involve
a journal of mathematics

Linear dependency for the difference in exponential regression

Indika Sathish and Diawara Norou

mathematical sciences publishers

2010 vol. 3, no. 2

INVOLVE 3:2(2010)

Linear dependency for the difference in exponential regression

Indika Sathish and Diawara Norou

(Communicated by Kenneth S. Berenhaut)

In the field of reliability, a lot has been written on the analysis of phenomena that are related. Estimation of the difference of two population means have been mostly formulated under the no-correlation assumption. However, in many situations, there is a correlation involved. This paper addresses this issue. A sequential estimation method for linearly related lifetime distributions is presented. Estimations for the scale parameters of the exponential distribution are given under square error loss using a sequential prediction method. Optimal stopping rules are discussed using concepts of mean criteria, and numerical results are presented.

1. Introduction

In recent years, there has been a great deal of interest in looking at parameters and characterization of linearly related lifetime distributions, and more specifically of exponential types distributions. In the literature, estimation of the parameters using the sequential prediction method can be found in many areas such as statistical sciences, industrial quality control, communication science, computer simulations, genetics and many more. The sequential analysis method is carried out to determine improvements on the estimators and reduce noises related to the lifetime distributions. However, in many cases, when a pair of distributions are considered, the assumption of independence is assumed. There are contexts in which the assumption of independence is not realistic, such as in [Carpenter et al. 2006]. This paper extends the results that are proposed by including a correlation in estimating the difference parameter between two exponentially distributed functions. It is organized as follows. In Section 2, we present the basic results and the problem of interest. In Section 3, we present the sequential analysis method for the estimation of the difference of the scale parameters. Many works, such as [Mukhopadhyay and Hamdy 1984], have addressed the estimation of the difference

MSC2000: primary 60-00; secondary 60-08.
Keywords: survival, stopping rule, bivariate exponential.

of the location parameters of two distributions. Lai [2001] gives a thorough review of the sequential analysis technique along with challenges. The sections that follow are about the stopping rule technique and simulations.

2. Preliminaries and problem of interest

We consider the class of exponential family type probability distributions on the real line from [McCullagh and Nelder 1983]. The class is defined by the family of densities $\mathcal{G}$ with respect to the Lebesgue measure as

$$f(x;\theta,\varphi)=\exp\left\{\frac{\theta T(x)-b(\theta)}{a(\varphi)}+c(x,\varphi)\right\}, \tag{1}$$

where

- $f\in\mathcal{G}$,
- φ is a constant scale parameter, typically called the nuisance parameter,
- θ is a location parameter,
- $a(\varphi)$ and $c(x,\varphi)$ are specific functions of the scale parameter, and
- $b(\theta)$ and $T(x)$ are functions of the location parameter and variable x, respectively.

In fact, this exponential family density in (1) is a reformulation of the form given in [McCullagh and Nelder 1983] as they simplify $T(x)$ in (1) to simply x. Also, the expression (1) generalizes the exponential family type of distributions as described in [Terbeche et al. 2005] in the sense that

- if φ is known, (1) is the linear exponential family with canonical parameter θ;
- if φ is unknown, (1) may be used as a 2-parameter exponential family type.

As described in [McCullagh and Nelder 1983], this family includes the normal, exponential, gamma, and Poisson types of distributions. In this setting,

$$U=U(\theta)=\frac{\partial\log L(\theta,x)}{\partial\theta}=\frac{\partial f(x,\theta)/\partial\theta}{f(x,\theta)} \tag{2}$$

is the score function. Note that

- $E(U)=0$,
- $\mathrm{Var}(U)=E(U^2)=-E(\partial U/\partial\theta)=I(\theta)$, known as Fisher's information.

In the exponential family case, as in (1),

$$l(\theta,\varphi,x)=\log L(\theta,\varphi,x)=\frac{\theta T(x)-b(\theta)}{a(\varphi)}+c(x,\varphi),$$
$$U=\frac{\partial l}{\partial\theta}=\frac{T(x)-\partial b(\theta)/\partial\theta}{a(\varphi)},$$
$$E(U)=0 \implies E\big(T(x)\big)=\frac{\partial b(\theta)}{\partial\theta}=b'(\theta).$$

Based on some index set I, we now consider two classes of exponential families of random variables called $\mathbf{X}=(X_i)_{i\in I}$ and $\mathbf{Y}=(Y_i)_{i\in I}$, with densities

$$f(x_i;\theta,\varphi)=\exp\left\{\frac{\theta T(x_i)-b(\theta)}{a(\varphi)}+c(x_i,\varphi)\right\}, \tag{3}$$

$$f(y_i;\tilde{\theta},\tilde{\varphi})=\exp\left\{\frac{\tilde{\theta}T(y_i)-\tilde{b}(\tilde{\theta})}{\tilde{a}(\tilde{\varphi})}+\tilde{c}(y_i,\tilde{\varphi})\right\}. \tag{4}$$

in the classes $\mathscr{G}_X$ and $\mathscr{G}_Y$, with the linear relationship

$$Y_i=aX_i+Z_i, \tag{5}$$

where $i\in I$, a is a fixed positive constant, and the Z_i are unknown random variables whose means are of interest.

The set I is an index countable set that could be finite or infinite. The linear relation described in (5) of association of random variables is not new, but is still a challenging problem. In fact, many authors [Carpenter et al. 2006; Iyer et al. 2002; 2004] have suggested its importance in applications.

Our goal is to estimate the parameter

$$\lambda=E_{\tilde{\theta}}[T(\mathbf{Y})]-aE_{\theta}[T(\mathbf{X})]\ , \tag{6}$$

with square error loss. When $a=1$, this equation reduces to the difference between two dependent exponential family of distributions. The dependence concept is the innovation here as in many cases independence is assumed, even if it is known that there is great cost associated with that independence assumption.

3. Sequential analysis

We use the sequential estimation procedure to estimate the mean of the difference of two exponential families distributions with conjugate priors of the gamma or Bernoulli or Poisson types. This procedure helps address the problem in the small sample size case, maintaining a high power . The approach we use is Bayesian and we assume that $\pi_1(\theta)$ and $\pi_2(\tilde{\theta})$ are the conjugate priors given by

$$\pi_1(\theta)\propto\exp[t(\mu_1\theta-\varphi(\theta))],\qquad \pi_2(\tilde{\theta})\propto\exp[s(\mu_2\tilde{\theta}-\tilde{\varphi}(\tilde{\theta}))].$$

This is not a new idea; Diaconis and Ylvisaker [1979] adopted this alternative to the maximum likelihood estimation regarding the parameter θ as a random variable with prior distribution, and the inference was based on the posterior distribution. They used this setting in the exponential family with conjugate prior distribution of the parameter θ given as

$$\pi(\theta) = \frac{\exp\{t(\mu\theta - \phi(\theta))\}}{\int \exp\{t(\mu\theta - \phi(\theta))\} d\theta}, \tag{7}$$

where $\theta \in \Theta$, t can be thought as prior sample size, and μ is the mean parameter. See also [Annis 2007].

In that regard, we see that $\mu_1 = E_{\pi_1}[\varphi'(\theta)]$ and $\mu_2 = E_{\pi_2}[\tilde{\varphi}'(\tilde{\theta})]$ are prior estimators of $E_\theta[T(\mathbf{X})]$ and $E_{\tilde{\theta}}[T(\mathbf{Y})]$, respectively.

Hence, following an idea from [Terbeche et al. 2005], the Bayes estimate of λ, based on a random sample of size n of $X_1, X_2, \dots, X_n$ of $\mathbf{X}$, and $Y_1, Y_2, \dots, Y_n$ of $\mathbf{Y}$ is given by

$$\begin{aligned}\hat{\lambda} = \hat{\lambda}(\mathbf{X}, \mathbf{Y}) &= \hat{\lambda}(X_1, \dots, X_n, Y_1, \dots, Y_n) \\ &= E[\lambda | X_1, \dots, X_n, Y_1, \dots, Y_n] \\ &= E[\tilde{b}'(\tilde{\theta}) | Y_1, \dots, Y_n] - aE[b'(\theta) | X_1, \dots, X_n],\end{aligned}$$

where

$$E[b'(\theta)|X_1, \dots, X_n] = \frac{n\bar{T}_n^{\mathbf{X}} + t\mu_1}{n+t}, \quad E[\tilde{b}'(\tilde{\theta})|Y_1, \dots, Y_n] = \frac{n\bar{T}_n^{\mathbf{Y}} + s\mu_2}{n+s}, \tag{8}$$

with

$$\bar{T}_n^{\mathbf{X}} = \frac{T(X_1) + \ldots + T(X_n)}{n}, \quad \bar{T}_n^{\mathbf{Y}} = \frac{T(Y_1) + \ldots + T(Y_n)}{n}.$$

Hence,

$$\hat{\lambda} = \frac{n\bar{T}_n^{\mathbf{Y}} + s\mu_2}{n+s} - a\frac{n\bar{T}_n^{\mathbf{X}} + t\mu_1}{n+t}. \tag{9}$$

The asymptotic estimate for the parameter as $n \longrightarrow \infty$ is

$$\hat{\lambda} = \bar{T}_n^{\mathbf{Y}} - a\bar{T}_n^{\mathbf{X}}. \tag{10}$$

A criteria for stopping the estimation of λ is developed. When $t = s$,

$$\hat{\lambda} = \frac{n(\bar{T}_n^{\mathbf{Y}} - a\bar{T}_n^{\mathbf{X}}) + t(\mu_2 - a\mu_1)}{n+t} = \frac{n}{n+t}(\bar{T}_n^{\mathbf{Y}} - a\bar{T}_n^{\mathbf{X}}) + \frac{t}{n+t}(\mu_2 - a\mu_1).$$

When $t = s = n$,

$$\hat{\lambda} = \frac{(\bar{T}_n^{\mathbf{Y}} - a\bar{T}_n^{\mathbf{X}}) + (\mu_2 - a\mu_1)}{2}. \tag{11}$$

In the sequential analysis idea, the sample size is not predetermined. Hence, a natural question to ask is when is the sample size large enough to make conclusions.

4. Stopping rules

The Bayes risk of the estimate $\hat{\lambda}$ of λ with respect to the prior $\pi(\theta)$ in (7) is

$$r(\theta, \hat{\lambda}) = E[R(\theta, \hat{\lambda})],$$

where $R(\theta, \hat{\lambda}) = E[L(\theta, \hat{\lambda})]$ and $L(\theta, \hat{\lambda}) = (\lambda - \hat{\lambda})^2$ is the loss function.

In this setting, the Bayes risk is given by

$$\begin{aligned}
r(\pi_1, \pi_2) &= r(\hat{\lambda}(\mathbf{X}, \mathbf{Y})) \\
&= E_{(\mathbf{XY})}[E_{\lambda|(\mathbf{X},\mathbf{Y})}(\hat{\lambda}(\mathbf{X}, \mathbf{Y}) - \lambda)^2] \\
&= E_{(\mathbf{X},\mathbf{Y})}[\mathrm{Var}(\lambda|(\mathbf{X}, \mathbf{Y}))] \\
&= E_{(\mathbf{X},\mathbf{Y})}\big[\mathrm{Var}(\tilde{b}'(\tilde{\theta}) - ab'(\theta)|(\mathbf{X}, \mathbf{Y}))\big] \\
&= E_{(\mathbf{X},\mathbf{Y})}\big[\mathrm{Var}(\tilde{b}'(\tilde{\theta})) + a^2 \mathrm{Var}(b'(\theta)) - 2a\rho\sqrt{\mathrm{Var}(\tilde{b}'(\tilde{\theta}))}\sqrt{\mathrm{Var}(b'(\theta))}\big],
\end{aligned}$$

and the upper bound is achieved using the idea of Equation (4) in [Terbeche et al. 2005]. It is given by

$$r(\pi_1, \pi_2) = E_{\mathbf{Y}}\left[E_{\tilde{\theta}|\mathbf{Y}}\left|\frac{\tilde{b}''(\tilde{\theta})}{n+s}\right|\right] + a^2 E_{\mathbf{X}}\left[E_{\theta|\mathbf{X}}\left|\frac{b''(\theta)}{n+t}\right|\right], \tag{12}$$

with equality achieved in (12) when $\rho = \mathrm{corr}(\tilde{b}'(\tilde{\theta}), b'(\theta)) = \mathrm{corr}(\mathbf{X}, \mathbf{Y}) \geq 0$ is minimized.

Considering the loss function

$$L(\lambda, \hat{\lambda}, n) = (\lambda - \hat{\lambda})^2 + cn, \tag{13}$$

where c can be looked at as the cost of sampling, and the decision rule $\Delta = (\tau, \delta)$, where $\tau = \tau_n(\mathbf{x}, \mathbf{y})$ is the stopping rule and $\delta = \delta_n(\mathbf{x}, \mathbf{y})$ is the decision rule, we have that the Bayes risk to minimize from a suitable sample size n obtained sequentially given by

$$\begin{aligned}
r(\tau, \pi_1, \pi_2) &= E_{(\mathbf{X},\mathbf{Y},\tau)}\left[\frac{U_n}{n+t} + \frac{V_n}{n+s} - 2a\rho\sqrt{[\mathrm{Var}(\tilde{b}'(\tilde{\theta}))]}\sqrt{[\mathrm{Var}(b'(\theta))]} + cn\right] \\
&= E_{(\mathbf{Y},\tau)}\left[\frac{U_n}{n+t}\right] + E_{(\mathbf{X},\tau)}\left[\frac{V_n}{n+s}\right] \\
&\qquad + E_{(\mathbf{X},\mathbf{Y},\tau)}\big[-2a\rho\sqrt{[\mathrm{Var}(\tilde{b}'(\tilde{\theta}))]}\sqrt{[\mathrm{Var}(b'(\theta))]} + cn\big],
\end{aligned}$$

where $U_n = E_{\mathbf{Y},\tau}|\tilde{b}''(\tilde{\theta})|$ and $V_n = E_{\mathbf{X},\tau}|b''(\theta)|$.

Using ideas in [Terbeche et al. 2005] to achieve the upper bound in (12), the stopping rule criteria can be expressed as if

$$U_n \leq c(n+t)^2 \quad \text{or} \quad V_n \leq c(n+s)^2,$$

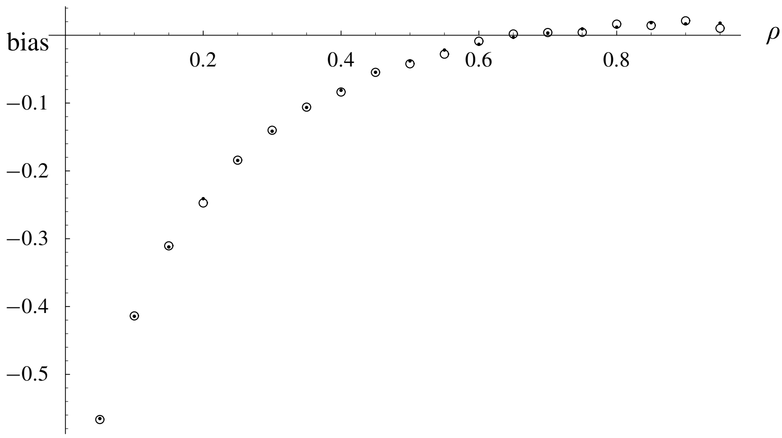

Figure 1. Graph of the bias from ρ for $c = 0$.

then take another pair of observations. Otherwise, stop the collection process. That is the estimation of the difference of the two exponential distributions can be evaluated from the available informative sample. In other words, the stopping variable is defined by the quantity

$$n \geq \min\left\{\sqrt{\frac{U_n}{c}} - t, \sqrt{\frac{V_n}{c}} - s\right\}. \tag{14}$$

In order to study the optimized stopping rule in (14) and its efficiency, a numerical simulation technique is provided in Section 5. We consider two exponentially related distributions with gamma priors.

5. Simulation

We have described a methodology to compare the mean difference between two exponential distributions that are linearly related. In this section, we show an example of a simulation data of the related bivariate exponential distribution with the different values of the correlations ρ.

Since we consider two dependent random variables, we create one exponential random variable and create the other one with the desired correlation ρ. We generate sample data of size 50. We assume a coefficient of linear relationship $a = 1$ of simultaneous occurrence as described in [Marshall and Olkin 1967], and $c = 0$ and $c = 0.25$ in (13) over 5000 runs. The simulation was carried out using SAS.

The results of the two figures show that data does not need to be large to achieve convergence. The pattern is the same regardless of the number of runs. Figures 1 and 2 give the bias of the mean difference for $c = 0$ and $c = 0.25$, respectively. The convergence is justified by the maximal error we allowed to reach based on the stopping rule, when the data generation and bias are computed at three and five

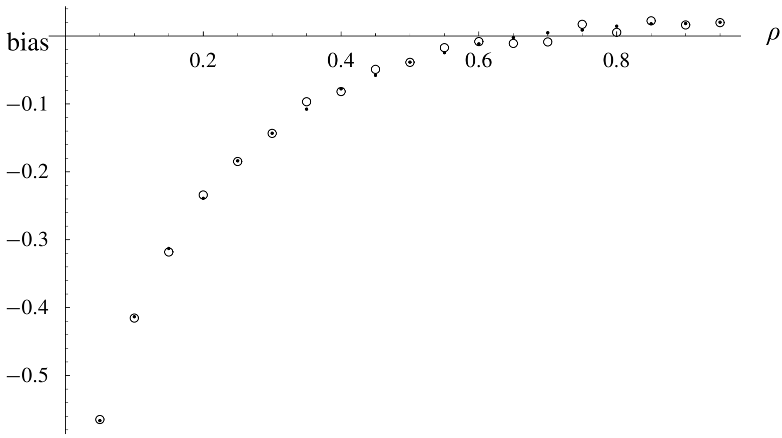

Figure 2. Graph of the bias from ρ for $c = 0.25$.

decimal places (circles and dots, respectively). The algorithm performs very well even when the sample size is small, showing great robustness.

The resulting plot of the bias is very helpful in explaining the effectiveness of the estimator. When the correlation is present, this new estimator should be considered. Furthermore, the choice of the cost of resampling c does not affect significantly in the error estimation. Setting $c = 0.25$ as in Figure 2 shows the same trend as for Figure 1. The risk is then minimized considerably when the correlation is significant.

6. Conclusion

The proposed sequential parametric procedure in the estimation of the difference of two exponential distribution is quite useful and relevant. This sequential estimation for the bivariate distributions of the exponential type families is used to get an estimate of the mean difference. It is more efficient in terms of bias.

Acknowledgements

The authors are grateful to the referees for their detailed suggestions, comments and insights, which improved the quality of the paper considerably.

References

[Annis 2007] D. H. Annis, "A note on quasi-likelihood for exponential families", *Statist. Probab. Lett.* **77**:4 (2007), 431–437. MR 2339048 Zbl 1108.62022

[Carpenter et al. 2006] M. Carpenter, N. Diawara, and Y. Han, "A new class of bivariate survival and reliability models", *Amer. J. Math. Management Sci.* **26**:1-2 (2006), 163–184. MR 2007m:62144 Zbl 1154.62337

[Diaconis and Ylvisaker 1979] P. Diaconis and D. Ylvisaker, "Conjugate priors for exponential families", *Ann. Statist.* **7**:2 (1979), 269–281. MR 80f:62016 Zbl 0405.62011

[Iyer and Manjunath 2004] S. K. Iyer and D. Manjunath, "Correlated bivariate sequences for queueing and reliability applications", *Comm. Statist. Theory Methods* **33**:2 (2004), 331–350. MR 2045319 Zbl 1066.62054

[Iyer et al. 2002] S. K. Iyer, D. Manjunath, and R. Manivasakan, "Bivariate exponential distributions using linear structures", *Sankhyā Ser. A* **64**:1 (2002), 156–166. MR 1968380

[Lai 2001] T. L. Lai, "Sequential analysis: some classical problems and new challenges", *Statist. Sinica* **11**:2 (2001), 303–408. MR 2002d:62001 Zbl 1037.62081

[Marshall and Olkin 1967] A. W. Marshall and I. Olkin, "A multivariate exponential distribution", *J. Amer. Statist. Assoc.* **62** (1967), 30–44. MR 35 #6241 Zbl 0147.38106

[McCullagh and Nelder 1983] P. McCullagh and J. A. Nelder, *Generalized linear models*, Monographs on Statistics and Applied Probability **37**, Chapman & Hall, London, 1983. MR 85k:62161 Zbl 0588.62104

[Mukhopadhyay and Hamdy 1984] N. Mukhopadhyay and H. I. Hamdy, "On estimating the difference of location parameters of two negative exponential distributions", *Canad. J. Statist.* **12**:1 (1984), 67–76. MR 86b:62139 Zbl 0543.62061

[Terbeche et al. 2005] M. Terbeche, B. O. Oluyede, and A. Barbour, "On sequential and fixed designs for estimation with comparisons and applications", *SORT* **29**:2 (2005), 217–233. MR 2208558 Zbl 05633857

Received: 2009-11-09 Revised: 2010-03-26 Accepted: 2010-04-18

sindika@odu.edu
Old Dominion University,
Department of Mathematics and Statistics,
4700 Elkhorn Avenue, Norfolk, VA 23529, United States

ndiawara@odu.edu
Old Dominion University,
Department of Mathematics and Statistics,
4700 Elkhorn Avenue, Norfolk, VA 23529, United States
http://www.odu.edu/~ndiawara

involve

pjm.math.berkeley.edu/involve

See inside back cover or http://pjm.math.berkeley.edu/involve for submission instructions.

The subscription price for 2010 is US $100/year for the electronic version, and $120/year (+$20 shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94704-3840, USA.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94720-3840 is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFLOW™ from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

http://www.mathscipub.org

A NON-PROFIT CORPORATION

Typeset in LaTeX

involve

2010 vol. 3 no. 2

1944-4176(2010)3:2;1-F