Institute of Mathematical Statistics

LECTURE NOTES — MONOGRAPH SERIES

Separate Optimum Estimating Function for the Ruled Exponential Family

T. Yanagimoto
Institute of Statistical Mathematics
Tokyo

Y. Hiejima
Graduate University for Advanced Studies
Tokyo

ABSTRACT

A general family of distributions with mutually orthogonal parameters is introduced. The mean parameter is estimated by using the score function, and the dispersion parameter by using the projected estimating function of the score function. Both the estimating functions attain the minimum of the sensitivity criterion due to Godambe (1960, 1976).

Key Words: Negative binomial distribution, orthogonality, separate estimating function, unbiasedness

1 Introduction

For many distributions in common practice, the parameter has two types of components. One of them represents the mean, and the other the dispersion. The mean is usually estimated by using the score function, which is essentially free from the remaining component. This fact is a theoretical background of the familiar generalized linear model (GLM).

The aims of the present paper are to define a family of distributions having the above properties in a general way, and also to discuss the separate estimation of each component of the parameter.

2 Ruled exponential family

Consider first a density function of a random variable x on R^n. For simplicity we will not distinguish a random variable and a sample of size 1, unless any confusion is anticipated. Let $t(= t(x))$ be a statistic on R^s with $s < n$, and

m a point in R^s. Define a general family of probability density functions having the common support as

$$\mathcal{F}_m = \{q(x) \mid E(t \mid q(x)) = m \text{ and } E(e^{\beta t} \mid q(x)) \text{ exists for } \beta \in B\} \quad (2.1)$$

where the open set $B \subset R^s$ includes 0 and it may depend on the density function $q(x)$. Consider a subfamily of $\mathcal{F}_m$ as $\mathcal{F}_m(\Delta) = \{q(x;\delta)|\delta \in \Delta\}$ where $\Delta \subset R^r$ with $p = r + s \leq n$. Let $B(\delta)$ be the parameter space of β given δ, and define the parameter space of $\theta = (\beta, \delta)$ as $\Theta = \{(\beta, \delta)|\delta \in \Delta,\ \beta \in B(\delta)\}$. Then the ruled exponential family is defined as follows.

Definition. Let t be an s-dimensional statistic and the family $\mathcal{F}_m(\Delta)$ be a subfamily of $\mathcal{F}_m$ in (2.1). Define $\mathcal{P}(\Theta)$ with $\theta = (\beta, \delta)$ as

$$\mathcal{P}(\Theta) = \{p(x;\ \beta,\ \delta) \mid p(x;\ \beta,\ \delta) = \exp(\beta t)q(x;\delta)/\kappa(\beta,\delta),\quad \theta \in \Theta\} \quad (2.2)$$

where $\kappa(\beta, \delta)$ is the normality constant. We will call this the ruled exponential family.

The function, $\kappa(\beta, \delta_0)$ for a fixed δ_0, is the moment generating function of $q(x;\delta_0)$. For convenience we suppressed the point m in the notation of the family $\mathcal{P}(\Theta)$. This is because the point concerns the family only marginally; in many familiar examples the family does not depend on m at all. The name of the family comes from the ruled surface in geometry, where a line in the ruled surface corresponds with an exponential family $\mathcal{P}(\Theta(\delta_0))$ with $\Theta(\delta_0) = \{(\beta, \delta_0)|(\beta, \delta_0) \in \Theta\}$.

Let $\mu = E\{t \mid p(x;\ \beta,\ \delta)\}$. Then we can employ another parametrization $\theta = (\mu, \delta)$. Since this parametrization is more convenient, it will be used throughout. Another regularity condition is that the components of the parameter, μ and δ, are variable independent, that is, $\Theta = M \otimes \Delta$ where M and Δ are the parameter spaces of μ and δ. When this condition is satisfied, the family (2.2) does not depend on the choice of m.

Example 2.1. Consider the exponential family of distributions having the density function $p(x;\theta) = \exp\{\beta(\theta)t - b(\theta) + a(x)\}$. Consider also the common partition $t' = (t_1', t_2')$, $\beta(\theta) = (\beta_1(\theta),\ \beta_2(\theta))$, and $\mu(\theta) = (\mu_1(\theta), \mu_2(\theta))$ with $\mu(\theta) = E(t' \mid p(x;\theta))$. Let $\theta_1 = \mu_1(\theta)$ and $\theta_2 = \beta_2(\theta)$. Then (θ_1, θ_2) is orthogonal as in Huzurbazar (1956). An exponential family is obviously a ruled exponential family by setting $t = t_1$, $\mu = \mu_1(\theta)$ and $\delta = \beta_2(\theta)$. Consider a subfamily $\{p(x;\ (\mu, \delta^\dagger)) \mid \delta^\dagger \in \tau(\Delta)\}$ for a smooth function $\tau(\cdot) : R^r \to R^k$ with $k < r$. Then the family is the ruled exponential family, while it is not always an exponential family but a curved exponential family.

3 Some properties

The simple structure of the ruled exponential family yields properties useful for constructing estimators. Write the log-likelihood function $l(x;\theta)(= l(x;\ \mu,\ \delta))$ and its partial derivatives as $l_\mu(x;\ \mu,\ \delta)$ and $l_\delta(x;\ \mu,\ \delta)$. Recall that $\mathcal{P}(\Theta(\delta_0))$ is the exponential family with the sufficient statistic t free from δ_0. Thus the following proposition is derived from the theory of the exponential family.

Proposition 1. i) The conditional distribution of x given t is free of μ.
ii) The statistic t is complete for μ in $\mathcal{P}(\Theta(\delta_0))$.
iii) $l_\mu(x;\ \mu,\ \delta) = V^{-1}(\mu,\delta)(t-\mu)$ where $V(\mu,\delta) = \mathrm{Var}(t)$.

Theorem 3.6 (Amari 1985) states that the e-geodesic between $p(x;\ \mu,\ \delta^*)$ and $p(x;\ \mu^*,\ \delta^*)$ and the m-geodesic between $p(x;\ \mu,\ \delta)$ and $p(x;\ \mu,\ \delta^*)$ intersect orthogonally at $p(x;\ \mu,\ \delta^*)$. The subfamily $\mathcal{P}(\Theta(\delta_0))$ for every δ_0 is e-flat, and $\mathcal{P}(\Theta(\mu_0))$ for every μ_0 is a subspace of the m-flat space $\mathcal{F}_m$. Thus we obtain

Proposition 2. The components μ and δ are orthogonal, that is,

$$D((\mu,\ \delta),\ (\mu^*,\ \delta^*)) = D((\mu,\ \delta),\ (\mu,\delta^*)) + D((\mu,\ \delta^*),\ (\mu^*,\ \delta^*)) \qquad (3.1)$$

for any $(\mu,\ \delta)$ and $(\mu^*,\ \delta^*)$ where $D(\theta,\ \theta^*) = E\{\log p(x;\ \theta)/p(x;\ \theta^*) \mid p(x;\ \theta)\}$.

Two existing families in the literature are closely related to the ruled exponential family. They are the generalized power series distribution on the nonegative integers in Patil (1964) and the discrete exponential dispersion model in Jorgensen (1987). The former covers the ruled exponential family, but any study on the structure of the family is not done. It is shown that the latter is covered by the ruled exponential family.

4 Separate estimating function

We begin with discussing a 'separate estimating function' under a general condition before pursuing that in the exponential model. The regularity conditions on an estimating function $g(x;\ \theta)$, $x \in R^n$, $\theta \in \Theta \subset R^p$ in Godambe (1976) will be assumed.

Consider the common partition $g(x;\ \theta) = (g_1(x;\ \theta),\ g_2(x;\ \theta))$ and $\theta = (\theta_1,\ \theta_2)$ where the dimensions are s and r $(s + r = p)$, respectively. We call an estimating function *separate*, if $g_1(x;\ \theta)$ and $g_2(x;\ \theta)$ depend only on θ_1 and θ_2, respectively. A practical way to make an estimating function separate is that $\tilde{g}_1(x;\ \theta_1) = g(x;\ \theta_1,\ \hat{\theta}_2(\theta_1))$ where $\hat{\theta}_2(\theta_1)$ is the solution of $g_2(x;\ \theta_1,\ \theta_2) = 0$, and $\tilde{g}_2(x;\ \theta_2)$ is defined similarly. This treatment is

employed in yielding the profile likelihood estimating function. The derived separate estimating function, however, is usually biased, as emphasized in Yanagimoto and Yamamoto (1993).

Another conventional way is to discuss $g_1(x;\, \theta_1,\, \theta_{20})$and $g_2(x;\, \theta_{10},\, \theta_2)$ for a fixed $\theta_0 = (\theta_{10}, \theta_{20})$. The estimating function $g_1(x;\, \theta_1,\, \theta_{20})$ is unbiased at $\theta \in \Theta(\theta_{20}) = \{\theta \mid \theta_2 = \theta_{20},\, \theta \in \Theta\}$, but is not unbiased globally. A projection method of $g_1(x;\, \theta_1,\, \theta_{20})$ on the space of (globally) unbiased estimating functions is developed in Amari and Kawanabe (in press).

Now the score function for μ is written as $l_\mu(x;\, \mu,\, \delta) = V^{-1}(\mu, \delta)(t - \mu)$, and is essentially separate from δ. In fact it is equivalent with $t - \mu$. Lindsey (1995) called $(\mu,\, \delta)$ estimation orthogonal, when the MLE of μ is free from δ. On the other hand the score function for δ is not separate from μ. Let μ_0 be an arbitrary value. Then $l_\delta(x;\, \mu_0,\, \delta)$ is unbiased only at $\theta \in \Theta(\mu_0)$. Proposition 1 (i) and (ii) yield the optimality of $lc_\delta(x;\, \delta,\, |\, t)$.

Proposition 3. i) The estimating function $l_\mu(x;\, \mu,\, \delta_0)$ is optimum for every δ_0, attaining the Cramer-Rao bound.

ii) For every μ_0 the projection of $l_\delta(x;\, \mu_0,\, \delta)$ on the space of unbiased estimating functions is $lc_\delta(x;\, \delta \mid t)$, which is free from μ_0.

It is shown that both the estimating functions in Proposition 3 attain the minimum of the sensitivity criterion by Godambe (1960, 1976). Note that the estimating function $lc_\delta(x;\, \delta \mid t)$ does not depend on μ at all. Thus the two components are estimated in a separate way. Note also that the two estimating functions are orthogonal (Godambe 1991).

5 Examples

The following example introduces a family of possible usefulness in practice.

Example 5.1. The beta density function is written as $\Pi x_i^{p-1}(1 - x_i)^{q-1}/\mathrm{Be}(p, q)$ with the support (0,1). The family of beta density functions is an exponential family with the sufficient statistics, $\log x$ and $\log(1 - x)$. The sample mean is known to perform favorably, as an estimator of the population mean $p/(p + q)$. Let $a(0 < a < 1)$ be a constant, and set $p = a\delta$ and $q = (1 - a)\delta$. Consider the density function

$$p(x;\, \beta,\, \delta) = \Pi \frac{x_i^{a\delta-1}(1 - x_i)^{(1-a)\delta-1} e^{\beta x}}{\mathrm{Be}(a\delta,\, (1 - a)\delta) M(a\delta;\, \delta;\, \beta)}$$

where $M(\cdot;\, \cdot;\, \cdot)$ is the confluent hypergeometric function. The family of these density functions is the curved exponential family and also the ruled

exponential family. Thus the sample mean is an efficient estimator of the mean.

It is possible to extend a ruled exponential family to that having an infinite dimensional component δ, and also to that having infinite dimensional statistic t.

Example 5.2. Consider the n-dimensional point process $x(t)' = (x_1(t), \ldots, x_n(t))$, $0 < t < T$, having the intensity function

$$\prod \lambda_i(t) = \prod \tilde{\lambda}(t) \exp \delta z_i(t)$$

where $\tilde{\lambda}(t)$ is a positive intensity function and $z(t)' = (z_1(t), \ldots, z_n(t))$ be a covarite such that all the components are not identical. Write $\lambda(t) = \tilde{\lambda}(t) \Sigma \exp \delta z_i(t)$. Then the density function is expressed as

$$\prod_i [\{ \prod_{t \in I_i} \lambda_i(t) \} \exp - \int_0^T \lambda_i(s) \ ds \ ]$$

$$= [\prod_{t \in I} \lambda(t) \exp - \int_0^T \lambda(s) ds] \prod_i \{ \prod_{t \in I_i} \exp \delta z_i(t) / \sum_{i=1}^n \exp \delta z_i(t) \}$$

where $\lambda(t) = \Sigma \lambda_i(t)$, $I_i = \{t \mid x_i(t) = 1, \ 0 < t < T\}$ and $I = \cup I_i$. Set $x_T(t) = \Sigma x_i(t)$, that is, the superimpose. Then the intensity function of $x_T(t)$ is $\Sigma \lambda_i(t) = \lambda(t)$, which is free from δ. Let $\beta(t) = \log\{\lambda(t)/\lambda_0(t)\}$. By regarding the inner product $< \beta(t), x_T(t) >$ as $\sum_{i \in I} \beta(t)$, we can show that $\lambda(t)$ and δ are orthogonal in the sense of (3.1).

References

Amari, S-I. (1985). *Differential-Geometrical Methods in Statistics.* Springer, Berlin.

Amari, S-I. and Kawanabe, M. (1997). Information geometry of estimating functions in semiparametric statistical models, *Bernoulli*, **3**, 29-54.

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.*, **31**, 1208-1211.

Godambe, V.P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, **63**, 277-284.

Godambe, V.P. (1991). Orthogonality of estimating functions and nuisance parameters. *Biometrika*, **78**, 143-151.

Huzurbazar, V.S. (1956). Sufficient statistics and orthogonal parameters. *Sankhya*, **17**, 217-220.

Jorgensen, B. (1987). Exponential dispersion models (with discussion). *J. Roy. Statist. Soc., Ser. B*, **49**, 127-162.

Lindsey, J.K. (1995). *Parametric Statistical Inference*, Clarendon Press, Oxford.

Patil, G.P. (1964). Estimation of the generalized power series distribution with two parameters and its application to binomial distribution. In *Contribution to Statistics* (ed. by C.R. Rao), 335-344. Pregibon Press, Oxford.

Yanagimoto, T. and Yamamoto, E. (1991). The role of unbiasedness in estimating equations. *In: Estimating Function* (ed. by V.P. Godambe). Oxford University Press, New York, 89-101.