Technical Report 1231
July 1988

Smoothed Regression Estimates with Pearson Noise

Roger W. Johnson

NAVAL OCEAN SYSTEMS CENTER

San Diego, California 92152-5000

E. G. SCHWEIZER, CAPT, USN
Commander

R. M. HILLYER
Technical Director

ADMINISTRATIVE INFORMATION

This report summarizes work performed during fiscal year 1988 in the Architecture and Applied Research Branch, Naval Ocean Systems Center (NOSC), Code 421.

Released by
M. C. Mudurian, Acting Head
Architecture and Applied
Research Branch

Under authority of
J. A. Salzmann, Head
Information Systems
Division

ACKNOWLEDGMENTS

The author appreciates the support of Dr. Cooper and Dr. Gordon, who directed the NOSC independent research (IR) program and approved the research leading to this report. The IR program is funded by the Office of Naval Research. The author also thanks Vic Monteleon and John Salzmann for their line-management support of this work.

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION/AVAILABILITY OF REPORT
2b DECLASSIFICATION/DOWNGRADING SCHEDULE	Approved for public release, distribution is unlimited.
4 PERFORMING ORGANIZATION REPORT NUMBER(S) NOSC TR 1231	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION Naval Ocean Systems Center 6b OFFICE SYMBOL (If applicable)	7a NAME OF MONITORING ORGANIZATION
6c ADDRESS (City, State and ZIP Code) San Diego, CA 92152-5000	7b ADDRESS (City, State and ZIP Code)
8a NAME OF FUNDING/SPONSORING ORGANIZATION Office of Chief of Naval Research Independent Research Program 8b OFFICE SYMBOL (If applicable) OCNR-10P	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c ADDRESS (City, State and ZIP Code) Arlington, VA 22217-5000	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 61152N PROJECT NO: RR00N00 TASK NO: AGENCY ACCESSION NO: DN308 052

11 TITLE (Include Security Classification)

SMOOTHED REGRESSION ESTIMATES WITH PEARSON NOISE

12 PERSONAL AUTHOR(S)

R. W. Johnson

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Final	FROM 1987 TO 1988	July 1988	55

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Bayes estimate, Regression, Dominant estimate, Simultaneous estimation, Pearson variate, Squared error loss

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

Li and Hwang (1984) examined estimates for a regression problem which are a compromise between the naive, raw data estimate and a nonparametric estimate. They developed such compromise, or "smoothed," estimates assuming Gaussian noise. In this report, we develop smoothed regression estimates in the more general case in which Pearson noise is present. Estimates are established having good Bayes risk and ordinary risk.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a NAME OF RESPONSIBLE PERSON R. W. Johnson	22b TELEPHONE (Include Area Code) (619) 553-4005 22c OFFICE SYMBOL Code 421

DD FORM 1473, 84 JAN

83 APR EDITION MAY BE USED UNTIL EXHAUSTED
ALL OTHER EDITIONS ARE OBSOLETE

CONTENTS

1. INTRODUCTION

In this report, we examine the regression problem considered by Li and Hwang (1984). A number of important Navy problems may be cast in the form of regression problems. Villalobos and Wahba (1987), for instance, note that this is the case with the task of estimating posterior probabilities in classification problems. Whereas Li and Hwang (1984) consider the errors in their regression problem to be normally distributed, however, we will allow for a more general class of error distributions to accommodate problems in which this normality assumption is not satisfied.

Suppose that observations $y_1, y_2, \ldots, y_n$ are made at levels $x_1, x_2, \ldots, x_n$ with

$$y_j = s(x_j) + \varepsilon_j \qquad (1.1)$$

where the function s is unknown and the ε_j are independent random errors having mean zero. Using vector notation, we may write (1.1) as

$$y = \mu + \varepsilon \qquad (1.1')$$

where $y = (y_1, \ldots, y_n)^t$, $\mu = (\mu_1, \ldots, \mu_n)^t = (s(x_1), \ldots, s(x_n))^t$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t$. Note that the observed vector y is a simple estimate of the unknown vector μ.

Li and Hwang (1984) consider estimates $\hat{\mu}$ of μ of the form

$$\hat{\mu} = (1-c)y + cM_n y \qquad (1.2)$$
$$= y - c(I-M_n)y$$

where $c=c(y)$ is a real-valued function of y, M_n is a specified $n \times n$ matrix, and I is the $n \times n$ identity matrix. Such an estimate may be viewed as a compromise between the estimate y of μ and the estimate $M_n y$ of μ. If $c=0$, then $\hat{\mu}=y$. If $c=1$, then $\hat{\mu}=M_n y$. For each value of c the estimate $\hat{\mu}$ lies upon the line passing through the points y and $M_n y$ (see Fig. 1.1). If $0 \leq c \leq 1$, then $\hat{\mu}$ lies on the line segment between y and $M_n y$. The matrix M_n in the estimate $M_n y$ will usually result as a consequence of adopting some nonparametric approach to the estimation of μ. See section 3 of Li and Hwang (1984) for examples of the choice of M_n. (Errata for Li and Hwang (1984) are given in Appendix A.)

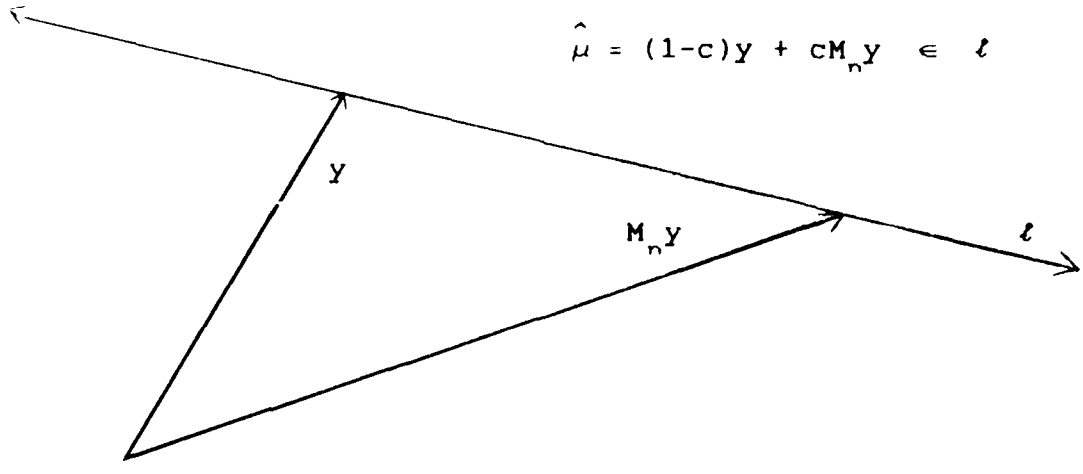

Figure 1.1. Geometry of the Li and Hwang (1984) estimate.

Relying on a result of Stein (1981), Li and Hwang (1984) exhibited good choices of c when the errors have identical normal distributions. Specifically, choices of c are given so that the estimate $\hat{\mu}$ dominates y as an estimate of μ with respect to squared error loss. (See Appendix B for a short review of standard decision theory terms.)

The main purpose of this document is to present good estimates of the form given in (1.2) which allow the errors to have distributions which are not necessarily normal.

2. PEARSON RANDOM VARIABLES

As mentioned in the introduction, Li and Hwang (1984) relied on a result of Stein (1981) to establish choices of c in (1.2) so that $\hat{\mu}$ dominates y as an estimator of μ when the errors are independent, identically distributed normals. Stein (1981) established his result with the aid of an identity which is satisfied for normal random variables. Specifically, if X is a normal random variable with mean θ and variance σ^2, then for any suitable function h

$$E\ (X-\theta)h(X) = \sigma^2\ E\ h'(X) \qquad (2.1)$$

where E denotes the expectation operator. Since an identity of this sort holds for random variables having distributions in the Pearson (1895) class (see Hudson (1978), Johnson (1984), or Haff and Johnson (1986a)), which includes the normal, we suppose that the errors in our regression problem have Pearson distributions.

Specifically, we assume that the errors are independent, with ε_j having probability density function f (w), where

$$f_j'(w) = \frac{\theta_j - w}{\beta_{j0} + \beta_{j1}w + \beta_{j2}w^2}\ f_j(w).$$

We say that ε_j has a Pearson density with parameters θ_j, β_{j0}, β_{j1}, β_{j2}, respectively. For future reference, let

$$a_j(w) = \frac{\beta_{j0} + \beta_{j1}w + \beta_{j2}w^2}{1 - 2\beta_{j2}} \qquad (2.2)$$

and

$$b_j(w) = \int^{w} \frac{1}{a_j(t)}\, dt. \qquad (2.3)$$

Note that the b_j are only specified to within some arbitrary constant of integration. Estimates of μ, yet to be presented, will involve the functions a_j and b_j.

Examples of random variables having Pearson densities are listed in Table 2.1. For these densities, the Pearson parameterization and the functions $a(\cdot)$ and $b(\cdot)$ are listed in Table 2.2.

We now state an extension of (2.1) to the Pearson family.

Theorem 2.1: Let X be a Pearson random variable with density f on the interval (r,s) and $a(\cdot)$ defined by (2.2). If $h(\cdot)$ is a differentiable function such that

$$\lim_{x \to r} a(x)h(x)f(x) = \lim_{x \to s} a(x)h(x)f(x) = 0, \qquad (2.4)$$

then

$$E\,(X-\upsilon)h(X) = E\,a(X)h'(X) \qquad (2.5)$$

where $\upsilon=(\theta+\beta_1)/(1-2\beta_2)$ provided these expectations exist.

Proof: See Hudson (1978), Johnson (1984), or Haff and Johnson (1986a) for a proof using integration by parts. ■

Table 2.1. Examples of Pearson densities.

Name Notation	Density Mean	Variance
Normal $N(\theta,\sigma^2)$	$\dfrac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-(x-\theta)^2/2\sigma^2\right], \quad -\infty < x < \infty$ $E\,X = \theta$	$\text{Var } X = \sigma^2$
Beta $B(\alpha,\beta)$	$\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 < x < 1 \quad (\alpha,\beta > 0)$ $E\,X = \dfrac{\alpha}{(\alpha+\beta)}$	$\text{Var } X = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Gamma $\Gamma(\alpha,\beta)$	$\dfrac{\beta^\alpha x^{\alpha-1} \exp(-\beta x)}{\Gamma(\alpha)}, \quad x > 0 \quad (\alpha,\beta > 0)$ $E\,X = \alpha/\beta$	$\text{Var } X = \alpha/\beta^2$
Reciprocal $I\Gamma(\alpha,\beta)$	$\dfrac{\beta^\alpha \exp(-\beta/x)}{x^{\alpha+1}\,\Gamma(\alpha)}, \quad x > 0 \quad (\alpha,\beta > 0)$ $E\,X = \dfrac{\beta}{(\alpha-1)}, \quad \alpha > 1$	$\text{Var } X = \dfrac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \quad \alpha > 2$
T $t(\alpha,\theta,\sigma^2)$	$\dfrac{\Gamma((\alpha+1)/2)}{(\alpha\pi\sigma^2)^{1/2}\,\Gamma(\alpha/2)} \left[1 + \dfrac{(x-\theta)^2}{\alpha\sigma^2}\right]^{-(\alpha+1)/2}, \quad -\infty < x < \infty \quad (\alpha > 0)$ $E\,X = \theta, \quad \alpha > 1$	$\text{Var } X = \dfrac{\alpha\sigma^2}{(\alpha-2)}, \quad \alpha > 2$

(contd)

Table 2.1. Continued.

Name Notation	Density Mean	Variance
F $F(\alpha,\beta)$	$\dfrac{\Gamma((\alpha+\beta)/2)\ \alpha^{\alpha/2}\ \beta^{\beta/2}}{\Gamma(\alpha/2)\ \Gamma(\beta/2)}\ x^{\alpha/2-1}\ (\beta+\alpha x)^{-(\alpha+\beta)/2}, \quad x > 0 \quad (\alpha,\beta > 0)$	
	$E\ X = \dfrac{\beta}{(\beta-2)}, \quad \beta > 2$	$\text{Var}\ X = \dfrac{2\beta^2(\alpha+\beta-2)}{\alpha(\beta-2)^2(\beta-4)}, \quad \beta > 4$
Power	$\theta\ k^{-\theta}\ x^{\theta-1}, \quad 0 < x < k \quad (\theta > 0)$	
	$E\ X = \dfrac{k\theta}{(\theta+1)}$	$\text{Var}\ X = \dfrac{k^2\theta}{(\theta+2)(\theta+1)^2}$
Pareto	$\theta\ k^{\theta}\ x^{-(\theta+1)}, \quad x > k > 0 \quad (\theta > 0)$	
	$E\ X = \dfrac{k\theta}{(\theta-1)}, \ \theta > 1$	$\text{Var}\ X = \dfrac{k^2\theta}{(\theta-1)^2(\theta-2)}, \quad \theta > 2$
Pearson Type IV	$\propto \exp\left[\dfrac{Q'(\theta)}{\beta_2 k}\ \arctan\left(\dfrac{Q'(x)}{k}\right)\right] Q(x)^{-(1/2\beta_2)}, \quad -\infty < x < \infty$ where $Q(x)=\beta_0+\beta_1 x+\beta_2 x^2$ and $k^2=4\beta_0\beta_2-\beta_1^2 > 0$	
	$m = E\ X = \dfrac{\theta+\beta_1}{1-2\beta_2}, \ \beta_2 < 1/2$	$\text{Var}\ X = \dfrac{Q(m)}{1-3\beta_2}, \ \beta_2 < 1/3$

Table 2.2. Some Pearson parameterizations.

Density	$(\theta, \beta_0, \beta_1, \beta_2)$	$a(w)$	$b(w)$
$N(\theta, \sigma^2)$	$(\theta, \sigma^2, 0, 0)$	σ^2	w/σ^2
$B(\alpha, \beta)$	$\left(\frac{(\alpha-1)}{(\alpha+\beta-2)}, 0, \frac{1}{(\alpha+\beta-2)}, \frac{-1}{(\alpha+\beta-2)}\right)$		
		$\frac{w(1-w)}{(\alpha+\beta)}$	$(\alpha+\beta)\ \ln\left[\frac{w}{(1-w)}\right]$
$\Gamma(\alpha, \beta)$	$((\alpha-1)/\beta, 0, 1/\beta, 0)$	w/β	$\beta \ln w$
$I\Gamma(\alpha, \beta)$	$(\beta/(\alpha+1), 0, 0, 1/(\alpha+1))$	$w^2/(\alpha-1)$	$-(\alpha-1)/w$
$t(\alpha, \theta, \sigma^2)$	$\left(\theta, \frac{\alpha\sigma^2+\theta^2}{(\alpha+1)}, \frac{-2\theta}{(\alpha+1)}, \frac{1}{(\alpha+1)}\right)$		
		$\frac{\alpha\sigma^2 + (w-\theta)^2}{(\alpha-1)}$	$\frac{(\alpha-1)}{(\alpha\sigma^2)^{1/2}} \arctan\left[\frac{(w-\theta)}{(\alpha\sigma^2)^{1/2}}\right]$
$F(\alpha, \beta)$	$\left(\frac{\beta(\alpha-2)}{\alpha(\beta+2)}, 0, \frac{2\beta}{\alpha(\beta+2)}, \frac{2}{(\beta+2)}\right)$		
		$\frac{2w}{(\beta-2)}(w+\beta/\alpha)$	$-\frac{\alpha(\beta-2)}{4} \ln(1+2/(\alpha w))$
Power	$\left(0, 0, 0, \frac{-1}{(\theta-1)}\right)$	$\frac{-w^2}{(\theta+1)}$	not useful

(contd)

Table 2.2. Continued.

Density	$(\theta,\beta_0,\beta_1,\beta_2)$	a(w)	b(w)
Pareto	$(0,0,0,\frac{1}{(\epsilon+1)})$	$\frac{w^2}{(\theta-1)}$	not useful
Pearson Type IV	$(\theta,\beta_0,\beta_1,\beta_2)$ with $k^2=4\beta_0\beta_2-\beta_1^2 > 0$	$\frac{\beta_0+\beta_1 w+\beta_2 w^2}{(1-2\beta_2)}$	$\frac{2(1-2\beta_2)}{k}\arctan\left[\frac{Q'(w)}{k}\right]$

Note that Theorem 2.1 reduces correctly in the event X is a normal random variable. *For, from* Table 2.2, $\upsilon=(\theta+\beta_1)/(1-2\beta_2)=\theta$ and $a(x)=x$. Substituting into (2.5) we obtain (2.1) provided h satisfies the conditions of Theorem 2.1.

An understanding of υ and $a(\cdot)$ in Theorem 2.1 is given by the following result:

Corollary 2.1: If Theorem 2.1 is satisfied with $h(x)=1$ and $h(x)=x$, then

$$E\ X = \upsilon = (\theta+\beta_1)/(1-2\beta_2) \quad \text{and}$$

$$\text{Var } X = E\ a(X) = (\beta_0+\beta_1\upsilon+\beta_2\upsilon^2)/(1-3\beta_2).$$

Hence, in this case, υ is the mean of X and a(X) is an unbiased estimate of the variance of X.

Proof: Set $h(X)=1$ and $(X-\upsilon)$, respectively. ■

Higher order moments of Pearson variables may be found with the aid of a recurrence formula derived from (2.5) by setting $h(x)=(x-\upsilon)^n$. The first four moments, for example, may be used to determine bounds on tail probabilities. See Appendix C for details.

Throughout this report, we use the fact that if ε_j is a Pearson random variable having mean zero, then $y_j=\mu_j+\varepsilon_j$ is a Pearson random variable having mean μ_j. This is a consequence of the following result:

Theorem 2.2: If U is a Pearson random variable with parameters m,r,s,t, then V=eU+f is a Pearson random variable with parameters em+f, $e^2r-efs+f^2t$, es-2ft, t. Furthermore

$$a_V(v) = e^2\, a_U(u) \quad \text{and}$$
$$b_V(v) = b_U(u)/e.$$

Proof: See Kaskey, Krishnaiah, Kolman, and Steinberg (1980) for the proof that V is a Pearson random variable with the given parameters. The expressions for $a_V(v)$ and $b_V(v)$ follow by direct calculation from (2.2) and (2.3). ■

3. Bayes Estimates

Our problem is to choose c so that $\hat{\mu}=y-c(I-M_n)y$ (recall (1.2)) is a good estimate of μ when y is a vector of independent Pearson random variables having mean μ. In this section we approach this problem from a Bayesian perspective. In particular, we suppose that μ has a prior distribution $\pi(\mu)$. As the optimal c is best understood in terms of the Bayes estimate, call it δ^B, of μ, we begin by stating δ^B.

Theorem 3.1: Suppose that Y is a n×1 vector of independent Pearson random variables for which $a_1(y_1),\ .\ .\ .,a_n(y_n)$, given by (2.2), are completely specified. Then the Bayes estimate of μ with respect to squared error loss, provided it exists, is given componentwise by

$$\delta_i^B(y) = y_i + a_i'(y_i) + a_i(y_i)\ \frac{d\ \ln\ f(y)}{dy_i} \qquad (3.1)$$

where

$$f(y) = \int \prod_{i=1}^{n} f_i(y_i|\mu_i)\ \ d\pi(\mu) \qquad (3.2)$$

is the marginal density of y.

Proof: See Johnson (1984), p. 31, or Haff and Johnson (1986a), p. 46. ■

Example 3.1: Suppose that Y_i, i=1,. . .,n, are independent $N(\mu_i,\sigma^2)$ random variables with σ^2 known. Then, from Table 2.2, $a_i(y_i)=\sigma^2$. Assume that μ_i, i=1,. . .,n, are independent $N(\gamma_i,\tau^2)$ random variables with the γ_i and τ^2 known. A standard calculation (see, for example, Berger (1985), pp. 127-128) shows that f(y), the marginal distribution of y, is the multivariate normal density with mean $\gamma = (\gamma_1,. . .,\gamma_n)^t$ and covariance matrix $(\sigma^2+\tau^2)I$. Substituting $a_i(y_i)$ and f(y) into (3.1) we obtain

$$\delta_i^B(y) = y_i + 0 - \sigma^2 \frac{(y_i-\gamma_i)}{(\sigma^2+\tau^2)}$$

$$= (1-r)y_i + r\gamma_i$$

where $r=\sigma^2/(\sigma^2+\tau^2)$. Note that $0 \le r \le 1$, so that the Bayes estimate of μ_i lies between γ_i and y_i. Also note in this example that δ_i^B depends on y only through y_i. In general, δ_i^B may depend upon all of the components of y.

Example 3.2: Suppose that Y_i, i=1,. . .,n, are independent $I\Gamma(\alpha_i,\beta_i)$ random variables with the α_i known. So, from Table 2.2, $a_i(y_i)=y_i^2/(\alpha_i-1)$. Also assume the improper prior

$$\pi(\beta) = \prod_{i=1}^{n} (\beta_i)^{r_i}$$

for β. Some calculation (c.f. Example 3.4 of Haff and Johnson (1986a)) reveals

$$f(y) = \prod_{i=1}^{n} \frac{\Gamma(\alpha_i + r_i)}{\Gamma(\alpha_i)} y_i^{r_i}$$

so that the formal Bayes estimate is

$$\delta_i^B(y) = \frac{(\alpha_i + r_i + 1)}{(\alpha_i - 1)} y_i$$

In Theorem 3.2, which follows, we present the Bayes estimate of μ among the class of estimates $\hat{\mu}=y-c(I-M_n)y$ with respect to squared error loss. Such a Bayes estimate may be referred to as a "restricted Bayes" estimate for we restrict ourselves to looking at estimates of a given form. In contrast, δ^B given by (3.1), may be thought of as the "unrestricted Bayes" estimate of μ.

Before stating Theorem 3.2, we provide a heuristic derivation of the restricted Bayes estimate of the form (1.2). Consider Figure 3.1. Pictured are the estimates y, $M_n y$, and δ^B (given by (3.1)) of μ. If we are going to restrict our attention to those estimates of μ which lie along the line ℓ through y and $M_n y$, then our Bayesian perspective leads us to say the best estimate of μ will be that point on ℓ nearest δ^B. So we desire $\delta^B - \hat{\mu}$ to be orthogonal to ℓ. This orthogonality implies

$$(\delta^B - \hat{\mu})'(y - M_n y) = 0$$

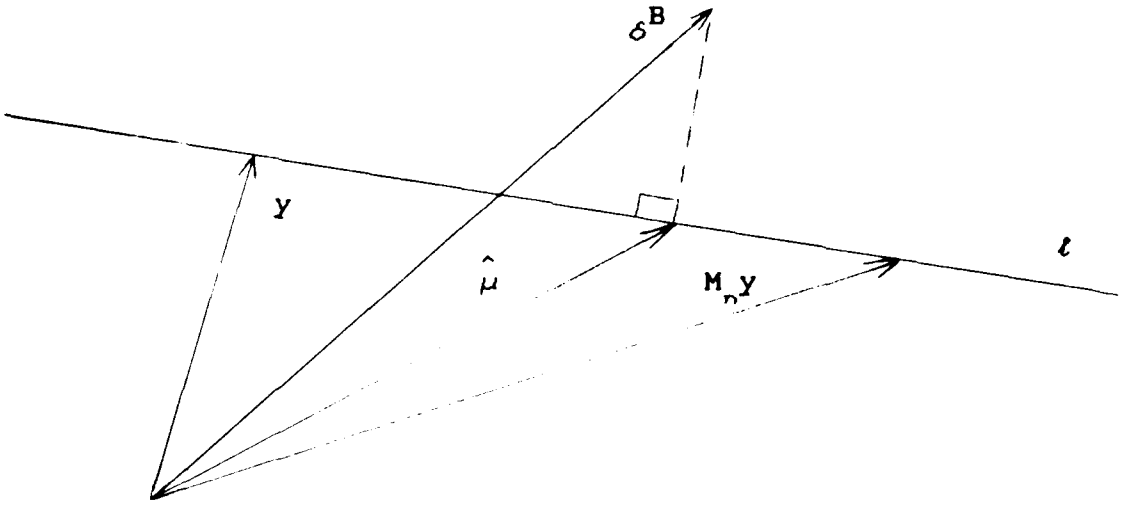

Figure 3.1. Geometry of the restricted Bayes estimate.

which, writing out $\hat{\mu}$ and simplifying, gives

$$c = c^* = \frac{(y-\delta^B)^t (I-M_n)y}{\|(I-M_n)y\|^2} \qquad (3.3)$$

where $\|\cdot\|^2$ *denotes the Euclidean norm,* $\|w\|^2 = w^t w$.

To summarize, the choice $c = c^*$ yields an estimate $\hat{\mu}$ of μ along the line connecting y and $M_n y$, which minimizes the Euclidean distance of $\hat{\mu}$ from the Bayes estimate δ^B. Sufficient conditions under which $\hat{\mu}$ with $c = c^*$ is the Bayes estimator with respect to the class of estimators $\hat{\mu}$ are now given in Theorem 3.2:

Theorem 3.2: Suppose Y is an $n\times 1$ vector of independent Pearson random variables having finite second moments. Let $\lambda(y) = (\lambda_1(y),\ldots,\lambda_n(y))^t = c^*(I-M_n)y$, where c^* is given by (3.3). If (2.4) holds componentwise for $h(y)=\lambda_i(y)$, and (3.4) holds, then the Bayes estimator of the form

$$\hat{\mu} = y - c(I-M_n)y$$

is given with $c=c^*$ in (3.3).

Proof: As the conditions of Theorem 2.5 of Haff and Johnson (1986a) are satisfied, we may apply this result to obtain

$$R(\hat{\mu},\mu) = R(Y,\mu) + E\,[-2a^t\nabla\lambda + \lambda^t\lambda\,]$$

where $a = a_n(y) = (a_1(y_1),\ldots,a_n(y_n))^t$ and $\nabla\lambda = (\partial\lambda_1/\partial y_1,\ldots,\partial\lambda_p/\partial y_p)$. Therefore

$$r(\hat{\mu}) = r(Y) + \int_\mu \left[\int_y (-2a^t\nabla\lambda + \lambda^t\lambda)\, f(y|\mu)\, dy \right] d\pi(\mu),$$

where $f(y|\mu)$ denotes the integrand in (3.2). Now, supposing

$$\int_y (|2a^t\nabla\lambda| + \lambda^t\lambda)\, f(y|\mu)\, dy < \infty \qquad (3.4)$$

we may apply Fubini's Theorem (see, for example, Rudin (1974), p. 150) to obtain

$$r(\hat{\mu}) = r(Y) + \int_y (-2a^t\nabla\lambda + \lambda^t\lambda)\, f(y)\, dy$$

where $f(y)$ is given by (3.2). Integrating by parts, we find

$$\int_y a^t\nabla\lambda\, f(y)\, dy = \int_y (y-\delta^B)^t\lambda\, f(y)\, dy\,, \qquad (3.5)$$

the boundary terms vanishing as a consequence of the assumption that (2.4) holds for each λ_i. Therefore

$$r(\hat{\mu}) = r(Y) + \int_y [-2(y-\delta^B)^t\lambda + \lambda^t\lambda]\, f(y)\, dy.$$

Up until now the computations have been performed for any λ which satisfies the necessary assumptions. Taking $\lambda = cDy$ where $D \equiv (I-M_n)$, we find

$$r(\hat{\mu}) = r(Y) + \int_y [-2c(y-\delta^B)^tDy + c^2\|Dy\|^2]\, f(y)\, dy. \qquad (3.6)$$

Denoting the dependence of $\hat{\mu}$ on c by writing $r(\hat{\mu})=r(\hat{\mu}(c))$, we find

$$r(\hat{\mu}(c+\psi)) - r(\hat{\mu}(c)) = \int_y \psi\, [2(c-c^*)+\psi]\, \|Dy\|^2\, f(y)\, dy$$

for $\psi=\psi(y)$. Consequently

$$r(\hat{\mu}(c^*+\psi)) = r(\hat{\mu}(c^*)) + \int_y \psi^2 \|Dy\|^2 f(y)\, d(y) \qquad (3.7)$$

$$\geq r(\hat{\mu}(c^*))$$

so that $c=c^*$ is the choice of c minimizing the Bayes risk. In order for $c=c^*$ to be the unique choice of c (up to an equivalence class of functions whose members are equal a.e.), we implicitly assume that $\|Dy\|^2 f(y) > 0$ a.e. with respect to Lebesgue measure over the region of integration. ■

Theorem 2.5 of Haff and Johnson (1986a), used in the proof of Theorem 3.2, is an easy extension of Theorem 2.1.

In the sequel, we will let μ^* denote the estimate $\hat{\mu}$ with $c=c^*$. Substitution of δ^B into μ^* in the previous Examples 3.1 and 3.2 is easily accomplished.

Example 3.1 (continued): In this case, we have

$$\mu^* = y - \frac{\sigma^2}{(\sigma^2+\tau^2)} \frac{(y-\gamma)^t(I-M_n)y}{\|(I-M_n)y\|^2}(I-M_n)y$$

Example 3.2 (continued): Here

$$\mu^* = y + \frac{y^tQ(I-M_n)y}{\|(I-M_n)y\|^2}(I-M_n)y$$

where Q is the diagonal matrix, $Q=\mathrm{diag}((r_1+2)/(\alpha_1-1),\ .\ .\ .,\ (r_n+2)/(\alpha_n-1))$.

Note that if $\delta^B = y - g(y)(I-M_n)y$, then $\mu^*=\delta^B$. In other words, if the Bayes estimate lies on the line ℓ (recall Figure 3.1), then the restricted Bayes estimate is equal to the Bayes estimate.

Using (3.7) of Theorem 3.2, we may state the amount of improvement in Bayes risk of the estimate μ^* over any other estimate on ℓ. As an example, we state the improvement over the estimate y of μ in the following corollary:

Corollary 3.1: The improvement in Bayes risk of the estimate μ^* over the estimate y of μ is

$$\int_y [c^*]^2 \|(I-M_n)y\|^2 f(y)\, dy.$$

Proof: Let $\psi = -c^*$ in (3.7) so that the left-hand side of this equation is the Bayes risk of y. ■

In Theorem 3.2, we presented the optimal Bayes estimate of the form $\mu=y-c(I-M_n)y$ where c is a *function*, $c=c(y)$: $\mathbb{R}^n \to \mathbb{R}^1$. In Theorem 3.3, we present the optimal Bayes estimate in the event that we restrict c to being a constant.

Theorem 3.3: Given the setting and assumptions of Theorem 3.2, the Bayes estimate of the form

$$\hat{\mu} = y - c(I-M_n)y,$$

where c is a constant, is given with

$$c = \bar{c} = \frac{E^Y(Y-\delta^B)'(I-M_n)Y}{E^Y\|(I-M_n)Y\|^2} = \frac{\int_y (y-\delta^B)'(I-M_n)y\, f(y)\, dy}{\int_y \|(I-M_n)y\|^2\, f(y)\, dy} \qquad (3.8)$$

$$= \frac{E^Y\, tr[A(I-M_n)]}{E^Y\|(I-M_n)Y\|^2} = \frac{\int_y tr[A(I-M_n)]\, f(y)\, dy}{\int_y \|(I-M_n)y\|^2\, f(y)\, dy}, \qquad (3.9)$$

where $A=diag(a_1(Y_1),\ldots,a_n(Y_n))$ is a diagonal matrix and tr denotes the trace operator.

Proof: Differentiating (3.6) with respect to c, we obtain

$$\frac{dr(\hat{\mu}(c))}{dc} = -2\int_y (y-\delta^B)'Dy\, f(y)dy + 2c\int_y \|Dy\|^2\, f(y)dy$$

where $D=(I-M_n)$. Noting that this derivative is zero for $c=\bar{c}$ as given in the integral expression of (3.8), and

$$\frac{d^2r(\hat{\mu}(c))}{dc^2} = 2\int_y \|Dy\|^2\, f(y)dy > 0,$$

we see that (3.8) holds. That (3.9) holds is a consequence of rewriting the numerator integral of (3.8) by using (3.5) with $\lambda=(I-M_n)y$. ■

Comparing (3.3) with the first expression for $\bar{c}$ in (3.8) we see that $\bar{c}$ may be viewed as an approximation to c^*. In particular, taking the expected value of the numerator of (3.3) and dividing by the expected value of the denominator of (3.3) we obtain $\bar{c}$. The expectations here are with respect to the marginal density of Y as given in (3.2).

In the sequel, we will let $\bar{\mu}$ denote the estimate $\hat{\mu}$ with $c=\bar{c}$. Note from (3.8) that $\bar{\mu}=\delta^B$ when $\delta^B=Y-c(I-M_n)Y$ for some constant c.

Example 3.1 (continued): Using (3.9) we find $\bar{c}= E^Y \text{ tr}[AD]/E^Y \|Dy\|^2$, where $D=(I-M_n)$ and $A=\text{diag}(a_1(Y_1),\ .\ .\ .,a_n(Y_n))=\sigma^2 I$. It follows that

$$\bar{c} = \frac{\sigma^2 \text{ tr } D}{(\sigma^2+\tau^2) \text{ tr } D'D + \|D\gamma\|^2},$$

the denominator expectation evaluated by using Theorem 4.6.1 on p. 139 of Graybill (1976). Finally

$$\bar{\mu} = y - \frac{\sigma^2 \text{ tr } D}{(\sigma^2+\tau^2) \text{ tr } D'D + \|D\gamma\|^2} Dy$$

where $D=(I-M_n)$.

Example 3.2 (continued): In this example $\bar{c}$ is undefined as the expectations involved do not exist. This is a result of having placed an improper prior on β.

As done in Corollary 3.1 for μ^*, we may compute the improvement in Bayes risk with $\bar{\mu}$.

Corollary 3.2: The improvement in Bayes risk of the estimate $\bar{\mu}$ over the estimate y of μ is

$$[\bar{c}]^2 \int_y \|(I-M_n)y\|^2 f(y)\, dy.$$

Proof: Substitute $\bar{c}$ for c in (3.6). ■

Because the class of estimates $\hat{\mu}=y-c(I-M_n)y$, where $c=c(y)$ is a function of y contains that in which c is a constant, μ^* will outperform $\bar{\mu}$ in terms of Bayes risk. Also, by design, both μ^* and $\bar{\mu}$ outperform the estimates Y and M_nY in terms of Bayes risk. To summarize:

$$r(\mu^*) \le r(\bar{\mu}) \le \min \left\{ r(Y),\ r(M_nY) \right\}.$$

We can, of course, look at estimates of the form (1.2) for restrictions on $c=c(y)$ other than those already chosen. So far we have taken a look at the two extremes of such restrictions. The estimate μ^* resulted in having placed

no restrictions on c, and the estimate $\bar{\mu}$ resulted in having restricted c to being a constant. In the theorem which follows, we look at estimates of the form (1.2) with $c(y)= d/\|(I-M_n)y\|^2$, where d is a constant. The resulting Bayes estimate restricted to this class of choices of (1.2) will be useful later in understanding estimates which have good (ordinary) risk.

Theorem 3.4: Given the setting and assumptions of Theorem 3.2 with $\lambda = \|(I-M_n)y\|^{-2}(I-M_n)y$, the Bayes estimate of the form

$$\hat{\mu} = y - \frac{d}{\|(I-M_n)y\|^2}(I-M_n)y$$

where d is a constant, is given with

$$d = d^* = \frac{E^Y\left[\frac{\text{tr } AD}{\|DY\|^2} - \frac{2Y^tD^tDADY}{\|DY\|^4}\right]}{E^Y\|DY\|^{-2}} \qquad (3.10)$$

$$= \frac{E^Y\left[\frac{(Y-\delta^B)^tDY}{\|DY\|^2}\right]}{E^Y\|DY\|^{-2}} \qquad (3.11)$$

where $A=\text{diag}(a_1(Y_1),\ldots,a_n(Y_n))$ is a diagonal matrix, $D=(I-M_n)$, and tr denotes the trace operator.

Proof: Similiar to the proofs of Theorems 3.2 and 3.3, and thus is omitted. ■

Comment: Note that if we remove the expectations in (3.11) the estimate $\hat{\mu}$ becomes that of Theorem 3.2.

In the special case that Y is a vector of independent $N(\mu_i,\sigma^2)$ random variables, we have $A=\sigma^2 I$ and (3.10) becomes

$$d^* = \sigma^2 \left[\text{tr } D - 2 \frac{E^Y \left[\frac{Y^t D^t DDY}{\|DY\|^4} \right]}{E^Y \|DY\|^{-2}} \right]. \qquad (3.12)$$

Since

$$\lambda_{min}(\tilde{D}) \le (DY)^t D(DY)/\|DY\|^2 \le \lambda_{max}(\tilde{D}),$$

where $\tilde{D}=(D+D^t)/2$, we may write upper and lower bounds for (3.12). Namely,

$$\sigma^2[\text{tr } D - 2\lambda_{max}(\tilde{D})] \le d^* \le \sigma^2[\text{tr } D - 2\lambda_{min}(\tilde{D})]. \qquad (3.13)$$

By Theorem 4.2 of the next section $\hat{\mu}$ dominates Y with d^* equal to the lower bound of (3.13).

Note that if D is an idempotent matrix (i.e., $D^2=D$), then (3.12) becomes

$$d^*=\sigma^2[\text{tr } D - 2] \qquad (3.14)$$

which, by Theorem 9.1.5 of Graybill (1983), p. 300, is also

$$= \sigma^2[\text{rank } D - 2]. \qquad (3.14')$$

4. DOMINANT ESTIMATES

In this section, we digress from our main discussion regarding estimates of the form of (1.2) to present two dominance results. Both of these dominance results generalize work done by Stein (1973, 1981) under the assumption of normality. The first dominance result states sufficient conditions on the marginal density (3.2) under which the Bayes estimate, δ^B, dominates the estimate Y of μ. This result was proved in Haff and Johnson (1986a). The second dominance result also looks at estimates which improve upon the estimate Y of μ. It was derived independently by Johnson (in an unpublished work), and Chou (1988). Actually, the result of Johnson is somewhat more general; compare Theorem 3.1 of Chou (1988) with Theorem 4.2 below.

Before stating the first dominance result, we present some notation to be used throughout this section. For a vector $Y=(Y_1,\ .\ .\ .,Y_n)^t$ of independent Pearson random variables, let

$$g(y) \equiv f(y) \prod_{i=1}^{n} a_i(y_i) \qquad (4.1)$$

where f(Y) is given by (3.2) and the $a_i(y_i)$ are given by (2.2). Also let

$$\nabla_B = (\partial/\partial b_1,\ \partial/\partial b_2,\ .\ .\ .,\ \partial/\partial b_n)^t,$$

and

$$B = B(Y) = (b_1(Y_1), \ldots, b_n(Y_n))^t,$$

where the $b_i = b_i(y_i)$ are given by 2.3. With this notation, we may write the Bayes estimate of μ more simply. In particular, we may rewrite (3.1) as $\delta^B = Y + \nabla_B \log g(Y)$. Finally, let

$$\nabla_B^2 h \equiv \sum_{i=1}^{n} \frac{\partial^2 h}{\partial b_i^2}.$$

Theorem 4.1: Suppose that Y is an n×1 vector of independent Pearson random variables satisfying the conditions of Theorem 3.1. Let $\lambda(y)=(\lambda_1(y),\ldots,\lambda_n(y))^t = \nabla_B \log g(y)$, where g is defined by (4.1). If (2.4) holds componentwise for $h(y)=\lambda_i(y)$ and $E_\mu^Y \lambda^t\lambda < \infty$, then

$$R(\delta^B,\mu) = R(Y,\mu) + 4\, E_\mu^Y \nabla_B^2 [g(Y)^{1/2}]/[g(Y)^{1/2}]. \qquad (4.2)$$

Consequently, when dealing with squared error loss, δ^B dominates Y as an estimate of μ if

$$\nabla_B^2 [g(Y)^{1/2}] < 0. \qquad (4.3)$$

Proof: Noting that $\delta^B = Y + \nabla_B \log g(Y) = Y + \lambda(Y)$, apply Theorem 2.5 of Haff and Johnson (1986a) to obtain

$$R(\delta^B,\mu) = R(Y,\mu) + E\,[2a^t\nabla\lambda + \lambda^t\lambda]$$

(c.f. the proof of Theorem 3.2). Rewriting the expression

in square brackets, we have

$$R(\delta^B,\mu) = R(Y,\mu) + 4\,E\sum_{i=1}^{n}\frac{\dfrac{\partial^2}{\partial b_i^2}\exp\left[1/2\int\lambda_i\,db_i\right]}{\exp\left[1/2\int\lambda_i\,db_i\right]}.$$

This reduces to (4.2) with $\lambda_i=\partial \log g/\partial b_i$. Finally, when (4.3) holds, the above expectation is negative so that $R(\delta^B,\mu) < R(Y,\mu)$. ■

We now state our second dominance result.

Theorem 4.2: Suppose Y is an $n\times 1$ vector of independent Pearson random variables having finite second moments. Let $\lambda(y)=c(y)DB(y)$, where D is a specified $n\times n$ matrix of constants and $c(y):\mathbb{R}^n\to\mathbb{R}^1$ remains to be specified. If (2.4) holds componentwise for $h(y)=\lambda_i(y)$ and $E_\mu^Y\lambda'\lambda < \infty$, then

$$\tilde{\mu} \equiv Y - c(Y)DB(Y) \quad \text{dominates} \quad Y$$

as an estimate of μ with respect to squared error loss for . . .

(i) Symmetric D when

$$c(y) = \{B'[(\text{tr } D)I - 2D]^{-1}D^2B\}^{-1}$$

and the largest eigenvalue of D, $\lambda_{max}(D)$, is less than $(\text{tr } D)/2$.

(ii) Arbitrary D when

$$c(y) = \frac{(\text{tr } D) - 2\lambda^*}{\|DB\|^2}$$

and $\lambda^* = \lambda^*(D) \equiv \lambda_{max}((D+D^t)/2)$ is less than $(\text{tr } D)/2$.

Comment: It should be noted that the dominant estimates $\tilde{\mu}$ in the above theorem are not of the form (1.2) unless $B(Y)$ is a scalar multiple of Y. This only happens when Y is a vector of normal variates. When Y is a vector of normal variates, note that $\tilde{\mu}$ with c in case (ii) is of the form $\hat{\mu}$, the restricted Bayes estimate, given in Theorem 3.4 with $D=(I-M_n)$.

Proof: Note that $c(y)=(B^tNB)^{-1}$ with a symmetric N for each of the two choices of c in the Theorem. Applying Theorem 2.5 of Haff and Johnson (1986a) and using the symmetry of N, we find

$$\begin{aligned}\Delta R &\equiv R(Y,\mu) - R(\tilde{\mu},\mu) \\ &\equiv E\,\|Y-\mu\|^2 - E\,\|(Y - c(Y)DB(Y)) - \mu\|^2 \\ &= E\left[\frac{-4B^tD^tNB}{(B^tNB)^2} + \frac{2(\text{tr } D)}{B^tNB} - \frac{\|DB\|^2}{(B^tNB)^2}\right]. \qquad (4.4)\end{aligned}$$

We desire to show $\Delta R>0$ for cases (i) and (ii) above.

Case (i): Suppose $D=D^t$. Simple algebra gives

$$\Delta R = E\left[\frac{B^t(\,2[(\mathrm{tr}\,D)I - 2D]N - D^2\,)B}{(B^tNB)^2}\right].$$

Taking $N = [(\mathrm{tr}\,D)I - 2D]^{-1}D^2/\gamma$, we obtain

$$\Delta R = \gamma(2-\gamma)\,E\,\|DB\|^2/(B^tNB)^2 > 0$$

for $0 < \gamma < 2$, with the greatest improvement in ΔR coinciding with $\gamma=1$. This completes the proof of case (i).

Case (ii): Taking $N = D^tD/\gamma$, (4.4) becomes

$$\Delta R = E\left[\frac{-4\gamma}{\|DB\|^2}\left[\frac{B^tD^tD^tDB}{\|DB\|^2}\right] + \frac{\gamma\,[2(\mathrm{tr}\,D) - \gamma]}{\|DB\|^2}\right].$$

But

$$\begin{aligned}\frac{B^tD^tD^tDB}{\|DB\|^2} &\le \max_{\|z\|=1} z^tDz \\ &= \max_{\|z\|=1} z^t((D+D^t)/2)z \\ &\equiv \lambda_{max}((D+D^t)/2) \\ &\equiv \lambda^*.\end{aligned}$$

So, assuming $\gamma>0$

$$\Delta R \ge E\left[\frac{-4\gamma}{\|DB\|^2}\lambda^* + \frac{\gamma\,[2(\mathrm{tr}\,D) - \gamma]}{\|DB\|^2}\right]$$

$$= \gamma\,(2[(\text{tr } D) - 2\lambda^*] - \gamma)\; E\,\|DB\|^{-2}$$

$$> 0$$

for $0 < \gamma < 2[(\text{tr } D)-2\lambda^*]$. The right-hand side is maximized with $\gamma=(\text{tr } D)-2\lambda^*$. This completes the proof of case *(ii)*. ■

Under the assumption of normality, case (i) of Theorem 4.2 was established by Stein (1981; p. 1142) and case (ii) was established by Li and Hwang (1984; proposition 1, p. 892).

One of the assumptions of Theorem 4.1 is that $E\,\lambda'\lambda$ be finite. We give sufficient conditions for this to be the case in the following result:

Theorem 4.3: The quantity $E\,\lambda'\lambda = E\,c(Y)\|DB\|^2$ is finite in case (i) if D is positive definite. It is finite in case (ii) if D is of full rank.

Proof: With $c(y)=(B'NB)^{-1}$

$$E\,\lambda'\lambda \;=\; E\left[\frac{\|DB\|^2}{(B'NB)^2}\right].$$

Case (i): Note that the matrices $[(\text{tr } D)I-2D]^{-1}$ and D^2 commute and are symmetric. Applying Theorem 10.6.8, p. 322, of Mirsky (1972), there exists an orthogonal matrix P such that

$$P^t[(\text{tr } D)I - 2D]^{-1}P = Q$$

and

$$P^tD^2P = R$$

where Q and R are diagonal matrices. Hence

$$E\ \lambda^t\lambda = E\ \frac{(P^tB)^tR(P^tB)}{[(P^tB)^tQR(P^tB)]^2}$$

$$\leq \frac{(\max r_i)}{(\min q_i r_i)^2}\ E\ \|B\|^{-2}\ .$$

In the previous expression, the r_i are the diagonal entries of R, and the q_i are the diagonal entries of Q. Since D is positive definite, the r_i are positive. Also, $\lambda_{max}(D) < (\text{tr } D)/2$ implies that the q_i are positive. Since $\lambda_{max}(D) < (\text{tr } D)/2$ implies $n \geq 3$ (use the fact that the trace of a matrix is equal to the sum of its eigenvalues), it suffices to show that $E\ \|B\|^{-2}$ is finite for $n \geq 3$ to complete the proof of case (i).

Case (ii): There exists a matrix P, by Theorem 10.3.4 of Mirsky (1972), such that $P^tD^tDP = R$, where R is a diagonal matrix. Consequently

$$E\ \lambda^t\lambda = [(\text{tr } D)-2\lambda^*]^2\ E\ \|DB\|^{-2}$$

$$= [(\text{tr } D)-2\lambda^*]^2\ E\ [(P^tB)^tR(P^tB)]^{-1}$$

$$= \frac{[(\text{tr } D)-2\lambda^*]^2}{(\min r_i)}\ E\ \|B\|^{-1}.$$

By Theorem 12.22 of Graybill (1983), D of full rank insures that none of the non-negative eigenvalues r_i of D^tD are zero. Again, since $\lambda^* < (\text{tr } D)/2$ implies $n \geq 3$, it suffices to show that $E \|B\|^{-2}$ is finite for $n \geq 3$ to complete the proof of case *(ii)*.

So, to prove the theorem in its entirety, it remains to show that $E \|B\|^{-2}$ is finite for $n \geq 3$. From (1.2) of Haff and Johnson (1986a) we obtain

$$E \|B\|^{-2} = \int \cdots \int \|B\|^{-2} f(y|\mu)\, dy_1 \cdots dy_n$$

where $f(y|\mu)$ is

$$\prod_{i=1}^{n} \varphi_i(\mu_i)\, a_i(y_i)^{-1} \exp(\mu_i \int a_i(y_i)^{-1} dy_i - \int y_i a_i(y_i)^{-1} dy_i).$$

Noting that $|dy_i/db_i| = a_i(y_i)$ the change of variables $x_i = b_i(y_i)$ gives

$$E \|B\|^{-2} = \int \cdots \int \|x\|^{-2} \prod_{i=1}^{n} \exp(\mu_i x_i - \psi_i(\mu_i)) k_i(x_i)\, dx_1 \cdots dx_n.$$

Since this integral is bounded over the region $\|x\|^2 > \delta$ (by $1/\delta$), it remains to show that the above integral is finite over the region $\|x\|^2 < \delta$ when $n \geq 3$. Now rewrite the integral in terms of the polar coordinates

$$x_1 = r \cos \theta_1$$
$$x_2 = r \sin \theta_1 \cos \theta_2$$

$$
\begin{aligned}
x_3 &= r \sin\theta_1 \sin\theta_2 \cos\theta_3 \\
&\vdots \\
x_{n-2} &= r \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{n-3} \cos\theta_{n-2} \\
x_{n-1} &= r \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{n-3} \sin\theta_{n-2} \cos\theta_{n-1} \\
x_n &= r \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{n-3} \sin\theta_{n-2} \sin\theta_{n-1}
\end{aligned}
$$

where $0 < \theta_i < \pi$ for i=1,2,. . .,n, and $0 < \theta_{n-1} < 2\pi$. The Jacobian, J, in this case is

$$J = r^{n-1} \prod_{i=1}^{n} (\sin\theta_i)^{n-i-1}.$$

Noting that $\|x\|^2=r^2$ the transformed integrand becomes $r^{n-1}/r^2 = r^{n-3}$ times a function bounded in the sphere $r^2 < \delta$ (the k_i are bounded in the sphere if we assume a continuous density and (2.4) holds componentwise with h(x)=1). Consequently, $E\,\|B\|^{-2}$ is finite for $n \geq 3$. This completes the proof. ■

We illustrate Theorem 4.2 with two examples. For ease of presentation, we choose D=I in each. With this selection of D, we have $c(y)=(n-2)/\|B\|^2$ in both case (i) and case (ii), giving

$$\tilde{\mu} = Y - \frac{(n-2)}{\|B\|^2} B \qquad (4.5)$$

For $\tilde{\mu}$ to dominate Y, we require $1 = \lambda_{max}(I) < (\text{tr } I)/2 = n/2$ (i.e., $n \geq 3$). As a third example, the interested reader may wish to consult section 5 of Stein (1981). Here, in the normal case, Stein considers the choice of the weight in a three-term symmetric moving average.

Example 4.1 (James and Stein (1961)): Suppose Y is an $n \times 1$ vector of independent $N(\mu_i, \sigma^2)$ random variables. From Table 2.2, $B = B(Y) = Y/\sigma^2$. Hence, for $n \geq 3$,

$$\tilde{\mu} = Y - \frac{(n-2)\sigma^2}{\|Y\|^2} Y = \left[1 - \frac{(n-2)\sigma^2}{\|Y\|^2} \right] Y$$

dominates Y as an estimate of μ with respect to squared error loss. Recalling that the components of B are determined only up to a constant (see (2.3) and the discussion which follows), we may generalize the above by taking $B = B(Y) = (Y-v)/\sigma^2$, where v is any specified $n \times 1$ vector of constants. In particular, for $n \geq 3$,

$$\tilde{\mu} = Y - \frac{(n-2)\sigma^2}{\|Y-v\|^2} (Y-v) \qquad (4.6)$$

dominates Y. From (4.6) we see that the estimate $\tilde{\mu}$ corrects the estimate Y by an amount $-[(n-2)\sigma^2/\|Y-v\|^2] \cdot (Y-v)$. For $n \geq 3$, the ith component of this correction term is negative if $Y_i > v_i$, is zero if $Y_i = v_i$, and is positive if $Y_i < v_i$. Consequently, we may view the estimate (4.6) componentwise as modifying the estimate Y_i by moving it toward (and, in

some cases, beyond) v_i. In practice, this estimate performs best when v_i is close to μ_i, i=1,. . .,n.

Example 4.2 (Johnson (1984)): Suppose Y is an n×1 vector of independent random variables whose *ith* component has a $B(\alpha_i,\beta_i)$ distribution with $s_i=\alpha_i+\beta_i$ known, but α_i and β_i unknown. From Table 2.2, we may take $b_i(y_i)$, the *ith* component of B, to be s_i (ln $[y_i/(1-y_i)]$ - ln $[v_i/(1-v_i)]$), where v_i is *any* constant, $0 < v_i < 1$. With this choice of B, (4.5) dominates Y as an estimate of $\mu=(\alpha_1/s_1,. . .,\alpha_n/s_n)$ for $n \geq 3$. As in the previous example, $\tilde{\mu}$ may be thought of as modifying the estimate Y_i by moving it toward v_i.

5. SUMMARY AND CONCLUSIONS

In this report, we have presented estimates of the mean μ of a vector Y of independent Pearson random variables.

The Pearson class of random variables, which includes several well-known variates such as the normal, was introduced in Section 2. With the notation defined in (2.2) and (2.3), Tables 2.1 and 2.2 summarized the salient features of particular Pearson variates. Throughout the report, theoretical results were illustrated by a variety of different Pearson random variables.

In Section 3, we examined estimates of μ of the form

$$\hat{\mu} = y - cDy \tag{5.1}$$

where $D=(I-M_n)y$. These estimates may be thought of as a compromise between a raw data estimate y of μ, and a nonparametric estimate $M_n y$ of μ. We determined the choice of c, a real-valued function of y, yielding the smallest Bayes risk for $\hat{\mu}$. Specifically, this choice of c was found to be

$$c = c^* = \frac{(y-\delta^B)^t Dy}{\|Dy\|^2} \tag{5.2}$$

where δ^B is the Bayes estimate of μ. This was derived by both geometric and analytic arguments. We also determined

the optimal choice of c when c was assumed to be of a particular functional form. If $c(y)=d/\|Dy\|^2$, for instance, then $d=d^*$ given by (3.10), or the equivalent (3.11), yields the best performance in terms of Bayes risk.

Unfortunately, we were unable to determine c for which $\hat{\mu}$ dominates Y in our Pearson setting except in the normal case. We hope that the two dominance results for estimates not of the form (5.1) (recall Theorem 4.1 and Theorem 4.2) will aid in finding such a c. In particular, the sufficient condition (4.3) given for δ^B to dominate Y may help establish simple conditions on δ^B in (5.2) so that $\hat{\mu}$ dominates Y. Also, perhaps, $c = c^*$ might be approximated to yield a dominant estimate $\hat{\mu}$.

BIBLIOGRAPHY

Berger, J. (1980). *Statistical Decision Theory: Foundations, Concepts, and Methods.* Springer-Verlag, New York.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis,* 2nd edition. Springer-Verlag, New York.

Bhattacharyya, B. (1987). "One sided Chebyshev inequality when the first four moments are known." *Communications in Statistics: Theory and Methods,* 16, 2789-2791.

Chernoff, L. (1981). "A note on an inequality involving the normal distribution." *The Annals of Probability,* 9, 533-535.

Chou, J. (1988). "An identity for multidimensional continuous exponential families and its applications." *Journal of Multivariate Analysis,* 24, 129-142.

Elderton, W., and Johnson, N. (1969). *Systems of Frequency distributions.* Cambridge University Press, London.

Graybill, F. (1976). *Theory and Application of the Linear Model.* Wadsworth Publishing, Belmont, California.

Graybill, F. (1983). *Matrices with Applications in Statistics.* Wadsworth Publishing, Belmont, California.

Haff, L., and Johnson, R. (1986a). "The superharmonic condition for simultaneous estimation of means in exponential families." *The Canadian Journal of Statistics,* 14, 43-54.

Haff, L., and Johnson, R. (1986b). "A note on simultaneous estimation of Pearson modes." Unpublished manuscript.

Hudson, H. (1978). "A natural identity for exponential families with applications in multiparameter estimation." *The Annals of Statistics,* 6, 473-484.

James, W., and Stein, C. (1961). "Estimation with quadratic loss." *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* 1, University of California Press, 361-380.

Johnson, R. (1984). "Simultaneous estimation of generalized Pearson means." PhD dissertation, Department of Math, University of California, San Diego.

Johnson, R. (1987). "Simultaneous estimation of binomial N's." *Sankhyā: The Indian Journal of Statistics*, 49, 264-267.

Kaskey, G., Krishnaiah, P., Kolman, B. and Steinberg, L. (1980). "Transformations to normality." *Handbook of Statistics*, 1, 321-341. North Holland.

Klaassen, C. (1985). "On an inequality of Chernoff." *The Annals of Probability*, 13, 966-974.

Li, K. (1986). "Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing." *The Annals of Statistics*, 14, 1101-1112.

Li, K., and Hwang, J. (1984). "The data-smoothing aspect of Stein estimates." *The Annals of Statistics*, 12, 887-897.

Maritz, J. (1970). *Empirical Bayes Methods*. Methuen, London.

Mirsky, L. (1972). *An Introduction to Linear Algebra*. Oxford University Press, London.

Morris, C. (1983). "Natural exponential families with quadratic variance functions: Statistical theory." *The Annals of Statistics*, 11, 515-529.

Ord, J. (1967). "On a system of discrete distributions." *Biometrika*, 54, 649-656.

Ord, J. (1968). "The discrete Student's t distribution." *The Annals of Mathematical Statistics*, 39, 1513-1516.

Pearson, K. (1895). "Memoir on a skew variation in homogeneous material." *Philosophical Transactions of the Royal Society*, A, 186, 343-414.

Rudin, W. (1974). *Real and Complex Analysis*, 2nd edition. McGraw-Hill, Inc., New York.

Shinozaki, N. (1984). "Simultaneous estimation of location parameters under quadratic loss." *The Annals of Statistics*, 12, 322-335.

Stein, C. (1973). "Estimation of the mean of a multivariate normal distribution." *Proceedings of the Prague Symposium on Asymptotic Statistics*, 345-381.

Stein, C. (1981). "Estimation of the mean of a multivariate normal distribution." *The Annals of Statistics*, 9, 1135-1151.

Stone, C. (1977). "Consistent nonparametric regression." *The Annals of Statistics*, 5, 595-645.

Villalobos, M., and Wahba, G. (1987). "Inequality-constrained multivariate smoothing splines with application to the estimation of posterior probabilities." *Journal of the American Statistical Association*, 82, 239-248.

Appendix A

ERRATA FOR LI AND HWANG (1984)

In this appendix, we list some minor errors in Li and Hwang (1984). The substance of their results are unaffected by these corrections.

Under their Theorem 1 (all references to theorems and equations in this appendix refer to Li and Hwang (1984)), the right-hand side of the equality (2.9) should read

$$(1 + o_p(1))n^{-1}\|M_n y - f_n\|^2 + o_p(En^{-1}\|M_n y - f_n\|^2).$$

On the second line following (2.13)

$$(1+x)^{-1} \leq 1-x \text{ for } x > 0$$

should be replaced by

$$(1+x)^{-1} \geq 1-x \text{ for } x > 0.$$

In the line following (2.14), replace

$$+ 2(2n^{-1} + 3(n^{-1}\text{tr } M_n)^{1/2})n^{-1}\|A_n y\|^2$$

with

$$+ 2(2n^{-1} + 3(n^{-1}\text{tr } M_n^2)^{1/2})n^{-1}\|A_n y\|^2.$$

The line below (2.20), we read "Finally (2.17) follows from (2.16),(2.6) and (2.20)." We also need the fourth moment of the ε_i to exist here.

On the right-hand side of the equality (2.21), replace

$$o_p(n^{-1}\text{tr } M_n^2 + n^{-1}\|A_n f_n\|^2 + n^{-1}\|M_n y - f_n\|)^2$$

with

$$o_p(n^{-1}\text{tr } M_n^2 + n^{-1}\|A_n f_n\|^2 + n^{-1}\|M_n y - f_n\|^2).$$

Four lines below (2.25), the inequality

$$\leq (n^{-1}\text{tr } M_n)^2 + mn^{-2}\text{tr } M_n^2$$

is not necessarily true. It suffices to have

$$\leq (n^{-1}\text{tr } M_n)^2 + 2mn^{-2}\text{tr } M_n^2$$

instead.

Two lines above (2.26), replace

$$o(En^{-1}\|M_n y - f_n\|^2) = o_p(n^{-1}\|M_n y - f_n\|^2)$$

with

$$o_p(En^{-1}\|M_n y - f_n\|^2) = o_p(n^{-1}\|M_n y - f_n\|^2).$$

On the second line from the end of the proof of Theorem 1, on page 891, replace

$$= 2(m + 2)\lambda(M_n^2)(\text{tr } M_n^2)^{-1}(E\|M_n y - f_n\|^2)^2$$

with

$$= 2(m + 2)\lambda(M_n^2)(\text{tr } M_n^2)^{-1}(En^{-1}\|M_n y - f_n\|^2)^2.$$

Appendix B

DECISION THEORY TERMINOLOGY

In this appendix, we review some standard terminology used in decision theory.

Let $Y = (Y_1, \ldots, Y_n)^t$ be a vector of independent random variables with the ith coordinate having a density of $f(y_i|\mu_i)$. We understand $f(y_i|\mu_i)$ to denote a family of densities indexed by the parameter μ_i. Also, let $EY = (EY_1, \ldots, EY_n)^t = \mu$. To estimate μ by $\phi = \phi(Y) = (\phi_1(Y), \ldots, \phi_n(Y))^t$ we will use the *squared error loss* function L where

$$L(\phi,\mu) \equiv (\phi-\mu)^t(\phi-\mu)$$

$$= \sum_{i=1}^{n} (\phi_i(Y) - \mu_i)^2.$$

The *expected loss* or *risk*, $R(\phi,\mu)$, incurred in estimating μ by ϕ is then given by

$$R(\phi,\mu) \equiv E_\mu^X L(\phi,\mu)$$

$$= \int_{\mathbb{R}^n} L(\phi,\mu) f(y|\mu)\, dy$$

where $f(y|\mu) \equiv \prod_{i=1}^{n} f(y_i|\mu_i)$ and $dy \equiv \prod_{i=1}^{n} dy_i$.

The superscripts of the expectation symbol E denote the random variables with respect to which we are taking the expectation. The subscripts of E denote fixed parameters. Such superscripts and subscripts are suppressed when these are clear from the context.

We will say that $\phi_1(Y)$ *dominates* $\phi_2(Y)$ *in estimating* μ *with respect to squared error loss* provided

$$R(\phi_1,\mu) \leq R(\phi_2,\mu)$$

for all μ, with strict inequality for some μ. The phrase *with respect to squared error loss*, since it is understood, will generally be suppressed (we use parentheses to enclose such phrases in what follows) Loss functions other than squared error loss, of course, could be used.

In the event there does not exist an estimator $\phi = \phi(Y)$ which dominates a particular estimator $\phi^* = \phi^*(Y)$, we call ϕ^* an *admissible* estimator. An estimator which is not admissible is *inadmissible*.

One basis of comparison between two estimators $\phi_1 = \phi_1(Y)$ and $\phi_2 = \phi_2(Y)$ can be made by examining how large their risks may become as we vary μ. In particular, we may prefer ϕ_1 to ϕ_2 if

$$\sup_{\mu} R(\phi_1,\mu) < \sup_{\mu} R(\phi_2,\mu)$$

and call an estimator *minimax* if it minimizes this supremum. That is, ϕ^* is minimax if

$$\sup_{\mu} R(\phi^*,\mu) = \inf_{\phi} \sup_{\mu} R(\phi,\mu).$$

If we have prior information about μ in the form of a probability distribution $\pi(\mu)$ for μ, then estimators may be compared on the basis of their *Bayes risk*. The Bayes risk, $r = r(\phi) = r(\phi,\pi)$, of an estimator ϕ is given by a weighted average of the risk. In particular

$$r(\phi) = r(\phi,\pi) \equiv \int_{\mathbb{R}^n} R(\phi,\mu)\, d\pi(\mu).$$

Note that the case of the letter r distinguishes whether we are dealing with the (ordinary) risk or Bayes risk. We say $\phi^* = \phi^*(Y)$ is a *Bayes estimate of* μ *(with respect to the prior distribution* π*)* if

$$r(\phi^*,\pi) = \min_{\phi} r(\phi,\pi). \qquad \text{(B.1)}$$

In the above discussion on Bayes estimates, we assume that $\pi(\mu)$ is a probability distribution. That is, we assume $\int d\pi(\mu) = 1$. Yet, even when $\int d\pi(\mu) = \infty$ we may still find a solution to (B.1). The prior in this case is called an *improper* prior and the resulting estimate in called a *formal Bayes* estimate.

We will, at times, restrict our attention to a particular class, $\mathcal{C}$, of estimates, ϕ, over which we will take the above minimum. In this event, ϕ^* is a *Bayes estimate with respect to the class* $\mathcal{C}$ *(and prior distribution* π*)*. Such a Bayes estimate may be spoken of as a *restricted* Bayes estimate.

Appendix C

A ONE-SIDED CHEBYSHEV INEQUALITY WHEN THE FIRST FOUR MOMENTS ARE KNOWN

Below, we recall a Theorem of Bhattacharyya (1987) which gives a bound for the tail probability of a random variable whose first four moments are known.

Theorem C.1 (Bhattacharyya (1987)): Let X be a random variable with mean μ, and let σ^2, μ_3, μ_4 be the second, third, and fourth central moments, respectively. Also let $s=\mu_3/\sigma^3$ and $k=\mu_4/\sigma^4$. For every non-negative t satisfying $t^2-st-1>0$

$$P(X-\mu \geq t\sigma) \leq \frac{k-s^2-1}{(k-s^2-1)(1+t^2) + (t^2-st-1)^2} . \quad (C.1)$$

For a Pearson random variable with parameters θ, β_0, β_1, and β_2, we have

$$\mu = EX = (\theta+\beta_0)/(1-2\beta_2)$$

$$\sigma^2 = E(X-\mu)^2 = Q(\mu)/(1-3\beta_2)$$

$$\mu_3 \equiv E(X-\mu)^3 = 2Q'(\mu)Q(\mu)/[(1-3\beta_2)(1-4\beta_2)]$$

$$\mu_4 \equiv E(X-\mu)^4 = 3Q(\mu)[2Q'(\mu)^2+(1-4\beta_2)Q(\mu)]/[(1-3\beta_2)(1-4\beta_2)(1-5\beta_2)]$$

where $Q(\mu) = \beta_0 + \beta_1\mu + \beta_2\mu^2$, provided these moments exist and Theorem 2.1 holds for $h(x)=x^n$, n=0,1,2,3.

Example C.1: If X is $N(\theta,\sigma^2)$, then we find, by using Table 2.2 and the above moment relations that $\mu_3=0$ and $\mu_4=3\sigma^4$. Hence, inequality (C.1) holds for $t > 1$, with $s=0$ and $k=3$.

Example C.2: If X is $\Gamma(\alpha,\beta)$, then we find, by using Table 2.2 and the above moment relations, that $\mu=\alpha/\beta$, $\sigma^2=\alpha/\beta^2$, $\mu_3=2\alpha/\beta^3$, and $\mu_4=3\alpha(\alpha+2)/\beta^4$. Consequently, the inequality (C.1) holds for $t > [1 + (\alpha+1)^{1/2}]/\alpha^{1/2}$, with $s=2/\alpha^{1/2}$ and $k=3+6/\alpha$.

Appendix D

BOUNDS FOR THE VARIANCE OF A FUNCTION OF A PEARSON RANDOM VARIABLE

Klaassen (1985) presents upper and lower bounds for the variance of a function, G, of an arbitrary random variable. For continuous random variables, the bounds involve derivatives of G, while for discrete random variables, the bounds involve differences of G. Klaassen's result generalizes the result established by Chernoff (1981) in the case where the random variable is normally distributed.

In this appendix, we apply the work of Klaassen to the Pearson class of densities.

Theorem D.1: Let X be a Pearson random variable on (r,s) with finite variance σ^2 satisfying (2.4) with h(x)=1. Then

$$[E\ a(X)g(X)]^2/\sigma^2 \leq \text{Var } G(X) \leq E\ [a(X)g(X)^2] \qquad (D.1)$$

where g(X)=G'(X).

Proof: Apply Theorems 2.1 and 3.1 of Klaassen (1985), with μ=Lebesgue measure, $\chi(x,y)= 1_{(b,x)}(y) - 1_{(x,b)}(y)$, where b= $EX=(\theta+\beta_1)/(1-2\beta_2)$, $h(x)=(1-2\beta_2)$, and $H(x)=(1-2\beta_2)x-(\theta+\beta_1)$. ■

Example D.1: If X is normal with mean μ and variance σ^2, then $a(x)=\sigma^2$ (see Table 2.2), and (D.1) becomes

$$\sigma^2 \, E\,[g(X)]^2 \;\le\; \text{Var } G(X) \;\le\; \sigma^2 \, E\,[g(X)^2].$$

Example D.2: If X is a beta variate, $B(\alpha,\beta)$, then $\sigma^2 = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$ (see Table 2.1) and $a(x)=x(1-x)/(\alpha+\beta)$ (see Table 2.2) so that (D.1) becomes

$$\frac{(\alpha+\beta+1)}{\alpha\beta} E[X(1-X)g(X)]^2 \le \text{Var } G(X) \le \frac{1}{(\alpha+\beta)} E\,[X(1-X)g(X)^2].$$

We now apply Klaassen's result to discrete Pearson random variables. These random variables are defined on p. 83 of Johnson (1984). Some examples appear in Table D.1.

Theorem D.2: Let X be a discrete Pearson random variable on $\{N_0, \ldots, N_1\}$ with finite variance σ^2. Then

$$[E\, d(X)g(X)]^2/\sigma^2 \;\le\; \text{Var } G(X) \;\le\; E\,[d(X)g(X)^2]$$

where

$$d(x) = a(x) - (x-\mu),$$

$$a(x) = (\beta_0 + \beta_1 x + \beta_2 x^2)/(1-2\beta_2),$$

$$\mu = EX = (\theta+\beta_1-1)/(1-2\beta_2),$$

$$g(x) \equiv G(x+1) - G(x)$$

provided

$$\lim_{x \to N_i} a(x)f(x) = 0,$$

for $i=0$ when $N_0 = -\infty$ and for $i=1$ when $N_1 = \infty$.

Proof: Apply Theorems 2.1 and 3.1 of Klaassen (1985) with μ =counting measure, $\chi(x,y) = 1_{(([\nu],x))}(y) - 1_{((x,[\nu]))}(y) - (\nu-[\nu])\, 1_{((([\nu]))}(y)$ where $[\nu]$ denotes the integer part of ν and $\nu=EX=(\theta+\beta_1-1)/(1-2\beta_2)$, $h(x)=(1-2\beta_2)$, and $k(x)=(1-2\beta_2)x-(\theta+\beta_1-1)$. ■

Table D.1. Some discrete Pearson random variables.

Name	Probability Distribution	$(\theta,\beta_0,\beta_1,\beta_2)$
Poisson	$\dfrac{e^{-\lambda}\lambda^y}{y!}$, $y=0,1,2,\ldots$	$(\lambda,0,1,0)$
Binomial	$\binom{n}{y} p^y q^{n-y}$, $y=0,1,\ldots,n$	$(p(n+1),0,q,0)$ where $q=(1-p)$
Negative Binomial	$\binom{r+y-1}{y} p^y q^r$, $y=\ldots,-1,0,1,\ldots$	$((r-1)p/q,0,1/q,0)$ where $q=(1-p)$
Discrete t (Ord (1968))	$\propto \prod_{i=0}^{k} [(y+a+i)^2 + b^2]^{-1}$, $y=\ldots,-1,0,1,\ldots$ $0 \le a \le 1$, $0 < b^2 < \infty$ k a non-negative integer	$((1-k-2a)/2, [(a+k)^2+b^2]/2(k+1), [2(a+k)+1]/2(k+1), 1/2(k+1))$