ABSTRACT

Title of dissertation: ESTIMATION THEORY OF A LOCATION PARAMETER IN SMALL SAMPLES

Tinghui Yu
Doctor of Philosophy, 2008

Dissertation directed by: Professor Abram Kagan
Department of Mathematics, Statistics Program

The topic of this thesis is estimation of a location parameter in small samples. Chapter 1 is an overview of the general theory of statistical estimates of parameters, with a special attention on the Fisher information, Pitman estimator and their polynomial versions. The new results are in Chapters 2 and 3 where the following inequality is proved for the variance of the Pitman estimator t_n from a sample of size n from a population $F(x-\theta)$: $n\mathrm{Var}(t_n) \geq (n+1)\mathrm{Var}(t_{n+1})$ for any $n \geq 1$, only under the condition $\int x^2 dF(x) < \infty$ (even the absolute continuity of F is not assumed). The result is much stronger than the known $\mathrm{Var}(t_n) \geq \mathrm{Var}(t_{n+1})$. Among other new results are (i) superadditivity of $1/\mathrm{Var}(t_n)$ with respect to the sample size: $1/\mathrm{Var}(t_{m+n}) \geq 1/\mathrm{Var}(t_m) + 1/\mathrm{Var}(t_n)$, proved as a corollary of a more general result; (ii) superadditivity of $\mathrm{Var}(t_n)$ for a fixed n with respect to additive perturbations; (iii) monotonicity of $\mathrm{Var}(t_n)$ with respect to the scale parameter of an additive perturbation when the latter belongs to the class of self-decomposable random variables. The technically most difficult result is an inequality for $\mathrm{Var}(t_n)$, which is a

stronger version of the classical Stam inequality for the Fisher information. As a corollary, an interesting property of the conditional expectation of the sample mean given the residuals is discovered. Some analytical problems arising in connection with the Pitman estimators are studied. Among them, a new version of the Cauchy type functional equation is solved. All results are extended to the case of polynomial Pitman estimators and to the case of multivariate parameters. In Chapter 4 we collect some open problems related to the theory of location parameters.

ESTIMATION THEORY OF A LOCATION PARAMETER IN SMALL SAMPLES

by

Tinghui Yu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2008

Advisory Committee:
Professor Abram Kagan, Chair/ Advisor
Professor Leonid Koralov
Professor Eric Slud
Professor Paul Smith
Professor Prakash Narayan/ Dean's Representative

Foreword

In this thesis, I worked on a series of interesting mathematical problems arising from a bunch of classical, in other words well studied, topics of statistics and information theory. Some of the solutions to those problems may lead to immediate applications such as the estimate of information contained in a sensoring network. Some of them can be considered of mere methodological interest so far. I hope to find a position in the "real world" for these theories in my future research.

On the other hand, some of the problems discussed in this thesis are far from closed. For example, in Chapter 3 I studied the characteristic function of those multivariate distributions admitting a linear Pitman estimator for a univariate location parameter. It is not clear how large a family is associated with these characteristic functions of a very special form. It seems difficult but very likely to work out some nontrivial solutions on these problems. I am glad that this thesis is opening another gate to me besides leaving me a title of PhD degree.

After finishing the editorial work, I think for a moment and decided not to dedicate this thesis to anybody because in this world there are too many people I love so deeply that I can not afford to dedicate anything of mine to only a few of them. My parents are always in the first position to receive my honor and gratitude. Next is my advisor, Professor Kagan, who has led me into this field and given me almost every piece of motivations throughout the work of this thesis. I am not a smart student. I always need supports on my back either in studies or researches. To me, Professor Kagan is the best source of knowledge and strength. Without

the help of all these professors including Grace Yang, Paul Smith and Eric Slud, it would have been impossible for me to survive these years in the University of Maryland and finally finish successfully. Moreover, during the writing of this thesis, I owe much, in material or spirit, to Yu-Ru Huang, Yabing Mai, Linbao Zhang, Bo Li, Weiran Sun, Denise Sam, Ning Jiang and Ziliang Li. There are too many other names I should mention here to express my gratitude for their friendship and love. I will keep their names in my mind.

Tinghui Yu

College Park, Maryland

April, 13 2008

Table of Contents

Chapter 1

Introduction

The topic of the thesis is estimation of a location parameter in small samples. Though a location parameter is rather special, many results proved for location parameter families can be extended to the linear regression setups that are of practical importance. Moreover, due to its special character, the theory of estimation of a location parameter in small samples can be developed much further than that of a general parameter. Besides having a mathematical statistical interest, the results obtained for a location parameter can show the direction of research for estimating a general parameter, both in small samples and asymptotically.

1.1 Basic concepts of statistical estimation of parameters

To make the presentation self-contained, here basic concepts of parameter estimation theory are briefly discussed. For details see Bahadur's notes (2002) and Lehmann (1983).

Let the data be represented as a random element $\mathbf{X}$ with values in a measurable space $(\mathfrak{X}, \mathscr{A})$, where $\mathscr{A}$ is a σ-algebra. The distribution P_θ of $\mathbf{X}$ is assumed to belong to a family $\mathscr{P} = \{P_\theta, \theta \in \Theta\}$ parameterized by an abstract valued parameter. In parametric (as opposed to nonparametric) problems, Θ is usually a subset of $\mathbb{R}^s$. The triple $\{\mathfrak{X}, \mathscr{A}, \mathscr{P}\}$ is called a statistical model.

The goal is to use the data $\mathbf{X}$ to get information on the unknown value of θ. For example, if $\gamma : \Theta \mapsto \mathbb{R}$ is a parametric function, what is the value of $\gamma(\theta)$?

A statistic $T(\mathbf{X})$ used as an "approximation" for $\gamma(\theta)$ is called an *estimator* of $\gamma(\theta)$. Certainly, there are "good" and "poor" estimators. The former are distinguished from the latter by a *loss function* $L : T(\mathfrak{X}) \times \gamma(\Theta) \mapsto [0, +\infty)$ that measures the loss incurred from the approximation of $\gamma(\theta)$ by $T(\mathbf{X})$. The expectation of the loss is called the *risk* of $T(\mathbf{X})$ as an estimator of $\gamma(\theta)$:

$$R_L(T(\mathbf{X}), \gamma(\theta)) = E_\theta[L(T(\mathbf{X}), \gamma(\theta))] = \int_{\mathfrak{X}} L(T(\mathbf{x}), \gamma(\theta)) dP_\theta(\mathbf{x})^1.$$

An estimator T_1 of $\gamma(\theta)$ is better than T_2 with respect to the loss L if

$$R_L(T_1(\mathbf{X}), \gamma(\theta)) \leq R_L(T_2(\mathbf{X}), \gamma(\theta)), \quad \forall \theta \in \Theta. \tag{1.1}$$

An estimator T_1 is *optimal* in class $\mathscr{T}$ of estimators of $\gamma(\theta)$ if (1.1) holds for any $T_2 \in \mathscr{T}$. An estimator $T_2(\mathbf{X})$ of $\gamma(\theta)$ is called *admissible* if there is no $T_1(\mathbf{X})$ such that (1.1) holds with a strict inequality for at least one $\theta \in \Theta$. Admissibility per se does not justify using an estimator. For example, the constant estimator $T(\mathbf{X}) = \gamma(\theta_0)$ that completely ignores the data can be admissible with respect to some loss functions. Optimality in a natural class plus admissibility is what statisticians are looking for.

Thus, the goal is not admissibility per se but admissibility of an estimator $T(\mathbf{X})$ of $\gamma(\theta)$ possessing some desirable properties, for example *unbiasedness*:

$$E_\theta T(\mathbf{X}) = \gamma(\theta), \quad \forall \theta \in \Theta.$$

[1] If no confusion in the notation should arise, from now on the integration region $\mathfrak{X}$ will be omitted without special note.

In this thesis, we consider the case when $\mathfrak{X} = (\mathbb{R}^m)^n$, $\mathscr{A}$ is the σ-algebra of Borel sets in $\mathbb{R}^{mn}$, $P_\theta(A) = \int_A dF_\theta(x_1)\dots dF_\theta(x_n)$, $\forall A \in \mathscr{A}$, where F_θ is a distribution on $\mathbb{R}^m$. In other words, the measure P_θ is generated by a sample $\mathbf{X} = (X_1, \dots, X_n)$ from an m-variate population F_θ.

1.2 Fisher information, Sufficiency and Cramér-Rao inequality

Assume Θ is an open set. If P_θ, $\theta \in \Theta \subset \mathbb{R}$, are absolutely continuous with respect to a measure μ on $(\mathfrak{X}, \mathscr{A})$, with densities

$$\frac{dP_\theta}{d\mu} = p(x;\theta),$$

such that the *Fisher score*

$$J(X;\theta) = \frac{\partial \ln p(X;\theta)}{\partial \theta}$$

is well defined with

$$I_X(\theta) = \mathrm{Var}(J(X;\theta)) = \int_{-\infty}^{+\infty} |J(x;\theta)|^2 p(x;\theta) d\mu < +\infty,$$

then $I_X(\theta)$ is called the *Fisher information* on θ contained in X. Under mild regularity type conditions, (see e.g. Kagan, Linnik and Rao (1973)), it has the following properties justifying the name "information":

(i) *Additivity.* If for all θ, X_1, X_2 are independent random elements taking values in $(\mathfrak{X}_1, \mathscr{A}_1)$ and $(\mathfrak{X}_2, \mathscr{A}_2)$ respectively with densities $p_1(x_1, \theta)$, $p_2(x_2, \theta)$ and

$$\mathbf{X} = (X_1, X_2) \in (\mathfrak{X}, \mathscr{A}) = (\mathfrak{X}_1 \times \mathfrak{X}_2, \mathscr{A}_1 \otimes \mathscr{A}_2)$$

with density $p(\mathbf{x};\theta) = p_1(x_1;\theta)p_2(x_2;\theta)$, then

$$I_{\mathbf{X}}(\theta) = I_{X_1}(\theta) + I_{X_2}(\theta). \tag{1.2}$$

In particular, if $\mathbf{X} = (X_1, \ldots, X_n)$ is a sample from a population with density $p(x;\theta)$, then $I_{\mathbf{X}}(\theta) = nI_{X_1}(\theta)$, where $I_{X_1}(\theta)$ is, as the notation indicates, the Fisher information on θ contained in a single observation X_1.

(ii) *Monotonicity.* Let $T : (\mathfrak{X}, \mathscr{A}) \mapsto (\mathfrak{T}, \mathscr{B})$ be a statistic with density $q(t;\theta)$ with respect to a measure ν on $(\mathfrak{T}, \mathscr{B})$. Note that if $P_\theta << \mu$ then the distributions Q_θ of T are absolutely continuous with respect to ν, where $\nu(B) = \mu(T^{-1}B)$, $\forall B \in \mathscr{B}$. In this case, the Fisher information in T never exceeds that in X

$$I_T(\theta) \leq I_X(\theta), \quad \theta \in \Theta.$$

A statistic S is called *sufficient* for a family $\mathscr{P} = \{P_\theta, \theta \in \Theta\}$ (or briefly for θ when it is clear what family $\mathscr{P}$ is under consideration), if for any bounded measurable function φ there exists a version $\tilde{\varphi}$ of the conditional expectation $E_\theta(\varphi|S)$ such that $\tilde{\varphi}$ is a statistic (i.e., does not involve θ):

$$\tilde{\varphi} = E_\theta(\varphi|S). \tag{1.3}$$

If the family $\mathscr{P}$ is dominated, (1.3) is equivalent to the factorization

$$p(\mathbf{x};\theta) = R(S(\mathbf{x});\theta)h(\mathbf{x}). \tag{1.4}$$

This is a classical theorem due to Halmos and Savage (1949).

If S is sufficient, it preserves the Fisher information in the data:

$$I_S(\theta) = I_{\mathbf{X}}(\theta), \quad \theta \in \Theta. \tag{1.5}$$

If $p(\mathbf{x};\theta) > 0$ for all $x \in \mathfrak{X}$, $\theta \in \Theta$ the converse holds: if a statistic T preserves the Fisher information, T is sufficient (for $\mathscr{P}$). Without positivity of $p(\mathbf{x};\theta)$ this is not

true, as shown in Kagan and Shepp (2005). They constructed some distributions with density $p(\mathbf{x};\theta)$ vanishing at a single point $x = x(\theta)$. Under this setting one can find an insufficient statistic T preserving the Fisher information in $\mathbf{X}$.

(iii) *Reparametrization formula.*

If $g : \Theta \mapsto \Theta'$ for some open Θ' is a differentiable into mapping, then for $\eta = g(\theta)$, the Fisher information (on η) $I_X(\eta)$ is well defined and satisfies

$$I_X(\theta) = |g'(\theta)|^2 I_X(\eta)|_{\eta=g(\theta)}. \tag{1.6}$$

(iv) *Cramér-Rao inequality.*

If $T(\mathbf{X})$ is an unbiased estimator of $\gamma(\theta)$, a differentiable function of the parameter. Then under mild regularity conditions, such as $\frac{\partial}{\partial\theta}\int T(x)p(x;\theta)dx = \int T(x)\frac{\partial}{\partial\theta}p(x;\theta)dx$, the Cauchy-Schwarz inequality implies that

$$\sqrt{\mathrm{Var}(J(\mathbf{X};\theta))\mathrm{Var}(T(\mathbf{X}))} \geq \mathrm{Cov}(J,T) = -\gamma'(\theta).$$

Rearranging the terms, one gets the Cramér-Rao lower bound:

$$\mathrm{Var}(T) \geq \frac{[\gamma'(\theta)]^2}{\mathrm{Var}(J)} = \frac{[\gamma'(\theta)]^2}{I_X(\theta)}. \tag{1.7}$$

An unbiased estimator $T(\mathbf{X})$ of $\gamma(\theta)$ for which (1.7) becomes an equality for all $\theta \in \Theta$ is called an *efficient estimator* or $\gamma(\theta)$. For a sample $\mathbf{X} = (X_1, \ldots, X_n)$ of size n from a population with density $p(x;\theta)$ and information $I_{\mathbf{X}}(\theta)$, the Cramér-Rao inequality by virtue of (i) takes the form

$$\mathrm{Var}(T(\mathbf{X})) \geq \frac{(\gamma'(\theta))^2}{nI_{X_1}(\theta)}.$$

The multivariate versions of the above properties are straightforward.

1.3 Estimation in large samples

Large sample theory deals with asymptotic properties of estimators as n goes to infinity. Let $X_1, X_2, \ldots, X_n, \ldots$ be independent identically distributed random variables with distribution function $F_\theta(x)$, $\theta \in \Theta \subset \mathbb{R}$ and let $\tilde{\theta}_n = \tilde{\theta}_n(X_1, \ldots, X_n)$ be an estimator of θ based on $X_1, \ldots, X_n$, $n = 1, 2, \ldots$. A sequence of estimators $\{\tilde{\theta}_n, n = 1, 2, \ldots\}$ of θ is called *consistent* if

$$\tilde{\theta}_n \to \theta, \tag{1.8}$$

in P_θ-probability as $n \to \infty$, and is called *strongly consistent* if the convergence in (1.8) is with P_θ-probability one. Consistency is the minimal requirement demanded from an estimator in large samples. Of more interest for statistical inference is the limiting distribution of properly normalized estimators.

If the likelihood of X_i is given by a density $p(x;\theta)$ with $\theta \in \Theta$ an open set in $\mathbb{R}^s$, a maximum likelihood estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is defined as

$$\arg\max_{\theta\in\Theta} \prod_{i=1}^{n} p(X_i, \theta),$$

and is usually obtained as a solution of the equation

$$\sum_{i=1}^{n} \frac{\partial \ln p(X_i, \theta)}{\partial \theta} = 0, \theta \in \Theta.$$

Under regularity type conditions (for conditions close to the minimal ones, see Ibragimov and Has'minskii (1981)),

$$\sqrt{n}(\hat{\theta}_n - \theta) \longrightarrow_d N(0, 1/I_{X_1}(\theta)). \tag{1.9}$$

In view of (1.9), an asymptotic estimator $\tilde{\theta}_n$, $n = 1, 2, \ldots$ with the property

$$\sqrt{n}(\tilde{\theta}_n - \theta) \longrightarrow_d N(0, 1/I_{X_1}(\theta))$$

is called *asymptotically efficient*. If $\sqrt{n}(\tilde{\theta}_n - \theta) \longrightarrow_d N(0, \sigma^2(\theta))$, the ratio

$$\text{aseff}_\theta\{\tilde{\theta}_n\} = \frac{1/I_{X_1}(\theta)}{\sigma^2(\theta)}$$

is called the *asymptotic efficiency* of $\{\tilde{\theta}_n, n = 1, 2, \ldots\}$. In "regular" cases, $\text{aseff}_\theta\{\tilde{\theta}_n\}$ is bounded above by 1.

The convergence in (1.9) does not guarantee the existence of $\text{Var}(\hat{\theta}_n)$. Ibragimov and Has'minskii showed that under some additional conditions (beyond those guaranteeing asymptotic normality)

$$E_\theta(\hat{\theta}_n - \theta)^2 = \frac{1}{nI(\theta)}(1 + o(1)), \tag{1.10}$$

and

$$\text{Var}_\theta(\hat{\theta}_n) = \frac{1}{nI(\theta)}(1 + o(1)). \tag{1.11}$$

Certainly, (1.10) and (1.11) imply

$$\text{Bias}_\theta(\hat{\theta}_n) = E_\theta(\hat{\theta}_n - \theta) = o(1/\sqrt{n}),$$

whence the MLE is known as an *asymptotically unbiased* estimator.

1.4 Estimation in small samples

Estimation theory in small samples concentrates mainly on the structure of the uniformly minimum variance unbiased estimators (known by their abbreviation

UMVUE). If $S : (\mathfrak{X}, \mathscr{A}) \mapsto (\mathfrak{T}, \mathscr{B})$ is sufficient for a family $\mathscr{P} = \{P_\theta, \theta \in \Theta\}$, the classical Rao-Blackwell theorem says that any estimator $T(\mathbf{X})$ (with finite second moment) of a parameter function $\gamma(\theta)$ can be improved by an estimator $\tilde{T}$ depending on $\mathbf{X}$ only through S:

$$\tilde{T}(S) = E_\theta(T|S). \tag{1.12}$$

Sufficiency guarantees that $\tilde{T}$ is still a statistic. The operation (1.12), known as Rao-Blackwellization, plainly preserves unbiasedness, that is,

$$E_\theta(T) = \gamma(\theta) \Longrightarrow E_\theta(\tilde{T}) = \gamma(\theta), \text{ and } \mathrm{Var}_\theta(T) \geq \mathrm{Var}_\theta(\tilde{T}), \theta \in \Theta.$$

Rao-Blackwellization describes the structure of all UMVUEs for families possessing complete sufficient statistics.

A sufficient statistic T is called *complete* if $E_\theta[h(T)] = 0$ implies $h(T) \equiv 0$ with P_θ-probability one for all $\theta \in \Theta$. If a family $\mathscr{P}$ possesses a complete sufficient statistic, it is easy to see that a complete sufficient statistic is unique as a statistic. In other words, two complete sufficient statistics generate the same σ-algebra. That is, if T_1, T_2 are two complete sufficient statistics for $\mathscr{P}$ then there exists a one to one measurable function g such that

$$P_\theta[T_1 = g(T_2)] = 1, \text{ for almost all } \theta \in \Theta.$$

In this case, the class of UMVUEs coincides with the class of statistics with finite second moments that are functions of the complete sufficient statistic, that is, measurable with respect to the σ-algebra generated by T.

Bahadur (1955) showed that if a family $\mathscr{P} = \{P_\theta, \theta \in \Theta\}$ is such that every parameter function $\gamma(\theta)$ that possesses an unbiased estimator with finite variance

also possesses a UMVUE, then $\mathscr{P}$ has a complete sufficient statistic. For a general family $\mathscr{P}$, some parameter functions possess UMVUEs while the others do not. For a class of families with *incomplete* sufficient statistics, the structure of the UMVUEs is known in a few cases (Kagan and Konikov (2005)).

1.5 Location parameter families

A family $\{F(x;\theta), \theta \in \mathbb{R}\}$ of distributions on $\mathbb{R}$ depends on a scalar valued location parameter if

$$F(x;\theta) = F(x-\theta).$$

The main feature here is that the distribution function (or the density $F' = p$ when it exists) is not a function of two arguments x and θ as in the case of a general univariate parameter, but of a single argument $x - \theta$. This fact allows to develop the theory of estimation of θ much farther than in the general case. Many results obtained for the univariate location parameter can be extended to the case of multivariate location parameter families

$$\{F(\mathbf{x}-\boldsymbol{\theta}), \boldsymbol{\theta} \in \mathbb{R}^s\},$$

where $\mathbf{x} \in \mathbb{R}^s$ and $\boldsymbol{\theta} \in \mathbb{R}^s$.

Also of interest is the case of a multivariate $\mathbf{X}$ and univariate θ when the family under consideration is

$$F(\mathbf{x}-\theta) = F(\mathbf{x}-\theta\cdot\mathbf{1}) = F(x_1-\theta,\ldots,x_n-\theta),$$

where $\mathbf{1}$ is the multivariate column vector of ones.

1.6 Equivariance and the Pitman estimators

The concept of equivariance is due to Pitman (1938). Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sample from population $F(x - \theta)$. A statistic $T(\mathbf{X})$ is called *equivariant* if

$$T(\mathbf{X} + c) = T(\mathbf{X}) + c, \ \forall c \in \mathbb{R}. \tag{1.13}$$

Using an equivariant estimator of θ means that the estimator does not depend on the choice of the origin.

The class of equivariant estimators is rather large. Examples are the sample mean $\bar{X}$, any convex combination of the order statistics, the MLE and even any single observation X_i, $i = 1, 2, \ldots$. Plainly, a statistic $T(\mathbf{X})$ is equivariant if and only if it can be written as

$$T(\mathbf{X}) = T_0(\mathbf{X}) + g(R_n) \tag{1.14}$$

for some Borel function $g : \mathbb{R}^n \mapsto \mathbb{R}$, where $T_0(\mathbf{X})$ is an arbitrarily chosen equivariant estimator and $R_n = (r_1, \ldots, r_n) = (X_1 - \bar{X}, \ldots, X_n - \bar{X})$ is the vector of *residuals*.

Notice that the residuals $r_k = X_k - \bar{X}$, $k = 1, \ldots, n$ are linearly dependent so that $R_n \in \mathbb{R}^n$ is actually a vector of $(n - 1)$ functionally independent components. There are different (but equivalent) forms of representing the residuals, for example $\{X_1 - X_n, X_2 - X_n, \ldots, X_{n-1} - X_n\}$. They all lead to the same σ-algebra.

One may wish to find the (uniformly) best estimator among the class of equivariant estimators. It certainly depends on the choice of the loss function. Equivariance is a property that partitions the sample space $\mathbb{R}^n$ into equivalence classes such that within each class the data points are not distinguishable in estimating θ

up to constant shifts. It seems natural to consider those loss functions satisfying $L(T(\mathbf{X}+c);\theta+c) = L(T(\mathbf{X});\theta)$ for any constant c. Such a loss function must be in the form

$$L(T(\mathbf{X});\theta) = L(T(\mathbf{X})-\theta). \tag{1.15}$$

In this case, the risk of an equivariant estimator is a constant in θ:

$$\begin{aligned} R_T(\theta) &= \int\cdots\int_{\mathbb{R}^n} L(T(\mathbf{x})-\theta)dF(x_1-\theta)\cdots dF(x_n-\theta) \\ &= \int\cdots\int_{\mathbb{R}^n} L[T(\mathbf{x}-\theta)]dF(x_1-\theta)dF(x_n-\theta) \\ &= \int\cdots\int_{\mathbb{R}^n} L[T(\mathbf{u})]dF(u_1)\cdots dF(u_n) = R_T. \end{aligned}$$

The last equality here is due to a change in the variable $\mathbf{u} = \mathbf{x} - \theta\cdot\mathbf{1}$.

Under natural loss functions, comparing two equivariant estimators reduces to comparing two numbers. Under mild conditions, there exists a best (with respect to L) equivariant estimator attaining the infimum of the loss L. It is called the *Pitman estimator* with respect to L.

We shall be dealing mainly with the quadratic loss:

$$L_Q(T(\mathbf{X}),\theta) = (T(\mathbf{X})-\theta)^2.$$

Other frequently used loss functions are Laplacian loss:

$$L_A(T(\mathbf{X}),\theta) = |T(\mathbf{X})-\theta|,$$

and the confidence loss:

$$L_C(T(\mathbf{X});\theta) = \mathbf{I}_{\{|x|>\Delta\}}(T(\mathbf{X})-\theta),$$

where $\mathbf{I}_A(\cdot)$ is the indicator function of the set A and Δ is a positive number of the user's choice.

One can easily see that if $\mu_2 = \int x^2 dF(x) < \infty$, then the Pitman estimator $t_n = t_n(\mathbf{X})$ with respect to the quadratic loss can be written as

$$t_n(\mathbf{X}) = T_0(\mathbf{X}) - E_0(T_0(\mathbf{X})|R_n), \tag{1.16}$$

where E_0 denotes the (conditional) expectation calculated when $\theta = 0$, and T_0 is an arbitrary equivariant estimator as in (1.14). The uniqueness of $t_n(\mathbf{X})$ is obvious from its definition when the conditional expectation is considered as a projection (in L^2 sense) onto the space of square integrable functions of R_n.

For $L_A(T-\theta) = |T-\theta|$ and $\int |x| dF(x) < \infty$, the Pitman estimator can be written as

$$t_n(\mathbf{X}) = T_0(\mathbf{X}) - \text{Median}_0(T_0(\mathbf{X})|R_n),$$

where $\text{Median}_0(T_0(\mathbf{X})|R_n)$ is the median of the conditional distribution of $T_0(\mathbf{X})$ given R_n calculated when $\theta = 0$.

For $L_C(T;\theta) = \mathbf{I}_{|x|>\Delta}(T-\theta)$, with $\Delta > 0$ given, the Pitman estimator can be represented as

$$t_n(\mathbf{X}) = T_0(\mathbf{X}) - u_0(T_0(\mathbf{X})|R_n),$$

where $u_0(T_0(\mathbf{X})|R_n)$ can be any number chosen according to the value of R_n such that the conditional probability $P\{T_0 \in (u_0 - \Delta, u_0 + \Delta)|R_n\}$ is maximized. Since the loss function L_C is not convex, the choice of u_0 is usually not unique.

All these representations for different loss functions are special cases of definition (1.14). They assume very few or even no conditions on the moments. Even

independence of the observations are not required, not to mention any smoothness of the distribution function F. If the density function $F' = p$ of X_i exists , then the Pitman estimator under the quadratic loss (1.16) can be written as

$$t_n(\mathbf{X}) = \frac{\int_{-\infty}^{+\infty} t \prod_{i=1}^{n} p(X_i - t)dt}{\int_{-\infty}^{+\infty} \prod_{i=1}^{n} p(X_i - t)dt}, \tag{1.17}$$

indicating that t_n is the Bayesian estimator of θ corresponding to the (improper) uniform prior on θ.

Simple explicit representations of t_n are known in a few cases.

Example 1 *Gaussian distribution with density* $p(x) = \exp[-(x-\theta)^2/2\sigma^2]/\sqrt{2\pi\sigma^2}$. *The conditional distribution of* X_n *given* $R_n = (r_1, \ldots, r_n)$ *is normal:*

$$X_n | R_n \sim N\left(-\sum_{i=1}^{n-1} \frac{r_i}{n}, \frac{\sigma^2}{n}\right).$$

Thus under all three loss functions(L_Q, L_A *and* L_C*), the Pitman estimator remains the same*

$$t_n(\mathbf{X}) = \bar{X}. \quad \square$$

Example 2 *Exponential distribution with density* $p(x) = \lambda e^{-\lambda(x-\theta)}\mathbf{I}_{\{x>\theta\}}$. *Given the value of* R_n, X_n *has a shifted exponential distribution with the (conditional) density* $p_{X_n|R_n}(x|R_n) = n\lambda e^{-n\lambda(x-\mu)}\mathbf{I}_{\{x>\mu\}}$, *where* $\mu = \max(0, -r_1, \ldots, -r_{n-1})$. *Under the quadratic loss*

$$t_n(\mathbf{X}) = X_{(1)} - \frac{1}{n\lambda},$$

where $X_{(1)}$ *is the minimal observation from the sample* $\mathbf{X}$ *of size* n. *Under the Laplace loss,*

$$t_n(\mathbf{X}) = X_{(1)} - \frac{\ln 2}{n\lambda}.$$

Under the confidence loss with an arbitrary choice of $\Delta > 0$,

$$t_n(\mathbf{X}) = X_{(1)} - \Delta. \quad \square$$

Example 3 *Uniform distribution with density* $p(x) = \mathbf{I}_{(\theta,\theta+1)}(x)$. *The conditional distribution of* X_n *given* R_n *is again uniform on the interval* $(\max(0, -r_1, \ldots, -r_{n-1}), \min(1, 1-r_1, \ldots, 1-r_{n-1}))$. *Under either the quadratic or the Laplace loss, the Pitman estimators are the same:*

$$t_n(\mathbf{X}) = \frac{X_{(1)} + X_{(n)}}{2} - \frac{1}{2}.$$

Under the confidence loss, the estimator is not unique if Δ *is small. For instance, if* $\Delta \leq 1/2 - (X_{(n)} - X_{(1)})/2$, *a version of the Pitman estimator is*

$$t_n(\mathbf{X}) = X_{(1)} - \Delta.$$

In fact, given the sample $\mathbf{X}$, *any value in the interval* $[X_{(n)} - 1 + \Delta, X_{(1)} - \Delta]$ *can be taken as a version of the Pitman estimator under* L_C. *All these estimators achieve the minimal* L_C *risk within the class of equivariant estimators.* $\square$

The concepts of equivariance and Pitman estimators extend to the multivariate case naturally. Let $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ be a sample from an s-variate population with distribution $F(x_1 - \theta_1, \ldots x_s - \theta_s)$, depending on an s-variate location parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)$. An s-variate statistic $\mathbf{T}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is called *equivariant* if for any constant vector $\mathbf{c} \in \mathbb{R}^s$, $\mathbf{T}_n(\mathbf{X}_1 + \mathbf{c}, \ldots, \mathbf{X}_n + \mathbf{c}) = \mathbf{T}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n) + \mathbf{c}$. For instance, $\bar{\mathbf{X}} = (\bar{X}_1, \ldots, \bar{X}_s)^{\mathrm{T}} = (1/n)(\sum_k X_{k1}, \ldots, \sum_k X_{ks})^{\mathrm{T}}$ is an equivariant estimator of $\boldsymbol{\theta}$.

Define the *joint residual*

$$\mathbf{R}_n = (R_{1n}, \ldots, R_{sn}) = (X_{11} - \bar{X}_1, \ldots, X_{ns} - \bar{X}_s) \in \mathbb{R}^{sn}.$$

Any equivariant estimator $\mathbf{T}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ can be written into a form like (1.14):

$$\begin{aligned} \mathbf{T}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n) &= \bar{\mathbf{X}} + [T_n(\mathbf{X}_1, \ldots, \mathbf{X}_n) - \bar{\mathbf{X}}] \\ &= \bar{\mathbf{X}} + g(\mathbf{X}_1, \ldots, \mathbf{X}_n), \end{aligned} \tag{1.18}$$

where $g(\mathbf{X}_1, \ldots, \mathbf{X}_n) \in \mathbb{R}^s$ is an shift invariant vector function. That is, for any shift $\mathbf{c}$:

$$\begin{aligned} & g(\mathbf{X}_1 + \mathbf{c}, \ldots, \mathbf{X}_n + \mathbf{c}) \\ = \; & \mathbf{T}_n(\mathbf{X}_1 + \mathbf{c}, \ldots, \mathbf{X}_n + \mathbf{c}) - (\bar{\mathbf{X}} + \mathbf{c}) = \mathbf{T}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n) - \bar{\mathbf{X}} \\ = \; & g(\mathbf{X}_1, \ldots, \mathbf{X}_n). \end{aligned}$$

In light of the above invariant property of the function g, one sees that g is a measurable function of $\mathbf{R}_n$ by simply letting $\mathbf{c} = \bar{\mathbf{X}}$:

$$g(\mathbf{X}_1, \ldots, \mathbf{X}_n) = g(\mathbf{X}_1 - \bar{\mathbf{X}}, \ldots, \mathbf{X}_n - \bar{\mathbf{X}}).$$

Hence definition (1.18) can be rewritten as

$$\mathbf{T}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n) = \bar{\mathbf{X}} + g(\mathbf{R}_n). \tag{1.19}$$

The Pitman estimator minimizes the covariance function of the estimators in the class of equivariant estimators. As usual, we write $A < B$ for matrices A and B if $B - A$ is positive definite. In (1.19), the covariance matrix depends only on the function g if the population distribution of the samples $\mathbf{X}_i$ is considered known up to only a location shift. Then it is easy to prove that under the least square criterion the unique choice of g minimizing the covariance matrix of $\mathbf{T}_n$ must be the projection of $\bar{\mathbf{X}}$ onto the space of square integrable functions of $\mathbf{R}_n$, that is,

$E_0[\bar{\mathbf{X}}|\mathbf{R}_n]$. Accordingly, a representation of the vector Pitman estimator similar to (1.16) is:

$$\begin{aligned} \mathbf{t}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n) &= \bar{\mathbf{X}} - E_0[\bar{\mathbf{X}}|\mathbf{R}_n] \\ &= \begin{pmatrix} \bar{X}_1 - E_0[\bar{X}_1|\mathbf{R}_n] \\ \vdots \\ \bar{X}_s - E_0[\bar{X}_s|\mathbf{R}_n] \end{pmatrix}. \end{aligned} \tag{1.20}$$

The conditional expectations are all calculated given the joint residual $\mathbf{R}_n$. Unless all the components of $\mathbf{X}_i$ are independent, the Pitman estimator of the i-th component θ_i depends not only on the i-th components $(X_{1i}, \ldots, X_{ni})$ of the data, but also on the complete data set $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$. It is interesting to see that the ordering of the multivariate equivariant estimators in terms of their covariance matrices is linear. But the same relation defined on all (unbiased) estimators of the location parameter $\boldsymbol{\theta}$ is not linear.

Sometimes statisticians are only interested in estimating a linear function of $\boldsymbol{\theta}$, not the whole location parameter $\boldsymbol{\theta}$. Hence we need an equivariant estimator for the p-variate ($p \leq s$) vector $A\boldsymbol{\theta} \in \mathbb{R}^p$, where $A \in \mathbb{M}_{p\times s}$ is a given matrix specifying the linear transformation of interest. For any constant vector $\mathbf{c} \in \mathbb{R}^s$, an equivariant ($p$-variate) estimator $\hat{\mathbf{T}}$ of $A\boldsymbol{\theta}$ must satisfy

$$\hat{\mathbf{T}}_n(\mathbf{X}_1 + \mathbf{c}, \ldots, \mathbf{X}_n + \mathbf{c}) = \hat{\mathbf{T}}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n) + A\mathbf{c}.$$

Starting from the definition, it is easy to prove the following formula for the Pitman estimator of $A\boldsymbol{\theta}$:

$$\hat{\mathbf{t}}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n) = A\bar{\mathbf{X}} - E_0(A\bar{\mathbf{X}}|\mathbf{R}_n). \tag{1.21}$$

If $\mathbf{t}_n$ is the best equivariant estimator of $\boldsymbol{\theta}$, it is not surprising to find out

$$\hat{\mathbf{t}}_n = A\mathbf{t}_n.$$

In the class of equivariant estimators, optimality is invariant under linear transformations provided the loss is location invariant.

1.7 Fisher information on a location parameter

If X is a random variable with cdf $F(x-\theta)$ such that $F^{'} = p$ exists and p is differentiable with $\int (p'(x)/p(x))^2 p(x)dx < \infty$, then the Fisher information on θ contained in X (briefly, the information on θ in X) is defined as

$$I_X(\theta) = E_\theta[J(X)^2] = \int_{-\infty}^{+\infty} \frac{p'(x-\theta)^2}{p(x-\theta)}dx = \int_{-\infty}^{+\infty} \frac{p'(u)^2}{p(u)}du, \tag{1.22}$$

with a change of variable $u = x - \theta$ in the last equality. Plainly, the right hand side of the equality is a quantity depending on the functional form of the distribution F, but not on the parameter θ. Since it does not depend on θ, we shall write I_X instead of $I_X(\theta)$.

If F is not absolutely continuous or if it is but the density p is such that $p'(x)/p(x)$ is not square integrable with respect to $p(x)dx$, then I_X is set equal to infinity. Under such a definition, it remains unclear if a sample $(X_1, \ldots, X_n)$ from a population $F(x-\theta)$ with $I_X = +\infty$ allows the construction of an estimator of θ such that

$$\sqrt{n}(\tilde{\theta}_n - \theta) \longrightarrow_d N(0, \sigma^2)$$

with arbitrarily small σ^2.

For instance, let $F(x)$ be the cdf of the uniform distribution on $(0,1)$ with density $p(x) = \mathbf{I}_{(0,1)}(x)$. Singularity occurs only at the end points of the support of p. The Lebesgue integral in (1.22) assigns a finite number to the Fisher information in spite of such a singularity on the set $\{0,1\}$ of measure zero. On the other hand, the MLE of θ is $\hat{\theta} = X_{(n)} - 1$, where $X_{(n)}$ is the largest observation from a sample of size n. One can easily calculate that

$$\mathrm{Var}(\sqrt{n}\hat{\theta}) = n^2/[(n+1)^2(n+2)] \to 0.$$

For any finite number I_X, there is a sufficiently large n admitting the inequality $\mathrm{Var}(\sqrt{n}\hat{\theta}) < 1/I_X$, which is in the inverse direction of the Cramér-Rao inequality. Without regularity conditions, the Cramér-Rao inequality may not hold.

We turn to another example. Let $F(x)$ be the cdf of the Laplace (double exponential) distribution whose density is $p(x) = e^{-|x|}/2$. Note that $p(x)$ is smooth except when $x = 0$. Now (1.22) sets $I_X = 1$, while the MLE of θ is the sample median $X_{(n/2)}$ with $\mathrm{Var}(\sqrt{n}X_{(n/2)}) \to 1 = 1/I_X$. We have consistency with the Cramér-Rao inequality here.

Both of these examples have a density p such that p'/p is undefined only on a zero measure subset of $\mathrm{supp}\{p\}$. But we may need to assume different values for their Fisher information if we want to generalize their statistical interpretation consistently. A better definition for I_X should explain how a statistician can benefit from observing a random variable with infinite information on θ. The following definition is due to Huber (1964):

$$I_X = \sup_{\psi \in C_c^1(\mathbb{R})} \frac{[\int \psi'(x)dF(x)]^2}{\int \psi^2(x)dF(x)}, \tag{1.23}$$

where $C_c^1(\mathbb{R})$ is the space of continuously differentiable functions with compact support. The equivalence between (1.23) and (1.22) is established by the following theorem in Huber (1981):

Theorem 1.7.1 *Let I_X be defined by (1.23). The following statements are equivalent:*

(i) $I_X < +\infty$,

(ii) F has an absolutely continuous density p, and $(\ln p(X))' = p'(X)/p(X) \in L_F^2$, i.e., $\int (p'(x)/p(x))^2 dF(x) < +\infty$.

In either case, $I_X = \int (p'(x)/p(x))^2 p(x)dx$.

If $I_X = +\infty$, from Huber's definition there follows the existence of $\psi(x)$ such that the quotient in (1.23) can be made arbitrarily large. Therefore, in the theory of estimating functions we can find an estimator associated with $\psi(x)$ such that arbitrarily small asymptotic variance can be achieved. The statistical corollary of this fact will be discussed in Section 1.11.

To analyze Huber's definition (1.23), one may define a linear functional on $C_c^1(\mathbb{R})$:

$$A\psi = -\int \psi'(x)p(x)dx, \quad \forall \psi \in C_c^1(\mathbb{R}).$$

By definition, the L^2 operator norm of this functional is $\|A\|_F^2 = I_X$. One may extend the integration by parts formula to the case in which p is not necessarily absolutely continuous but only weakly differentiable. In particular, if there exists a *generalized function* $p'(x)$ satisfying the following equalities:

$$A\psi = -\int \psi'(x)p(x)dx = \int \psi(x)p'(x)dx, \tag{1.24}$$

then $p'(x)$ is called the weak derivative of $p(x)$ (see Ziemer (1989)). If p is absolutely continuous, $p'(x)$ coincide with the common definition of its derivative almost everywhere (with respect to the Lebesgue measure). Particularly, on the set of measure zero where p is not differentiable, p' can be set equal to any finite number. The above definition imposes no difficulty in defining the Lebesgue integral

$$\begin{aligned} |A\psi|^2 &= \left| \int \psi(x) p'(x) dx \right|^2 \\ &= \left| \int_{\text{supp}\{p\}} \psi(x) \frac{p'(x)}{p(x)} dF(x) \right|^2 \leq \|\psi(X)\|_F^2 \int_{\text{supp}\{p\}} \left(\frac{p'(x)}{p(x)} \right)^2 dF(x). \end{aligned}$$

Therefore $I_X < +\infty$ if the Fisher score $J(X)$ belongs to L_F^2. We will demonstrate this idea with the two examples we just mentioned earlier in this section.

Example 4 *Let $p(x) = e^{-|x|}/2$ be the PDF of the Laplace distribution. It can be easily verified that $p(x)$ is absolutely continuous with only one singular point at $x = 0$. A version of its weak derivative is:*

$$p'(x) = \begin{cases} -e^{-x}/2 & x > 0 \\ 0 & x = 0 \\ e^{x}/2 & x < 0. \end{cases} \tag{1.25}$$

By definition (1.24), $A\psi = \int_{-\infty}^{0} \psi(x) e^x/2 dx - \int_0^{+\infty} \psi(x) e^{-x}/2 dx$. One may calculate the norm of A associated with the Laplace distribution:

$$\begin{aligned} \|A\|_F^2 &= \sup \frac{\left(\int_{-\infty}^{0} (\psi(x) e^x/2) dx - \int_0^{+\infty} (\psi(x) e^{-x}/2) dx \right)^2}{\int_{-\infty}^{+\infty} (\psi^2(x) e^{-|x|}/2) dx} \\ &\leq \sup \frac{(\int_{-\infty}^{0} (|\psi(x)| e^x/2) dx + \int_0^{+\infty} (|\psi(x)| e^{-x}/2) dx)^2}{E[\psi^2(X)]} = \sup \frac{(E|\psi(X)|)^2}{E[\psi^2(X)]} \leq 1. \end{aligned}$$

The last inequality is due to the Cauchy-Schwarz inequality. Moreover, (1.22) assigns the value 1 to I_X. Therefore $I_X = 1 < +\infty$ by Theorem 1.7.1. □

If p is not absolutely continuous, then p' has to be defined with the concept of generalized functions.

Example 5 *Let X be a random variable uniformly distributed on the unit interval with density $p(x) = \mathbf{I}_{(0,1)}(x)$, where $\mathbf{I}_{(0,1)}$ is the indicator function of the set $(0,1)$. One version of the weak derivative of p is $p'(x) = \delta_0(x) - \delta_1(x)$, where $\delta_t(x)$ is the Dirac delta function with a shift t. Then (1.24) becomes:*

$$A\psi = \int_{-\infty}^{+\infty} \psi(x)[\delta_0(x) - \delta_1(x)]dx = \psi(0) - \psi(1).$$

Accordingly, the Fisher information contained in a uniformly distributed random variable is $I_X = \sup_\psi((A\psi)^2/E(\psi^2)) = \sup_\psi((\psi(0) - \psi(1))^2/E(\psi^2)$. This functional is obviously not bounded even if we only consider those $\psi(x)$ with support on $(-1/2, 1/2)$ and $E[\psi^2] = 1$. By (1.23), $I_X = +\infty$, as the density p is not absolutely continuous. It is consistent with the Cramér-Rao inequality when we consider the asymptotic variance of the MLE:

$$\mathrm{Var}(\sqrt{n}\hat{\theta}) = \frac{n^2}{(n+1)^2(n+2)} \to 0 = \frac{1}{I_X}. \quad \square$$

The following remark is due to Vershik (see Fintushal (1975)). Under regularity type conditions, for any arbitrary smooth function ψ,

$$\int \psi(x)J(x)dF(x) = -\int \psi'(x)dF(x).$$

With an inner product $< f, g >= \int f(x)g(x)dF(x)$ defined on the space $L_F^2(X)$, the above equality can be rewritten as

$$< \psi(X), J(X) >= - < D\psi(X), 1 >= - < \psi(X), D^*1 >, \quad \forall \psi \in L_F^2, \qquad (1.26)$$

where D is the operator of differentiation and D^* its adjoint. The equations indicate that $J(X) = -D^*1$, where 1 is the constant function defined on $\mathfrak{X}$. Following the Cauchy-Schwarz inequality, (1.26) indicates

$$< D\psi(X), 1 >^2 =< \psi(X), D^*1 >^2 \leq \|\psi(X)\|^2 \|D^*1\|^2$$

$$\Rightarrow \frac{(\int D\psi(x) dF(x))^2}{\|\psi(X)\|^2} \leq \|D^*1\|^2,$$

Comparing (1.26) with (1.23), we see that $I_X \leq \|D^*1\|^2$. At the same time, the Cauchy-Schwarz inequality guarantees the equality sign hold for some ψ, so that we can conclude $I_X = \|D^*1\|$.

In summary, the Fisher information is finite if the constant function 1 belongs to the domain of D^*, i.e., $D^*1 \in L_F^2(X)$. Since D is an unbounded operator on $L_F^2(X)$, this condition is a restriction on F. Shlyakhtenko (2005) adopted this definition when proving a parallel version of the monotonicity of Shannon's Entropy in the free probability theory.

1.8 Stam inequality

For independent random variables X and Y, let I_X, I_Y and I_{X+Y} be the Fisher information in the observations shifted by an unknown parameter θ: $X + \theta$, $Y + \theta$ and $X + Y + \theta$, respectively. Due to independence of X and Y, $X + Y$ is "more random" than X, so that one can expect an inequality

$$I_{X+Y} \leq I_X.$$

To sketch a rigorous proof of the inequality, suppose that the data is presented in a pair: $(X+\theta, Y)$. Then by the monotonicity of the Fisher information,

$$I_{X+\theta+Y} = I_{X+Y} \leq I_{X+\theta,Y} = I_X,$$

where the independent component Y provides no information for estimating θ.

A much less trivial inequality is due to Stam(1959):

$$\frac{1}{I_{X+Y}} \geq \frac{1}{I_X} + \frac{1}{I_Y}. \tag{1.27}$$

For iid X and Y, (1.27) becomes $I_X \geq 2I_{X+Y}$. Barron and Madiman (2006) generalized the latter inequality. Let $X_1, X_2, \ldots$ be a sequence of iid random variables with finite information I_{X_1}. Under rather strong regularity type conditions, they proved that

$$nI_{X_1+\ldots+X_n} = I_{(X_1+\ldots+X_n)/\sqrt{n}} \text{ decreases with } n.$$

For a small sample version of this result, see Chapter 2.

1.9 Properties of the Pitman estimators

In this section, we consider the Pitman estimators with respect to the quadratic loss L_Q.

First, it is obvious from (1.16) that t_n is an unbiased estimator of θ.

Second, if $\mu_2 = \int |x|^2 dF(x) < \infty$, then for any $n \geq 1$

$$\mathrm{Var}(t_n) \leq \mathrm{Var}(t_{n-1}). \tag{1.28}$$

It is almost trivial, since t_{n-1} is an equivariant estimator from a sample of size n, while t_n is the optimal equivariant estimator. Notice that for traditional estimators

of a general parameter such as the MLE, analogs of (1.28), that is, monotonicity in the sample size of a reasonable measure of accuracy, are not known.

If a sample comes from a population $F(x-\theta)$ with absolutely continuous F, $F' = p$ exists almost everywhere and there exists an equivariant estimator $T_0(\mathbf{X})$ with $E[T_0(\mathbf{X})^3] < +\infty$ (in particular, if $\int |x|^3 \mathrm{d}F(x) < \infty$). Stein (1959) proved that $t_n(\mathbf{X})$ is admissible (actually, a little more general result was proved). Brown (1966) made a detailed discussion on the admissibility of $t(\mathbf{X})$. He proved the admissibility of $t_n(\mathbf{X})$ for different loss functions, when the best equivariant estimator is unique.

Some characterization results are proved for the Pitman estimator. For $n \geq 3$, $t_n = \bar{X}$ if and only if F is Gaussian, which is known as the Kagan-Linnik-Rao (KLR) Theorem (see Kagan *et al.* (1973)). For $n = 2$, $t_n = \bar{X}$ for any symmetric F. If $X_{(1)}$, $X_{(n)}$ are the minimum and maximum order statistics from a sample of size $n \geq 3$, then under some regularity type conditions on F, $t_n = (X_{(1)} + X_{(n)})/2$ if and only if F is the distribution function of a uniform random variable on $(-a, a)$ for any $a > 0$ (see Bondesson (1974)).

From the Bayesian representation (1.17), if S is sufficient for the family of densities $\{\prod_{i=1}^n p(x_i - \theta), \theta \in \mathbb{R}\}$, then t_n depends on $\mathbf{X} = (X_1, \ldots, X_n)$ only through S. Certainly, this also follows from Stein's result about admissibility cited above. If t_n is not a function of S, then $E_\theta(t_n|S)$ is an estimator uniformly better than t_n contradicting the admissibility of t_n. Besides, it is of some (at least, methodological) interest to find out if the Pitman estimator depends on the data only through sufficient statistics where F is not assumed absolutely continuous.

The above results hold in small samples. In large samples, as $n \to \infty$, t_n

behaves like the MLE, though in small samples it is better since the MLE is an equivariant estimator.

As shown in Ibragimov and Has'minskii (1981), under very mild regularity type conditions

$$\sqrt{n}[t_n(\mathbf{X}) - \theta] \longrightarrow_d N(0, 1/I_{X_1}). \tag{1.29}$$

Moreover, if we assume the additional condition $\int_{-\infty}^{+\infty} |x|^\delta dF(x) < +\infty$ for some $\delta > 0$, then

$$E_\theta[\sqrt{n}(t_n(\mathbf{X}) - \theta)]^2 = \frac{1}{I_{X_1}}(1 + o(1)). \tag{1.30}$$

1.10 Polynomial Pitman estimators

Pitman estimators are defined with conditional expectations, which are not quite convenient for computations. In this section, we are going to look at a simplified polynomial version of Pitman estimators. Assume that for some integer $k \geq 1$,

$$\mu_{2k} = \int |x|^{2k} dF(x) < \infty. \tag{1.31}$$

Let $\Lambda_k = \text{span}\{r_1^{d_1} \dots r_n^{d_n} |$ the d_i are nonnegative integers, $d_1 + \dots + d_n \leq k\}$ be the space of polynomials of degree k in the residuals

$$R_n = (r_1, \dots, r_n) = (X_1 - \bar{X}, \dots, X_n - \bar{X}).$$

Then Λ_k is a finite dimensional subspace of the Hilbert space $L_F^2(X_1, \dots, X_n)$ of all functions $\psi(X_1, \dots, X_n)$ with

$$\|\psi\|_F^2 = E_\theta|\psi(X_1, \dots, X_n)|^2 = \int \cdots \int \psi^2 dF(x_1 - \theta) \dots dF(x_n - \theta) < \infty,$$

with standard inner product

$$< \psi_1, \psi_2 >= E_0(\psi_1 \psi_2).$$

In Kagan (1966) a version of t_n, called the *polynomial* (of degree k) *Pitman estimator* was introduced as

$$t_n^{(k)} = \bar{X} - \hat{E}_0(\bar{X}|\Lambda_k), \tag{1.32}$$

where $\hat{E}_0(\cdot|\Lambda_k)$ is the operator of projection into the Hilbert subspace Λ_k, while assuming $\theta = 0$.

Similar to t_n, $t_n^{(k)}$ is the best polynomial of degree k equivariant estimator of θ. Since $t_n^{(k)}$ is an equivariant estimator,

$$\mathrm{Var}(t_n^{(k)}) \geq \mathrm{Var}(t_n).$$

In return, $t_n^{(k)}$ depends not on the entire function F but only on its first $2k$ moments.

Under the same condition (1.31), one can define the *polynomial Fisher score* $J^{(k)}$ and hence the *polynomial Fisher information*. Denote by $L_F^2(X)$ the space of functions $\psi(X)$ (of one argument) with $\int |\psi(x)|^2 dF(x) < \infty$ and standard inner product

$$< \psi_1, \psi_2 >= E_0[\psi_1(X)\psi_2(X)].$$

Let $\mathbf{P}_k = \mathrm{span}\{X^j | j = 0, \ldots, k\}$ be the subspace of $L_F^2(X)$ of all polynomials of degree k, set

$$J^{(k)}(X) = \hat{E}_\theta(J(X)|\mathbf{P}_k). \tag{1.33}$$

Clearly, $E_\theta[J^{(k)}(X)] \equiv 0$, and the expected square

$$I_X^{(k)} = E_\theta[J^{(k)}(X)]^2 = E_0[J^{(k)}(X)]^2. \tag{1.34}$$

now has the meaning (and properties) of the Fisher information. The polynomial score depends only on the first $2k$ moments of F and can be defined without reference to $J(X)$, and without assuming absolute continuity of F:

$$< \psi_k(X), J^{(k)}(X) >= - < \psi_k'(X), 1 >, \quad \forall \psi_k(X) \in \mathbf{P}_k. \tag{1.35}$$

The properties of the polynomial Pitman estimators $t_n^{(k)}$ are similar to those of t_n:

(i) $t_n^{(k)}$ is an unbiased estimator of θ,

(ii) $\mathrm{Var}(t_n^{(k)}) \leq \mathrm{Var}(t_{n-1}^{(k)})$,

(iii) $\mathrm{Var}(t_n^{(k)}) \leq \mathrm{Var}(t_n^{(k-1)})$,

(iv) $t_n^{(k)} = \bar{X}$ for $n \geq 3$ if and only if the first $(k+1)$ moments of the distribution function $F(x)$ coincide with the corresponding moments of some normal distribution:

$$\mu_\ell = \begin{cases} (\ell-1)(\ell-3)\cdots 1 \cdot \sigma^\ell & \text{if } \ell \text{ is odd, } 1 \leq \ell \leq k+1 \\ 0 & \text{if } \ell \text{ is even, } 1 \leq \ell \leq k+1 \end{cases}$$

In fact $\bar{X}$ is an admissible estimator of θ in the class of polynomial equivariant estimators of order up to k if and only if the above condition is satisfied.

The asymptotic behavior of $t_n^{(k)}$ was studied in Kagan et. al. (1973). Assuming that $\mu_{2k} < +\infty$ and $F(x)$ has at least k points of increase,

$$\sqrt{n}(t_n^{(k)} - \theta) \xrightarrow{d} N(0, 1/I_{X_1}^{(k)}), \quad \text{as } n \to \infty. \tag{1.36}$$

Also in this case,

$$E_\theta(\sqrt{n}(t_n^{(k)} - \theta))^2 = \frac{1}{I_{X_1}^{(k)}}(1 + o(1)). \tag{1.37}$$

There is a useful modification of $t_n^{(k)}$. Namely, let $m_j = \sum_{i=1}^n (X_i - \bar{X})^j/n$ denote the sample central moment of order j ($m_0 = 1, m_1 = 0$). Set

$$\tau_n^{(k)}(\mathbf{X}) = \bar{X} - \sum_{j=0}^{k} A_{j,n} m_j. \tag{1.38}$$

where $A_{j,n}$ are the optimal coefficients, i.e.,

$$\mathrm{Var}_0(\bar{X} - \sum A_{j,n} m_j) = \min_{a_{0n},\dots,a_{kn}} \mathrm{Var}(\bar{X} - \sum a_{jn} m_j).$$

Certainly,

$$\tau_n^{(k)} = \bar{X} - \hat{E}_0(\bar{X}|\mathrm{span}(1, m_2, \dots, m_k)).$$

In small samples, $\mathrm{Var}(\tau_n^{(k)}) \geq \mathrm{Var}(t_n^{(k)})$. However, $\tau_n^{(k)}$ has much simpler structure than $t_n^{(k)}$ and moreover, asymptotically it behaves like $t_n^{(k)}$ (see Kagan (1986)). Under the same condition as required for (1.36),

$$\sqrt{n}(\tau_n^{(k)} - \theta) \longrightarrow_d N(0, 1/I_{X_1}^{(k)}), \quad n \to \infty. \tag{1.39}$$

Besides, under no extra conditions

$$E_\theta(\sqrt{n}(\tau_n^{(k)} - \theta))^2 = \frac{1}{I_{X_1}^{(k)}}(1 + \mathrm{o}(1)). \tag{1.40}$$

1.11 Estimating functions and the Fisher information

Let X be a random element with values in $(\mathfrak{X}, \mathscr{A})$ and a probability distribution P_θ, $\theta \in \Theta \subset \mathbb{R}$. A function $\psi(x;\theta)$ is called an *estimating function* for $\mathscr{P} = \{P_\theta, \theta \in \Theta\}$ (or briefly for θ if $\mathscr{P}$ is assumed) if

(i) $E_\theta \psi(X;\theta) \equiv 0$, $\forall \theta \in \Theta$,

(ii) $E_\theta[\frac{\partial}{\partial\theta}\psi(X;\theta)] \neq 0$, $\forall \theta \in \Theta$,

(iii) $\mathrm{Var}_\theta \psi(X;\theta) = E_\theta[\psi(X;\theta)]^2 < \infty$, $\forall \theta \in \Theta$.

The concept is due to Godambe (1970) and is justified by the following analysis.

If $(X_1, \ldots, X_n)$ is a sample from $\mathscr{P}$, then under mild regularity type conditions, the *estimating equation*

$$\sum_{i=1}^{n} \psi(X_i;\theta) = 0$$

has a solution

$$\check{\theta}_n = \check{\theta}(X_1, \ldots, X_n)$$

such that

$$\sqrt{n}(\check{\theta} - \theta) \longrightarrow_d N(0, \sigma^2_\psi(\theta)),\ n \to \infty$$

where $\sigma^2_\psi(\theta) = \mathrm{Var}_\theta[\psi(X;\theta)]/[E_\theta(\frac{\partial}{\partial\theta}\psi(X;\theta))]^2$. Its reciprocal $I_\psi(\theta) = 1/\sigma^2_\psi(\theta)$ is called the information associated with ψ.

If P_θ are absolutely continuous with respect to a measure μ on $(\mathfrak{X}, \mathscr{A})$ with density $p(x;\theta) = dP_\theta/d\mu$ and the Fisher information I_X is finite, then the Fisher score $J(X;\theta)$ is an estimating function. By definition, the MLE is a solution of the corresponding estimating equation $\sum_{i=1}^n J(X_i,\theta) = 0$. The information associated with the estimating function J is the Fisher information on θ in X: $I_J(\theta) = I_X(\theta)$. From the asymptotic optimality of the MLE, it follows that $I_J(\theta) \geq I_\psi(\theta)$ for any estimating function ψ. In other words, $J(X;\theta)$ maximizes the information in the theory of estimating functions.

Let now X be a random variable with distribution function $F(x-\theta)$ depending on a location parameter. If $\psi(x;\theta) = \psi(x-\theta)$ is an estimating function, the

corresponding estimating equation

$$\sum_{i=1}^{n} \psi(X_i - \theta) = 0 \tag{1.41}$$

generates an equivariant estimator $\hat{\theta}_n$. Hence (1.41) is called an *equivariant estimating equation* and ψ in such a special form is an *equivariant estimating function.*

Following the asymptotic analysis on the general estimating function theory, we now have

$$\sqrt{n}(\hat{\theta}_n - \theta) \longrightarrow_d N(0, \sigma_\psi^2),$$

where $\sigma_\psi^2 = E_0[\psi^2(X)]/[E_0(\psi^{'}(X))]^2$ does not depend on θ. The information associated with ψ is $I_\psi = 1/\sigma_\psi^2 = [E_0(\psi^{'}(X))]^2/E_0[\psi^2(X)]$.

Comparing I_ψ with (1.23), we see that $I_X = \sup_\psi I_\psi$. If $\sup_\psi I_\psi = \infty$, then there exists an (equivariant) estimating function $\tilde{\psi}(x-\theta)$ with arbitrarily large value of

$$I_{\tilde{\psi}} = \frac{E_\theta[\tilde{\psi}'(X-\theta)]^2}{\mathrm{Var}_\theta \tilde{\psi}(X-\theta)} = \frac{E_0[\tilde{\psi}'(X)]^2}{\mathrm{Var}_0 \tilde{\psi}(X)}.$$

It means that there is an estimating equation $\sum \tilde{\psi}(X_i - \theta) = 0$, the solution of which leads to an estimator $\tilde{\theta}(X_1, \ldots, X_n)$ with arbitrarily small asymptotic variance. In other words, for observations with infinite Fisher information there exists an estimator (obtained by a regular method of estimating equations) with asymptotic variance $\leq \epsilon^2/n$ for any $\epsilon > 0$.

Assume that for some integer $k \geq 1$, $\mu_{2k} = \int x^{2k} dF < \infty$. One can consider the (equivariant) *polynomial estimating functions* $\psi_k(x-\theta) \in \mathbf{P}_k$. Here

$$\psi_k(x) = a_0 + a_1 x + \ldots + a_k x^k$$

is a polynomial of degree k. The corresponding polynomial estimating equation $\sum_{i=1}^{n} \psi_k(X_i - \theta) = 0$ generates an equivariant estimator $\check{\theta}_n^{(k)}$ (in a neighborhood of the true θ), which is not a polynomial, such that

$$\sqrt{n}(\check{\theta}_n^{(k)} - \theta) \longrightarrow_d N\left(0, \frac{\mathrm{Var}_0[\psi_k(X)]}{E_0[\psi_k'(X)]^2}\right).$$

One can see that the optimal polynomial estimating function, i.e., that minimizes $\mathrm{Var}_0[\psi_k(X)]/E_0[\psi_k'(X)]^2$ over all k-th order polynomials ψ_k is the polynomial Fisher score $J^{(k)}(X)$ as defined in (1.35).

Theorem 1.11.1 *Suppose $\mu_{2k} < \infty$. Then*

$$\sup_{\psi_k \in \mathbf{P}_k} \frac{(\int \psi_k'(x) dF(x))^2}{\int \psi_k^2(x) dF(x)} = \|J^{(k)}(X)\|^2 = I_X^{(k)}. \tag{1.42}$$

Proof. Define a linear functional $A\psi_k = -\int \psi_k'(x) dF(x) = - < \psi_k', 1 >$ for arbitrary $\psi_k(x) \in \mathbf{P}_k$. Now compare it with definition (1.35). One can see $A\psi_k =< \psi_k, J^{(k)} >$, while $J^{(k)}$ is the only element in $\mathbf{P}_k$ that satisfies the equation. By Cauchy-Schwarz inequality,

$$|A\psi_k|^2 \leq \|\psi_k\|^2 \|J^{(k)}\|^2 = E[\psi_k^2(X)] I_X^{(k)}$$
$$\Rightarrow \frac{|A\psi_k|^2}{E(\psi_k^2)} \leq I_X^{(k)}$$

Take the supremum over all ψ_k's on both sides. Then (1.42) follows because the supremum of $I_X^{(k)}$ can be attained when $\psi_k = J^{(k)}$:

$$\frac{(\int (J^{(k)}(x))' dF(x))^2}{\int (J^{(k)}(x))^2 dF(x)} = \frac{[\int (J^{(k)}(x))^2 dF(x)]^2}{\int (J^{(k)}(x))^2 dF(x)} = I_X^{(k)}. \qquad \square$$

Notice that on the right hand side of (1.42), $I_X^{(k)}$ is the polynomial Fisher information in X. Its reciprocal $1/I_X^{(k)}$ is the asymptotic variance of $\sqrt{n}(t_n^{(k)} - \theta)$ as

indicated by (1.37). It is worthwhile to emphasize that the same number serves as the asymptotic variance of the estimators $\sqrt{n}(\check{\theta}_n^{(k)} - \theta)$ as well, though $\check{\theta}_n^{(k)}$ is not a polynomial estimator.

Chapter 2

Behavior of the Pitman estimators in small samples

In this chapter we will study some fine small sample properties of the Pitman estimators t_n with respect to quadratic loss.

Consider the setup of direct measurements when independent identically distributed observations $X_1, \ldots, X_n$ are of the form

$$X_i = \theta + \epsilon_i,\ i = 1, \ldots, n.$$

Here $\theta \in \mathbb{R}$ is a (location) parameter of interest and $\epsilon_1, \ldots, \epsilon_n$ are iid errors, $\epsilon_i \sim F$ so that $X_i \sim F(x - \theta)$.

Certainly, the setup of direct measurements is very special but many results obtained here can be extended to the setup of linear regression when independent (but not identically distributed) observations are of the form

$$X_i = a_{i1}\theta_1 + \ldots + a_{is}\theta_s + \epsilon_i,\ i = 1, \ldots, n$$

with $\boldsymbol{\theta} \in \mathbb{R}^s$ as a parameter and a design matrix (a_{ir}) assumed known.

Let $R_n = (X_1 - \bar{X}, \ldots, X_n - \bar{X})$ be the vector of residuals. As shown in (1.16), if $\mathrm{Var}(X_i) < \infty$ then $t_n = t_n(X_1, \ldots, X_n)$, the Pitman estimator of θ with respect to the quadratic loss, is

$$t_n = \bar{X} - E_0(\bar{X}|R_n).$$

As mentioned in Section 1.9, under some additional regularity type conditions, from

(1.30) it follows

$$\frac{1}{n\mathrm{Var}(t_n)} \longrightarrow I_{X_1},\ n \to \infty.$$

The two quantities I_{X_1} and $\mathrm{Var}(t_n)$ are closely connected, not only in large samples, but we shall see that in small samples, the properties of $\mathrm{Var}(t_n)$ are similar to those of the Fisher information.

2.1 Monotonicity of $n\mathrm{Var}(t_n)$

The first property to be proved is the *monotone decrease of* $n\mathrm{Var}(t_n)$ *in* n, which is much deeper than the decrease of $\mathrm{Var}(t_n)$. Notice in passing that, to the best of author's knowledge, there is no general proof of monotone decrease in n of $\mathrm{Var}(\hat{\theta}_n)$ for the MLE $\hat{\theta}_n$. Let us start with an example.

Example 6 *If F is Gaussian $N(0, \sigma^2)$, $n\mathrm{Var}(t_n) = \sigma^2$ for all n and this constancy is a characteristic property of the Gaussian distribution.*

If F is the distribution function of an exponential distribution with parameter λ,

$$n\mathrm{Var}(t_n) = n\mathrm{Var}\left[X_{(1)} - \frac{1}{n\lambda}\right] = \frac{1}{n\lambda^2}.$$

If F is the distribution function of a uniform distribution on $(0, 1)$,

$$n\mathrm{Var}(t_n) = n\mathrm{Var}\left[\frac{X_{(1)} + X_{(n)}}{2} - \frac{1}{2}\right] = \frac{n}{2(n+1)(n+2)}. \quad \square$$

Our proof of the monotonicity of $n\mathrm{Var}(t_n)$, is based on a fundamental lemma whose idea goes back to Hoeffding (1948), Efron and Stein (1981) and in the stated form by Artstein, Ball, Barthe and Noar (2004).

Let $\mathfrak{S}$ be an arbitrary collection of subsets of the index set $\{1, 2, \ldots, n\}$. Set

$$r(\mathfrak{S}) = \max_{i \in \{1,\ldots,n\}} \sum_{\mathbf{s} \in \mathfrak{S}} \mathbf{I}_{\mathbf{s}}(i), \tag{2.1}$$

where $\mathbf{I}_{\mathbf{s}}$ is the indicator of the index set $\mathbf{s}$. In other words, $r(\mathfrak{S})$ is the maximum number of times an index $i \in \{1, \ldots, n\}$ appears in the elements of $\mathfrak{S}$. For example, if $\mathfrak{S}$ is the collection of all unordered sets $\mathbf{s}$ of m $(1 \leq m \leq n)$ elements from $\{1, \ldots, n\}$, then for any fixed i there are exactly $\binom{n-1}{m-1}$ elements in $\mathfrak{S}$ containing this index i, hence $r(\mathfrak{S}) = \binom{n-1}{m-1}$.

Lemma 2.1.1 *(Artstein et al. (2004)) Let $X_1, \ldots, X_n$ be independent random variables and $\mathfrak{S}$ a given collection of subsets of $\{1, \ldots, n\}$. Suppose that with every $\mathbf{s} \in \mathfrak{S}$, a measurable function $\psi_{\mathbf{s}} = \psi_{\mathbf{s}}(X_i; i \in \mathbf{s})$ with $E(\psi_{\mathbf{s}}^2) < \infty$ is associated. Then for any probability distribution $\{w_{\mathbf{s}} | \sum_{\mathbf{s} \in \mathfrak{S}} w_{\mathbf{s}} = 1\}$ on the set $\mathfrak{S}$, we have*

$$\mathrm{Var}\left(\sum_{\mathbf{s} \in \mathfrak{S}} w_{\mathbf{s}} \psi_{\mathbf{s}}\right) \leq r(\mathfrak{S}) \sum_{\mathbf{s} \in \mathfrak{S}} w_{\mathbf{s}}^2 \mathrm{Var}(\psi_{\mathbf{s}}). \tag{2.2}$$

If the maximum in the right hand side of (2.1) is attained at only one index i, the equality sign in (2.2) holds only if for all $\mathbf{s} \in \mathfrak{S}$, $\psi_{\mathbf{s}}$ is additively decomposable, that is, $\psi_{\mathbf{s}} = \sum_{i \in \mathbf{s}} \phi_{\mathbf{s}i}(X_i)$ for some measurable functions $\phi_{\mathbf{s}i}$.

The following proof of Lemma 2.1.1 is due to Madiman and Barron (2006), where it is called *Variance Drop Lemma*. First of all we need another lemma.

Lemma 2.1.2 *Let ξ be a random variable and let η_1, η_2 be random elements with $E|\xi| < \infty$. Suppose that (ξ, η_1) and η_2 are independent. Then*

$$E(\xi | \eta_1, \eta_2) = E(\xi | \eta_1) \quad \text{a.s.} \tag{2.3}$$

Proof. See Shao (2003), page 41. □

Proof of Lemma 2.1.1: The proof is divided into three parts.

(i) We need an orthogonal decomposition of the $\psi_{\mathbf{s}}$. For this purpose, fix $\mathbf{s}$ for a moment, and set $E_{\mathbf{s}\backslash\mathbf{t}}(\cdot) = E(\cdot|X_i; i \in \mathbf{s}\backslash\mathbf{t})$ for any $\mathbf{t}$, a subset of $\mathbf{s}$. Notice that for independent $X_1, \ldots, X_n$, these conditional expectation operators are commutative. For any $\mathbf{t}_1 \subset \mathbf{s}$, $\mathbf{t}_2 \subset \mathbf{s}$ and any function φ with $E|\varphi| < \infty$,

$$\begin{aligned} E_{\mathbf{s}\backslash\mathbf{t}_1}\{E_{\mathbf{s}\backslash\mathbf{t}_2}[\varphi(X_1, \ldots, X_n)]\} &= E_{\mathbf{s}\backslash\mathbf{t}_2}\{E_{\mathbf{s}\backslash\mathbf{t}_1}[\varphi(X_1, \ldots, X_n)]\} \\ &= E_{\mathbf{s}\backslash(\mathbf{t}_1\cup\mathbf{t}_2)}[\varphi(X_1, \ldots, X_n)]. \end{aligned}$$

Note that the conditional expectation operators can be viewed as projections onto some properly defined Hilbert spaces. It is known that projections are commutative if and only if their composites are also projections, as the above equation implies. To verify the equation, the following two properties of the conditional expectation are used. First, for any function φ and index sets $\mathbf{s}$, $\mathbf{s}'$ with $\mathbf{s} \subset \mathbf{s}'$

$$E[E(\varphi|X_i; i \in \mathbf{s}')|X_i; i \in \mathbf{s}] = E(\varphi|X_i; i \in \mathbf{s}).$$

Second, if $\varphi = \varphi(X_i; i \in \mathbf{s})$ for some index set $\mathbf{s}$ then

$$E(\varphi|X_j, X_i; i \in \mathbf{s}) = E(\varphi|X_i; i \in \mathbf{s}),$$

for any $j \notin \mathbf{s}$ by Lemma 2.1.2.

Combining the conditional expectation operators in the proper order, one can rewrite $\psi_{\mathbf{s}}$ as

$$\begin{aligned}\psi_{\mathbf{s}} &= \prod_{j\in\mathbf{s}}[E_{\mathbf{s}\setminus\{j\}} + (I - E_{\mathbf{s}\setminus\{j\}})]\psi_{\mathbf{s}} \\ &= \sum_{\mathbf{t}\subset\mathbf{s}}\left[\prod_{j\notin\mathbf{t}} E_{\mathbf{s}\setminus\{j\}}\prod_{j\in\mathbf{t}}(I - E_{\mathbf{s}\setminus\{j\}})\right]\psi_{\mathbf{s}} \\ &= \sum_{\mathbf{t}\subset\mathbf{s}}\left[E_{\mathbf{s}\setminus\mathbf{t}}\prod_{j\in\mathbf{t}}(I - E_{\mathbf{s}\setminus\{j\}})\right]\psi_{\mathbf{s}},\end{aligned}$$

where I is the identity operator. On setting

$$\phi_{\mathbf{st}} = \left[E_{\mathbf{s}\setminus\mathbf{t}}\prod_{j\in\mathbf{t}}(I - E_{\mathbf{s}\setminus\{j\}})\right]\psi_{\mathbf{s}},$$

one obtains a decomposition of the function $\psi_{\mathbf{s}}$:

$$\psi_{\mathbf{s}} = \sum_{\mathbf{t}\subset\mathbf{s}}\phi_{\mathbf{st}}, \tag{2.4}$$

where $\phi_{\mathbf{st}}$ corresponding to different $\mathbf{t}$ are orthogonal (uncorrelated). Indeed, for any two distinct index sets $\mathbf{t}_1, \mathbf{t}_2 \subset \mathbf{s}$, there is at least one index j distinguishing them from each other, that is, j is in exactly one of $\mathbf{t}_1$, $\mathbf{t}_2$. Without loss in generality, let $j \in \mathbf{t}_1$, $j \notin \mathbf{t}_2$. Then

$$\mathrm{Cov}(\phi_{\mathbf{st}_1}, \phi_{\mathbf{st}_2}) = \mathrm{Cov}[E_{\mathbf{s}\setminus\{j\}}\phi_{\mathbf{st}_1}, (I - E_{\mathbf{s}\setminus\{j\}})\phi_{\mathbf{st}_2}] = 0.$$

Thus, (2.4) is an orthogonal decomposition. Even further, for index sets $\mathbf{s}_1 \neq \mathbf{s}_2$, one may have orthogonal decompositions for $\psi_{\mathbf{s}_1}$ and $\psi_{\mathbf{s}_2}$, respectively:

$$\psi_{\mathbf{s}_1} = \sum_{\mathbf{t_1}\subset\mathbf{s}_1}\phi_{\mathbf{s}_1\mathbf{t}_1},\ \psi_{\mathbf{s}_2} = \sum_{\mathbf{t}_2\subset\mathbf{s}_2}\phi_{\mathbf{s}_2\mathbf{t}_2}.$$

Then following the same argument,

$$\mathrm{Cov}(\phi_{\mathbf{s}_1\mathbf{t}_1}, \phi_{\mathbf{s}_2\mathbf{t}_2}) = 0, \tag{2.5}$$

for any $\mathbf{t}_1,\ \mathbf{t}_2$ with $\mathbf{t}_1 \neq \mathbf{t}_2$.

(ii) Turn now to the proof of (2.2). The left hand side is

$$\begin{aligned}
\mathrm{Var}\left(\sum_{\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}\psi_{\mathbf{s}}\right) &= \mathrm{Var}\left[\sum_{\mathbf{s}\in\mathfrak{S}}\left(w_{\mathbf{s}}\sum_{\mathbf{t}\subset\mathbf{s}}\phi_{\mathbf{st}}\right)\right] \\
&= \mathrm{Var}\left[\sum_{\mathbf{t}}\sum_{\mathbf{s}\supset\mathbf{t},\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}\phi_{\mathbf{st}}\right] \\
&= \sum_{\mathbf{t}}\mathrm{Var}\left[\sum_{\mathbf{s}\supset\mathbf{t},\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}\phi_{\mathbf{st}}\right]. \qquad (2.6)
\end{aligned}$$

The last line is due to the mutual orthogonality of $\phi_{\mathbf{st}}$ with fixed $\mathbf{s}$ and different $\mathbf{t}$ indices, as shown in (2.5). Notice that for any fixed $\mathbf{t}$, there are at most $r(\mathfrak{S})$ different $\mathbf{s}$ from $\mathfrak{S}$ containing $\mathbf{t}$. Therefore there are at most $r(\mathfrak{S})$ terms in the inner sum of (2.6).

The inequality $(EX)^2 \leq E(X^2)$ implies

$$\begin{aligned}
&\left(\frac{1}{k}\sum_{i=1}^{k} Y_i\right)^2 \leq \frac{1}{k}\sum_{i=1}^{k} Y_i^2 \\
&\Rightarrow \left(\sum_{i=1}^{k} Y_i\right)^2 \leq k\sum_{i=1}^{k} Y_i^2. \qquad (2.7)
\end{aligned}$$

On setting $k = r(\mathfrak{S})$ and $Y_{\mathbf{s}} = w_{\mathbf{s}}(\phi_{\mathbf{st}} - E\phi_{\mathbf{st}})$, (2.7) becomes:

$$\begin{aligned}
&\left[\sum_{\mathbf{s}\supset\mathbf{t},\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}(\phi_{\mathbf{st}} - E\phi_{\mathbf{st}})\right]^2 \leq r(\mathfrak{S})\sum_{\mathbf{s}\supset\mathbf{t},\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}^2(\phi_{\mathbf{st}} - E\phi_{\mathbf{st}})^2 \qquad (2.8) \\
&\Rightarrow \mathrm{Var}\left[\sum_{\mathbf{s}\supset\mathbf{t},\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}\phi_{\mathbf{st}}\right] \leq r(\mathfrak{S})\sum_{\mathbf{s}\supset\mathbf{t},\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}^2\mathrm{Var}(\phi_{\mathbf{st}}).
\end{aligned}$$

Taking the summation over all $\mathbf{t}$, one gets an upper bound for (2.6), thus completing the proof of (2.2):

$$\begin{aligned}
\mathrm{Var}\left[\sum_{\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}\psi_{\mathbf{s}}\right] &\leq \sum_{\mathbf{t}} r(\mathfrak{S}) \sum_{\mathbf{s}\supset\mathbf{t},\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}^2 \mathrm{Var}(\phi_{\mathbf{st}}) \qquad (2.9)\\
&= r(\mathfrak{S}) \sum_{\mathbf{t}} \sum_{\mathbf{s}\supset\mathbf{t},\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}^2 \mathrm{Var}(\phi_{\mathbf{st}})\\
&= r(\mathfrak{S}) \sum_{\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}^2 \sum_{\mathbf{t}\subset\mathbf{s}} \mathrm{Var}(\phi_{\mathbf{st}})\\
&= r(\mathfrak{S}) \sum_{\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}^2 \mathrm{Var}(\psi_{\mathbf{s}}). \qquad (2.10)
\end{aligned}$$

(iii) Now we shall prove that additive decomposability of the functions $\psi_{\mathbf{s}}$ is necessary for the equality sign in (2.2) to hold. Suppose that for an $\mathbf{s}'$, the corresponding choice of function $\psi_{\mathbf{s}'}$ is not additively decomposable. Then on the right hand side of (2.4), there must exist some $\mathbf{t}'$ consisting of at least two indices. Fixing this choice of $\mathbf{t}'$, the inner sum of (2.6) runs over all $\mathbf{s}$ containing $\mathbf{t}'$. Due to the assumption that $r(\mathfrak{S})$ is attained at only one index, that is, no two distinct indices can be found simultaneously in $r(\mathfrak{S})$ elements of $\mathfrak{S}$. This implies that the number of terms in the sum is equal to some number m, $m < r(\mathfrak{S})$. Through (2.7), one sees

$$\mathrm{Var}\left(\sum_{\mathbf{s}\supset\mathbf{t}',\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}\phi_{\mathbf{st}'}\right) \leq m \sum_{\mathbf{s}\supset\mathbf{t}',\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}^2 \mathrm{Var}(\phi_{\mathbf{st}'}) < r(\mathfrak{S}) \sum_{\mathbf{s}\supset\mathbf{t}',\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}^2 \mathrm{Var}(\phi_{\mathbf{st}'}).$$

Since the inequality is strict, the equality sign in (2.9) can not hold. This proves the last claim of the Lemma. □

It is worthwhile to make a few remarks:

Remark 1. In Lemma 2.1.1, $X_1, \ldots, X_n$ are assumed independent, not necessarily identically distributed.

Remark 2. The inequality (2.2) is stronger than what the Cauchy-Schwarz in-

equality directly implies in this setting. For $\mathfrak{S} = \{\mathbf{s}|\mathbf{s} \subset \{1, \ldots, n\}, |\mathbf{s}| = m\}$, the collection of all index sets of size m, $0 < m < n$, one has

$$r(\mathfrak{S}) = \binom{n-1}{m-1} < |\mathfrak{S}| = \binom{n}{m}.$$

Compare (2.2) with the Cauchy-Schwarz inequality. The former provides a tighter bound on $\mathrm{Var}(\sum_{\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}\psi_{\mathbf{s}})$:

$$\mathrm{Var}\left(\sum_{\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}\psi_{\mathbf{s}}\right) \leq r(\mathfrak{S}) \sum_{\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}^2 \mathrm{Var}(\psi_{\mathbf{s}}) < |\mathfrak{S}| \sum_{\mathbf{s}\in\mathfrak{S}} w_{\mathbf{s}}^2 \mathrm{Var}(\psi_{\mathbf{s}}).$$

Remark 3. The inequality (2.2) is actually a property of projection operators in Hilbert spaces. Suppose $\{e_1, \ldots, e_n\}$, $n \geq 1$, is an orthonormal basis set, the span of which is a Hilbert space with some properly defined inner product. If $\{v_1, \ldots, v_k\}$ is a set of arbitrary vectors chosen from the space $\mathrm{span}\{e_1, \ldots, e_n\}$, then there is an obvious inequality

$$\left\| \sum_{i=1}^{k} v_i \right\|^2 \leq k \sum_{i=1}^{k} \|v_i\|^2.$$

Following the idea of Lemma 2.1.1, the above inequality can be sharpened if v_i's do not depend on the whole set of bases. As in the Lemma, given a family of index sets $\mathfrak{S} = \{\mathbf{s}|\mathbf{s} \subset \{1, \ldots, n\}\}$, and the vectors $v_{\mathbf{s}} \in \mathrm{span}\{e_i|i \in \mathbf{s}\}$, then the inequality becomes

$$\left\| \sum_{\mathbf{s}\in\mathfrak{S}} v_{\mathbf{s}} \right\|^2 \leq r(\mathfrak{S}) \sum_{\mathbf{s}\in\mathfrak{S}} \|v_{\mathbf{s}}\|^2.$$

We turn now to the main result of this section.

Theorem 2.1.1 *Let t_n, $n = 1, 2, \ldots$, be the Pitman estimators of θ from a sample of size n from a population $F(x - \theta)$. If for some m, $\mathrm{Var}(t_m) < \infty$, then for all*

$n \geq m$,

$$n\mathrm{Var}(t_n) \geq (n+1)\mathrm{Var}(t_{n+1}). \tag{2.11}$$

For $n \geq 2$, the equality sign holds if and only if F is Gaussian.

Proof. Let $(X_1, \ldots, X_n)$ be a sample from a population $F(x-\theta)$ and assume that $\mathrm{Var}(t_m) < \infty$ for some m. Denote by $t_{n\backslash j}$, $j = 1, \ldots, n$ the Pitman estimator of θ from $(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n)$ (so that, for example $t_{n\backslash 1}$ is the Pitman estimator from $(X_2, \ldots, X_n)$). Plainly, all $t_{n\backslash j}$ are equidistributed with t_{n-1} and, in particular, $\mathrm{Var}(t_{n\backslash j}) = \mathrm{Var}(t_{n-1})$.

Following Hoeffding's idea to define an equivariant U-statistic, the estimator $\sum_{j=1}^{n} t_{n\backslash j}/n$ is equivariant and thus

$$\mathrm{Var}(t_n) \leq \frac{1}{n^2}\mathrm{Var}\left(\sum_{j=1}^{n} t_{n\backslash j}\right). \tag{2.12}$$

Let now $\mathfrak{S}$ be the collection of all the index sets of size $n-1$. Then $r(\mathfrak{S}) = n-1$ and by virtue of Lemma 2.1.1,

$$\mathrm{Var}\left(\sum_{j=1}^{n} t_{n\backslash j}\right) \leq (n-1)\sum_{j=1}^{n}\mathrm{Var}(t_{n\backslash j}) = n(n-1)\mathrm{Var}(t_{n-1}).$$

Combining this with (2.12) proves (2.11).

When F is Gaussian,

$$n\mathrm{Var}(t_n) = (n+1)\mathrm{Var}(t_{n+1}) = \mathrm{Var}(X_1).$$

Conversely, suppose that for some $n \geq m$, $n\mathrm{Var}(t_n) = (n+1)\mathrm{Var}(t_{n+1})$. By part (iii) of the proof of Lemma 2.1.1, the $t_{n\backslash j}$ are additively decomposable for all $j = 1, \ldots, n$:

$$t_{n\backslash j} = \phi_1(X_1) + \ldots + \phi_{j-1}(X_{j-1}) + \phi_{j+1}(X_{j+1}) + \ldots + \phi_n(X_n).$$

As a Pitman estimator, $t_{n\backslash j}$ is symmetric in the X_i's, and since it is equivariant, the ϕ_i are linear: $\phi_j(X_j) = t_1(X_j)/(n-1) = X_j/(n-1)$. In (2.12) we also need

$$t_n = \frac{1}{n}\sum_{j=1}^{n} t_{n\backslash j} = \bar{X}, \text{ with probability 1}$$

for the assumption to hold because the Pitman estimator t_n is unique. The Kagan-Linnik-Rao Theorem asserts that for sample size $n \geq 3$, $t_n = \bar{X}$ only when F is Gaussian, and for $n = 2$, it holds true trivially for any symmetric F. □

If F is a distribution with finite variance, the condition of Theorem 2.1.1 is fulfilled for $m = 1$ (and thus for any m). But $\mathrm{Var}(t_n) < \infty$ holds for many populations with infinite second moment, for example, Cauchy. Note that in Theorem 2.1.1 even absolute continuity of F is not required, not to mention the finiteness of the Fisher information.

One can classify F as regular if

$$\lim_{n\to\infty} n\mathrm{Var}(t_n) > 0$$

and nonregular if the limit (which always exists) is zero. Under very mild conditions on F, if the Fisher information I_{X_1} is finite,

$$n\mathrm{Var}(t_n) \searrow \frac{1}{I_{X_1}}, \text{ as } n \to \infty.$$

Port and Stone (1974) proved that in case of $I_{X_1} = \infty$, the same relation still holds. According to the above result, the convergence here is monotone.

The generalization of (2.11) to the multivariate case is straightforward. Let $(\mathbf{X}_1, \mathbf{X}_2, \ldots)$ be an infinite iid sequence of s-vectors from a population with unknown location parameters $F(x_1 - \theta_1, \ldots, x_s - \theta_s)$ with finite covariance matrix $\mathrm{Var}(\mathbf{X}_1)$.

Given an arbitrary constant s-vector $\mathbf{v} = (v_1, \ldots, v_s)^T$, the Pitman estimator of the linear function $v_1\theta_1 + \ldots + v_s\theta_s$ is defined as in (1.21):

$$\hat{t}_n = \mathbf{v}^T \mathbf{t}_n,$$

where $\mathbf{t}_n$ is the s-variate Pitman estimator of $(\theta_1, \ldots, \theta_s)^T$. By virtue of Theorem 2.1.1,

$$n\mathrm{Var}(\hat{t}_n) = n\mathrm{Var}(\mathbf{v}^T\mathbf{t}_n) = n\mathbf{v}^T\mathrm{Var}(\mathbf{t}_n)\mathbf{v}$$

decreases with n. Since $\mathbf{v}$ is arbitrary, the conclusion of Theorem 2.1.1 for $\hat{t}_n$ implies

$$n\mathrm{Var}(\mathbf{t}_n) \geq (n+1)\mathrm{Var}(\mathbf{t}_{n+1}), \tag{2.13}$$

where the inequality is understood in the sense of positive definiteness.

Return to the discussion on the univariate θ setup. Under the condition $\int x^{2k} dF(x) < \infty$ for some integer $k \geq 1$, the above arguments are extended word for word to the polynomial Pitman estimators $t_n^{(k)}$. The polynomial Pitman estimator $t_{n\backslash j}^{(k)}$ of degree k from $(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n)$ is equidistributed with $t_{n-1}^{(k)}$ and thus

$$\mathrm{Var}(t_{n\backslash j}^{(k)}) = \mathrm{Var}(t_{n-1}^{(k)}).$$

The above inequalities hold with t_n, t_{n-1}, $t_{n\backslash j}$ replaced with $t_n^{(k)}$, $t_{n-1}^{(k)}$, $t_{n\backslash j}^{(k)}$. Hence, $n\mathrm{Var}(t_n^{(k)})$ decreases with n. The family of estimators under consideration grows with k. For fixed n, $\mathrm{Var}(t_n^{(k)})$ decrease with k. Therefore, for any increasing infinite sequence $\{k_1, k_2, \ldots\}$, we have $n\mathrm{Var}(t_n^{(k_n)})$ monotonically decrease with n.

The above proof of monotonicity is due to the fact that the classes in which t_n and $t_n^{(k)}$ are optimal are rather large. To illustrate this, consider a simplified version

of the polynomial Pitman estimator (1.38):

$$\tau_n^{(k)} = \bar{X} - \hat{E}_0(\bar{X}|1, m_2, \ldots, m_k),$$

where $m_j = (1/n)\sum_1^n (X_i - \bar{X})^j$. Of the two estimators $\tau_n^{(k)}$ and $\sum_{j=1}^n \tau_{n\backslash j}^{(k)}/n$, the latter is a polynomial equivariant estimator but not of the form $\bar{X} - \sum_{j=0}^k a_{j,n} m_j$. The inequality (2.12) is not guaranteed in this case. It seems not very likely that $n\mathrm{Var}(\tau_n^{(k)})$ monotonically decreases in n.

The proof of the last claim of Theorem 2.1.1 used the fact that t_n $(n \geq 3)$ is additively decomposable if and only if the underlying distribution is Gaussian. Combined with Theorem 2.1.1, this characterization property of the Gaussian distribution can be refined.

Corollary 2.1.1 *Let both* $\mathbf{X} = (X_1, \ldots, X_n)$ *and* $\mathbf{Y} = (Y_1, \ldots, Y_m)$, $n,\ m \geq 2$, *be independent samples from the same population distribution* $F(x - \theta)$. *Then the Pitman estimator from the pooled sample* $(\mathbf{X}, \mathbf{Y})$ *is a linear combination of the Pitman estimators from* $\mathbf{X}$ *and* $\mathbf{Y}$, *i.e.,*

$$t_{\mathbf{X},\mathbf{Y}} = w_1 t_{\mathbf{X}} + w_2 t_{\mathbf{Y}}, \quad w_1 + w_2 = 1$$

if and only if F *is Gaussian.*

Proof. For sufficiency, simply notice that when F is Gaussian

$$t_{\mathbf{X}} = \bar{X}, \quad t_{\mathbf{Y}} = \bar{Y},$$

and

$$t_{\mathbf{X},\mathbf{Y}} = \frac{n}{n+m}\bar{X} + \frac{m}{n+m}\bar{Y}.$$

To prove the necessity, we start with the following fact

$$\mathrm{Var}(t_{\mathbf{X}}) \geq \frac{n+m}{n}\mathrm{Var}(t_{\mathbf{X},\mathbf{Y}}), \quad \mathrm{Var}(t_{\mathbf{Y}}) \geq \frac{n+m}{m}\mathrm{Var}(t_{\mathbf{X},\mathbf{Y}}), \tag{2.14}$$

by Theorem 2.1.1. Combined together they imply

$$w_1^2\mathrm{Var}(t_{\mathbf{X}}) + w_2^2\mathrm{Var}(t_{\mathbf{Y}}) \geq \left[w_1^2\frac{n+m}{n} + (1-w_1)^2\frac{n+m}{m}\right]\mathrm{Var}(t_{\mathbf{X},\mathbf{Y}}).$$

The right hand side of the inequality is maximized with respect to w_1 when

$$w_1 = n/(m+n)$$

such that

$$w_1^2\mathrm{Var}(t_{\mathbf{X}}) + w_2^2\mathrm{Var}(t_{\mathbf{Y}}) \geq \mathrm{Var}(t_{\mathbf{X},\mathbf{Y}}).$$

The equality sign holds only if both inequalities in (2.14) become equalities. By Theorem 2.1.1 F is Gaussian. □

The coefficients (w_1, w_2) are not specified in the statement of the corollary. Nevertheless, the choice that minimizes $\mathrm{Var}(w_1 t_{\mathbf{X}} + w_2 t_{\mathbf{Y}})$ is uniquely determined by the variances of $t_{\mathbf{X}}$ and $t_{\mathbf{Y}}$:

$$w_1 = \frac{1/\mathrm{Var}(t_{\mathbf{X}})}{1/\mathrm{Var}(t_{\mathbf{X}}) + 1/\mathrm{Var}(t_{\mathbf{Y}})}.$$

When F is Gaussian, the above equality gives $w_1 = n/(n+m)$.

2.2 Superadditivity of $1/\mathrm{Var}(t_n)$

In this section an inequality for $1/\mathrm{Var}(t_n)$ is proved, a special case of which is a small sample version of additivity of the Fisher information. Additivity of the

Fisher information (see (1.2)) means that for independent random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ with distributions $F_1(\mathbf{x} - \theta \cdot \mathbf{1})$ and $F_2(\mathbf{x} - \theta \cdot \mathbf{1})$,

$$I_{\mathbf{X}_1,\mathbf{X}_2} = I_{\mathbf{X}_1} + I_{\mathbf{X}_2}.$$

Consider a more general setup: let $\mathbf{X}_k = (X_{k1}, \ldots, X_{kn_k})$, $k = 1, \ldots, N$, be independent samples from populations $F_k(x - \theta)$ with a common (but unknown) location parameter θ. For an arbitrary index set $\mathbf{s} \subset \{1, \ldots, N\}$, denote by $t_{\mathbf{s}}$ the Pitman estimator of θ constructed from the pooled sample $\{\mathbf{X}_i,\ i \in \mathbf{s}\}$, and by $t_{[N]}$ the Pitman estimator from the complete data set $(X_{11}, \ldots, X_{Nn_N})$. The monotonicity of $\mathrm{Var}(t_{\mathbf{s}})$ in its index $\mathbf{s}$ is obvious:

$$\mathrm{Var}(t_{\mathbf{s}_1}) \leq \mathrm{Var}(t_{\mathbf{s}_2}), \quad \mathbf{s}_1 \subset \mathbf{s}_2,$$

so that trivially the best equivariant estimator of θ from the data is $t_{[N]}$. The following result is much stronger.

Theorem 2.2.1 *Suppose $\mathfrak{S}$ is an arbitrary collection of index sets. Then*

$$\frac{1}{\mathrm{Var}(t_{[N]})} \geq \frac{1}{r(\mathfrak{S})} \sum_{\mathbf{s} \in \mathfrak{S}} \frac{1}{\mathrm{Var}(t_{\mathbf{s}})}. \tag{2.15}$$

Note that no regularity condition is required.

Proof. Set $\psi_{\mathbf{s}} = t_{\mathbf{s}}$ in Lemma 2.1.1. For any choice of weights $\sum_{\mathbf{s}} w_{\mathbf{s}} = 1$, one has

$$r(\mathfrak{S}) \sum_{\mathbf{s}} w_{\mathbf{s}}^2 \mathrm{Var}(t_{\mathbf{s}}) \geq \mathrm{Var}\left(\sum_{\mathbf{s}} w_{\mathbf{s}} t_{\mathbf{s}} \right). \tag{2.16}$$

To minimize the expression on the left, one needs to choose the weights

$$w_{\mathbf{s}} = \frac{\pi_{\mathbf{s}}}{\sum_{\mathbf{c} \in \mathfrak{S}} \pi_{\mathbf{c}}}$$

where $\pi_{\mathbf{s}} = 1/\mathrm{Var}(t_{\mathbf{s}})$. Hence the inequality becomes

$$r(\mathfrak{S})\frac{1}{\sum_{\mathbf{s}} 1/\mathrm{Var}(t_{\mathbf{s}})} \geq \mathrm{Var}\left(\sum_{\mathbf{s}} w_{\mathbf{s}} t_{\mathbf{s}}\right).$$

Finally, note that $\sum_{\mathbf{s}} w_{\mathbf{s}} t_{\mathbf{s}}$ is an equivariant estimator of θ from the complete data set $(\mathbf{X}_1, \ldots, \mathbf{X}_N)$, and thus its variance is not less than $\mathrm{Var}(t_{[N]})$:

$$\mathrm{Var}\left(\sum_{\mathbf{s}} w_{\mathbf{s}} t_{\mathbf{s}}\right) \geq \mathrm{Var}(t_{[N]}).$$

Combining the last two inequalities gives (2.15). □

For disjoint index sets, for example $\mathbf{s}_1 = \{1\}$, ..., $\mathbf{s}_N = \{N\}$, (2.15) reduces to

$$\frac{1}{\mathrm{Var}(t_{[N]})} \geq \frac{1}{\mathrm{Var}(t_{\mathbf{X}_1})} + \ldots + \frac{1}{\mathrm{Var}(t_{\mathbf{X}_N})}.$$

It is a small sample counterpart of the additivity of the Fisher information (1.2).

When all the samples come from the same population $F_k(x-\theta) = F(x-\theta)$, $k = 1, \ldots, N$, (2.15) means superadditivity of $1/\mathrm{Var}(t_n)$ with respect to the sample size.

Corollary 2.2.1 *Let t_n be the Pitman estimator from a sample of size n from population $F(x-\theta)$. If $\mathrm{Var}(t_1) < \infty$, in particular, if $\int x^2 dF(x) < \infty$, then for any n_1 and n_2 with $n_1,\ n_2 \geq 2$*

$$\frac{1}{\mathrm{Var}(t_{n_1+n_2})} \geq \frac{1}{\mathrm{Var}(t_{n_1})} + \frac{1}{\mathrm{Var}(t_{n_2})}. \tag{2.17}$$

The equality sign holds if and only if F is Gaussian.

Proof. Inequality (2.17) is a special case of (2.15), when $N = 2$ and the two samples $\mathbf{X}_1 = (X_1, \ldots, X_{n_1})$, $\mathbf{X}_2 = (X_{n_1+1}, \ldots, X_{n_1+n_2})$ are independently drawn from the same population.

For the equality sign in (2.17) to hold, the equality sign must hold in (2.16):

$$\mathrm{Var}[w_1 t_{\mathbf{X}_1} + w_2 t_{\mathbf{X}_2}] = \mathrm{Var}(t_{\mathbf{X}_1+\mathbf{X}_2}),$$

where $w_1 = \mathrm{Var}(t_{n_2})/(\mathrm{Var}(t_{n_1}) + \mathrm{Var}(t_{n_2}))$, $w_2 = 1 - w_1$. By virtue of the uniqueness of the Pitman estimator,

$$w_1 t_{\mathbf{X}_1} + w_2 t_{\mathbf{X}_2} = t_{\mathbf{X}_1+\mathbf{X}_2}$$

with probability 1. According to Corollary 2.1.1, this equation holds only if F is Gaussian.

Conversely, for samples from a Gaussian population with variance σ^2 and $\mathrm{Var}(t_n) = \sigma^2/n$ so that

$$\frac{1}{\mathrm{Var}(t_{n_1+n_2})} = \frac{n_1}{\sigma^2} + \frac{n_2}{\sigma^2} = \frac{1}{\mathrm{Var}(t_{n_1})} + \frac{1}{\mathrm{Var}(t_{n_2})}. \quad \square$$

The same arguments work for the polynomial Pitman estimators. Indeed, from the definition of $t_n^{(k)}$ one can see that it is the best in the class of estimators

$$\bar{X} + q(X_1 - \bar{X}, \ldots, X_n - \bar{X})$$

where $q(u_1, \ldots, u_n)$ is a polynomial of degree k.

Both $t_m^{(k)}$ and $t_n^{(k)}$ are equivariant polynomials and thus for w_1, w_2 satisfying $w_1 + w_2 = 1$, one has

$$\mathrm{Var}(t_{m+n}^{(k)}) \leq w_1^2 \mathrm{Var}(t_m^{(k)}) + w_2^2 \mathrm{Var}(t_n^{(k)})$$

so that the optimal choice of w_1, w_2 leads to

$$\frac{1}{\mathrm{Var}(t_{m+n}^{(k)})} \geq \frac{1}{\mathrm{Var}(t_m^{(k)})} + \frac{1}{\mathrm{Var}(t_n^{(k)})}.$$

The results are easily extended to the case of a multivariate parameter. If V_n is the covariance matrix of the Pitman estimator of a multivariate parameter $\boldsymbol{\theta}$ from a sample of size n with distribution $F(\mathbf{x}-\boldsymbol{\theta})$, then

$$V_{m+n}^{-1} \geq V_m^{-1} + V_n^{-1}.$$

Note that commutativity of V_m and V_n is not assumed. Moreover, the covariance matrix V_n is invertible for any n if the information matrix $I_{\mathbf{X}_1}$ is well defined as the Cramér-Rao inequality indicates:

$$nV_n \geq I_{\mathbf{X}_1}^{-1} > \mathbf{0}.$$

2.3 Additive perturbations

We first compare the variance of the Pitman estimators from samples from populations $F_1(x-\theta), \ldots, F_N(x-\theta)$ and $F(x-\theta)$ where $F = F_1 * \ldots * F_N$, and then turn to some generalizations of this setup.

2.3.1 Superadditivity of $\mathrm{Var}(t_n)$

First we discuss the superadditivity of $\mathrm{Var}(t_n)$ with respect to the "addition in the samples". Let F_1, F_2 be distribution functions with finite second moments; set $F = F_1 * F_2$. Denote by t'_n, t''_n the Pitman estimators from independent samples $(X'_1, \ldots, X'_n)$, $(X''_1, \ldots, X''_n)$ with distributions $F_1(x-\theta)$, $F_2(x-\theta)$ respectively, and by t_n the Pitman estimator from a sample $(X_1, \ldots, X_n)$ from population $F(x-\theta)$. We have the following result:

Theorem 2.3.1 *Suppose* $\int x^2 dF(x) < +\infty$. *For any* $n \geq 2$

$$\mathrm{Var}(t_n) \geq \mathrm{Var}(t'_n) + \mathrm{Var}(t''_n). \tag{2.18}$$

Proof. To simplify the notations, we define the residuals

$$\begin{aligned} R_n &= (X_1 - \bar{X}, \ldots, X_n - \bar{X}), \\ R'_n &= (X'_1 - \bar{X}', \ldots, X'_n - \bar{X}'), \\ R''_n &= (X''_1 - \bar{X}'', \ldots, X''_n - \bar{X}''). \end{aligned}$$

From (1.16),

$$\mathrm{Var}(t_n) = \mathrm{Var}(\bar{X}) - \mathrm{Var}\{E_0[\bar{X}|R_n]\}.$$

The variances do not depend on θ so that one may assume $\theta = 0$.

$$\begin{aligned} \mathrm{Var}(t_n) &= \mathrm{Var}(\bar{X}) - \mathrm{Var}\{E_0[\bar{X}|R_n]\} \\ &= \mathrm{Var}(\bar{X}') + \mathrm{Var}(\bar{X}'') - \mathrm{Var}\{E_0[\bar{X}|R_n]\}, \\ \mathrm{Var}(t'_n) &= \mathrm{Var}(\bar{X}') - \mathrm{Var}\{E_0[\bar{X}'|R'_n]\}, \\ \mathrm{Var}(t''_n) &= \mathrm{Var}(\bar{X}'') - \mathrm{Var}\{E_0[\bar{X}''|R''_n]\}. \end{aligned}$$

Since $F = F_1 * F_2$,

$$X_1 - \bar{X} \stackrel{d}{=} X'_1 - \bar{X}' + X''_1 - \bar{X}'', \ldots, X_n - \bar{X} = X'_n - \bar{X}' + X''_n - \bar{X}'',$$

that is, $R_n = R'_n + R''_n$. Here $\stackrel{d}{=}$ means "equidistributed". The σ-algebra generated by R_n, $\mathscr{R} = \sigma\{R_n\} = \sigma\{R'_n + R''_n\} = \sigma\{X_1 - \bar{X}, \ldots, X_n - \bar{X}\}$, is a subalgebra of $\tilde{\mathscr{R}} = \sigma\{R'_n, R''_n\} = \sigma\{X'_1 - \bar{X}', X''_1 - \bar{X}'', \ldots, X'_n - \bar{X}', X''_n - \bar{X}''\}$.

That is why

$$E_0(\bar{X}|\mathscr{R}) = E_0\{E_0(\bar{X}|\tilde{\mathscr{R}})|\mathscr{R}\}.$$

By virtue of a well known property of the conditional expectation,

$$\mathrm{Var}[E_0(\bar{X}|\mathscr{R})] \leq \mathrm{Var}[E_0(\bar{X}|\tilde{\mathscr{R}})].$$

Applying Lemma 2.1.2 first to

$$\xi = \bar{X}', \eta_1 = R_n', \eta_2 = R_n''$$

and second to

$$\xi = \bar{X}'', \eta_1 = R_n'', \eta_2 = R_n''$$

we get

$$\begin{aligned}
\mathrm{Var}[E_0(\bar{X}|R_n)] &= \mathrm{Var}[E_0(\bar{X}' + \bar{X}''|\mathscr{R})] \\
&\leq \mathrm{Var}[E_0(\bar{X}' + \bar{X}''|\tilde{\mathscr{R}})] \\
&= \mathrm{Var}[E_0(\bar{X}'|R_n', R_n'')] + \mathrm{Var}[E_0(\bar{X}''|R_n', R_n'')] \\
&= \mathrm{Var}[E_0(\bar{X}'|R_n')] + \mathrm{Var}[E_0(\bar{X}''|R_n'')],
\end{aligned}$$

since R_n' and R_n'' are independent. Hence,

$$\begin{aligned}
\mathrm{Var}(t_n) &\geq \mathrm{Var}(\bar{X}') + \mathrm{Var}(\bar{X}'') - \mathrm{Var}[E_0(\bar{X}'|R_n')] - \mathrm{Var}[E_0(X''|R_n'')] \\
&= \mathrm{Var}(t_n') + \mathrm{Var}(t_n''). \quad \square
\end{aligned}$$

This is a result for small samples that requires only the existence of the second moments of the observations (even the absolute continuity of distributions is not assumed).

It is worthwhile to notice that (2.18) follows directly from a general property of the Pitman estimator of a multivariate parameter. Suppose the random elements remain the same, that is, $\epsilon_i' \sim F_1$, $\epsilon_i'' \sim F_2$, but the location parameters in the first and second samples are different:

$$\begin{aligned} X_i' &= \theta_1 + \epsilon_i',\ \epsilon_i' \sim F_1, \\ X_i'' &= \theta_2 + \epsilon_i'',\ \epsilon_i'' \sim F_2. \end{aligned}$$

Let t_n' be the Pitman estimator of θ_1 from the sample $(X_1', \ldots, X_n')$, and let t_n'' be the Pitman estimator of θ_2 from $(X_1'', \ldots, X_n'')$. By definition (1.20) one can easily see that (t_n', t_n'') is the Pitman estimator of the bivariate parameter (θ_1, θ_2). That is, (t_n', t_n'') minimizes the covariance matrix in the class. By virtue of (1.21) $t_n' + t_n''$ is the best equivariant estimator of the sum $\theta_1 + \theta_2$, while t_n is an equivariant estimator of $\theta_1 + \theta_2$ from the same data $\{(X_1', X_1''), \ldots, (X_n', X_n'')\}$. Thus (2.18) follows immediately:

$$\mathrm{Var}(t_n' + t_n'') = \mathrm{Var}(t_n') + \mathrm{Var}(t_n'') \leq \mathrm{Var}(t_n).$$

The Pitman estimator of $\theta_1 + \theta_2$ is unique with probability 1. If the equality sign in the above inequality holds true so that

$$\mathrm{Var}[t_n'(X_1', \ldots, X_n') + t_n''(X_1'', \ldots, X_n'')] = \mathrm{Var}(t_n(X_1' + X_1'', \ldots, X_n' + X_n'')),$$

then

$$t_n'(X_1', \ldots, X_n') + t_n''(X_1'', \ldots, X_n'') = t_n(X_1' + X_1'', \ldots, X_n' + X_n'') \text{ a.s.} \qquad (2.19)$$

This is a Cauchy type functional equation. If the equality sign holds for all real values X_i' and X_i'', then the traditional theory of functional equations guarantees that the

functions t'_n, t''_n and t_n are all linear in their arguments (see Aczél (1966)). However, the Cauchy type functional equations for random variables have special features. Let X, Y be independent random variables having continuous distributions. The question of whether an equation

$$f(X) + g(Y) = h(X + Y) \tag{2.20}$$

holding with probability 1 where f, g and h are measurable functions implies

$$P\{f(X) = a_1 X + b_1\} = 1,\ P\{g(Y) = a_2 Y + b_2\} = 1$$

for some constants $a_1,\ b_1,\ a_2,\ b_2$ has a negative answer.

Indeed, let ξ be a uniform random variable on $(0, 1)$. Consider its dyadic expression

$$\xi = \sum_{k=1}^{\infty} \frac{\xi_k}{2^k},$$

where $\xi_1, \xi_2, \ldots$ are independent binary random variables with $P(\xi_k = 0) = P(\xi_k = 1) = 1/2$. Now set

$$X = \sum_{k \text{ even}} \frac{\xi_k}{2^k},\ Y = \sum_{k \text{ odd}} \frac{\xi_k}{2^k}.$$

Then X and Y are independent random variables with continuous (though singular) distributions and they both are functions of $X + Y = \xi$ (X and Y are *strong* components of ξ in terminology of Hoffmann-Jorgensen *et al.* (2007)). Thus, for any measurable functions f and g, the relation (2.20) holds.

On the other hand, if both X and Y have almost everywhere positive (with respect to the Lebesgue measure) densities and if f, g and h are measurable locally integrable functions, then the equation (2.20) has only linear solutions f, g (and certainly h). The proof is as follows.

Under the above assumptions, one has

$$f(x) + g(y) = h(x + y) \tag{2.21}$$

almost everywhere with respect to the Lebesgue measure. Take a smooth function $k(x)$ with compact support, multiply both sides of (2.21) by $k(x)$ and integrate over x. Then

$$\int_{-\infty}^{+\infty} f(x)k(x)dx + g(y)\int_{-\infty}^{+\infty} k(x)dx = \int_{-\infty}^{+\infty} h(x+y)k(x)dx = \int_{-\infty}^{+\infty} h(u)k(u-y)du,$$

where the right hand side is continuous in y. Thus, $g(y)$ is continuous, and so is $f(x)$, implying that (2.21) holds everywhere. It becomes the classical Cauchy equation that has only linear solutions. This idea is due to Hillel Furstenberg.

Returning to (2.19) and noticing that $E|t'_n| < \infty$, $E|t''_n| < \infty$, one concludes that if F_1 and F_2 are absolutely continuous with positive densities, then for almost all fixed values $X'_2 = x'_2$, ..., $X''_n = x''_n$

$$t'_n(X'_1, x'_2, \ldots, x'_n) + t''_n(X''_1, x''_2, \ldots, x''_n) = t_n(X'_1 + X''_1, x'_2 + x''_2, \ldots, x'_n + x''_n)$$

with probability 1 in terms of X'_1 and X''_1. This implies

$$t'_n(X'_1, x'_2, \ldots, x'_n) = CX'_1 + D',$$

$$t''_n(X''_1, x''_2, \ldots, x''_n) = CX''_1 + D'',$$

for some constants C, $D' = D'(x'_2, \ldots, x'_n)$ and $D'' = D''(x''_2, \ldots, x''_n)$. Due to the symmetry within t'_n and t''_n with respect to the arguments, the above equations imply the linearity of these Pitman estimators. Thus, (2.19) holding for $n \geq 3$ characterizes the Gaussian distributions F_1 and F_2.

Similar results hold for the polynomial versions of the Pitman estimators (see Section 1.10). Assuming that for some positive integer k,

$$\mu_{2k} = \int x^{2k} dF_i(x) < \infty, \ i = 1, 2.$$

The polynomial degree k Pitman estimators are defined as

$$t_n^{(k)} = \bar{X} - \hat{E}_0(\bar{X}|\Lambda_k(R_n))$$

where $\hat{E}_0(\cdot|\Lambda_k(R_n))$ is the projector into the space of all polynomials of degree k in the residuals R_n with obvious changes for $t_n^{(k)'}$ and $t_n^{(k)''}$. With these definitions, we see the following facts:

(i) $\Lambda_k(R_n) = \Lambda_k(R_n' + R_n'') \subset \Lambda_k(R_n', R_n'')$.

Hence for any random variable ξ with $E|\xi|^2 < \infty$, its projection into the smaller (Hilbert) space has a smaller norm:

$$\mathrm{Var}[\hat{E}_0(\xi|\Lambda_k(R_n' + R_n''))] \leq \mathrm{Var}[\hat{E}_0(\xi|\Lambda_k(R_n', R_n''))]. \tag{2.22}$$

(ii) Let ξ be a random variable such that the pair (ξ, R_n') is independent of R_n''. Then

$$\hat{E}_0[\xi|\Lambda_k(R_n', R_n'')] = \hat{E}_0[\xi|\Lambda_k(R_n')]. \tag{2.23}$$

This is a linear analog of Lemma 2.1.2 and its proof follows directly from the definition of the projection.

Based on (i), (ii), the key step in the proof of Theorem 2.3.1 can be repeated:

$$\begin{aligned}
\mathrm{Var}[\hat{E}_0(\bar{X}|\Lambda_k(R_n))] &= \mathrm{Var}[\hat{E}_0(\bar{X}' + \bar{X}''|\Lambda_k(R_n))] \\
&\leq \mathrm{Var}[\hat{E}_0(\bar{X}' + \bar{X}''|\Lambda_k(R_n', R_n''))] \\
&= \mathrm{Var}[E_0(\bar{X}'|\Lambda_k(R_n'))] + \mathrm{Var}[E_0(\bar{X}''|\Lambda_k(R_n''))],
\end{aligned}$$

resulting in

$$\mathrm{Var}(t_n^{(k)}) \geq \mathrm{Var}(t_n^{(k)'}) + \mathrm{Var}(t_n^{(k)''})$$

and

$$\frac{1}{I_{X_1}^{(k)}} \geq \frac{1}{I_{X_1'}^{(k)}} + \frac{1}{I_{X_1''}^{(k)}}$$

where $I^{(k)}$ is the polynomial Fisher information defined in (1.34).

2.3.2 Another proof of the Stam inequality

Let us see what it gives in large samples assuming that the Fisher information $I_{X'},\ I_{X''}$ and I_X on θ in $X_i',\ X_i''$ and X_i, respectively, is finite. According to (1.30),

$$\begin{aligned}
\mathrm{Var}(t_n') &= \frac{1}{I_{X'}}(1+o(1)),\\
\mathrm{Var}(t_n'') &= \frac{1}{I_{X''}}(1+o(1)),\\
\mathrm{Var}(t_n) &= \frac{1}{I_X}(1+o(1)).
\end{aligned}$$

Thus, combining this with the result of Theorem 2.3.1, one gets the Stam inequality (1.27): for independent $X',\ X''$ and $X = X' + X''$

$$\frac{1}{I_X} \geq \frac{1}{I_{X'}} + \frac{1}{I_{X''}}.$$

The original proof of the Stam inequality is based on the following property of the Fisher score J: for independent X' and X''

$$J(X'|X'+X'') = J(X'+X''), \tag{2.24}$$

as one can see from

$$\begin{aligned}
E[J(X')e^{it(X'+X'')}] &= E[J(X')e^{itX'}]E[e^{itX''}] = -E[ite^{itX'}]E[e^{itX''}]\\
&= -E[ite^{it(X'+X'')}] = E[J(X'+X'')e^{it(X'+X'')}],
\end{aligned}$$

which holds for any t.

There is a simple and elegant proof by Zamir (1998) where the Stam inequality is obtained directly from the basic properties of the Fisher information (monotonicity, additivity and the reparametrization formula (1.6)).

Zamir's proof of (1.27): Let $w_1,\ w_2$ be positive numbers with $w_1 + w_2 = 1$ and take independent observations X_i' of the form

$$X_i' = w_i\theta + X_i,\ i = 1, 2$$

with $\theta \in \mathbb{R}$ as a parameter and $X_1,\ X_2$ independent. Due to the reparametrization formula (1.6),

$$I_{X_i'}(\theta) = w_i^2 I_{X_i}(\theta),\ i = 1, 2.$$

Consider now a statistic

$$T(X_1',\ X_2') = X_1' + X_2' = \theta + X_1 + X_2.$$

Due to monotonicity and additivity of the Fisher information,

$$I_{X_1+X_2} = I_{X_1'+X_2'} \leq I_{X_1'} + I_{X_2'} = w_1^2 I_{X_1} + w_2^2 I_{X_2}.$$

On choosing

$$w_i = \frac{1/I_{X_i}}{1/I_{X_1} + 1/I_{X_2}},\ i = 1, 2,$$

one immediately gets the Stam inequality

$$\frac{1}{I_{X_1+X_2}} \geq \frac{1}{I_{X_1}} + \frac{1}{I_{X_2}}. \quad \square$$

Kagan generalized the proof to the multivariate case when $\mathbf{X} = (X_1, \ldots, X_s)$ is an s-variate random vector with density $p(\mathbf{x} - \boldsymbol{\theta}) = p(x_1 - \theta_1, \ldots, x_s - \theta_s)$ depending

on an s-variate location parameter $\boldsymbol{\theta}\in\mathbb{R}^s$. The matrix $\tilde{I}_{\mathbf{X}}$ of Fisher information on $\boldsymbol{\theta}$ in $\mathbf{X}$ does not depend on $\boldsymbol{\theta}$:

$$\tilde{I}_{\mathbf{X}} = (I_{rq})_{r,q=1,\dots,s},\ I_{rq} = \int_{\mathbf{x}:p(\mathbf{x})>0} \frac{1}{p}\left(\frac{\partial p}{\partial x_r}\right)\left(\frac{\partial p}{\partial x_q}\right) d\mathbf{x}. \tag{2.25}$$

Note that $\tilde{I}_{\mathbf{X}}$ is positive definite (the matrix $\tilde{I}_{\mathbf{X}}(\theta)$ of Fisher information on a general s-variate parameter, not necessarily location, is non-negative definite).

Indeed, take a nonzero $\mathbf{c}\in\mathbb{R}^s$ and consider a random vector $\tilde{\mathbf{X}}$ with density $p(x_1-c_1\theta,\dots,x_s-c_s\theta)$. Plainly, $I_{\tilde{\mathbf{X}}}(\theta)=\mathbf{c}^T\tilde{I}_{\mathbf{X}}\mathbf{c}$ and due to monotonicity, one has $I_{\tilde{\mathbf{X}}}(\theta)\geq I_{\tilde{X}_r}(\theta)$. The density of the r-th component $\tilde{X}_r$ of $\tilde{\mathbf{X}}$ is $p_r(x_r-c_r\theta)$ so that $I_{\tilde{X}_r}(\theta)>0$ if $c_r\neq 0$. Hence $\tilde{I}_{\mathbf{X}}$ is positive definite.

Now let W_1, W_2 be $(s\times s)$ matrices with $W_1+W_2=I_s$, the $(s\times s)$ identity matrix. Set

$$\mathbf{X}_i' = W_i\theta + \mathbf{X}_i,\ i=1,2,$$

where $\mathbf{X}_1$, $\mathbf{X}_2$ are independent s-variate random vectors and $\boldsymbol{\theta}\in\mathbb{R}^s$. Using the basic properties of the Fisher information, one gets

$$\tilde{I}_{\mathbf{X}_1+\mathbf{X}_2} = \tilde{I}_{\mathbf{X}_1'+\mathbf{X}_2'} \leq \tilde{I}_{\mathbf{X}_1'}(\theta)+\tilde{I}_{\mathbf{X}_2'}(\theta) = W_1^{\mathrm{T}}\tilde{I}_{\mathbf{X}_1}W_1 + W_2^{\mathrm{T}}\tilde{I}_{\mathbf{X}_2}W_2.$$

Choosing

$$W_i = (\tilde{I}_{\mathbf{X}_i})^{-1}\{(\tilde{I}_{\mathbf{X}_1})^{-1}+(\tilde{I}_{\mathbf{X}_2})^{-1}\}^{-1},\ i=1,2,$$

results in

$$\tilde{I}_{\mathbf{X}_1+\mathbf{X}_2} \leq \{(\tilde{I}_{\mathbf{X}_1})^{-1}+(\tilde{I}_{\mathbf{X}_2})^{-1}\}^{-1}$$

whence, by taking the inverse of both sides, one obtains the multivariate Stam

inequality

$$(\tilde{I}_{\mathbf{X}_1+\mathbf{X}_2})^{-1} \geq (\tilde{I}_{\mathbf{X}_1})^{-1} + (\tilde{I}_{\mathbf{X}_2})^{-1}.$$

On analyzing Zamir's proof of the Stam inequality, one can see that actually the following two properties were used:

(i) The Fisher information on θ_i contained in an observation of X_i with density $p_i(x;\theta_i),\ i = 1,2$ does not depend on $\theta_i \in \Theta = (a,b),\ a \leq 0,\ b > 0$ (one needs $\alpha\Theta \subset \Theta$ for any $\alpha,\ 0 < \alpha < 1$),

$$0 < I_{X_i}(\theta_i) = I_i < \infty,\ i = 1,2.$$

This condition plainly holds in case of location parameters $\theta_1,\ \theta_2$ but it is much more general. If X has a density $p(x;\eta)$ and a new parameter θ is introduced by $\eta = g(\theta)$ so that $\tilde{p}(x;\theta) = p(x;g(\theta))$, then $I_X(\theta) = |g'(\theta)|^2 I_X(\eta)|_{\eta=g(\theta)}$, whence one can construct many families with constant Fisher information. For example, if X has a Poisson distribution with mean η, the reparametrization $\eta = C\theta^2$ stabilizes the information on θ.

(ii) The distribution of a statistic $T = T(X_1, X_2)$ depends on $(\theta_1,\ \theta_2)$ only through $\theta_1 + \theta_2$, i. e., if its distribution is given by a density $p(t;\theta_1,\ \theta_2)$, then

$$p(t;\theta_1,\ \theta_2) = p(t;\theta_1 + \theta_2),\ t \in \mathcal{T}.$$

If $p_i(x;\theta_i) = p_i(x - \theta_i),\ i = 1,2$ and $T(X_1 + X_2) = X_1 + X_2$, (ii) is plainly satisfied.

Theorem 2.3.2 *Assume $X_1,\ X_2$ are independent and conditions (i), (ii) are satisfied. Then the following Stam's type inequality holds for the Fisher information*

$I_T(\theta)$ on θ in T:

$$\frac{1}{I_T(\theta)} \geq \frac{1}{I_1} + \frac{1}{I_2}.$$

Proof. Take positive $w_1,\ w_2$ with $w_1 + w_2 = 1$ and set $\theta_1 = w_1\theta, \theta_2 = w_2\theta$. Then $\theta_1 + \theta_2 = \theta$. One has $I_{X_i}(\theta) = w_i^2 I_i,\ i = 1, 2$ and due to monotonicity and additivity of the Fisher information,

$$I_T(\theta) \leq I_{X_1}(\theta) + I_{X_2}(\theta) = w_1^2 I_1 + w_2^2 I_2.$$

Choosing

$$w_i = \frac{1/I_i}{1/I_1 + 1/I_2},\ i = 1, 2$$

leads to the claim of Theorem 2.3.2. □

The idea of Theorem 2.3.2 works in various setups, as the following example demonstrates.

Let independent random variables $X_1,\ X_2$ have densities $\theta_1 p_1(\theta_1 x),\ \theta_2 p_2(\theta_2 x)$ depending on scale parameters $\theta_1,\ \theta_2 \in \mathbb{R}_+$. If the distributions of X_1 and X_2 are concentrated on $\mathbb{R}_+$ or $\mathbb{R}_-$, the setup is reduced to that of location parameters. This assumption is not made here.

Let $T(X_1,\ X_2) = X_1 X_2$. It is easily seen that the distribution of T depends on $\theta_1,\ \theta_2$ only through the scale parameter $\theta = \theta_1\theta_2$,

$$p(t; \theta) = \theta p(\theta x).$$

Simple calculations show that

$$I_{X_i}(\theta_i) = \theta_i^{-2} I_{X_i}(1),\ i = 1, 2;$$

$$I_T(\theta) = \theta^{-2} I_T(1).$$

Now set $\theta_1 = \theta^{\gamma_1},\ \theta_2 = \theta^{\gamma_2}$ with $\gamma_i > 0,\ \gamma_1 + \gamma_2 = 1$. Then $\theta_1\theta_2 = \theta$ and

$$I_{X_i}(\theta) = (\gamma_i \theta^{\gamma_i - 1})^2 I_{X_i}(\theta_i) = \gamma_i^2 \theta^{-2} I_{X_i}(1),\ i = 1, 2.$$

One has

$$I_T(\theta) \leq I_{X_1}(\theta) + I_{X_2}(\theta)$$

whence

$$I_T(1) \leq \gamma_1^2 I_{X_1}(1) + \gamma_2^2 I_{X_2}(1).$$

Recently Madiman and Barron (2006) proved a much stronger version of the Stam inequality: for independent (not necessarily identically) distributed $X_1, \ldots, X_n$

$$\frac{1}{I_{X_1+\ldots+X_n}} \geq \frac{1}{\binom{n-1}{m-1}} \sum_{s \in S} \frac{1}{I_{\sum_{i \in s} X_i}}, \tag{2.26}$$

where S is the set of all combinations of m elements chosen from $\{1, \ldots, n\}$.

One of the corollaries of (2.26) is monotone decreasing in n of the information $I_{(X_1+\ldots+X_n)/\sqrt{n}} = n I_{X_1+\ldots+X_n}$ in the normalized sum of independent identically distributed $X_1, X_2, \ldots$. For $m = n - 1$ (2.26) becomes

$$\frac{1}{I_{X_1+\ldots+X_n}} \geq \frac{n}{n-1} \frac{1}{I_{X_1+\ldots+X_{n-1}}}$$

whence for those X_i's with finite Fisher information I_X,

$$(n-1) I_{X_1+\ldots+X_{n-1}} \geq n I_{X_1+\ldots+X_n},\ \text{as } n \to \infty, \tag{2.27}$$

a much stronger result than simple monotone decreasing in n of $I_{X_1+\ldots+X_n}$. Even further, the equality sign in (2.27) holds true if and only if the observations X_i are Gaussian.

As corollary of (2.26) the following facts are worth mentioning.

Fact 1. It provides an alternative ("Fisher information") proof of the Central Limit Theorem (See Barron (1986)).

Fact 2. Suppose $H(X)$ is the (Shannon's) entropy in a random variable X with $\mathrm{Var}(X) = \sigma^2$, and $X_t = X + \sqrt{t}Z$, where t is an arbitrary constant and $Z \sim N(0, \sigma^2)$ is independent of X. Then de Bruijn's identity holds (see, e.g., Madiman and Barron (2005))

$$H(X) = \frac{1}{2}\ln(2\pi e) - \frac{1}{2}\int_0^{+\infty}\left[I_{X_t} - \frac{1}{1+t}\right]dt.$$

Combine this with (2.26), one has the monotonicity in the entropy

$$H\left(\frac{X_1 + \ldots + X_n}{\sqrt{n}}\right) \geq H\left(\frac{X_1 + \ldots + X_{n-1}}{\sqrt{n-1}}\right).$$

This inequality with n replaced by 2^k, and $n-1$ by 2^{k-1}, $k = 2, 3, \ldots$ was proved in Shannon (1948). However, a proof for an arbitrary n was obtained only in Artstein, Ball, Barthe and Noar (2004).

2.3.3 Fisher score under additive perturbations

As mentioned in the last section, the Stam inequality is based on the relation (2.24)

$$J(X+Y) = E[J(X)|X+Y], \tag{2.28}$$

where $X \sim F_1(x-\theta)$ is independent of $Y \sim F_2(y)$, and $J(X)$, $J(X+Y)$ are the Fisher scores from F_1 and $F_1 * F_2$, respectively. In order to generalize the Stam inequality to the case of polynomial information, one needs to find a similar relation for the polynomial score.

Recall that in (1.33) the polynomial score of X is defined as the projection of the Fisher score $J(X)$ onto the polynomial space $\mathbf{P}_k(X) = \text{span}\{X^j | j = 0, \ldots, k\}$:

$$J^{(k)}(X) = \hat{E}[J(X)|\mathbf{P}_k(X)] = \sum_{i=0}^{k} a_i X^i,$$

so that for any integer n, $0 \leq n \leq k$

$$E[J^{(k)}(X) \cdot X^n] = -E\left[\frac{d}{dX} X^n\right].$$

Following (1.33), it is straightforward to define the polynomial score in the sum

$$J^{(k)}(X+Y) = \hat{E}[J(X+Y)|\mathbf{P}_k(X+Y)].$$

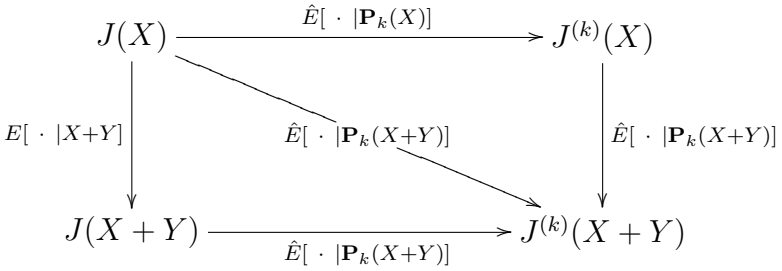
Kagan (2002) proved the polynomial version of (2.28)

$$J^{(k)}(X+Y) = \hat{E}[J^{(k)}(X)|\mathbf{P}_k(X+Y)]$$

by verifying the equality

$$E[J^{(k)}(X) \cdot (X+Y)^n] = E[J^{(k)}(X+Y) \cdot (X+Y)^n]$$

for any n, $0 \leq n \leq k$. All these relations can be demonstrated in the following diagram



Notice that in this diagram, the arrows are commutative.

It is interesting to see that the relation shown in the diagram does not only hold in the polynomial space, but is also true in many other finite dimensional spaces. For instance, in the space spanned by trigonometric functions up to order k

$$\mathbf{T}_k(X) = \text{span}\{1, \sin X, \cos X, \sin 2X, \cos 2X, \ldots, \sin kX, \cos kX\},$$

the trigonometric score is defined as

$$J_T^{(k)}(X) = \hat{E}[J(X)|\mathbf{T}_k(X)] = \sum_{i=0}^{k} a_i \sin(iX) + b_i \cos(iX), \tag{2.29}$$

so that $E[J_T^{(k)}(X)] = 0$ and for any integer n, $1 \leq n \leq k$

$$E[J_T^{(k)}(X) \cdot \sin(nX)] = -nE[\cos(nX)],$$
$$E[J_T^{(k)}(X) \cdot \cos(nX)] = nE[\sin(nX)].$$

According to definition (2.29), one may calculate

$$\begin{aligned}
& E[J_T^{(k)}(X) \cdot \sin n(X+Y)] \\
= \; & E[J_T^{(k)}(X)(\sin nX \cos nY + \cos nX \sin nY)] \\
= \; & E[J_T^{(k)}(X) \cdot \sin nX]E[\cos nY] + E[J_T^{(k)}(X) \cdot \cos nX]E[\sin nY] \\
= \; & -nE[\cos nX]E[\cos nY] + nE[\sin nX]E[\sin nY] \\
= \; & -nE[\cos n(X+Y)] \\
= \; & E[J_T^{(k)}(X+Y) \cdot \sin n(X+Y)],
\end{aligned}$$

and similarly

$$
\begin{aligned}
& E[J_T^{(k)}(X) \cdot \cos n(X+Y)] \\
= \; & E[J_T^{(k)}(X) \cdot \cos nX]E[\cos nY] - E[J_T^{(k)}(X) \cdot \sin nX]E[\sin nY] \\
= \; & nE[\sin nX]E[\cos nY] + nE[\cos nX]E[\sin nY] \\
= \; & nE[\sin n(X+Y)] \\
= \; & E[J_T^{(k)}(X+Y) \cdot \cos n(X+Y)],
\end{aligned}
$$

whence the trigonometric version of (2.28):

$$J_T^{(k)}(X+Y) = \hat{E}[J_T^{(k)}(X)|\mathbf{T}_k(X+Y)].$$

It is very likely that the Stam inequality holds for the information $I_{Trig} = \mathrm{Var}[J_T^{(k)}(X)]$.

Another example is the space $\mathbf{E}_k(X)$ spanned by functions of the form $e^{mX}\sin nX$ and $e^{mX}\cos nX$, with $0 \le m \le M,\ 0 \le n \le N$ for some fixed $M,\ N$. It is easy to verify the relation

$$\hat{E}[J(X+Y)|\mathbf{E}_k(X+Y)] = \hat{E}\{\hat{E}[J(X)|\mathbf{E}_k(X)]|\mathbf{E}_k(X+Y)\}.$$

2.3.4 A strong version of superadditivity of t_n

Let $\boldsymbol{\epsilon}_k = (\epsilon_{k1}, \ldots, \epsilon_{kn})$, $k = 1, \ldots, N$, be independent n-variate random vectors with distribution $F_k(x_1, \ldots, x_n)$ respectively. When these random vectors are shifted by an unknown location parameter θ, a statistician can observe

$$\mathbf{X}_k = (X_{k1}, \ldots, X_{kn}) = (\epsilon_{k1} + \theta, \ldots, \epsilon_{kn} + \theta),\ k = 1, \ldots, N.$$

Or even worse, one can only pick up the information after these sources are summed together. Let $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) \stackrel{d}{=} \sum_{k=1}^{N} \boldsymbol{\epsilon}_k \sim F(x_1, \ldots, x_n)$, then the observations

enter as

$$\mathbf{X} = (X_1, \ldots, X_n) = (\epsilon_1 + \theta, \ldots, \epsilon_n + \theta).$$

Assuming

$$\sigma_k^2 = \int (x_1 + \ldots + x_n)^2 dF_k(x_1, \ldots, x_n) < \infty,\ k = 1, \ldots, N$$

(then plainly $\sigma^2 = \int (x_1 + \ldots + x_n)^2 dF(x_1, \ldots, x_n) < \infty$), the Pitman estimator $t_{k,n} = t_{k,n}(X_{k1}, \ldots, X_{kn})$ of θ from the k-th sample can be written as

$$t_{k,n} = \bar{X}_k - E_0(\bar{X}_k | R_k),$$

where $\bar{X}_k = \frac{1}{n} \sum_{i=1}^n X_{ki}$, $R_k = (X_{k1} - \bar{X}_k, \ldots, X_{kn} - \bar{X}_k)$ is the vector of residuals and E_0 stands for the expectation taken for $\theta = 0$. One can easily see that

$$\mathrm{Var}(t_{k,n}) = \sigma_k^2/n^2 - \mathrm{Var}E_0(\bar{X}_k | R_k).$$

Similarly, for the Pitman estimator

$$t_n = \bar{X} - E_0(\bar{X} | R)$$

of θ from $(X_1, \ldots, X_n)$ one has

$$\mathrm{Var}(t_n) = \sigma^2/n^2 - \mathrm{Var}E_0(\bar{X} | R)$$

where $\bar{X}$ and R stand for the sample mean and vector of residuals from $(X_1, \ldots, X_n)$. For an index set $\mathbf{s} \subset \{1, \ldots, N\}$ we set

$$\bar{X}_{\mathbf{s}} = \sum_{k \in \mathbf{s}} \bar{X}_k,\ R_{\mathbf{s}} = \sum_{k \in \mathbf{s}} R_k \text{ (componentwise)} \tag{2.30}$$

and

$$t_{\mathbf{s},n} = \bar{X}_{\mathbf{s}} - E_0(\bar{X}_{\mathbf{s}} | R_{\mathbf{s}}).$$

(Plainly, (2.30) makes sense for any subset $\mathbf{u} \subset \{1, \ldots, N\}$; we will need this later). The latter is the Pitman estimator of θ from

$$X_{\mathbf{s}} = (X_{\mathbf{s}1}, \ldots, X_{\mathbf{s}n}) = \left(\sum_{k \in \mathbf{s}} \epsilon_{k1} + \theta, \ldots, \sum_{k \in \mathbf{s}} \epsilon_{kn} + \theta \right).$$

Theorem 2.3.3 *Under only the condition $\sigma_1^2 < \infty, \ldots, \sigma_N^2 < \infty$, for any $n \geq 1$ and $1 \leq m \leq N$,*

$$\mathrm{Var}(t_n) \geq \frac{1}{\binom{N-1}{m-1}} \sum_{\mathbf{s}} \mathrm{Var}(t_{\mathbf{s},n}). \tag{2.31}$$

where the summation on the right is extended over all unordered sets (combinations) $\mathbf{s}$ of m elements from $\{1, \ldots, N\}$.

Proof. To simplify notations, set $r = \binom{N-1}{m-1}$. One has

$$\begin{aligned} \mathrm{Var}(t_n) &= \sigma^2/n^2 - \mathrm{Var}[E_0(\bar{X}|R)] \\ &= \sum_{k=1}^{N} \sigma_k^2/n^2 - \mathrm{Var}\left[E_0\left(\sum_{k=1}^{N} \bar{X}_k | R \right) \right] \end{aligned}$$

for $\bar{X} = \sum_{k=1}^{N} \bar{X}_k$, and independence is assumed between the different groups. Also the right hand side of (2.31) can be decomposed as

$$\begin{aligned} \frac{1}{r} \sum_{\mathbf{s}} \mathrm{Var}(t_{\mathbf{s},n}) &= \frac{1}{r} \sum_{\mathbf{s}} [\sum_{k \in \mathbf{s}} \sigma_k^2/n^2 - \mathrm{Var}(E_0(\bar{X}_{\mathbf{s}}|R_{\mathbf{s}}))] \\ &= \frac{1}{r} \sum_{\mathbf{s}} \sum_{k \in \mathbf{s}} \sigma_k^2/n^2 - \frac{1}{r} \sum_{\mathbf{s}} \mathrm{Var}[E_0(\bar{X}_{\mathbf{s}}|R_{\mathbf{s}})] \\ &= \sum_{k=1}^{N} \sigma_k^2/n^2 - \frac{1}{r} \sum_{\mathbf{s}} \mathrm{Var}[E_0(\bar{X}_{\mathbf{s}}|R_{\mathbf{s}})] \end{aligned} \tag{2.32}$$

The third equality is due to the fact that each vector $\mathbf{X}_k$, $k = 1, \ldots, N$, is used in exactly $r = \binom{N-1}{m-1}$ Pitman estimators $t_{\mathbf{s},n}$ (in other words, each number k, $1 \leq k \leq N$ appears in exactly r combinations $\mathbf{s}$ of m elements from $\{1, \ldots, N\}$.

From (2.32) one sees that (2.31) is equivalent to

$$\mathrm{Var}\left[E_0\left(\sum_{k=1}^{N}\bar{X}_k|R\right)\right] \le \frac{1}{r}\sum_{\mathbf{s}}\mathrm{Var}[E_0(\bar{X}_{\mathbf{s}}|R_{\mathbf{s}})]. \tag{2.33}$$

On setting $\psi_{\mathbf{s}} = E_0(\bar{X}_{\mathbf{s}}|R_{\mathbf{s}})$ and $w_{\mathbf{s}} = 1/\binom{N}{m}$ for all $\mathbf{s}$ and noticing that $\psi_{\mathbf{s}}$ so defined depends only on $\mathbf{X}_k$, $k \in \mathbf{s}$, one has by virtue of Lemma 2.1.1

$$r\sum_{s}\mathrm{Var}[E_0(\bar{X}_s|R_s)] \ge \mathrm{Var}[\sum_{s}E_0(\bar{X}_s|R_s)]. \tag{2.34}$$

Denote by $\bar{\mathbf{s}}$ the complement of $\mathbf{s}$ in $\{1,\ldots,N\}$. Then $R_{\mathbf{s}}$ and $R_{\bar{\mathbf{s}}}$ depend on disjoint sets $\{\mathbf{X}_k, k \in \mathbf{s}\}$ and $\{\mathbf{X}_l, l \in \bar{\mathbf{s}}\}$ of independent random vectors $\mathbf{X}_1,\ldots,\mathbf{X}_N$ and thus are independent.

By virtue of Lemma 2.1.2,

$$\psi_{\mathbf{s}} = E_0(\bar{X}_{\mathbf{s}}|R_{\mathbf{s}}, R_{\bar{\mathbf{s}}}).$$

From the definition (2.30) of n-variate vectors $R_{\mathbf{s}}$ and $R_{\bar{\mathbf{s}}}$ one has $R = R_{\mathbf{s}} + R_{\bar{\mathbf{s}}}$. Now due to a well known property of the conditional expectation,

$$E_0(\bar{X}_{\mathbf{s}}|R) = E_0[E_0(\bar{X}_{\mathbf{s}}|R_{\mathbf{s}}, R_{\bar{\mathbf{s}}})|R] = E_0[E_0(\bar{X}_{\mathbf{s}}|R_{\mathbf{s}})|R].$$

Since for any random variable ξ and random element η

$$\mathrm{Var}(\xi) \ge \mathrm{Var}(E(\xi|\eta),$$

the previous relation results in

$$
\begin{aligned}
\mathrm{Var}[\sum_{\mathbf{s}} E_0(\bar{X}_{\mathbf{s}}|R_{\mathbf{s}})] &\geq \mathrm{Var}[E_0(\sum_{\mathbf{s}} E_0(\bar{X}_{\mathbf{s}}|R_{\mathbf{s}})|R)] \\
&= \mathrm{Var}[E_0(\sum_{\mathbf{s}} E_0(\bar{X}_{\mathbf{s}}|R_{\mathbf{s}}, R_{\bar{\mathbf{s}}})|R)] \\
&= \mathrm{Var}[\sum_{\mathbf{s}} E_0(\bar{X}_{\mathbf{s}}|R)] \\
&= \mathrm{Var}[E_0(\sum_{\mathbf{s}} \bar{X}_{\mathbf{s}}|R)] \\
&= \mathrm{Var}[E_0(r\sum_{k=1}^{N} \bar{X}_k|R)] \\
&= r^2\mathrm{Var}[E_0(\bar{X}|R)]. \qquad (2.35)
\end{aligned}
$$

The first three equalities follow from the properties of the conditional expectation discussed above and the fourth is due to the fact that each k appears in exactly r combinations $\mathbf{s}$. Combining (2.34) with (2.35) gives (2.33). □

It is of special interest to study the simpler setting where independence is also assumed within the $\mathbf{X}_k$'s. Let $\mathbf{X}_k = (X_{k1}, \ldots, X_{kn})$ be a sample of iid observations from $G_k(x-\theta)$ with $\mathrm{Var}(X_{k1}) < \infty$, $k = 1, \ldots, N$. Combining Theorem 2.3.3 with Ibragimov and Has'minskii's asymptotic formula (1.30) for the variance of the Pitman estimators, one gets from (2.33) the strong version (2.26) of the Stam inequality.

Particularly when $G_1 = \ldots = G_N = H$, we have the monotonicity of $\mathrm{Var}(t_n^{(N)})$ with respect to the group number N, in contrast to (2.11) whose monotonicity is with respect to the sample size n.

Corollary 2.3.1 *Let $\sigma_H^2 = \int x^2 dH(x) < \infty$. If $t_n^{(N)}$ is the Pitman estimator of θ*

from a sample of size n from $H^{*N}(x-\theta)$ *where* $H^{*N} = H * \cdots * H$*, then*

$$\frac{\mathrm{Var}[t_n^{(N)}]}{N} \geq \frac{\mathrm{Var}[t_n^{(N-1)}]}{N-1}. \tag{2.36}$$

Here n and N are independent parameters. The inequality (2.36) may be considered a small sample version of inequality (2.27) for the Fisher information.

Proof of Corollary. Choose $m = N-1$ in Theorem 1. Under the conditions of the corollary, $\mathrm{Var}(t_{\mathbf{s},n})$ are the same for all N combinations $\mathbf{s}$ of $N-1$ elements so that (2.33) becomes

$$\mathrm{Var}[t_n^{(N)}] \geq \frac{N}{N-1}\mathrm{Var}[t_n^{(N-1)}]. \quad \square$$

Assuming $\int |x|^\delta dH(x) < +\infty$ for a positive δ, (1.30) asserts

$$n\mathrm{Var}[t_n^{(N)}] \stackrel{n\to\infty}{\longrightarrow} \frac{1}{I_{X_{11}+\ldots+X_{N1}}}.$$

For $n \to \infty$, (2.36) becomes Madiman and Barron's inequality (2.27). That is, the monotonicity of Fisher information

$$(N-1)I_{X_{11}+\ldots+X_{(N-1),1}} \geq NI_{X_{11}+\ldots+X_{N1}}.$$

Now that $(X_{11}, \ldots, X_{Nn})$ is a set of iid data, for any n and N, one has

$$\begin{aligned}
\mathrm{Var}[t_n^{(N)}] &= \mathrm{Var}\left[\sum_{k=1}^N \bar{X}_k - E_0\left(\sum_{k=1}^N \bar{X}_k | R_1 + \ldots + R_N\right)\right] \\
&= \mathrm{Var}\left[\sum_{k=1}^N \bar{X}_k\right] - \mathrm{Var}\left[E_0\left(\sum_{k=1}^N \bar{X}_k | R_1 + \ldots + R_N\right)\right] \\
&= N\sigma_H^2/n - \mathrm{Var}[NE_0(\bar{X}_1 | R_1 + \ldots + R_N)].
\end{aligned}$$

Apply this calculation to (2.36), one easily sees a dissipative property of the conditional expectation.

Corollary 2.3.2 *Suppose* $\sigma_H^2 = \int x^2 dH(x) < \infty$. *For arbitrary* $N > 1$

$$(N-1)\mathrm{Var}[E_0(\bar{X}_1|R_1 + \ldots + R_{N-1})] \geq N\mathrm{Var}[E_0(\bar{X}_1|R_1 + \ldots + R_N)]. \quad (2.37)$$

Since $(\bar{X}_1, R_1, \ldots, R_{N-1})$ and $\mathbf{X}_N$, from which R_N is defined, are independent, monotonicity of $\mathrm{Var}(\bar{X}_1|R_1 + \ldots + R_N)$ follows from

$$\begin{aligned} \mathrm{Var}[E_0(\bar{X}_1|R_1 + \ldots + R_{N-1})] &= \mathrm{Var}[E_0(\bar{X}_1|R_1 + \ldots + R_{N-1}, R_N)] \\ &\geq \mathrm{Var}[E_0(X_{11}|R_1 + \ldots + R_N)]. \end{aligned}$$

Corollary 2.3.2 is much stronger than this.

2.3.5 Variance of $t_{X'+\lambda X''}$ as a function of λ

Let X', X'' be independent random variables. Set $X = X' + \lambda X''$ and denote by $F(x;\lambda)$ the distribution function of X. Hence X'' is viewed as a source of noise with a scalar λ. We are interested in studying the behavior of the variance of the Pitman estimator of θ as a function of λ.

Let $(X_1, \ldots, X_n)$ be a sample from population $F(x-\theta;\lambda)$ with $\theta \in \mathbb{R}$ as a parameter, $\lambda > 0$ known, and construct the Pitman estimator $t_{n,\lambda}$ of θ from $(X_1, \ldots, X_n)$. One would expect that $\mathrm{Var}(t_{n,\lambda})$ monotonically decreases as a function of λ on $(-\infty, 0)$ and increases on $(0, \infty)$. We can prove this for the so called self-decomposable X''. It seems likely that the property does not hold for arbitrary X'' (even when X' is Gaussian) though at the moment we do not have an example.

Recall that a random variable Y is *self-decomposable* if for any c, $0 < c < 1$,

Y is equidistributed with $cY + Z_c$, written

$$Y \stackrel{d}{=} cY + Z_c$$

where Z_c is independent of Y. If $f(t)$ is the characteristic function of Y, self-decomposability is equivalent to a factorization

$$f(t) = f(ct)g_c(t)$$

where $g_c(t)$ is a characteristic function. All random variables having stable distributions are self-decomposable. A self-decomposable random variable is necessarily infinitely divisible. Lukacs (1970) gave necessary and sufficient conditions for self-decomposability in terms of the Lévy spectral functions.

Theorem 2.3.4 *Let X' be an arbitrary random variable with $E(X')^2 < \infty$ and let X'' be a self-decomposable random variable with $E(X'')^2 < \infty$ independent of X'. Let $F(x;\lambda)$ be the distribution function of $X' + \lambda X''$. Then the variance* $\mathrm{Var}(t_{n,\lambda})$ *of the Pitman estimator of θ from a sample of size n from $F(x-\theta;\lambda)$ is increasing in λ on $(0,\infty)$ and decreasing on $(-\infty,0)$.*

The proof of Theorem 2.3.4 is similar to that of Theorem 2.3.1.

Proof. We start with the definition of the residuals

$$\begin{aligned} R'_n &= (X'_1 - \bar{X}', \ldots, X'_n - \bar{X}'), \\ R''_n &= (X''_1 - \bar{X}'', \ldots, X''_n - \bar{X}''). \end{aligned}$$

By definition (1.16)

$$t_{n,\lambda} = \bar{X}' + \lambda\bar{X}'' - E_0[\bar{X}' + \lambda\bar{X}''|R'_n + \lambda R''_n]$$

and

$$\mathrm{Var}(t_{n,\lambda}) = \mathrm{Var}(\bar{X}' + \lambda\bar{X}'') - \mathrm{Var}\{E_0[\bar{X}' + \lambda\bar{X}''|R'_n + \lambda R''_n]\}.$$

If $\lambda_1 > \lambda_2 > 0$, then $\lambda_2 = c\lambda_1$ for some c, $0 < c < 1$. Due to self-decomposability of X'', there exist random variables $Z_{c,1} \ldots, Z_{c,n}$ such that

$$X''_i - \bar{X}'' \stackrel{d}{=} c(X''_i - \bar{X}'') + (Z_{c,i} - \bar{Z}_c)$$

and the random variables $X'_1, \ldots, X'_n, X''_1, \ldots, X''_n, Z_{c,1}, \ldots, \bar{Z}_c$ are independent.

Define R_{Z_c} the vector of residuals for $\mathbf{Z}_c$ in the traditional way. The σ-algebra

$$\sigma\{R'_n + \lambda_1 R''_n\} = \sigma\{R'_n + \lambda_1 c R''_n + \lambda_1 R_{Z_c}\}$$

is smaller than σ-algebra

$$\sigma\{R'_n + \lambda_1 c R''_n, R_{Z_c}\}$$

and thus

$$\mathrm{Var}\{E_0[\bar{X}' + \lambda_1\bar{X}''|R'_n + \lambda_1 R''_n]\} \leq \mathrm{Var}\{E_0[\bar{X}' + \lambda_1 c\bar{X}'' + \lambda_1\bar{Z}_c|R'_n + \lambda_1 c R''_n, R_{Z_c}]\}.$$

The rest of the proof is the same as that of Theorem 2.3.1. □

The result can be considered a small sample version of the following property of the Fisher information. Let $I_\lambda = I_{X'+\lambda X''}$ denote the Fisher information on θ in an observation of $\theta + X' + \lambda X''$. If X', X'' are independent, $I(X') < \infty$ and X'' self-decomposable, then I_λ as a function of λ increases monotonically on $(-\infty, 0)$ and decreases monotonically on $(0, \infty)$.

There is an example of nonself-decomposable X'' when I_λ is not monotone; in the example that follows X' is Gaussian and $I_{X''} = \infty$.

Example 7 *Let X' be from a Gaussian distribution $N(\theta, 1)$ and X'' be a Bernoulli random variable with $P(X'' = 1) = P(X'' = 0) = 1/2$. Then*

$$I_{X'} = 1,$$

$$I_{X'+\lambda X''} \longrightarrow 1, \lambda \to +\infty.$$

Since $I_{X'+\lambda X''}$ is not a constant in λ, the two relations indicate that it can not be monotone in λ. $\square$

Chapter 3

Multivariate observations with a univariate location parameter

In this chapter we discuss estimating $\theta \in \mathbb{R}$ based on samples $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ from an s-variate population $F(\mathbf{x} - \mathbf{1} \cdot \theta)$. Here $\mathbf{X}_i = [X_{i1}, \ldots, X_{is}]^T$, $\mathbf{1} = [1, \ldots, 1]^T$ so that $F(\mathbf{x} - \mathbf{1} \cdot \theta) = F(x_1 - \theta, \ldots, x_s - \theta)$.

Though some results are similar to those in Chapters 1, 2, some others differ from their counterparts obtained for univariate observations depending on a univariate location parameter. For the sake of simplicity of notations, we consider the case of $s = 2$. The generalization to higher dimensions is straightforward.

3.1 Linearity of the Pitman estimator

Let $(\mathbf{X}, \mathbf{Y}) = (X_1, Y_1; \ldots; X_n, Y_n)$ be a sample from population $F(x - \theta, y - \theta)$ with finite variances $E(X_1 + Y_1)^2 < \infty$. Notice that definition (1.16) does not require independence between the observations. Following this definition, the Pitman estimator of θ is

$$t_n(\mathbf{X}, \mathbf{Y}) = T_0(\mathbf{X}, \mathbf{Y}) - E_0[T_0(\mathbf{X}, \mathbf{Y}) | X_1 - \bar{Y}, \ldots, X_n - \bar{Y}, \ldots, Y_1 - \bar{X}, \ldots, Y_n - \bar{X}],$$

where $\bar{X} = \sum_i X_i/n$, $\bar{Y} = \sum_i Y_i/n$ and $T_0(\mathbf{X}, \mathbf{Y})$ is an arbitrary equivariant estimator of θ. Notice that the residual vector $R_{\mathbf{X},\mathbf{Y}} = (X_1 - \bar{Y}, \ldots, Y_n - \bar{X})$, as a statistic, is equivalent to $(X_1 - (\bar{X} + \bar{Y})/2, \ldots, Y_n - (\bar{X} + \bar{Y})/2)$. Also it is worth noticing that the σ-algebra $\sigma\{X_1 - \bar{Y}, \ldots, Y_n - \bar{X}\}$ is bigger than the one used in definition

(1.20) for multivariate parameters: $\sigma\{X_1 - \bar{X}, \ldots, X_n - \bar{X}, Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y}\}$.

On noticing that for a bivariate Gaussian $F(x, y)$, the Pitman estimator of θ is

$$t_n(X_1, \ldots, Y_n) = w_1\bar{X} + w_2\bar{Y},$$

where

$$\begin{aligned} w_1 &= \arg\min_{w_1} \mathrm{Var}[w_1\bar{X} + (1 - w_1)\bar{Y}] \\ &= \frac{\mathrm{Var}(Y_1) - \mathrm{Cov}(X_1, Y_1)}{\mathrm{Var}(X_1) + \mathrm{Var}(Y_1) - 2\mathrm{Cov}(X_1, Y_1)}, \text{ and} \\ w_2 &= 1 - w_1, \end{aligned}$$

and having in mind the problem to be discussed, let us represent $t_n(\mathbf{X}, \mathbf{Y})$ as

$$t_n(\mathbf{X}, \mathbf{Y}) = w_1\bar{X} + w_2\bar{Y} - E_0(w_1\bar{X} + w_2\bar{Y}|R_{\mathbf{X},\mathbf{Y}})$$

The first question to be answered about $t_n(\mathbf{X}, \mathbf{Y})$ is when it is linear in $X_1, \ldots, Y_n$. In other words, when does the relation

$$t_n(\mathbf{X}, \mathbf{Y}) = w_1\bar{X} + w_2\bar{Y} \tag{3.1}$$

hold? By definition, (3.1) is equivalent to the zero regression of $w_1\bar{X} + w_2\bar{Y}$ against the residual

$$E_0(w_1\bar{X} + w_2\bar{Y}|R_{\mathbf{X},\mathbf{Y}}) = 0. \tag{3.2}$$

In the univariate case it was answered by the KLR theorem (see Section 1.9) claiming that for $n \geq 3$ the relation

$$E(\bar{X}|X_1 - \bar{X}, \ldots, X_n - \bar{X}) = const$$

holds if and only if X_i is Gaussian (for $n = 2$, it holds for any symmetric X_i). In the bivariate case, by virtue of the KLR theorem, the relation (3.1) for $n \geq 3$ implies that $w_1X_i + w_2Y_i = Z_i$ is Gaussian. Indeed, since

$$\sigma\{Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z}\} \subset \sigma\{X_1 - \bar{Y}, \ldots, Y_n - \bar{X}\},$$

from (3.1) one has

$$E(\bar{Z}|Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z}) = const,$$

whence Z_i is Gaussian. But unless $X_i,\ Y_i$ are independent, Gaussianity of the linear transformation $w_1X_i + w_2Y_i$ does not imply (bivariate) Gaussianity of (X_i, Y_i). And as one sees in the following theorem, in the multivariate case linearity of the Pitman estimator is no longer a characteristic property of the Gaussian distribution.

Theorem 3.1.1 *Let $(X_1, Y_1; \ldots; X_n, Y_n)$ with $n \geq 3$ be a sample from population $F(x - \theta, y - \theta)$ with finite variances $E(X_1 + Y_1)^2 < \infty$. Denote the characteristic function of F by $\varphi(u_1, u_2) = \int \exp(iu_1x + iu_2y)dF(x, y)$. Then (3.1) or (3.2) holds if and only if the following relation holds in a neighborhood of $(0, 0)$*

$$\varphi(u_1, u_2) = \exp(Q(u_1, u_2) + V(w_2u_1 - w_1u_2)), \tag{3.3}$$

where V is an arbitrary function and Q is a quadratic form

$$Q(u_1, u_2) \propto [u_1\ u_2] \begin{bmatrix} -5w_1^2 + 4w_1^3 + 2w_1 & 3w_1^2 - 3w_1 + 1 \\ 3w_1^2 - 3w_1 + 1 & -4w_1^3 + 7w_1^2 - 4w_1 + 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}. \tag{3.4}$$

Proof. Without loss of generality, we may assume $\theta = 0$ throughout the proof. For any $\mathbf{t}, \mathbf{s} \in \mathbb{R}^n$, (3.2) is equivalent to

$$E\{(w_1\bar{X} + w_2\bar{Y}) \exp(i(\sum_j t_j(X_j - \bar{Y}) + \sum_j s_j(Y_j - \bar{X})))\} = 0.$$

Rewrite the left hand side by separating the independent variables

$$
\begin{aligned}
& E\{(w_1\bar{X}+w_2\bar{Y})\exp(i(\sum_j t_j(X_j-\bar{Y})+\sum_j s_j(Y_j-\bar{X})))\} \\
= \;& E\{(w_1\bar{X}+w_2\bar{Y})\exp(i(\sum_j X_j(t_j-\bar{s})+\sum_j Y_j(s_j-\bar{t})))\} \\
= \;& \frac{1}{n}\sum_k E\{w_1X_k\exp(i(\sum_j X_j(t_j-\bar{s})+\sum_j Y_j(s_j-\bar{t})))+ \\
& . \qquad\qquad w_2Y_k\exp(i(\sum_j X_j(t_j-\bar{s})+\sum_j Y_j(s_j-\bar{t})))\} \\
= \;& \frac{1}{n}\sum_k\{E[w_1X_k\exp(i(X_k(t_k-\bar{s})+Y_k(s_k-\bar{t})))+ \\
& \qquad\qquad w_2Y_k\exp(i(X_k(t_k-\bar{s})+Y_k(s_k-\bar{t})))]\prod_{j\neq k}\varphi(t_j-\bar{s},s_j-\bar{t})\} \\
= \;& \frac{1}{n}\prod_j\varphi(t_j-\bar{s},s_j-\bar{t})\sum_k\frac{1}{\varphi(t_k-\bar{s},s_k-\bar{t})}\{w_1E[X_k\exp(i(X_k(t_k-\bar{s})+ \\
& \qquad\qquad Y_k(s_k-\bar{t})))]+w_2E[Y_k\exp(i(X_k(t_k-\bar{s})+Y_k(s_k-\bar{t})))]\} \\
= \;& \frac{1}{n}\prod_j\varphi(t_j-\bar{s},s_j-\bar{t})\sum_k[w_1\partial_1\ln\varphi(t_k-\bar{s},s_k-\bar{t})+w_2\partial_2\ln\varphi(t_k-\bar{s},s_k-\bar{t})]/i,
\end{aligned}
$$

where the subscripts of the partial differentiation denote the component to which the differentiation is applied.

By assumption the distribution F admits finite second moments, and the characteristic function does not vanish in a neighborhood of $(0,0)$. The function $\ln\varphi$ is at least twice continuously differentiable. The above calculation implies

$$
\sum_{k=1}^{n} w_1\partial_1\ln\varphi(t_k-\bar{s},s_k-\bar{t})+w_2\partial_2\ln\varphi(t_k-\bar{s},s_k-\bar{t})=0,\ \forall\mathbf{s},\mathbf{t}. \tag{3.5}
$$

Denote $h(a,b)=w_1\partial_1\ln\varphi(a,b)+w_2\partial_2\ln\varphi(a,b)$. (3.5) becomes a Cauchy type functional equation after a substitution $a_k=t_k-\bar{s}$, $b_k=s_k-\bar{t}$:

$$
\sum_{k=1}^{n} h(a_k,b_k)=0, \tag{3.6}
$$

for any $(a_1, \ldots, b_n)$ with $\sum_k a_k + \sum_k b_k = 0$. To solve for the function h, we notice the following facts:

(i) $\sum_{k=1}^{n} h(0,0) = 0 \;\Rightarrow\; h(0,0) = 0$.

(ii) Fix all $b_k = B$. Then (3.6) implies

$$\sum_{k=1}^{n} h(A_k - B, B) = 0,$$

for any $(A_1, \ldots, A_n)$ with $\sum_{k=1}^{n} A_k = 0$. When $n \geq 3$ and B fixed, this is a Cauchy functional equation (see Aczél (1966)). Its only (measurable) solution is linear in A.

$$h(A - B, B) = C_B A,$$

for some constant C_B. Change the variables by taking $a = A - B$, $b = B$. We have a solution to (3.6)

$$h(a, b) = C_b(a + b),$$

though the functional form of C_b is to be determined. By virtue of the symmetry in the notations a and b, we have

$$h(a, b) = C_b(a + b) = C_a(a + b),$$

where C_a is supposedly a function of a. The above equation holds true for all a, b. Hence $C_a \equiv C_b$ is a constant independent of the choices of a and b.

Combined with (i) and (ii), (3.5) indicates

$$w_1 \partial_1 \ln \varphi(u_1, u_2) + w_2 \partial_2 \ln \varphi(u_1, u_2) = C(u_1 + u_2), \;\; \forall u_1, u_2 \in \mathbb{R}$$

for some constant C. The PDE can be solved by making the following substitutions

$$\begin{cases} x_1 = w_1 u_1 + w_2 u_2 \\ x_2 = w_2 u_1 - w_1 u_2 \end{cases},$$

which is a rotation of the axis such that the direction of the derivative $w_1\partial_1 + w_2\partial_2$ lies on the x_1 axis:

$$\partial_{x_1} \ln \varphi \left(\frac{w_1 x_1 + w_2 x_2}{w_1^2 + w_2^2}, \frac{w_2 x - w_1 x_2}{w_1^2 + w_2^2} \right) = C[(w_1 + w_2)x_1 + (w_2 - w_1)x_2].$$

Take x_2 as a parameter and solve the first order ODE. Finally we have a formula for the characteristic function

$$\begin{aligned} & \ln \varphi(u_1, u_2) \\ = \; & C[\frac{(w_1 u_1 + w_2 u_2)^2}{2} + (w_2 - w_1)(w_1 u_1 + w_2 u_2)(w_2 u_1 - w_1 u_2)] + V(w_2 u_1 - w_1 u_2). \end{aligned}$$

The first term is a quadratic form in u_1, u_2. Simply expand the quantity within the parentheses, and substitute $w_2 = 1 - w_1$. Then (3.4) follows immediately. □

Remark 1. If φ does not vanish, then under this assumption the representation (3.3) holds not only in a neighborhood of $(0, 0)$ but also for every (u_1, u_2).

Remark 2. We pointed out that (3.1) implies Gaussianity of $w_1 X_i + w_2 Y_i$. Certainly, this follows directly from Theorem 3.1.1 since

$$\begin{aligned} E[e^{it(w_1 X_1 + w_2 X_2)}] &= \varphi(w_1 t, w_2 t) \\ &= \exp[Q(w_1, w_2)t^2 + V(w_2 w_1 t - w_1 w_2 t)] \\ &= \exp[Q(w_1, w_2)t^2], \end{aligned}$$

which is a characteristic function of a Gaussian distribution.

Remark 3. Not every function V makes (3.3) a characteristic function. Trivially, $V \equiv 1$ is such a counterexample. It is necessary to have $V(0) = 0$ in order that

$\varphi(0,0) = 1$. Beyond this, we hardly have any criterion in pointing out the complete class of distributions defined by (3.3). The characteristic functions of bivariate Gaussian random vectors are plainly of the form (3.3). But there are non-Gaussian random vectors whose characteristic functions are of this form. Here is a simple example.

Example 8 *Let $Z_1 \sim N(0, \sigma_1^2)$ be independent of $Z_2 \sim N(0, \sigma_2^2)$, and W a non-Gaussian random variable independent of the pair (Z_1, Z_2). Suppose W has a characteristic function $\varphi_W(t) = e^{V(t)}$. Then the random vector $(Z_1 + W, Z_2 - W)$ has a joint characteristic function*

$$\varphi(u_1, u_2) = \exp(-\frac{1}{2}(u_1^2 + u_2^2) + V(u_1 - u_2)).$$

Theorem 3.1.1 asserts that a sample of random copies of this vector with a location shift θ admits a linear Pitman estimator with $w_1 = w_2 = 1/2$. $\square$

Naturally one may try to break the characteristic function into a Gaussian component $\exp(Q(u_1, u_2))$, which needs a little technical modification to become a characteristic function of a Gaussian distribution, and an independent additive noise $\exp(V(w_2 u_1 - w_1 u_2))$. However, these are not all the distributions with their characteristic functions in the form of (3.3). There may exist some characteristic functions of the form (3.3) which can not be represented as $\exp(Q'(u_1, u_2))f(\mathbf{u})$, such that Q' is a nonpositive definite quadratic form and f is a valid characteristic function.

Remark 4. A few comments were given on the quadratic form Q in the proof. We put aside all the details with the fact that $\varphi(u_1, u_2)$ has to be a characteristic

function. Denote $Q(u_1, u_2) = \sum_{i,j=1,2} q_{ij} u_i u_j$. One sees from (3.4) that

$$q_{11} = C \cdot w_1(4w_1^2 - 5w_1 + 2),$$
$$det \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} = -C^2(2w_1 - 1)^2(2w_1^2 - 2w_1 + 1)^2.$$

For all choices of C and w_1, Q is non definite. We need to modify the definition of Q and V in order to justify the normal characteristic in the factor $\exp(Q(u_1, u_2))$. For some constant C', we can write

$$\begin{aligned} & Q(u_1, u_2) + V(w_2u_1 - w_1u_2) \\ = \; & [Q(u_1, u_2) - C'(w_2u_1 - w_1u_2)^2] + [C'(w_2u_1 - w_1u_2)^2 + V(w_2u_1 - w_1u_2)] \\ \stackrel{\Delta}{=} \; & Q'(u_1, u_2) + V'(w_2u_1 - w_1u_2). \end{aligned}$$

Now that there are three independent variables C, C' and w_1 in the definition of the quadratic form Q', on choosing them in an appropriate way one can get an arbitrary (particularly nonpositive definite) quadratic form Q'.

It is also interesting to see the following fact about Q:

$$\begin{aligned} \frac{q_{22} - q_{12}}{q_{22} + q_{11} - 2 * q_{12}} &= w_1 \\ &= \frac{\mathrm{Var}(Y_1) - \mathrm{Cov}(X_1, Y_1)}{\mathrm{Var}(X_1) + \mathrm{Var}(Y_1) - 2\mathrm{Cov}(X_1, Y_1)}. \end{aligned} \tag{3.7}$$

The coefficient w_1 (and accordingly w_2) is simultaneously defined by the the population variance and the "partial variance" from the Gaussian component in a similar way. This observation leads to the answer to a characterization problem in the next section.

Remark 5. Generally for arbitrary $s \geq 2$, let $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ be a sample of s-variate

observations, $\mathbf{X}_i = (X_{i1}, \ldots, X_{is})^T$, $i = 1, \ldots, n$, and hence $\bar{X}_i = \sum_{k=1}^{n} X_{ki}/n$. One can generalize Theorem 3.1.1 and show that

$$t_n(\mathbf{X}_1, \ldots, \mathbf{X}_n) = w_1\bar{X}_1 + \ldots + w_s\bar{X}_s,$$

for some $w_1, \ \ldots, \ w_s$ with $w_1 + \ldots + w_s = 1$ if and only if the characteristic function of $\mathbf{X}_i$ is in the following form:

$$\varphi(u_1, \ldots, u_s) = \exp(\mathbf{u}^T Q\mathbf{u} + V(\mathbf{u}^T[\mathbf{b}_1, \ldots, \mathbf{b}_{s-1}])),$$

for some symmetric $s \times s$ matrix Q and measurable function V. In particular, $\mathbf{b}_i$, $i = 1, \ldots, s-1$, are s-variate vectors chosen according to $\mathbf{w} = (w_1, \ldots, w_s)^T$ such that they are mutually orthogonal and $[\mathbf{w}, \mathbf{b}_1, \ldots, \mathbf{b}_{s-1}]$ is a projection matrix.

Remark 6. In Kagan and Rao (2005), a problem similar to (3.1) was studied. Let $(X_1, Y_1; \ldots; X_n, Y_n)$, $n \geq 3$ be a sample from $F(x - \theta, y)$ with finite second moment $\int x^2 dF(x) < \infty$. The components Y_i are ancillary with respect to of θ. They studied the linear condition of the Pitman estimator

$$t_n(\mathbf{X}, \mathbf{Y}) = \bar{X} - C\bar{Y} - C_0, \tag{3.8}$$

where C and C_0 are constants. They showed that (3.8) implies the following equation for the characteristic function of (X_i, Y_i):

$$u_1\varphi(u_1, u_2) = C_1\frac{\partial\varphi(u_1, u_2)}{\partial u_1} + C_2\frac{\partial\varphi(u_1, u_2)}{\partial u_2} + iC_3\varphi(u_1, u_2),$$

for some constants C_1, C_2 and C_3. Moreover, if F is absolutely continuous and the density function f is differentiable, then

$$f(x, y) = \exp(A_1x + A_2x^2 + Bxy + C(y)),$$

where A_1, A_2 and B are constants and C is an arbitrary function.

3.2 Some related characterization problems for the linear Pitman estimator

Nice results of characterization problems, particularly those associated with the Gaussian distribution, are abundant in the classic univariate setting. It turns out that those distributions defined by (3.3) play the same role in the multivariate setting as the univariate Gaussian distribution does in one dimension.

Given a univariate sample $(X_1, \ldots, X_n)$, $n \geq 3$, it is well known that the independence between the sample mean $\bar{X}$ and the residual R_X characterizes the Gaussian distribution. The following corollary of Theorem 3.1.1 is an analog of this fact in the multivariate setting.

Corollary 3.2.1 *Let $(X_1, Y_1; \ldots; X_n, Y_n)$, $n \geq 3$, be a sample from $F(x-\theta, y-\theta)$. Then the linear combination $w_1\bar{X} + w_2\bar{Y}$, $w_1 + w_2 = 1$ is independent of the residual $R_{\mathbf{X},\mathbf{Y}}$ if and only if the characteristic function of F is in the form (3.3).*

Proof. Necessity of the condition trivially follows the fact that the independence between $w_1\bar{X} + w_2\bar{Y}$ and $R_{\mathbf{X},\mathbf{Y}}$ implies the zero regression equation (3.2) and hence (3.3) by Theorem 3.1.1. We only prove the sufficiency.

Assume $\theta = 0$ throughout this proof. Then look at the joint characteristic function of $w_1\bar{X} + w_2\bar{Y}$ and $R_{\mathbf{X},\mathbf{Y}}$

$$\begin{aligned} & E[\exp(i((w_1\bar{X} + w_2\bar{Y})r + \sum_k (X_k - \bar{Y})t_k + (Y_k - \bar{X})s_k))] \\ = \; & E[\exp(i(\sum_k (\frac{w_1}{n}r + t_k - \bar{s}) + Y_k(\frac{w_2}{n}r + s_k - \bar{t})))] \\ = \; & \prod_{k=1}^{n} \varphi\left(\frac{w_1}{n}r + t_k - \bar{s}, \frac{w_2}{n}r + s_k - \bar{t}\right), \end{aligned} \tag{3.9}$$

where φ as defined in (3.3) is the characteristic function of F. Then $w_1\bar{X} + w_2\bar{Y}$ is independent of $R_{\mathbf{X},\mathbf{Y}}$ if and only if the above expression can be factored into a product of two functions in r and $(\mathbf{t}, \mathbf{s})$ respectively.

Denote $u_k = w_1 r/n + t_k - \bar{s}$, $v_k = w_2 r/n + s_k - \bar{t}$. Then (3.9) can be written explicitly according to (3.3):

$$\prod_{k=1}^{n} \exp(Q(u_k, v_k) + V(w_2 u_k - w_1 v_k)). \tag{3.10}$$

Note that $w_2 u_k - w_1 v_k = w_2(w_1 r/n + t_k - \bar{s}) - w_1(w_2 r/n + s_k - \bar{t})$ is independent of r. Hence $\prod_k \exp\{V(w_2 u_k - w_1 v_k)\}$ is plainly a function of the pair $(\mathbf{t}, \mathbf{s})$. (3.10) can be factored properly if and only if there are no cross terms between r and $(\mathbf{t}, \mathbf{s})$ in the quadratic form $Q(u_k, v_k)$. Expand the definition

$$\sum_k q_{11}\left(\frac{w_1}{n}r + t_k - \bar{s}\right)^2 + 2q_{12}\left(\frac{w_1}{n}r + t_k - \bar{s}\right)\left(\frac{w_2}{n}r + s_k - \bar{t}\right) + q_{22}\left(\frac{w_2}{n}r + s_k - \bar{t}\right)^2.$$

The only cross terms are

$$(q_{11}w_1 - q_{12}w_1 + q_{12}w_2 - q_{22}w_2)r\bar{t} + (q_{12}w_1 - q_{11}w_1 - q_{12}w_2 + q_{22}w_2)r\bar{s}.$$

Recall that w_1 is defined by the matrix Q in (3.7). Substitute both w_1 and w_2 with their definitions then both terms in the above expression vanish. The proof is complete. □

Next, we will show that sufficiency of a linear statistic $w_1\bar{X}+w_2\bar{Y}$ characterizes those distributions with characteristic function (3.3).

Corollary 3.2.2 *Suppose that $n \geq 2$ and the distribution of $w_1\bar{X}+w_2\bar{Y}$ is absolutely continuous. Then $w_1\bar{X} + w_2\bar{Y}$ is sufficient for θ only if F has a characteristic function (3.3).*

Proof. Sufficiency of $w_1\bar{X} + w_2\bar{Y}$ for θ means that the conditional distribution of $(X_1, Y_1; \ldots; X_n, Y_n)$ given $w_1\bar{X} + w_2\bar{Y}$ is independent of θ. In terms of the characteristic function, the conditional characteristic function of the sample is independent of θ:

$$E_\theta[e^{i(\mathbf{t}\cdot\mathbf{X}+\mathbf{s}\cdot\mathbf{Y})}|w_1\bar{X} + w_2\bar{Y}] = \phi_{\mathbf{t},\mathbf{s}}(w_1\bar{X} + w_2\bar{Y}), \tag{3.11}$$

for some measurable function $\phi_{\mathbf{t},\mathbf{s}}$. Setting $\xi_k = X_k - \theta$ and $\eta_k = Y_k - \theta$. Equality (3.11) leads to a functional equation for $\phi_{\mathbf{t},\mathbf{s}}$

$$\begin{aligned}
& \phi_{\mathbf{t},\mathbf{s}}(w_1\bar{\xi} + w_2\bar{\eta}) \\
= \;& E[\exp(i(\mathbf{t}\cdot\boldsymbol{\xi} + \mathbf{s}\cdot\boldsymbol{\eta}))|w_1\bar{\xi} + w_2\bar{\eta}] \\
= \;& E[.\exp(i(\mathbf{t}\cdot\mathbf{X} - \sum_k t_k\theta + \mathbf{s}\cdot\mathbf{Y} - \sum_k s_k\theta))|w_1\bar{X} + w_2\bar{Y} - \theta] \\
= \;& \exp(-i\sum_k (t_k + s_k)\theta)E[\exp(i(\mathbf{t}\cdot\mathbf{X} + \mathbf{s}\cdot\mathbf{Y}))|w_1\bar{X} + w_2\bar{Y}] \\
= \;& \exp(-i\sum_k (t_k + s_k)\theta)\phi_{\mathbf{t},\mathbf{s}}(w_1\bar{\xi} + w_2\bar{\eta} + \theta).
\end{aligned}$$

Substituting $u = w_1\bar{\xi} + w_2\bar{\eta}$ on both ends

$$\phi_{\mathbf{t},\mathbf{s}}(u)\exp(i\sum_k (t_k + s_k)\theta) = \phi_{\mathbf{t},\mathbf{s}}(u + \theta).$$

Fix the value of θ. The equality may hold only on a set of probability 1, and the exceptional set actually depends on θ. It was shown in Kagan *et al.* (1973) page 284 that if $w_1\bar{X} + w_2\bar{Y}$ has an absolutely continuous distribution, then there exist a set of probability 1 on which the above equality holds for all θ.

Fix u for some value in its domain, and let $v = u + \theta$. Due to the arbitrariness of θ, one may get the following relation for all $v \in \mathbb{R}$

$$\begin{aligned}\phi_{\mathbf{t},\mathbf{s}}(v) &= \phi_{\mathbf{t},\mathbf{s}}(u)\exp(i\sum(t_k+s_k)(v-u)) \\ &= C_{\mathbf{t},\mathbf{s}}\exp(i\sum(t_k+s_k)v),\end{aligned}$$

where $C_{\mathbf{t},\mathbf{s}} = \phi_{\mathbf{t},\mathbf{s}}(u)\exp\left(-i\sum(t_k+s_k)u\right)$ is a function of $(\mathbf{t},\mathbf{s})$.

The joint characteristic function of $(X_1,Y_1;\ldots;X_n,Y_n)$ and $w_1\bar{X}+w_2\bar{Y}$ is

$$\begin{aligned}& E[\exp(ir(w_1\bar{X}+w_2\bar{Y})+i(\mathbf{t}\cdot\mathbf{X}+\mathbf{s}\cdot\mathbf{Y}))] \\ =\ & E[\exp(\sum_k i(w_1r/n+t_k)X_k+i(w_2r/n+s_k)Y_k)] \qquad (3.12)\\ =\ & \prod_{k=1}^{n}\varphi\left(\frac{w_1r}{n}+t_k,\frac{w_2r}{n}+s_k\right),\end{aligned}$$

where φ is the characteristic function of the distribution F. On the other hand, rewrite the expectation in (3.12) by conditioning on $w_1\bar{X}+w_2\bar{Y}$:

$$\begin{aligned}& E[\exp(ir(w_1\bar{X}+w_2\bar{Y})+i(\mathbf{t}\cdot\mathbf{X}+\mathbf{s}\cdot\mathbf{Y}))] \\ =\ & E\{E[\exp(ir(w_1\bar{X}+w_2\bar{Y}))\exp(i(\mathbf{t}\cdot\mathbf{X}+\mathbf{s}\cdot\mathbf{Y}))|w_1\bar{X}+w_2\bar{Y}]\} \\ =\ & E[\exp(ir(w_1\bar{X}+w_2\bar{Y}))\phi(w_1\bar{X}+w_2\bar{Y})] \\ =\ & E[\exp(ir(w_1\bar{X}+w_2\bar{Y}))C_{\mathbf{t},\mathbf{s}}\exp(i(\sum_k t_k+s_k)(w_1\bar{X}+w_2\bar{Y}))] \\ =\ & C_{\mathbf{t},\mathbf{s}}E[\exp(i(r+\sum_k t_k+s_k)(w_1\bar{X}+w_2\bar{Y}))] \\ =\ & C_{\mathbf{t},\mathbf{s}}\varphi[w_1(r+\sum_k t_k+s_k)/n, w_2(r+\sum_k t_k+s_k)/n]^n. \qquad (3.13)\end{aligned}$$

Combining (3.12) and (3.13) gives a functional equation for φ

$$\prod_{k=1}^{n}\varphi\left(\frac{w_1r}{n}+t_k,\frac{w_2r}{n}+s_k\right) = C_{\mathbf{t},\mathbf{s}}\varphi\left[\frac{w_1}{n}\left(r+\sum_k t_k+s_k\right),\frac{w_2}{n}\left(r+\sum_k t_k+s_k\right)\right]^n. \qquad (3.14)$$

For shorter notations, we denote

$$u_k = \frac{w_1}{n} r + t_k,\ v_k = \frac{w_2}{n} r + s_k.$$

Hence

$$\frac{w_1}{n}\left(r + \sum_k t_k + s_k\right) = w_1\left(\bar{u} + \bar{v}\right),\ \frac{w_2}{n}\left(r + \sum_k t_k + s_k\right) = w_2\left(\bar{u} + \bar{v}\right).$$

In a neighborhood of $(0,0)$, φ does not vanish. Take logarithms on both sides of (3.14) and then differentiate with respect to r:

$$\begin{aligned} n\{w_1\partial_1 \ln\varphi[w_1(\bar{u}+\bar{v}), w_2(\bar{u}+\bar{v})] + w_2\partial_2 \ln\varphi[w_1(\bar{u}+\bar{v}), w_2(\bar{u}+\bar{v})]\} \\ = \sum_k [w_1\partial_1 \ln\varphi(u_k, v_k) + w_2\partial_2 \ln\varphi(u_k, v_k)]. \end{aligned}$$

Denote $h(u,v) = w_1\partial_1 \ln\varphi(u,v) + w_2\partial_2 \ln\varphi(u,v)$. We have a Cauchy type functional equation

$$nh[w_1(\bar{u}+\bar{v}), w_2(\bar{u}+\bar{v})] = \sum_{k=1}^{n} h(u_k, v_k),\ \forall u_1, \ldots, v_n. \tag{3.15}$$

By definition, $h(0,0) = 0$. Then fix $u_2 = \ldots = u_n = v_2 = \ldots = v_n = 0$. We have

$$h(u_1, v_1) = nh\left(w_1\frac{u_1+v_1}{n}, w_2\frac{u_1+v_1}{n}\right);$$

that is, for arbitrary (u,v), $h(u,v) = \tilde{h}(u+v)$ is a measurable function in $u+v$. It is sufficient to consider only the sums $z_k = u_k + v_k$, $k = 1, \ldots, n$. Then (3.15) becomes

$$n\tilde{h}(\bar{z}) = \sum_{k=1}^{n} \tilde{h}(z_k),\ \forall z_1, \ldots, z_n.$$

Differentiate both sides with respect to z_1. For arbitrary z_1 and $\bar{z}$ we have

$$\tilde{h}'(z_1) = \tilde{h}'(\bar{z}).$$

Therefore $\tilde{h}'$ =constant. We get a differential equation defining the characteristic function φ

$$w_1\partial_1 \ln \varphi(u,v) + w_2\partial_2 \ln \varphi(u,v) = \tilde{h}(u+v) = C(u+v),$$

for some constant C. It is the same differential equation as in the proof of Theorem 3.1.1, leading to the unique solution (3.3). □

The converse of Corollary 3.2.2 is proved in the next section when F has a positive density $p(x-\theta, y-\theta)$ and finite information $I_{X,Y}$. In Theorem 3.3.1, we will show that if F is given by (3.3), then the Fisher information in the linear function $I_{w_1\bar{X}+w_2\bar{Y}}$ is equal to that in the complete sample $I_{\mathbf{X},\mathbf{Y}}$. As mentioned in Chapter 1, under the assumption of positive density, only sufficient statistics preserve the Fisher information. Therefore $w_1\bar{X} + w_2\bar{Y}$ is sufficient if F is given by (3.3) and has a positive density.

3.3 Linearity of the Fisher score

Suppose F has an absolutely continuous density $p(x-\theta, y-\theta)$. Then the Fisher score is

$$J(X,Y) = \frac{\partial}{\partial\theta} \ln p(X-\theta, Y-\theta),$$

and hence the Fisher information is

$$I_{X,Y} = \mathrm{Var}[J(X,Y)^2] = \int\int \frac{[\partial_1 p(x,y) + \partial_2 p(x,y)]^2}{p(x,y)} dxdy.$$

Denote by $\tilde{I}_{X_1,Y_1}$ the matrix of Fisher information on (θ_1, θ_2) associated with an observation $(X, Y) \sim F(x - \theta_1, y - \theta_2)$, $(\theta_1, \theta_2) \in \mathbb{R}^2$. One can easily see that

$$I_{X,Y} = \mathbf{1}^T \tilde{I}_{X,Y} \mathbf{1}, \tag{3.16}$$

where $\mathbf{1} = (1, 1)^T$.

As shown in Section 1, if the Pitman estimator of θ from a sample $(X_1, Y_1; \ldots;$ $X_n, Y_n)$ from population $F(x - \theta, y - \theta)$ is linear:

$$t_n(\mathbf{X}, \mathbf{Y}) = w_1 \bar{X} + w_2 \bar{Y},$$

then is in a neighborhood of zero the characteristic function of F, of the form (3.3), and if the characteristic function does not vanish, it is of the form (3.3) for all t, s. Here a similar property of the Fisher score is proved.

Theorem 3.3.1 *The Fisher score in the setup $(X, Y) \sim F(x - \theta, y - \theta)$ with finite second moments $E(X + Y)^2 < \infty$ is linear. That is,*

$$J(X, Y) = \frac{w_1 X + w_2 Y - \theta}{c} \tag{3.17}$$

for some constant c if and only if F has a characteristic function (3.3).

Proof. Assuming $E(X+Y)^2 < \infty$, the characteristic function φ of F is differentiable. One has (for $\theta = 0$)

$$E[\exp(iu_1 X + iu_2 Y) \frac{w_1 X + w_2 Y}{c}] = -\frac{i}{c}[w_1 \partial_1 \varphi(u_1, u_2) + w_2 \partial_2 \varphi(u_1, u_2)]. \tag{3.18}$$

On the other hand, by virtue of a well known property of the Fisher score, the covariance between the Fisher score and an arbitrary random element is equivalent

to the negative derivative of the expected value of the same random element:

$$\begin{aligned} & E[\exp(iu_1X + iu_2Y)J(X,Y)] \\ = \; & -i(u_1+u_2)E(\exp(iu_1X+iu_2Y)) \\ = \; & -i(u_1+u_2)\varphi(u_1,u_2). \end{aligned} \tag{3.19}$$

Now (3.17) holds if and only if (3.18) and (3.19) are equal. Divide both equations by $\varphi(u_1, u_2)$. One gets a differential equation for φ:

$$w_1\partial_1 \ln\varphi(u_1,u_2) + w_2\partial_2 \ln\varphi(u_1,u_2) = c(u_1+u_2).$$

Recall that in the proof of Theorem 3.1.1, this same differential equation led to the characteristic function (3.3). □

The linear form in (3.17) is uniquely determined by the underlying distribution F, particularly by the covariance matrix of (X, Y). Since $w_1X + w_2Y$ is an unbiased estimator of θ, one has

$$E[(w_1X + w_2Y)J(X,Y)] = 1.$$

Combined with (3.17),

$$\begin{aligned} \{E[(w_1X+w_2Y)J(X,Y)]\}^2 &= \mathrm{Var}(w_1X+w_2Y)\mathrm{Var}[J(X,Y)] \\ &= \frac{[\mathrm{Var}(w_1X+w_2Y)]^2}{c^2} = 1. \end{aligned}$$

Hence the constant c is

$$c = \mathrm{Var}(w_1X + w_2Y),$$

so that from (3.17), one may calculate the Fisher information

$$I_{X,Y} = \mathrm{Var}[J(X,Y)] = \frac{1}{\mathrm{Var}(w_1X+w_2Y)}. \tag{3.20}$$

Moreover, if we consider an arbitrary linear estimator $\lambda_1 X + \lambda_2 Y$ with $\lambda_1 + \lambda_2 = 1$, then by the Cramér-Rao inequality and (3.20)

$$\mathrm{Var}(\lambda_1 X + \lambda_2 Y) \geq \frac{1}{I_{X,Y}} = \mathrm{Var}(w_1 X + w_2 Y).$$

It turns out that w_1 and w_2 depends on F as in (3.7):

$$(w_1, w_2) = \arg \min_{\lambda_1+\lambda_2=1} \mathrm{Var}(\lambda_1 X + \lambda_2 Y).$$

Next we will study the Fisher information associated with the linear function $w_1\bar{X} + w_2\bar{Y}$. For $(X, Y) \sim F(x - \theta, y - \theta)$, the distribution function of the sum $w_1\bar{X} + w_2\bar{Y} \sim H(z - \theta)$ depends on a location parameter, and the Fisher score from H is a projection of the score from F:

$$J(w_1 X + w_2 Y) = E[J(X, Y)|w_1 X + w_2 Y].$$

Assuming (3.17), $J(X, Y)$ is a function of $w_1 X + w_2 Y$. Therefore

$$J(w_1 X + w_2 Y) = J(X, Y)$$

and it leads to the following result

$$I_{X,Y} = \mathrm{Var}[J(X, Y)] = \mathrm{Var}[J(w_1 X + w_2 Y)] = I_{w_1 X + w_2 Y}. \tag{3.21}$$

When we have a sample of size n, $(X_1, Y_1; \ldots; X_n, Y_n)$, (3.21) becomes

$$I_{\mathbf{X},\mathbf{Y}} = nI_{X,Y} = I_{w_1\bar{X} + w_2\bar{Y}}$$

because $w_1 X + w_2 Y$ is normally distributed as Theorem 3.3.1 claims.

Conversely, (3.21) is not sufficient to characterize the linear score (3.17) or the distributions given by (3.3). $I_{X,Y} = I_{w_1 X + w_2 Y}$ implies that $J(X, Y)$ is a function of $w_1 X + w_2 Y$, but not necessarily linear.

Corollary 3.3.1 *Suppose* $(X, Y) \sim F(x-\theta, y-\theta)$ *with finite second moments* $E(X+Y)^2 < \infty$. *If*

$$I_{X,Y} = I_{w_1X+w_2Y}$$

and w_1X+w_2Y *is normally distributed, then the characteristic function of* F *satisfies (3.3) in a neighborhood of* $(0,0)$.

3.4 Different versions of the Stam inequalities

Let $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}$ be s-variate random vectors. If

$$\mathbf{X}_1 \sim F_1(\mathbf{x}-\boldsymbol{\theta}),\ \mathbf{X}_2 \sim F_2(\mathbf{x}-\boldsymbol{\theta}),\ \mathbf{X} \sim F(\mathbf{x}-\boldsymbol{\theta})$$

with a location vector $\boldsymbol{\theta} \in \mathbb{R}^s$, where $F = F_1 * F_2$, and $\tilde{I}_1$, $\tilde{I}_2$, $\tilde{I}$ are the matrices of Fisher information on $\boldsymbol{\theta}$ in $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}$ respectively, then the multivariate Stam inequality claims (see Section 2.3.2)

$$\tilde{I}^{-1} \geq \tilde{I}_1^{-1} + \tilde{I}_2^{-1}. \tag{3.22}$$

Suppose now that $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}$ are vectors whose distribution depends on a univariate location parameter θ. That is,

$$\mathbf{X}_1 \sim F_1(\mathbf{x}-\theta\cdot\mathbf{1}),\ \mathbf{X}_2 \sim F_2(\mathbf{x}-\theta\cdot\mathbf{1}),\ \mathbf{X} \sim F(\mathbf{x}-\theta\cdot\mathbf{1}).$$

The Fisher information on θ in $\mathbf{X}_1$ is

$$I_1 = \mathbf{1}^T\tilde{I}_1\mathbf{1},$$

and similarly for $\mathbf{X}_2$ and $\mathbf{X}$

$$I_2 = \mathbf{1}^T\tilde{I}_2\mathbf{1},\ I = \mathbf{1}^T\tilde{I}\mathbf{1}.$$

Repeating verbatim the proof of (1.27) in Section 2.3.2, one gets the Stam inequality

$$\frac{1}{\mathbf{1}^T\tilde{I}\mathbf{1}} \geq \frac{1}{\mathbf{1}^T\tilde{I}_1\mathbf{1}} + \frac{1}{\mathbf{1}^T\tilde{I}_2\mathbf{1}}. \tag{3.23}$$

Notice that this inequality differs from a special case of (3.22)

$$\mathbf{1}^T\tilde{I}^{-1}\mathbf{1} \geq \mathbf{1}^T\tilde{I}_1^{-1}\mathbf{1} + \mathbf{1}^T\tilde{I}_2^{-1}\mathbf{1}, \tag{3.24}$$

in light of the fact that

$$\frac{1}{\mathbf{1}^T\tilde{I}\mathbf{1}} \leq \mathbf{1}^T\tilde{I}^{-1}\mathbf{1}.$$

Indeed, due to the Cauchy-Schwarz inequality, for any vector $\mathbf{w} = (w_1, \ldots, w_s)^T$ with $\mathbf{w}^T\mathbf{w} = 1$ and symmetric positive definite matrix $\tilde{I}$

$$1 = \mathbf{w}^T\mathbf{w} = \mathbf{w}^T\tilde{I}^{1/2}\tilde{I}^{-1/2}\mathbf{w} \leq |\mathbf{w}^T\tilde{I}^{1/2}(\tilde{I}^{1/2})^T\mathbf{w}|^{1/2}|\mathbf{w}^T\tilde{I}^{-1/2}(\tilde{I}^{-1/2})^T\mathbf{w}|^{1/2},$$

whence

$$\frac{1}{\mathbf{w}^T\tilde{I}\mathbf{w}} \leq \mathbf{w}^T\tilde{I}^{-1}\mathbf{w}. \tag{3.25}$$

Here $\tilde{I}^{1/2}$ is the (unique) square root matrix of $\tilde{I}$. Both of these two matrices are positive definite because of the definiteness in $\tilde{I}$.

The inequality (3.25) has a simple statistical interpretation. Let the sample be from $F(x_1 - w_1\theta_1, \ldots, x_s - w_s\theta_s)$. Then $\tilde{I}^{-1}$ is the asymptotic variance matrix of the Pitman estimator for $(w_1\theta_1, \ldots, w_s\theta_s)$. In case of $\theta_1 = \ldots = \theta_s = \theta$, $\mathbf{w}^T t_n$ estimates θ with variance $\mathbf{w}^T\tilde{I}^{-1}\mathbf{w}$. On the other hand, if starting with the distribution $F(x_1 - w_1\theta, \ldots, x_s - w_s\theta)$, the information on θ is $\mathbf{w}^T\tilde{I}\mathbf{w}$. Hence the asymptotic variance of the Pitman estimator becomes $1/\mathbf{w}^T\tilde{I}\mathbf{w}$. In summary, (3.25) implies that there is no advantage to estimate a (location) parameter with a model of a dimension higher than necessary.

3.5 An analog of Huber's definition of the Fisher information

Our goal here is developing a definition of the Fisher information $I_{X,Y}$ on θ contained in (X, Y) with distribution $F(x-\theta, y-\theta)$ that does not require absolute continuity of F.

For $X \sim F(x-\theta)$, such a definition was suggested in Huber (1964) (see Section 1.7 (1.23)), who proved, in particular, that for $I_X < \infty$ F must be absolutely continuous. In the setup considered in this chapter, this is not true any more as the following example demonstrates.

Example 9 *Let ξ be an arbitrary random variable. Set*

$$(X, Y) = (\xi + \theta, \xi + \theta).$$

The random vector (X, Y) takes values on the diagonal $X = Y$ and, thus, is not absolutely continuous while the Fisher information on θ in (X, Y) is the same as that in X and is finite if the density $p(x)$ of ξ is such that

$$\int \left[\frac{p'(x)}{p(x)}\right]^2 p(x)dx < \infty. \quad \square$$

The definition of $I_{X,Y}$ for $(X, Y) \sim F(x-\theta, y-\theta)$ is inspired by Huber's idea of considering smooth estimating functions to which a little modification is added (borrowed from Port and Stone (1974)). Set

$$I_{X,Y} = \sup_{\psi \in C_c^1(\mathbb{R}^2)} \frac{\{E_0[\partial_X \psi(X, Y) + \partial_Y \psi(X, Y)]\}^2}{E_0[\psi(X, Y)]^2}, \tag{3.26}$$

where $C_c^1(\mathbb{R}^2)$ is the space of test functions, that is, the collection of continuously differentiable functions with compact support and $E_0\psi^2 > 0$. It is a dense subset of

the complete space of square integrable functions

$$L^2(F) = \{\psi(X,Y)|E[\psi(X,Y)^2] < \infty\}.$$

Let $\psi(X-\theta, Y-\theta) \in C_c^1(\mathbb{R}^2)$ be an estimating function for θ. If $(X_1, Y_1; \ldots;$ $X_n, Y_n)$ is a sample from population $F(x-\theta, y-\theta)$, the estimating equation

$$\sum_{k=1}^{n} \psi(X_k - \theta, Y_k - \theta) = 0 \tag{3.27}$$

has a solution $\tilde{\theta}_n = \tilde{\theta}_n(X_1, \ldots, Y_n)$ such that

$$\sqrt{n}(\tilde{\theta} - \theta) \longrightarrow_d N(0, \sigma_\psi^2),\ n \to \infty$$

where $1/\sigma_\psi^2 = \{E_0[\partial_X \psi(X,Y) + \partial_Y \psi(X,Y)]\}^2 / E_0[\psi(X,Y)]^2$ is, in a sense, the information associated with the estimating function ψ, and (3.26) is the information associated with the optimal estimating equation (its solutions are equivariant).

It is convenient to consider the following one-to-one transformation of (X,Y),

$$U = \frac{X+Y}{2},\ V = \frac{X-Y}{2}.$$

The distribution function of (U,V) is $H(u-\theta, v)$ so that V is ancillary for θ and this is the principal reason for replacing (X,Y) with (U,V). One can consider estimating functions $\phi(U-\theta, V)$ and estimating equations

$$\sum_{k=1}^{n} \phi(U_k - \theta, V_k) = 0,$$

whose solution behaves similarly to the solution of (3.27). There is one-to-one correspondence between estimating functions $\psi(X-\theta, Y-\theta)$ and $\phi(U-\theta, V)$ that

preserves the property of having compact supports. As a statistic, (U, V) is equivalent to (X, Y) so that

$$I_{X,Y} = I_{U,V} = \sup_{\phi \in C_c^1(\mathbb{R}^2)} \frac{\{E_0[\partial_U \phi(U,V)]\}^2}{E_0[\phi(U,V)]^2}. \tag{3.28}$$

Theorem 3.5.1 *Let (U, V) be a pair of random variables with a distribution function $H(u-\theta, v)$. The Fisher information $I_{U,V}$ on θ contained in the pair is defined as in (3.28). Then $I_{U,V} < \infty$ if and only if the conditional distribution $H_{U|V}(u|V)$ of U given V has an absolutely continuous density $p(u|V)$ with probability 1:*

$$H_{U|V}(du) = p(u|V)du \text{ a.s. } [\mathrm{H_V}], \tag{3.29}$$

and

$$\int \int_{\mathbb{R}^2} [\partial_u \ln p(u|v)]^2 dH_{U|V}(u,v) dH_V(v) < +\infty, \tag{3.30}$$

where H_V is the marginal distribution function of V. In this case,

$$I_{U,V} = E(I_{U|V}) = \int \int_{\mathbb{R}^2} [\partial_u \ln p(u|v)]^2 dH_{U|V}(u,v) dH_V(v). \tag{3.31}$$

Proof. In (3.28), it is sufficient to consider the supremum over all those ϕ's with fixed second moments $E_0[\phi(U,V)]^2 = 1$. Assuming (3.29) and (3.30), we have

$$\begin{aligned}
I_{X,Y} &= \sup_{E_0(\phi^2)=1} \left[\iint \partial_u \phi(u,v) H_{U,V}(dudv) \right]^2 \\
&= \sup_{E_0(\phi^2)=1} \left[\iint \partial_u \phi(u,v) p(u|v) du H_V(dv) \right]^2 \\
&= \sup_{E_0(\phi^2)=1} \left[\iint \phi(u,v) \partial_u p(u|v) du H_V(dv) \right]^2 \\
&= \sup_{E_0(\phi^2)=1} \left[\iint \phi(u,v) \frac{\partial_u p(u|v)}{p(u|v)} H_{U|V}(du) H_V(dv) \right]^2 \\
&\leq \iint \phi(u,v)^2 H_{U,V}(dudv) \iint (\partial_u \ln p(u|v))^2 H_{U,V}(dudv) < +\infty.
\end{aligned}$$

The third equality is obtained by integration by parts, and the fact that ψ has compact support. The inequality sign in the last row follows the Cauchy-Schwarz inequality. An equality sign holds, so that the supremum is attained, if and only if $\phi(U,V)$ is proportional to $\partial_U \ln p(U|V)$:

$$\phi(U,V) = \frac{\partial_U \ln p(U|V)}{\{E_0[\partial_U \ln p(U|V)]^2\}^{1/2}}.$$

Implying (3.31).

Conversely, define a linear operator A on the set of test functions $C_c^1(\mathbb{R}^2)$:

$$A\phi = \iint \partial_u \phi(u,v) H_{U,V}(dudv). \tag{3.32}$$

Consider the norm induced by the inner product

$$< \phi_1, \phi_2 >= \iint \phi_1(u,v)\phi_2(u,v) H_{U,V}(dudv).$$

Hence the information $I_{U,V}$ can be viewed as the operator norm of A:

$$\|A\|^2 = \sup_{\phi} \frac{|A\phi|^2}{\|\phi\|^2} = \sup_{\phi} \frac{[\iint \partial_u \phi(u,v) H_{U,V}(dudv)]^2}{\iint \phi(u,v)^2 H_{U,V}(dudv)}.$$

Provided $I_{U,V} < \infty$, A is a bounded linear operator defined on the dense subset $C_c^1(\mathbb{R}^2)$ of the Banach space $L^2(F)$. Extend A continuously onto the whole L^2 space. By the Riesz representation Theorem there exists (with probability 1) a unique function $g \in L^2(F)$ such that A can be written into an integral

$$A\phi = \iint \phi(u,v) g(u,v) H_{U,V}(dudv). \tag{3.33}$$

For almost all V, g is square integrable with respect to the conditional distribution $H_{U|V}$. It remains to show that $g(u,v) = \partial_u \ln p(u|v)$. On setting

$$f(u,v) = \int_{-\infty}^{u} g(t,v) H_{U|V}(dt). \tag{3.34}$$

Then proceed to check the following integral

$$\begin{aligned}
& \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \partial_u\phi(u,v)f(u,v)duH_V(dv) \\
= & \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \partial_u\phi(u,v)\int_{-\infty}^{u} g(t,v)H_{U|V}(dt)duH_V(dv) \\
= & \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{t} \partial_u\phi(u,v)g(t,v)duH_{U|V}(dt)H_V(dv) \\
= & \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \phi(t,v)g(t,v)H_{U,V}(dudv).
\end{aligned}$$

By (3.33), the last row becomes $A\psi$. Combining the above computation with the original definition (3.32), one obtains

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \partial_u\phi(u,v)f(u,v)duH_V(dv) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \partial_u\phi(u,v)H_{U,V}(dudv),$$

for an arbitrary test function ϕ. It implies the equivalence between the measures on which the integrals are taken on both sides:

$$f(u,v)duH_V(dv) = H_{U,V}(dudv) = H_{U|V}(du)H_V(dv).$$

Now (3.29) follows immediately from the above equation, implying

$$f(u,V) = p(u|V) \text{ a.s } [F_V].$$

By (3.33), $p(u|V)$ is absolutely continuous in u, such that g is explicitly defined

$$g(u,V) = \partial_u \ln p(u|V) \text{ a.s. } [H_V].$$

This completes the proof. □

Remark 1. Suppose (U,V) is a pair of random variables with an absolutely continuous joint density $p(u,v;\theta)$ depending on a general (not necessarily location)

parameter θ. If the Fisher information $I_{U,V}(\theta)$ on θ in the pair (U, V) is finite, i.e., the Fisher score $\partial_\theta \ln p(U, V; \theta)$ has finite variance, then simple calculations give

$$I_{U,V}(\theta) = I_V(\theta) + E(I_{U|V}(\theta|V)). \tag{3.35}$$

If $p(u, v; \theta) = p(u - \theta, v)$, then $I_V = 0$ since V is ancillary for θ. Thus, (3.31) is a special case of (3.35) when θ is a location parameter and the Fisher score is well defined. However, (3.31) also covers the case when the joint distribution of (U, V) is not absolutely continuous.

Remark 2. One may consider the original setting $(X, Y) = (U + V, U - V) \sim F_{X,Y}(x - \theta, y - \theta)$. Theorem 3.5.1 claims that the information $I_{X,Y}$ is finite only if the conditional distribution of $X + Y$ given $X - Y$ has (with probability 1) an absolutely continuous density whose logarithm is square integrable. It means that $I_{X,Y} < \infty$ does not necessarily imply absolute continuity of (X, Y) but the conditional distributions of (X, Y) on the straight lines parallel to the main diagonal $X = Y$ must be absolutely continuous (with respect to the linear Lebesgue measure) with probability 1.

This observation can be generalized to a higher dimensional setting. Suppose that $\mathbf{X} = (X_1, \ldots, X_n)$ is a random vector with distribution $F(x_1 - \theta, \ldots, x_n - \theta)$. Then the Fisher information is finite

$$I_{X_1,\ldots,X_n} < \infty$$

only if the conditional distributions of $(X_1, \ldots, X_n)$ on the straight lines parallel to the main diagonal

$$X_1 = X_2 = \ldots = X_n$$

have absolutely continuous densities admitting finite information with probability 1.

Kagan (unpublished manuscript) studied the case of $\mathbf{X} \sim N_n(\theta \cdot \mathbf{1}, \mathbf{V})$ with a singular variance-covariance matrix $\mathbf{V}$, and got a necessary and sufficient condition for $I_{\mathbf{X}} = \infty$, in which case there exists some unbiased constant estimator of θ. In terms of $\mathbf{V}$, $I_{\mathbf{X}} = \infty$ if and only if

$$\text{rank}(\mathbf{C}) < \text{rank}(\tilde{\mathbf{C}}), \tag{3.36}$$

where $\mathbf{C} = [c_{i,j}]_{i,j=1,\ldots,n}$ is the root matrix such that $\mathbf{V} = (\mathbf{C})^T\mathbf{C}$ and

$$\tilde{\mathbf{C}} = \begin{bmatrix} 1 & \ldots & 1 \\ c_{11} & \ldots & c_{n1} \\ & \vdots & \\ c_{1n} & \ldots & c_{nn} \end{bmatrix}.$$

Notice that (3.36) holds true if and only if the vector $[1, \ldots, 1]$ is not in the row space of $\mathbf{C}$, say, the probability of $\mathbf{X}$ concentrates on a hyperplane not parallel to the main diagonal. It is an example demonstrating Theorem 3.5.1

Remark 3. Port and Stone(1974) also dealt with the Fisher information $I_{U,V}$ as defined in (3.28). In order to eliminate the singularity, they started with the pair $(U+\sigma Z, V)$, where σ is a positive constant and Z is an independent standard normal random variable. By adding this smoothing factor, $U + \sigma Z$ always contains finite Fisher information on θ, and they proved that

$$\lim_{\sigma \to 0^+} I_{U+\sigma Z,V} = \sup_{\phi \in C_c^1(\mathbb{R}^2)} \frac{\{E_0[\partial_U \phi(U,V)]\}^2}{E_0[\phi(U,V)]^2}.$$

They also proved the following important properties of the information:

(i) *Additivity.* If $(U_i, V_i) \sim F_i(u-\theta, v)$, $i = 1, \ldots, n$, is an independent sequence of random vectors, then

$$I_{U_1,V_1,\ldots,U_n,V_n} = \sum_{i=1}^{n} I_{U_i,V_i}.$$

(ii) *Monotonicity.* If W is independent of (U, V) and does not depend on θ, then

$$I_{U,V,W} = I_{U,V}.$$

Additionally, if W is a measurable function of V, then

$$I_{U+W,V} = I_{U,V} \text{ and } I_{U,W} \leq I_{U,V}.$$

It remains an open problem to prove the classic monotonicity formula of the information. That is, if (U', V') is a measurable function of (U, V), then $I_{U',V'}(\theta) \leq I_{U,V}$. The inequality becomes an equality if (U', V') is a sufficient statistic of θ.

(iii) *Reparametrization formula.* If $c \neq 0$ and $(U', V') \sim H((u-\theta)/c, v)$, then $c^2 I_{U',V'} = I_{U,V}$, where $(U, V) \sim H(u-\theta, v)$. When the reparametrization is not linear, the formula is not known to us.

(iv) *Cramér-Rao inequality.* Notice that the information in (3.26) and (3.28) are both defined as the reciprocal of the variance of the optimal equivariant estimator. It remains an open problem whether the variance of an arbitrary, not necessarily equivariant, estimator is also bounded by the same quantity. Port and Stone proved the following inequality, which casts some light on the question: if $\mathrm{Var}(U) < \infty$ and $\hat{\theta}(\sigma)$ is an unbiased estimator of θ from the observation $(U + \sigma Z, V)$, then for any $\sigma > 0$

$$\mathrm{Var}[\hat{\theta}(\sigma)] \geq \frac{1}{I_{U+\sigma Z,V}}.$$

It is unknown whether a similar inequality holds true for the limiting case where $\sigma = 0$.

Chapter 4

Unsolved Problems

There are some open problems related to the above results. All those problems of certain interest will be formulated in this chapter.

1. If $X_1, \ldots, X_n$ is a sample of size n from $F(x-\theta)$ with an unknown location parameter θ, then there is no general proof of monotone decrease in n of $\mathrm{Var}(\hat{\theta}_n)$ for the MLE $\hat{\theta}_n$. Is there a proof in situations like F belongs to a (natural) exponential family?

2. If F is absolutely continuous, and $S(X_1, \ldots, X_n)$ is a sufficient statistic for θ, then one can easily prove that t_n is a measurable function of S. Give a proof of the same statement when F is not absolutely continuous.

3. Given a sample of size 1: $X \sim F(x-\theta)$, X is the Pitman estimator of θ. Is X also the UMVUE for θ? Moreover, one may wonder if X is a complete statistic for θ. By definition, X is complete sufficient for θ if and only if

$$\int_{-\infty}^{+\infty} f(x)dF(x-\theta) = 0, \ \forall\theta$$
$$\Rightarrow P_\theta\{f(X) = 0\} = 1 \ \forall\theta.$$

Is this statement true for any probability distribution F? If not, how can one characterize those F admitting such a statement? This analytical problem itself is of some independent interest in real function theory.

4. In the setup of linear regression, independent (but no longer identically

distributed) observations are of the form

$$X_i = a_{i1}\theta_1 + \ldots + a_{is}\theta_s + \epsilon_i,\ i = 1, \ldots, n$$

with a design matrix (a_{ir}) assumed known. Then $\theta \in \mathbb{R}^s$ can be considered, in a sense, a multivariate location parameter. Extend the idea of equivariance to the regression problems.

5. In Section 2.1, it is shown that if $\mathrm{Var}(t_n) < \infty$ for some n and $I_{X_1} < \infty$, then

$$n\mathrm{Var}(t_n) \longrightarrow \frac{1}{I_{X_1}},\ n \to \infty.$$

Does it still hold in the case of $I_X = \infty$?

6. In Corollary 2.1.1, the result is proved for samples $\mathbf{X}$ and $\mathbf{Y}$ from the same population distribution. Is it possible to be generalized to the case where $\mathbf{X}$ and $\mathbf{Y}$ are from two independent but not necessarily identically distributed populations?

7. In Theorem 2.3.1, it is proved that for independent samples $(X_1', \ldots, X_n')$ and $(X_1'', \ldots, X_n'')$

$$\mathrm{Var}[t_n(X_1' + X_1'', \ldots)] \geq \mathrm{Var}[t_n'(X_1', \ldots)] + \mathrm{Var}[t_n''(X_1'', \ldots)].$$

A stronger version of the same inequality (2.31) allows dependence between the samples in the following way:

$$X_k' = U_k + W_k,\ X_k'' = V_k + W_k,\ 1 \leq k \leq n,$$

where U_k, V_k and W_k are some independent random variables. Can we strengthen the inequality such that it requires only certain conditions in the moments of the samples?

8. In section 2.3.3, we studied the linear approximations of the Fisher score $J(X+Y)$ in some finite dimensional spaces. They possess a property similar to equation (2.28), which is the basis of the classical Stam inequality. Does the Stam inequality holds true for the information associated with these approximate scores?

9. In Theorem 2.3.4, it is proved that $\mathrm{Var}(t_{X'+\lambda X''})$ is monotone in λ when X'' is self-decomposable. Prove or disprove the statement when X'' is not self-decomposable.

When X'' is Gaussian, Port and Stone (1974) proved that

$$I_{X'+\lambda X''} \longrightarrow I_{X'},\ \lambda \to 0^+,$$

where $I_{X'}$ on the right hand side is Huber's information defined in (1.23). Try to analyze the expression when X'' is self-decomposable, but not necessarily Gaussian.

10. In Chapter 3, the characteristic function (3.3) defines a family of distributions. Describe these distributions in terms of their distribution functions.

11. In Remark 3, Section 3.5, there is a list of open problems associated with the information $I_{X,Y}$ defined in (3.26). $I_{X,Y}$ shares some common properties with the classical Fisher information. The proof to some of them are not obvious.

12. Extend the conclusions from this thesis to the case of a general parameter.

Bibliography

[1] Aczél, J. (1966) *Lectures on functional equations and their applications*, Academic Press, New York and London.

[2] Artstein, S., Ball, K.M., Barthe, F. and Noar, A. (2004) *Solution of Shannon's problem on the monotonicity of entropy*, J. Amer. Math. Soc. **17(4)**, 975-982.

[3] Bahadur, R.R. (1955) *A characterization of sufficiency*, Ann. Math. Stat., **26(2)**, 286-293.

[4] Bahadur, R.R. (edited by Stigler, S.M., Wang, W.H. and Xu D). (2002) *R.R. Bahadur's lectures on the theory of estimation*, IMS Lecture Notes-Monograph Series, Vol.31.

[5] Bahadur, R.R. (1957) *On unbiased estimates of uniformly minimum variance*, Sankhyā, **18**, 211-224.

[6] Balakrishnan, N. and Cohen, A.C. (1990) *Order statistics and inference - estimation methods*, Academic Press.

[7] Barron, A. (1986) *Entropy and the Central Limit Theorem*, Ann. Prob. **14(1)**, 336-342.

[8] Billingsley, P. (1986) *Probability and measure*, 2nd ed., Wiley, New York.

[9] Blachman, N. (1965) *The convolution inequality for entropy powers*, IEEE Trans. Inform. Theo., **IT-11**:267-271.

[10] Bondesson, L. (1974) *A characterization of the normal law*, Sankhyā, **A36**, 321-324.

[11] Brown, L.D. (1966) *On the admissibility of invariant estimators of one or more location parameters*, Ann. Math. Stat., **37(5)**, 1087-1136.

[12] Carlen, E.A. (1991) *Superadditivity of Fisher's information and logarithmic Sobolev inequalities*, J. Funct. Anal., **101(1)**: 194-211.

[13] Chu, J.T. and Hotelling, H. (1955) *The moments of the sample median* , Ann. Math. Stat., **26(4)**, 593-606.

[14] Cramér, H. (1946) *Mathematical methods of statistics*, Princeton University.

[15] Efron, B. and Stein, C. (1981) *The jackknife estimate of variance*, Ann. Stat., **9(3)**, 586-596.

[16] Fintushal, S.M. (1975) *Representation of Fisher information in terms of distribution moments*, Problems of Inform. Trans., **3**, 253-255.

[17] Godambe, V.P. (1991) *Estimating functions*, Oxford University Press.

[18] Halmos, P.R. and Savage, L.J. (1979) *Application of the Radon-Nikodym theorem to the theory of sufficient statistics*, Ann. Math. Stat., **20(2)**, 225-241.

[19] Hoeffding, W. (1948) *A class of statistics with asymptotically normal distribution*, Ann. Math. Stat., **19(3)**, 293-325.

[20] Hoffman-Jørgensen, J.H, Kagan, A.M., Pitt, L.D. and Shepp, L.A. (2007) *Strong decomposition of random variables*, J. of Theo. Prob., **20**, 211-220.

[21] Huber, P.J. (1981) *Robust statistics*, Wiley.

[22] Huber, P.J. (1964) *Robust estimation of a location parameter*, Ann. Math. Stat., **35(1)**, 73-101.

[23] Ibragimov, I.A. and Has'minskii, R.Z. *Asymptotic behavior of certain statistical estimates in the smooth case*, Teor. Veroyatn. Ee. Primen., **17**, 3 (1972); **18**, 1 (1973).

[24] Ibragimov, I.A. and Has'minskii, R.Z. (1981) *Statistical estimation. Asymptotic theory*, Springer-Verlag.

[25] Kagan, A.M. (1966) *On the estimation theory of location parameter*, Sankhyā, **A28**, 4, 335-352.

[26] Kagan, A.M. (1976) *Fisher information contained in a finite-dimensional linear space and a properly formulated version of the method of moments*, Problems of Inform. Trans., **12**, 25-42.

[27] Kagan, A.M. (1986) *A simple modification of Pitman estimates for a location parameter*. Theo. of Prob. and Its App., **30**, 598-603.

[28] Kagan, A.M. (2002) *An inequality for the Pitman estimators related to the Stam inequality*, Sankhyā, **A64**, 282-292.

[29] Kagan, A.M., Klebanov, Y.L. and Fintushal, S.M. (1974) *The asymptotic behavior of Pitman's polynomial estimators*, Zapiski Nauchn. Semin. LOMI, **43**.

[30] Kagan, A.M. and Konikov, M. (2006) *The structure of the UMVUEs from categorical data*, Theo. Prob. and Its App., **50(3)**, 466-473.

[31] Kagan, A.M., Linnik, Y.V. and Rao, C.R. (1973) *Characterization problems in Mathematical Statistics*, John Wiley, New York.

[32] Kagan, A.M. and Rao, C.R. (2005) *On estimation of a location parameter in presence of an ancillary component*, Theo. Prob. and Its App., **50**, 172-176.

[33] Kagan, A.M. and Shepp, L. (2005) *A sufficient paradox: an insufficient statistic preserving the Fisher information*, Amer. Stat., **59**, No.1, 54-56.

[34] Kagan, A.M., Landsman Z. and Rao, C.R. (2007) *Sub- and superadditivity à la Carlen of matrices related to the Fisher information*, J. of Stat. Planning and Inference, **137**, 291-298.

[35] Lehmann, E.L. (1983) *Theory of point estimation*, Springer-Verlag, New York.

[36] Lukacs, E. (1970) *Characteristic functions*, 2nd ed., Griffin, London.

[37] Madiman, M. and Barron, A. (2006) *Generalized entropy power inequalities and monotonicity properties of information*, Preprint.

[38] Naylor, A.W. and Sell, G.R. (2000) *Linear Operator Theory in Engineering and Science*, Springer.

[39] Pitman, E.J.G. (1938) *The estimation of location and scale parameters of a continuous population of any given form*, Biometrika, **30**, III-IV, 391-421.

[40] Stone, C.J. (1974) *Asymptotic properties of estimators of a location parameter*, Ann. of Stat., **2(6)**, 1127-1137.

[41] Port, S.C. and Stone, C.J. (1974) *Fisher information and the Pitman estimator of a location parameter*, Ann. of Stat., **2(2)**, 225-247.

[42] Prelov, V.V. and van der Meulen, E.C. (1995) *Asymptotics of Fisher information under weak perturbation*, Problems of Inform. Trans., **31(1)**, 14-22.

[43] Shannon, C.E. (1948) *A mathematical theory of communication*, Bell System Tech. J., **27**, 379-423.

[44] Shao, J. (2003) *Mathematical Statistics*, 2nd ed., Springer.

[45] Shlyakhtenko, D. (2005) *A free analogue of Shannon's problem on monotonicity of entropy*. Preprint at http://xxx.lanl.gov/abs/math.OA/0510103.

[46] Stam, A.J. (1959) *Some inequalities satisfied by the quantities of information of Fisher and Shannon*, Inform. and Control, **2**, 101-112.

[47] Stein, C. (1959) *The admissibility of Pitman's estimator for a single location parameter*, Ann. Math. Stat., **30(4)**, 970-979.

[48] Zamir, R. (1998) *A proof of the Fisher information inequality via a data processing argument*, IEEE Tran. Inform. Theo., **44(3)**, 1246-1250.

[49] Ziemer, W.P. (1989) *Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation*, Springer.