MACSYMA as a Tool for Statisticians

J. Michael Steele, Princeton University

ABSTRACT

Three concrete examples are used to show how symbolic computation can be of value in statistical research. Enough of the syntax of MACSYMA is discussed to provide an introduction to actual usage. The questions considered concern the relationship of conditional cumulants of homogeneous statistics, a method for proving limit theorems by studying zeros of polynomials, and the Legendre transformation of some probability densities. One question which is constantly considered is " When will MACSYMA seriously help ?"

1. Introduction

MACSYMA and other symbolic computational tools are now widely available. Still, the statistical community has not felt the full impact of these new tools and many people may not yet realize that the conduct of research in probability and statistics will be forever changed by their presence. The purpose of this report is to illustrate how even naive use of MACSYMA can be of substantial benefit in research activity which is far from being *a priori* computational.

The plan of the report is to first give a gentile introduction to the structure and syntax of MACSYMA, give a few hints for using MACSYMA comfortably in a UNIX environment, and then give three examples. The examples are not concocted but are just selected from the my own experience in using MACSYMA as one of the tools of day-to-day research. Finally, brief comments are given concerning the ways one can customize a MACSYMA environment to facilitate the types of computations common in probability and statistics.

2. Getting Started

The way to get started using MACSYMA is to fire up MACSYMA, open *An Introduction to MACSYMA* by Joel Moses and the MACSYMA Group of Symbloics, Inc.(1984), and start playing. This is fun and effective. In stark contrast, you can go to MACSYMA with a tough problem which you have not succeeded in doing by hand, try to use MACSYMA for the first time and end up meeting only frustration and likely failure.

In teaching MACSYMA to my friends, the hardest thing to get across always seems to be that MACSYMA is **just a tool.** A proper attitude is created by thinking of MACSYMA as being like integration by parts. Surely integration by parts is a magnificent tool. It shines in vital areas like the calculus of variations. It gives birth to the formulas of Gauss, Green, and Stokes. It is a weapon of first choice in many parts of asymptotics. But for all of these virtues, integration by parts will not solve all problems, and neither will MACSYMA. The first task of the beginner is to shun the inevitable disappointment of naive expectations and to articulate those areas of personal interest where MACSYMA has a realistic chance of success.

The scope of MACSYMA and its interactive nature are easy to see by looking over a brief but complete UNIX script of a trial session. In the first computations we review, the syntax should be obvious if one bears in mind that the c-lines are entered by the user and the d-lines are MACSYMA's responses. One small point is that ^ is used to denote exponentiation. The symbols (d1), (d2), etc. will appear on the left side on the terminal just like the c-labels. Their appearance on the right side in this report is an artifact of using the typesetting features of MACSYMA (check out the flag setting typeset:true).

(c1) diff(sin(x)/x,x,3);

$$\frac{3\sin(x)}{x^2} - \frac{6\sin(x)}{x^4} - \frac{\cos(x)}{x} + \frac{6\cos(x)}{x^3} \qquad \text{(d1)}$$

(c2) taylor(cos(x)^ 3,x,0,6);

$$1 - \frac{3x^2}{2} + \frac{7x^4}{8} - \frac{61x^6}{240} + \cdots \qquad \text{(d2)}$$

(c3) solve(2*x^ 2+x+2);

$$[x = -\frac{\sqrt{15}i + 1}{4}, x = \frac{\sqrt{15}i - 1}{4}] \qquad \text{(d3)}$$

(c5) quit();

This micro-session is enough to show some of the power in symbolic computation. One does not need MACSYMA to calculate the third derivative of sin(x)/x, or the Taylor series of $\cos(x)^3$ to four terms, but it is in fact easier and more reliable than doing it by hand. Definite and indefinite integrals are also available with a comparably natural syntax. The use of solve illustrates a slight variation on natural syntax ,but solve(2*x^ 2+x+3)=0 works just as well.

Before going to a real example, one should learn some quick syntax. First ";" terminates a command (or c) line. Also, [a,b,c] is a list containing a, b, and c. There are assignment operators ":=" which defines functions , and ":" for more typical assignment.

3. Some Practical Hints

There are some frustrations which most newcomers to MACSYMA seem to meet and which are easily overcome once "you know the trick". First, the beginner should avoid the MACSYMA editor. The best way to use MACSYMA is to work interactively only while exploring and playing. Once a serious problem is to be studied the MACSYMA commands should be put in a file , say comp.file, and executed in MACSYMA with the command **batch("comp.file")**

The benefits of this process are (1) If an error is made you can stop MACSYMA with ^Z and then just edit comp.file width your favorite editor, (2) The code you create becomes an asset which can be used in different contexts after the traditional pattern of "modify rather than write", and (3) The code can be conveniently quoted in papers, shared with others, etc.

Please do not be mislead. There are many other tools in MACSYMA which allow one to communicate with files and with structures within MACSYMA; but, for a while, these seem best left to *vertuosi.*

Two more practical hints ? Even if you do not wish to do ANY symbolic computation MACSYMA can be a Godsend for its FORTRAN and eqn facilities. In a nutshell FORTRAN(expression) produces the FORTRAN code to compute the specified expression. I find this a very pleasant way to avoid coding mistakes. In the same

way the typeset flag of MACSYMA permits you to get the eqn code of the resulting d-line. This is much more pleasant and precise than writing your own equations with eqn. The equations of this report were all done in this way.

4. Simplicity Though Diligence

One way in which MACSYMA can be of aid in research is simply by removing the difficulties of looking at the problem from many sides. Often one has an idea which can go unexplored because the calculations look too cumbersome. This inhibition emerges for two reasons. First, it takes conviction to go trough a tedious calculation which may not lead to a useful result, and second, the notion of usefulness is often linked to simplicity.

This block can be busted by MACSYMA in a number of cases. One from my experience concerns conditional cumulants of homogeneous statistics. The reason on can expect MACSYMA to be effective here is that the tasks required are just tedious, not ingenious. At differentiation, substitution, and rational simplification by expansion MACSYMA is as surefooted as a mountain goat.

5. Cumulant Correction Formulas

We will assume that f satisfies the homogeneity condition (1.1) and set

$$Z_n = f(X_1, X_2, \cdots, X_n) \tag{2.1}$$

where the X_i are independent and uniform on $[0,1]^d$. Further, we denote by Z'_n a random variable with the same definition as Z_n except that one of the X_i's is constrained to be on "forward" part of the boundary of $[0,1]^d$; that is, one of the X_i's is constrained to be in the set $B = [0,1]^d - [0,1)^d$.

If Y is any random variable the cumulant generating function of Y is defined by

$$\sum_{i=1}^{\infty} \kappa_i(Y) \frac{t^i}{i!} = \log E(e^{tY}) \tag{2.2}$$

and the numbers $\kappa_i(Y)$ are the cumulants (or semi-invariants) of Y. Naturally, $\kappa_1(Y)$ is the mean of Y, κ_2 is the variance, and these are the most important of the cumulants. The third and forth cumulants (the skewness and the kurtosis) are of interest principally because after the variance all of the cumulants of a normal distribution are zero. The cumulants of Z_n do not have to tend to zero for it to be asymptotically normal, but if the third and forth cumulants tend to zero it is a strong indication of asymptotic normality.

One of our goals in pursuing the simulation method studied here (and in particular with bothering about higher cumulants) is to obtain a computationally expedient method of assessing the possibility of asymptotic normality in the case of computationally important random variables like the length of the minimal spanning tree. In Ramey (1982) the asymptotic normality of the MST was proved under the hypothesis of an assumption which seems very difficult to verify (but which is still quite plausible). The asymptotic normality of the length of the MST remains an important open problem. In fact, there is no subadditive Euclidean functional for which normality has been proved yet it is quite likely to hold for the whole class (c.f. Beardwood, Halton, and Hammersley (1959) and Steele (1981)).

To begin the conditioning argument, we let

$$S = \min\{s : X_i \in [0,s]^d\}$$

thus, S is the size of the side of the smallest cube containing the data and the origin. We note that with probability one there is exactly one of the X_i on the forward part of the boundary of $[0,S]^d$. A key observation is that we then have that the distribution of Z_n equals the distribution of $S^p Z'_n$ where S and Z'_n are generated independently.

Now we can just calculate,

$$\exp \sum_{i=1}^{\infty} \kappa_i(Z_n) \frac{t^i}{i!} = E(E(e^{tZ_n} | S)) \tag{2.3}$$

$$= EE(e^{tS^p Z'_n} | S)) = E \exp \sum_{i=1}^{\infty} \kappa_i(Z'_n) S^{ip} \frac{t^i}{i!}$$

Next we expand both sides of the above using

$$\exp \sum_{i=1}^{\infty} a_i \frac{t^i}{i!} = 1 + a_1 + (a_1^2 + a_2)\frac{t^2}{2!}$$

$$+ (a_1^3 + 3a_2 a_1 + a_3)\frac{t^3}{3!} +$$

$$(a_1^4 + 6a_2 a_1^2 + 4a_3 a_1 + 3a_2^2 + a_4)\frac{t^4}{4!} + \cdots$$

and then complete taking the expectation by using the elementary fact that $E(S^\alpha) = dn(\alpha + dn)^{-1}$. On simplifying the algebra, one finds the following tidy expressions for the conditional cumulants (To save space we will write κ_i for $\kappa_i(Z_n)$:

$$\kappa_1(Z'_n) = \kappa_1\left(1 + \frac{p}{nd}\right) \tag{2.4a}$$

$$\kappa_2(Z'_n) = \kappa_2\left(1 + \frac{2p}{nd}\right) - \kappa_1^2 \frac{p^2}{n^2 d^2} \tag{2.4b}$$

$$\kappa_3(Z'_n) = \kappa_3\left(1 + \frac{3p}{nd}\right) - \frac{6\kappa_1\kappa_2 p^2}{n^2 d^2} + \frac{2\kappa_1^3 p^3}{d^3 n^3} \tag{2.4c}$$

and finally we have

$$\kappa_4(Z'_n) = \kappa_4\left(1 + \frac{4p}{dn}\right) - 12(\kappa_1\kappa_3 + \kappa_2^2)\frac{p^2}{n^2 d^2} \tag{2.4d}$$

$$+ \frac{24\kappa_1^2\kappa_2 p^3}{n^3 d^3} - \frac{6\kappa_1^4 p^4}{d^4 n^4}$$

The method we have used is a natural one, but there are two surprises. The first of these is that the homogeneity hypothesis is powerful enough to give such simple results. The second is that it turns out to be incredibly neater to give an expression for the conditional variances in terms of the unconditional ones rather than *viceversa*.

The idea of using a generating function approach to obtain conditional cumulant formulas is inherent in the definition of cumulants. It has been used often, and in particular the method is applied in Brillinger(1969) to obtain a general formula relating cumulant to conditional cumulants. Interestingly, that an attempt to obtain the identities (2.4a-d) directly from Brillinger's general formula is almost certainly doomed to failure because of the very much greater algebraic complexity of the expressions for the unconditional cumulants in terms of the conditional ones.

There is a second method for obtaining results relating conditional and unconditional cumulants which uses Mobius function of the lattice of partitions. Speed(1983) introduced this method and showed its attractiveness in a number of examples including an intriguing derivation

of Brillinger's formula. The Mobius function method shows great promise; but, at the present level of complexity, series inversion used here still seems to be the tool of choice.

Some comments on the formulas (2.4a-d) are in order before going on to the simulation method. In the first place, even the modest equation (2.4b) is able to provide some interesting lower bounds on the variance of certain subadditive Euclidean functionals. We will illustrate with the traveling salesman problem (TSP) in dimension 2 , but comparable results will obviously hold for the minimal spanning tree, minimal triangulation, and other functionals and other dimensions.

The only fact we require about the TSP is that if $f(X_1, X_2, \cdots, X_n)$ is the length of the shortest tour through the points X_i $1 \leqslant i \leqslant n$ then κ_1 is asymptotic to $\beta n^{\frac{1}{2}}$ as n tends to infinity (Beardwood, Halton, Hammersley (1959)). Since this f is homogeneous of order p=1, equation (2.4b) implies that κ_2 is $\Omega(\frac{1}{n})$. This modest bound is of interest because it is known that κ_2 is bounded for all n (Steele (1981)), but no better lower bound than the one just give is known. Naturally, it is expected for the TSP in d=2 that κ_2 converges to a positive constant as n goes to infinity.

6. Simplicity Though Diligence

One way in which MACSYMA can be of aid in research is simply by removing the difficulties of looking at the problem from many sides. Often one has an idea which can go unexplored because the calculations look too cumbersome. This inhibition emerges for two reasons. First, it takes conviction to go trough a tedious calculation which may not lead to a useful result, and second, the notion of usefulness is often linked to simplicity.

This block can be busted by MACSYMA in a number of cases. One from my experience concerns conditional cumulants of homogeneous statistics. The reason on can expect MACSYMA to be effective here is that the tasks required are just tedious, not ingenious. At differentiation, substitution, and rational simplification by expansion MACSYMA is as surefooted as a mountain goat.

One of the properties of mathematics (which also persists in statistics) is a strong preference for simple answers. Part of the process ot getting to understand the power of symbolic computation rest in comming to terms with complex results. What makes this novel is that one typically does not have to deal with complex results because once a calculation gets past a certain complexity threshold it is given up for lost. In a computational environment one can get extremely complex results and one faces the new challenge of using such results intelligently.

7. The Legendre Experience

This feature of the changing role of complexity is well illustrated by the enjoyable practice of computing Legendre transformations. For any function f(x) the Legendre transformation may be defined by

$$g(y) = max(xy - f(x))$$

The importance of this transformation in analysis comes in part from the relaxed inequality

$$xy \leqslant f(x) + g(y)$$

For example, the following famous transform relation is the key to Holders inequality,

$$xy \leqslant \frac{x^p}{p} + \frac{y^q}{q}$$

provided $\frac{1}{p}+\frac{1}{q} = 1$. In the same way that one can use Holder's inequality to determine the the dual space of the pth power integrable functions, one can start off with any Legendre transform pair.

A statistical problem of interest is the determination of all of those functions $h(\mu,\sigma)$ which have a finite integral when multiplied by the normal density $\phi(x;\mu,\sigma)$ and integrated with respect to μ and σ. To study this MACSYMA was used to calculate the (bivariate) Legendre transformation of ϕ with respect to μ and σ. The result takes a page to express. Still, this worm of an expression is probably the best available tool for determining which priors on μ and σ will guarantee a posterior distribution which is absolutely continuous with respect to Lebesgue measure.

For a discussion of the classical Legendre transformation (as it applies to Hamiltonian systems) as well as one version of a MACSYMA Legendre transformation one should consult Rand (1984).

8. Mixed Computations

The symbolic computational capacities of MACSYMA are augmented with numerical and graphical capabilities. This addition to the workbench permits the joint investigation of "formulas" and the exploration of their quantitative features.

One way this plays out to the investigators advantage is that it is possible to gain insight from much larger examples than would normally be possible. A useful example of this from my own experience come for the theory of random trees.

The detailed example is given in Steele(1985), but the role of MACSYMA is easy to review without those details. The generating function for the number of leaves of a random tree was derived in Renyi(1959) under the model of randomness which corresponds to the uniform distribution on the Pruffer codes (These codes provide a natural one-to-one correspondence between labeled unrooted trees and n-2 tuples of the integers $1 \leqslant k \leqslant n$.)

For the purpose of modeling, the uniform model on trees is quite limited , e.g. one can not escape the consequence of there being approximately $\frac{n}{e}$ leaves on the great majority of the trees. How much nicer for modeling purposes if we had a family of tree valued random variables whose expected number of trees could be chosen by the modeler ? There is such a family and a method for generating choices from that family given in Steele(1985).

One virtue of Renyi's distribution on random trees is that Renyi was able to prove a central limit theorem the number of leaves of a tree form his family (as n tends to infinity). To establish the corresponding result in the exponential family of random trees it appeared to be almost necessary to show that the generating functions given by Renyi had all real zeros. Naturally that opinion takes considerable motivation. Using MACSYMA it was possible to examine polynomials up to n=10 were it was discovered that for large n there were non-trivial complex roots. The plan of analysis was saved when it was found that by adding a stochasticly negligible term to Renyi's polynomial the roots could be forced back to the real line. This eventually lead to a proof of the cen-

tral limit theorem for the number of leaves of a tree from the exponential family. This is particularly satisfying because it plays back into the original motivation by making the mean number of leaves an essential parameter.

9. Concluding Remarks

This introduction to MACSYMA has been casual and discursive. The goal has been to motivate others to begin productive interaction with the tools which are now widely available. The hints which have been given are those that seem to be responsible and welcoming. There is a great deal of benefit to the statistical community which can be gained from symbolic computation and it is inevitable that current efforts will be made to look primative by those which will follow very shortly. It is especially intersting to imagine how our collective conception of what constitutes a useful answer will migrate in the next few years.

ACKNOWLEDGEMENT

This research was supported in part by NSF grant DMS-8414069. Results reproduced here were obtained in part by using MACSYMA, a large symbolic manipulation program developed at the MIT laboratory for computer Science and supported from 1975 to 1983 by NASA under grant NSG 1323, by ONR under grant N00014-77-C-0641, by DOE under grant ET-78-C-02-4687,by the USAF under F49620-79-C-020, and since 1982 by Symbolics, Inc. of Cambridge, Mass. MACSYMA is a trademark of Symbolics, Inc. Key Words: MACSYMA, symbolic computation, cumulants, Legendre transformation, central limit theorem. AMS Subject Classifications (1980): Major 68C20; Minor 62E30.

Author's Address: Program in Engineering Statistics , Engineering Quadrangle, Princeton University, Princeton N. J. 08544

REFERENCES

Beardwood, J. Halton, H. J. and Hammersley, J.M. (1959). The shortest path through many ponts. *Proc. Camb. Phil. Soc.* 55, 299-327.

Brillinger, D.R. (1969). The calculation of cumulants via conditioning. *Ann. Inst. Statist. Math.*, 21, 375-390.

Gladd, N. T. (1984) Symbolic Manipulation Techniques for Plasma Kinetic Theory Derivations *Journal of Computational Physics*, 56, 175-202.

Hochbaum, D. and Steele J. M. (1982). Steinhaus's geometric location problem for random samples in the plane, *Adv. Appl. Prob.*, 14, 56-67.

Karp, R. M. (1977). Probabalistic analysis of partitioning algorithms for the traveling salesman problem in the plane. *Math. Oper. Res.* 2 202-224.

MACSYMA Group of Symbolics, Inc.,Principal writer Joel Moses(1984), *An Introduction to MACSYMA* Symbolics, Inc. Cambridge Mass.

Mathlab Group (1983) *MACSYMA Reference Manual*, version Ten, Vol. I & II, Laboratory for Computer Science MIT, Cambridge, Mass.

Ramey, D. (1982) A Non-parametric Test of Bimodality with Application to Cluster Analysis, Department of Statistics,Yale University, New Haven Conn.

Rand, R. H. (1984) *Computer Algebra in Applied Mathematics: an Introduction to MACSYMA* Pitman Advanced Publishing Program, Boston.

Renyi, A. and Sulanke, R.(1964). Uber die konvexe Hulle von n zufallig gewahlten Punkten. II. *Z. Wahrscheinlichtkeitsteorie*, 3 138-147

Renyi, A.(1956). Some Remarks on the Theory of Trees. *Pub. Math. Inst. Hung. Acad. Sci.* 4, 73-85.

Speed, T. (1983). Cumulants and Partition Lattices. *Aust. J. Statist.* 25, 378-388

Steele, J. M. (1980). Subaditive Euclidean Functionals and nonlinear growth in geometric probability. *Ann. Probability.* 2 75-85.

Steele, J. M. (1985). Exponential Family of Trees. Technical Report, Program in Engineering Statistics and Management Science, Princeton University.