BMOM(2) and BMOM(IG) are nearly identical to BMOM(N) results, and thus we present only the graphs for BMOM(N) and TB.

Nonparametric Estimates of the Sampling Density of $\log K_{12}$ Using BMOM(N) and TB models. Generated Student-t Data with variance 4 and 4 degrees of freedom. 200 Observations used for estimation, 20 in Hold-out Sample. 2,500 Iterations.¹²

Figure 2a: Sampling Density of Log Odds

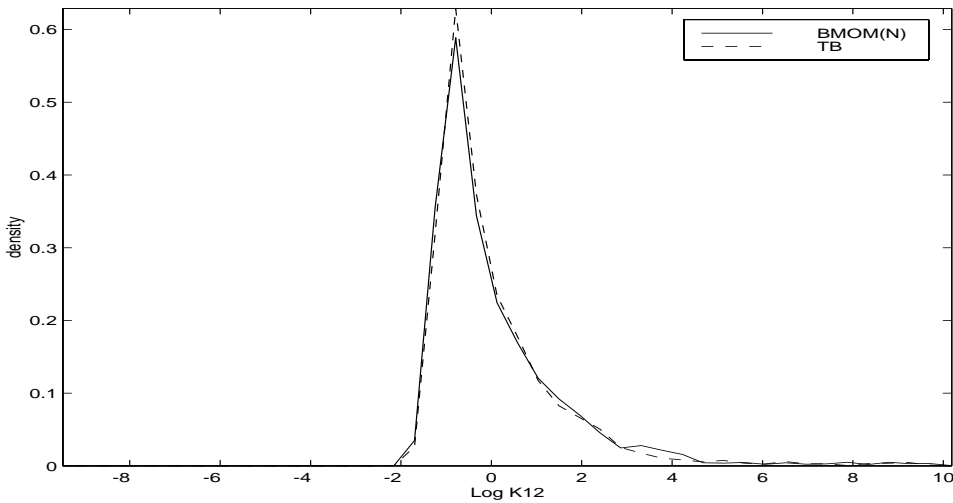
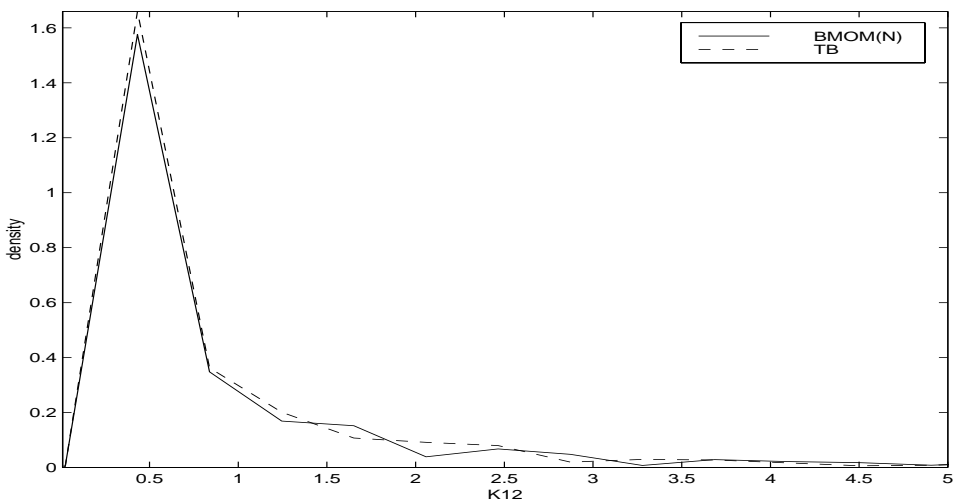


Figure 2b: Sampling Density of Odds



¹²For each draw from the true model, f_1 we compute the corresponding value for $\log K_{12}$. Given 2,500 of these draws, we then smooth them using a Gaussian kernel and fixed bandwidth. Note that results obtained from BMOM(2) and BMOM(IG) are nearly identical to BMOM(N) results, and thus we present only the graphs for BMOM(N) and TB.

References

- Barnard, G. Personal Communication, 1997.
- Bayes, T., “An Essay Towards Solving a Problem in the Doctrine of Chances,” *Phil. Trans. Royal. Soc.* 53 (1763), 370–418.
- Chaloner, K. and R. Brant, “A Bayesian Approach to Outlier Detection and Residual Analysis,” *Biometrika* 75 (1988), 651–659.
- Cover, T. and J. Thomas, *Elements of Information Theory* (New York: John Wiley & Sons, 1991).
- Currie, J., “The Geographic Extent of the Market: Theory and Application to the U.S. Petroleum Markets,” Ph.D. thesis, University of Chicago, 1996.
- DeGroot, M., *Optimal Statistical Decisions* (New York: McGraw-Hill, 1970).
- Diebold, F. and R. Lamb, “Why are Estimates of Agricultural Supply Response so Variable?,” *Journal of Econometrics* 76 (1997), 357–373.
- G. Judge, W.E. Griffiths, R. H. Carter, H. L. Lütkepohl and T.-C. Lee, *The Theory and Practice of Econometrics* (New York: John Wiley & Sons, Inc., 1985).
- Geweke, J., “Exact Inference in the Inequality Constrained Normal Linear Regression Model,” *Journal of Applied Econometrics* 1 (1986), 127–141.
- Golan, A., G. Judge, and D. Miller, *Maximum Entropy Econometrics* (Chichester: John Wiley & Sons, Ltd., 1996).
- Good, I., *Probability and the Weighting of the Evidence* (New York: Hafner, 1950).
- Green, E. and W. Strawderman, “A Bayesian Growth and Yield Model for Slash Pine Plantations,” *Journal of Applied Statistics* 23 (1996), 281–289.
- H.G. Dehling, T.K. Dijkstra, H.J. Guichelarr, W. Schaafsma, A.G.M. Steerneman, T.J. Wansbeek and J.T. van der Zee, “Structuring the Inferential Contest,” in D. Berry, K. Chaloner & J. Geweke, eds, *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner* (New York: John Wiley & Sons, 1996).
- Jaynes, E., “On the Rationale of Maximum-Entropy Methods,” *Proceedings of the IEEE* 70 (1982), 939–952.

- Jaynes, E., "Comment on 'Optimal Information Processing and Bayes' Theorem'," *American Statistician* 42 (1988), 280–281.
- Min, C-k. and A. Zellner, "Bayesian Analysis, Model Selection and Prediction," *Physics and Probability: Essays in honor of Edwin T. Jaynes* (1993), 195-206.
- Kullback, S., *Information Theory and Statistics* New York: Wiley, 1959.
- Mead, L. and N. Papanicolaou, "Maximum Entropy in the Problem of Moments," *Journal of Mathematical Physics* 25 (1984), 2404–2417.
- Palm, F.C. and A. Zellner, "To Combine or Not to Combine? Issues of Combining Forecasts," *Journal of Forecasting* 11 (1992), 687–701.
- Ryu, H., "Maximum Entropy Estimation of Density and Regression Functions," *Journal of Econometrics* 56 (1993), 397–440.
- Shore, J. and R. Johnson, "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy," *IEEE Transactions IT-26* (1980), 26–37.
- Soofi, E., "Capturing the Intangible Concept of Information," *Journal of the American Statistical Association* 89 (1994), 1243–1254.
- Soofi, E., "Information Theory and Bayesian Statistics," in D. Berry, K. Chaloner and J. Geweke, eds, *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner* (New York: John Wiley & Sons, 1996).
- van der Merve, A.J. and C.S. Viljoen, "Bayesian Method of Moments Analysis of the Seemingly Unrelated Regression Model," Presented at the South African Statistical Association Annual Meeting, November 1998.
- Zellner, A., *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley, 1971, reprinted in 1996 by Wiley Classics Library.
- , "Bayesian Analysis of Regression Errors," *Journal of the American Statistical Association* 70 (1975), 138–144.
- , "Optimal Information Processing and Bayes' Theorem," *American Statistician* 42 (1988), 278–284.
- , "Bayesian Methods and Entropy in Economics and Econometrics," in T. Grandy and L. Schick, eds, *Maximum Entropy and Bayesian Methods*, Dordrecht, Netherlands: Kluwer, 1991.

- , “The Finite Sample Properties of Simultaneous Equations’ Estimates and Estimators : Bayesian and Non-Bayesian Approaches,” in L. R. Klein, ed., *Journal of Econometrics: Annals Issue in honor of Carl Christ* 1998, and in Zellner 1997b.
- , “Bayesian Method of Moments/ Instrumental Variable (BMOM/IV) Analysis of Mean and Regression Models,” in J. Lee, W. Johnson and A. Zellner, eds, *Modeling and Prediction: Honoring Seymour Geisser*, Springer-Verlag, 1996 and in Zellner (1997b).
- , “The Bayesian Method of Moments (BMOM): Theory and Applications,” in T. Fomby and R. Hill, eds, *Advances in Econometrics: Applying Maximum Entropy to Econometric Problems* 12 (1997a), 85-105.
- , *Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers* (Chiltenham, U.K.: Edward Elgar Publishing Company, 1997b).
- and R. Highfield, R., “Calculation of Maximum Entropy Distributions and Approximation of Marginal Posterior Distributions,” *Journal of Econometrics* 37 (1988), 195–210.
- and B. Moulton, “Bayesian Regression Diagnostics with Applications to International Consumption and Income Data,” *Journal of Econometrics* 29 (1985), 187–211.
- and B. Sacks, “Bayesian Method of Moments (BMOM) Analysis of the Multiple Regression Model with Autocorrelated Errors,” H.G.B. Alexander Research Foundation, University of Chicago, 1996.