Moment-Type Estimation in the Exponential Family

By Roger R. Davidson and Daniel L. Solomon

Cornell University

BU-439-M December 1972

ABSTRACT

If $X, X_1, X_2, \ldots, X_n$ are independent real valued scalar random variables, either discrete or absolutely continuous and having common probability density function $f(\cdot|\underset{\sim}{\theta})$ where $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta \subset E^k$, then a method of moments estimator for $\underset{\sim}{\theta}$ based on $X_1, X_2, \ldots, X_n$ is a solution of the system of equations $m_j = E_{\underset{\sim}{\theta}} X^j$; $j = 1,2,\ldots,k$, where $m_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j$. The method is attributed to Karl Pearson but, although intuitively appealing, has little theoretical justification. In particular there are simple cases in which moment estimators are not even functions of a minimal sufficient statistic.

To insure that estimators be functions of a minimal sufficient statistic one might, instead of applying the principle to the raw moments, suggest setting the components of a minimal sufficient statistic equal to their expectations. This is sensible only if there is a minimal sufficient statistic of fixed dimension (independent of the sample size). In particular if $\{f(\cdot|\underset{\sim}{\theta}), \underset{\sim}{\theta} \in \Theta\}$ is a regular Koopman-Pitman-Darmois (exponential) family of distributions and Θ is the natural parameter space, so that we may write

$$f(x|\underset{\sim}{\theta}) = c(\underset{\sim}{\theta}) h(x) \, e^{\sum_{j=1}^{k} \theta_j t_j(x)}$$

then a minimal sufficient statistic for the family is

$$T = (T_1, T_2, \ldots, T_k) \equiv \left(\frac{1}{n} \sum_{i=1}^{n} t_1(X_i), \frac{1}{n} \sum_{i=1}^{n} t_2(X_i), \ldots, \frac{1}{n} \sum_{i=1}^{n} t_k(X_i) \right).$$

The estimation scheme suggested above would then seek $\underset{\sim}{\theta}$ satisfying

$$T_j = E_{\underset{\sim}{\theta}} T_j \equiv E_{\underset{\sim}{\theta}} t_j(X) , \qquad j = 1,2,\ldots,k .$$

We prove that under mild regularity conditions, a solution to this system is in fact a maximum likelihood estimator for $\underset{\sim}{\theta}$.

Although the theorem does not provide a mechanism for simplifying the calculation of maximum likelihood estimators, it does have pedagogic merit in that it strengthens the intuitive appeal of maximum likelihood estimation. Although the mathematics behind the proof is largely available in the literature, the authors have not seen the result stated in the present context.

Moment-Type Estimation in the Exponential Family

By Roger R. Davidson and Daniel L. Solomon

Cornell University

SUMMARY

Some aspects of the Pearson-Fisher controversy concerning the method of moments and the method of maximum likelihood are reviewed. In the multiparameter exponential family, a modification of the method of moments which requires the estimators to be functions of the minimal sufficient statistic is discussed. It is shown that these modified estimators are in fact the maximum likelihood estimators. Although the mathematics underlying the result is widely available in the literature, the authors have not seen it stated in the present context.

Keywords: ESTIMATION, EXPONENTIAL FAMILY, METHOD OF MAXIMUM LIKELIHOOD, METHOD OF MOMENTS

The oldest principle for estimation of the parameter(s) of a probability distribution is that embodied in the "method of moments". Suppose that $(X_1, X_2, \ldots, X_n)$ is a random sample from the distribution of a real valued (perhaps multivariate) random variable X (either discrete or absolutely continuous) having probability density function $f(\cdot|\underset{\sim}{\theta})$ where $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta \subset E^k$. Then an estimator for $\underset{\sim}{\theta}$ determined by the method of moments is a solution to the system of equations $E_{\underset{\sim}{\theta}} X^j = \Sigma X_i^j/n$, $j = 1,2,\ldots,k$.

Textbooks generally attribute the method of moments to Karl Pearson, crediting a series of papers beginning in 1894 (1894, 1898, 1902, and others), but Fisher (1937) notes that the "principle" of equating moments dates back to Bessel and Gauss and was largely developed by Thiele (1903).

Besides its intuitive appeal, the advantage of the method is that it often leads to comparatively simple calculations. Furthermore, the estimators inherit the large-sample properties of the sample moments and so under mild conditions, moment estimators are asymptotically normal with mean differing from the parameter values by terms of order n^{-1}. Unfortunately the large-sample variance (or multivariate analog) is in general not minimal so that the asymptotic efficiency is often considerably less than unity.

The significance of this weakness was hotly debated in the statistical literature, primarily by Pearson and Fisher. In his classic paper (1922)on the foundations of statistics, Fisher advocated the maximum likelihood method and demonstrated its superiority to the method of moments. The debate became vicious when a paper by Koshal (1933) offering suggestions for improvement of moment estimators by maximum likelihood techniques prompted a bitter attack by Pearson (1936) charging Koshal and Fisher with collusion. The paper was written shortly before Pearson's death and so the debate ended with Fisher's equally forceful reply (1937). The tenor of the debate can be seen in the following passage from that paper:

> Though the occasion of this paper is Pearson's attack on Koshal, it has been impossible to treat the matter in due perspective without a general criticism of methods originating with Pearson, which have been widely disseminated. The intrinsic worth of those methods has long appeared to me to have been gravely exaggerated. Pearson opens his paper with the italicized query "Wasting your time fitting curves by moments, eh?" thus expressing in his own words and style the scepticism with which he felt his procedures were being regarded by others. The question he raised seems to me not at all premature, but rather overdue.

That Fisher's response was not merely emotional is evidenced by the fact that the intensity of his feeling on the matter did not dissipate with time; for in 1950 Fisher wrote

> Pearson was an old man when it occurred to him to attack Koshal, but it would be a mistake to regard either the errors or the venom of that attack as a sign of failing powers. In both respects it is very much like what he had done repeatedly since the beginning of the century. If peevish intolerance of free opinion in others is a sign of senility, it is one which he had developed at an early age. Unscrupulous manipulation of factual material is also a striking feature of the whole corpus of Pearsonian writings, and in this matter some blame does seem to attach to Pearson's contemporaries for not exposing his arrogant pretensions.

A shortcoming of the method of moments which in our view is even more fundamental than its inefficiency is that it produces estimators which are not necessarily functions of a minimal sufficient statistic. This shortcoming is evidenced in some important special cases. For example, the moment estimator for the parameter of the Rayleigh distribution based on a random sample $(X_1,\ldots,X_n)$, $(n \geq 2)$ is a function of $\left(\Sigma X_i\right)^2$ whereas the minimal sufficient statistic is ΣX_i^2. For the two-parameter Gamma distribution $(n \geq 3)$ the moment estimators are functions of $\left(\Sigma X_i, \Sigma X_i^2\right)$ whereas the minimal sufficient statistic is $\left(\Sigma X_i, \Pi X_i\right)$.

To insure that estimators be functions of a minimal sufficient statistic, one might modify the method of moments by replacing the moment equations by equations in which the components of a minimal sufficient statistic are set equal to their expectations. This moment-type estimation procedure makes sense only if there is a minimal sufficient statistic of fixed dimension independent of the sample size; or equivalently if the family of underlying distributions is an exponential (Koopman-Pitman-Darmois) family. It is shown that this modification

of the method of moments produces estimators which are identical to the maximum likelihood estimators, thus partially reconciling the differences between Pearson and Fisher.

Specifically, the family of distributions is assumed to be of the form

$$f(x|\underset{\sim}{\theta}) = c(\underset{\sim}{\theta})h(x) \exp\left[\sum_{j=1}^{k} \theta_j t_j(x)\right] \tag{1}$$

where the set on which the density, or equivalently $h(x)$, is positive does not depend on $\underset{\sim}{\theta} = (\theta_1,\ldots,\theta_k)$. In this representation, $\underset{\sim}{\theta}$ is called the natural parameter and Θ, the space of values of $\underset{\sim}{\theta}$ for which (1) is a proper distribution is called the natural parameter space. The space Θ is convex (see Lehmann (1959), p. 51) and it is assumed that Θ contains an open set in E^k. Furthermore, it is assumed that the components of $\underset{\sim}{t} = (t_1(x),\ldots,t_k(x))$ are such that no linear combination of $t_1(x),\ldots,t_k(x)$ is constant for all x. The above two properties are nonrestrictive; for if either were not in effect then the exponent of (1) could be written in a form involving fewer than k components. One consequence of these assumptions is that each $\underset{\sim}{\theta} \in \Theta$ determines a distinct distribution in the family (1).

It is important to realize that the natural parameters are not necessarily in one-to-one correspondence with the parameters of interest. In particular, there are situations for which there are points in Θ which do not correspond to possible values of the parameters of interest. For example, the normal (τ,τ^2) distribution has a one-dimensional parameter of interest but a two-dimensional natural parameter. The natural parameter space associated with $\underset{\sim}{t} = (x,x^2)$ is the lower half-plane, $\Theta = [(-\infty,\infty) \times (-\infty,0)]$, but points $\underset{\sim}{\theta}$ in Θ which correspond to values of τ must lie on the curve $\theta_1^2 + \theta_2 = 0$. As a second example, the bivariate normal

$(0,0,1,1,\rho)$ distribution has a one-dimensional parameter of interest but a three-dimensional natural parameter. The natural parameter space associated with $\underset{\sim}{t} = (x^2, xy, y^2)$ is $\Theta = [(-\infty,0) \times (-\infty,\infty) \times (-\infty,0)]$, but points $\underset{\sim}{\theta}$ in Θ which correspond to values of ρ must have $\theta_1 = \theta_3$.

In the remainder of this study we restrict attention to situations in which the parameters of interest $\underset{\sim}{\eta}$ are in one-to-one correspondence with the natural parameters $\underset{\sim}{\theta}$. If $\underset{\sim}{X} = (X_1, \ldots, X_n)$ is a random sample from a distribution of the form (1), then a minimal sufficient statistic is

$$\underset{\sim}{T} = (T_1, \ldots, T_k) = \left(\frac{1}{n} \sum_{i=1}^{n} t_1(X_i), \ldots, \frac{1}{n} \sum_{i=1}^{n} t_k(X_i) \right) .$$

Under the moment-type estimation procedure suggested above we would estimate $\underset{\sim}{\theta}$ by $\underset{\sim}{\tilde{\theta}}$ satisfying

$$\underset{\sim}{T} = E_{\underset{\sim}{\theta}}(\underset{\sim}{T}) = E_{\underset{\sim}{\theta}}\left[\underset{\sim}{t}(X) \right] . \tag{2}$$

The parameters of interest, $\underset{\sim}{\eta} = g(\underset{\sim}{\theta})$, would then be estimated by $\underset{\sim}{\tilde{\eta}} = g(\underset{\sim}{\tilde{\theta}})$, where g is assumed to be one-to-one. The expectations $E_{\underset{\sim}{\theta}}(\underset{\sim}{t})$ are a familiar part of discussions of the exponential family. For instance, it is well known (see Lehmann (1959), p. 58) that

$$\begin{aligned} E_{\underset{\sim}{\theta}}\left[t_j(X) \right] &= - \partial \ell n \; c(\underset{\sim}{\theta}) / \partial \theta_j , & j &= 1, \ldots, k \\ \mathrm{Cov}_{\underset{\sim}{\theta}}\left[t_j(X), t_\ell(X) \right] &= - \partial^2 \ell n \; c(\underset{\sim}{\theta}) / \partial \theta_j \partial \theta_\ell , & j, \ell &= 1, \ldots, k \end{aligned} \tag{3}$$

where $c(\underset{\sim}{\theta})$ is the constant which ensures that (1) be a proper density.

Theorem. For the family of distributions (1), the moment-type estimator for $\underset{\sim}{\theta}$ obtained as a solution to the equations (2) is the maximum likelihood estimator.

Proof. The logarithm of the likelihood function is given by

$$\ell n\ L(\underset{\sim}{\theta}) = n\ \ell n\ c(\underset{\sim}{\theta}) + \sum_{j=1}^{k} \theta_j \sum_{i=1}^{n} t_j(x_i) - \sum_{i=1}^{n} \ell n\ h(x_i)$$

and the set of first and second-order derivatives are

$$\partial \ell n\ L(\underset{\sim}{\theta})/\partial\theta_j = n\ \partial \ell n\ c(\underset{\sim}{\theta})/\partial\theta_j + \sum_{i=1}^{n} t_j(x_i) \qquad j = 1,\ldots,k$$

$$\partial^2 \ell n\ L(\underset{\sim}{\theta})/\partial\theta_j\partial\theta_\ell = n\ \partial^2 \ell n\ c(\underset{\sim}{\theta})/\partial\theta_j\partial\theta_\ell\ , \qquad j,\ell = 1,\ldots,k$$

Under the assumption that no linear combination of $t_1(x),\ldots,t_k(x)$ is constant for all x, the variance-covariance matrix for $\underset{\sim}{t}$ is positive definite. It then follows from (3) that the equations

$$\partial \ell n\ L(\underset{\sim}{\theta})/\partial\theta_j = 0\ , \qquad j = 1,\ldots,k$$

admit a unique solution $\hat{\underset{\sim}{\theta}}$ which maximizes the likelihood. But from (3) it is seen that these equations are identical to (2).

The fact that for the family of distributions (1), the likelihood equations yield the unique maximum likelihood estimator for $\underset{\sim}{\theta}$, was first established by

Huzurbazar (1949). At one point in his argument, Huzurbazar suggests replacing $\underset{\sim}{T}$ by its expectation. However, this appears to have been used strictly as a technique of proof. To the best of our knowledge there has been no formal discussion of the use of equations (2) as an analogy to the method of moments.

To illustrate the result noted above, we consider the two-parameter family of Gamma distributions with probability densities

$$f(x;\alpha,\nu) = [\alpha^{\nu}/\Gamma(\nu)]e^{-\alpha x}x^{\nu-1}, \qquad x > 0,\ \alpha > 0,\ \nu > 0.$$

A natural parameterization is $\underset{\sim}{\theta} = (\alpha,\nu)$ with $\Theta = [(0,\infty) \times (0,\infty)]$ and $c(\alpha,\nu)$ $= \alpha^{\nu}/\Gamma(\nu)$, and the corresponding $\underset{\sim}{t} = (t_1(x), t_2(x)) = (-x, \ell nx)$. The moment-type estimators are obtained as a solution to equations (2), which with (3) reduce to

$$\bar{X} = E_{\underset{\sim}{\theta}}(X) = -\partial \ell n\ c(\alpha,\nu)/\partial\alpha = \nu/\alpha$$

and

$$(\overline{\ell nX}) = E_{\underset{\sim}{\theta}}(\ell nX) = -\partial \ell n\ c(\alpha,\nu)/\partial\nu = \psi(\nu)-\log(\alpha)$$

where $\psi(\nu) = d\ell n\Gamma(\nu)/d\nu$ is the digamma function. The above equations are precisely those obtained when maximizing the likelihood function (cf. Choi and Wette, (1969)). By contrast, the classical moment estimators are obtained as a solution to

$$\bar{X} = E(X) = \nu/\alpha \qquad \text{and} \qquad (\overline{X^2}) = E(X^2) = \nu(\nu+1)/\alpha^2 .$$

As observed earlier, this solution will not be a function of the minimal sufficient statistic.

When the k components of $\underset{\sim}{t} = (t_1(x), \ldots, t_k(x))$ are linearly independent homogeneous polynomials of degree at most k, the moment-type estimator for $\underset{\sim}{\theta}$ obtained as a solution to equations (2) is identical to the classical moment estimator. This result is noted in Fisher (1922, p. 356) in the special case when $k = 4$. It is clear that for families of the form (1) this is the only situation in which the moment-type estimator and the classical moment estimator of $\underset{\sim}{\theta}$ are the same.

REFERENCES

Choi, S. C. and Wette, R. (1969). Maximum likelihood estimation of the parameters of the gamma distribution and their bias. Technometrics, 11, 683-690.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Phil. Trans. Roy. Soc. London, Ser. A, 222, 309-368.

_____ (1937). Professor Karl Pearson and the method of moments. Ann. Eugenics, 7, 303-318.

_____ (1950). Contributions to Mathematical Statistics. New York: John Wiley.

Huzurbazar, V. S. (1949). On a property of distributions admitting sufficient statistics. Biometrika, 36, 71-74.

Koshal, R. S. (1933). Application of the method of maximum likelihood in the improvement of curves fitted by the method of moments. J. R. Statist. Soc. B, 96, 303-313.

Lehmann, E. L. (1959). Testing Statistical Hypotheses. New York: John Wiley.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. Phil. Trans. Roy. Soc. London, 185, 71-110.

Pearson, K. and Filon, L. N. G. (1898). Contributions to the mathematical theory of evolution, IV: On the probable errors of frequency constants and on the influence of random selection on variation and correlation. Phil. Trans. Roy. Soc. London, 191, 229-311.

Pearson, K. (1902). On the systematic fitting of curves to observations and measurements. Biometrika, 1, 265-303 and 2, 1-23.

_____ (1936). Method of moments and method of maximum likelihood. Biometrika, 28, 34-59.

Thiele, T. N. (1931). Theory of Observations. Reprint in Ann. Math. Statist., 2, 165-307, of the 1903 version.