IMS Lecture Notes–Monograph Series
Optimality: The Third Erich L. Lehmann Symposium
Vol. 57 (2009) 266–275

DOI: 10.1214/09-LNMS5716

Asymptotic Efficiency of Simple Decisions for the Compound Decision Problem

Eitan Greenshtein[1,*] **and Ya'acov Ritov**[2,*]

Duke University and Jerusalem, Israel

Abstract: We consider the compound decision problem of estimating a vector of n parameters, known up to a permutation, corresponding to n independent observations, and discuss the difference between two symmetric classes of estimators. The first and larger class is restricted to the set of all permutation invariant estimators. The second class is restricted further to simple symmetric procedures. That is, estimators such that each parameter is estimated by a function of the corresponding observation alone. We show that under mild conditions, the minimal total squared error risks over these two classes are asymptotically equivalent up to essentially $O(1)$ difference.

Contents

1. Introduction

Let $\mathcal{F} = \{F_\mu : \mu \in \mathcal{M}\}$ be a parameterized family of distributions. Let $Y_1, Y_2 \ldots$ be a sequence of independent random variables, where Y_i takes value in some space $\mathcal{Y}$, and $Y_i \sim F_{\mu_i}$, $i = 1, 2, \ldots$. For each n, we suppose that the sequence $\mu_{1:n}$ is known up to a permutation, where for any sequence $x = (x_1, x_2, \ldots)$ we denote the subsequence $x_s, \ldots, x_t$ by $x_{s:t}$. We denote by $\boldsymbol{\mu} = \boldsymbol{\mu}_n$ the set $\{\mu_1, \ldots, \mu_n\}$, i.e., $\boldsymbol{\mu}$ is $\mu_{1:n}$ without any order information. We consider in this note the problem of estimating $\mu_{1:n}$ by $\hat{\mu}_{1:n}$ under the loss $\sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2$, where $\hat{\mu}_{1:n} = \Delta(Y_{1:n})$. We assume that the family $\mathcal{F}$ is dominated by a measure ν, and denote the corresponding densities simply by $f_i = f_{\mu_i}$, $i = 1, \ldots, n$. The important example is, as usual, $F_{\mu_i} = N(\mu_i, 1)$.

Let $\mathcal{D}^S = \mathcal{D}_n^S$ be the set of all *simple symmetric decision functions* Δ, that is, all Δ such that $\Delta(Y_{1:n}) = (\delta(Y_1), \ldots, \delta(Y_n))$, for some function $\delta : \mathcal{Y} \to \mathcal{M}$. In particular, the best simple symmetric function is denoted by $\Delta_{\boldsymbol{\mu}}^S = (\delta_{\boldsymbol{\mu}}^S(Y_1), \ldots, \delta_{\boldsymbol{\mu}}^S(Y_n))$:

$$\Delta_{\boldsymbol{\mu}}^S = \operatorname*{arg\,min}_{\Delta \in \mathcal{D}_n^S} \mathrm{E}\, ||\Delta - \mu_{1:n}||^2,$$

[1] Department of Statistical Sciences, Duke University, Durham, NC 27708-0251, USA, e-mail: eitan.greenshtein@gmail.com

[2] Jerusalem, Israel, e-mail: yaacov.ritov@gmail.com

[*] Partially supported by NSF grant DMS-0605236, and an ISF Grant.

AMS 2000 subject classifications: Primary 62C25; secondary 62C12, 62C07.

Keywords and phrases: compound decision, simple decision rules, permutation invariant rules.

and denote

$$r_n^S = \mathrm{E}\,||\Delta_{\boldsymbol{\mu}}^S(Y_{1:n}) - \mu_{1:n}||^2,$$

where, as usual, $\|a_{1:n}\|^2 = \sum_{i=1}^n a_i^2$.

The class of simple rules may be considered too restrictive. Since the μs are known up to a permutation, the problem seems to be of matching the Ys to the μs. Thus, if $Y_i \sim N(\mu_i, 1)$, and $n = 2$, a reasonable decision would make $\hat\mu_1$ closer to $\mu_1 \wedge \mu_2$ as Y_2 gets larger. The simple rule clearly remains inefficient if the μs are well separated, and generally speaking, a bigger class of decision rules may be needed to obtain efficiency. However, given the natural invariance of the problem, it makes sense to be restricted to the class $\mathcal{D}^{PI} = \mathcal{D}_n^{PI}$ of all permutation invariant decision functions, i.e, functions Δ that satisfy for any permutation π and any $(Y_1, \dots, Y_n)$:

$$\Delta(Y_1, \dots, Y_n) = (\hat\mu_1, \dots, \hat\mu_n) \quad\iff\quad \Delta(Y_{\pi(1)}, \dots, Y_{\pi(n)}) = (\hat\mu_{\pi(1)}, \dots, \hat\mu_{\pi(n)}).$$

Let

$$\Delta_{\boldsymbol{\mu}}^{PI} = \operatorname*{arg\,min}_{\Delta \in \mathcal{D}^{PI}} \mathrm{E}\,||\Delta(Y^n) - \mu_{1:n}||^2$$

be the optimal permutation invariant rule under $\boldsymbol{\mu}$, and denote its risk by

$$r_n^{PI} = E||\Delta_{\boldsymbol{\mu}}^{PI}(Y_{1:n}) - \mu_{1:n}||^2.$$

Obviously $\mathcal{D}^S \subset \mathcal{D}^{PI}$, and whence $r_n^S \geq r_n^{PI}$. Still, ‘folklore’, theorems in the spirit of De Finetti, and results like Hannan and Robbins [5], imply that asymptotically (as $n \to \infty$) $\Delta_{\mu^n}^{PI}$ and $\Delta_{\mu^n}^S$ will have ‘similar’ mean risks: $r_n^S - r_n^{PI} = o(n)$. Our main result establishes conditions that imply the stronger claim, $r_n^S - r_n^{PI} = O(1)$.

To repeat, $\boldsymbol{\mu}$ is assumed known in this note. In the general decision theory framework the unknown parameter is the order of its member to correspond with $Y_{1:n}$, and the parameter space, therefore, corresponds to the set of all the permutations of $1, \dots, n$.

An asymptotic equivalence as above implies, that when we confine ourselves to the class of permutation invariant procedures, we may further restrict ourselves to the class of simple symmetric procedures, as is usually done in the standard analysis of compound decision problems. The later class is smaller and simpler.

The motivation for this paper stems from the way the notion of oracle is used in some sparse estimation problems. Consider two oracles, both *know* the value of $\boldsymbol{\mu}$. Oracle I is restricted to use only a procedure from the class $\mathcal{D}^{PI}$, while Oracle II is further restricted to use procedures from $\mathcal{D}^S$. Obviously Oracle I has an advantage, our results quantify this advantage and show that it is asymptotically negligible. Furthermore, starting with Robbins [7] various oracle-inequalities were obtained showing that one can achieve nearly the risk of Oracle II, by a ‘legitimate’ statistical procedure. See, e.g., the survey Zhang [10], for oracle-inequalities regarding the difference in risks. See also Brown and Greenshtein [2], and Wenuha and Zhang [11] for oracle inequalities regarding the ratio of the risks. However, Oracle II is limited, and hence, these claims may seem to be too weak. Our equivalence results, extend many of those oracle inequalities to be valid also with respect to Oracle I. We needed a stronger result than the usual objective that the mean risks are equal up to $o(1)$ difference. Many of the above mentioned recent applications of the compound decision notion are about sparse situations when most of the μs are in fact 0, the mean risk is $o(1)$, and the only interest is in total risk.

Let $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$ be some arbitrary ordering of $\boldsymbol{\mu}$. Consider now the Bayesian model under which $(\pi, Y_{1:n})$, π a random permutation, have a distribution given by

(1.1) π is uniformly distributed over $\mathcal{P}(1:n)$;

(1.2) Given π, $Y_{1:n}$ are independent, $Y_i \sim F_{\boldsymbol{\mu}_{\pi(i)}}$, $i = 1, \ldots, n$,

where for every $s < t$, $\mathcal{P}(s:t)$ is the set of all permutations of $s, \ldots, t$. The above description induces a joint distribution of $(M_1, \ldots, M_n, Y_1, \ldots, Y_n)$, where $M_i \equiv \boldsymbol{\mu}_{\pi(i)}$, for a random permutation π.

The first part of the following proposition is a simple special case of general theorems representing the best invariant procedure under certain groups as the Bayesian decision with respect to the appropriate Haar measure; for background see, e.g., Berger [1], Chapter 6. The second part of the proposition was derived in various papers starting with Robbins [7].

In the following proposition and proof, $\mathrm{E}_{\mu_{1:n}}$ is the expectation under the model in which the observations are independent, $Y_i \sim F_{\mu_i}$, and $\mathrm{E}_{\boldsymbol{\mu}}$ is the expectation under the above joint distribution of $Y_{1:n}$ and $M_{1:n}$. Note that under the latter model, for any $i = 1, \ldots, n$, marginally $M_i \sim \mathbb{G}_n$, the empirical measure defined by the vector $\boldsymbol{\mu}$, and conditional on $M_i = m$, $Y_i \sim F_m$.

Proposition 1.1. *The best simple and permutation invariant rules are given by*

(i) $\Delta^{PI}_{\boldsymbol{\mu}}(Y_{1:n}) = E_{\boldsymbol{\mu}}(M_{1:n}|Y_{1:n})$.
(ii) $\Delta^{S}_{\boldsymbol{\mu}}(Y_{1:n}) = (E_{\boldsymbol{\mu}}(M_1|Y_1), \ldots, E_{\boldsymbol{\mu}}(M_n|Y_n)$.
(iii) $r^S_n = r^{PI}_n + \mathrm{E}_{\boldsymbol{\mu}} \, \|\Delta^S_{\boldsymbol{\mu}} - \Delta^{PI}_{\boldsymbol{\mu}}\|^2$.

Proof. We need only to give the standard proof of the third part. First, note that by invariance $\Delta^{PI}_{\boldsymbol{\mu}}$ is an equalizer (over all the permutations of $\boldsymbol{\mu}$), and hence $\mathrm{E}_{\mu_{1:n}}(\Delta^{PI}_{\boldsymbol{\mu}} - \mu_{1:n})^2 = \mathrm{E}_{\boldsymbol{\mu}}(\Delta^{PI}_{\boldsymbol{\mu}} - M_{1:n})^2$. Also $\mathrm{E}_{\mu_{1:n}}(\Delta^{S}_{\boldsymbol{\mu}} - \mu_{1:n})^2 = \mathrm{E}_{\boldsymbol{\mu}}(\Delta^{S}_{\boldsymbol{\mu}} - M_{1:n})^2$. Then, given the above joint distribution,

$$\begin{aligned}
r^S_n &= \mathrm{E}_{\boldsymbol{\mu}} \, \|\Delta^S_{\boldsymbol{\mu}} - M_{1:n}\|^2 \\
&= \mathrm{E}_{\boldsymbol{\mu}} \, \mathrm{E}_{\boldsymbol{\mu}}\{\|\Delta^S_{\boldsymbol{\mu}} - M_{1:n}\|^2 | Y_{1:n}\} \\
&= \mathrm{E}_{\boldsymbol{\mu}} \, \mathrm{E}_{\boldsymbol{\mu}}\{\|\Delta^S_{\boldsymbol{\mu}} - \Delta^{PI}_{\boldsymbol{\mu}}\|^2 + \|\Delta^{PI}_{\boldsymbol{\mu}} - M_{1:n}\|^2 | Y_{1:n}\} \\
&= r^{PI}_n + \mathrm{E}_{\boldsymbol{\mu}} \, \|\Delta^S_{\boldsymbol{\mu}} - \Delta^{PI}_{\boldsymbol{\mu}}\|^2. \qquad \square
\end{aligned}$$

We now briefly review some related literature and problems. On simple symmetric functions, compound decision and its relation to empirical Bayes, see Samuel [9], Copas [3], Robbins [8], Zhang [10], among many other papers.

Hannan and Robbins [5] formulated essentially the same equivalence problem in testing problems, see their Section 6. They show for a special case an equivalence up to $o(n)$ difference in the 'total risk' (i.e., non-averaged risk). Our results for estimation under squared loss are stated in terms of the total risk and we obtain $O(1)$ difference.

Our results have a strong conceptual connection to De Finetti's Theorem. The exchangeability induced on $M_1, \ldots, M_n$, by the Haar measure, implies 'asymptotic independence' as in De Finetti's theorem, and consequently asymptotic independence of $Y_1, \ldots, Y_n$. Thus we expect $E(M_1|Y_1)$ to be asymptotically similar to $E(M_1|Y_1, \ldots, Y_n)$. Quantifying this similarity as n grows, has to do with the rate of convergence in De Finetti's theorem. Such rates were established by Diaconis and Freedman [4], but are not directly applicable to obtain our results.

After quoting a simple result in the following section, we consider in Section 3 the special important, but simple, case of two-valued parameter. In Section 4 we obtain a strong result under strong conditions. Finally, the main result is given in Section 5, it covers the two preceding cases, but with some price to pay for the generality.

2. Basic Lemma and Notation

The following lemma is standard in comparison of experiments theory; for background on comparison of experiments in testing see Lehmann [6], p. 86. The proof follows a simple application of Jensen's inequality.

Lemma 2.1. *Consider two pairs of distributions, $\{G_0, G_1\}$ and $\{\tilde G_0, \tilde G_1\}$, such that the first pair represents a weaker experiment in the sense that there is a Markov kernel $\mathbb{K}$, and $G_i(\cdot) = \int \mathbb{K}(y,\cdot)\, d\tilde G_i(y)$, $i = 1, 2$. Then*

$$\mathrm{E}_{G_0}\, \psi\biggl(\frac{dG_1}{dG_0}\biggr) \le \mathrm{E}_{\tilde G_0}\, \psi\biggl(\frac{d\tilde G_1}{d\tilde G_0}\biggr)$$

for any convex function ψ.

For simplicity denote $f_i(\cdot) = f_{\mu_i}(\cdot)$, and for any random variable X, we may write $X \sim g$ if g is its density with respect to a certain dominating measure. Finally, for simplicity we use the notation y_{-i} to denote the sequence $y_1, \dots, y_n$ without its i member, and similarly $\boldsymbol{\mu}_{-i} = \{\mu_1, \dots, \mu_n\} \setminus \{\mu_i\}$. Finally $f_{-i}(Y_{-j})$ is the marginal density of Y_{-j} under the model (1.1) conditional on $M_j = \mu_i$.

3. Two Valued Parameter

We suppose in this section that μ can get one of two values which we denote by $\{0, 1\}$. To simplify notation we denote the two densities by f_0 and f_1.

Theorem 3.1. *Suppose that either of the following two conditions holds:*

(i) *$f_{1-\mu}(Y_1)/f_\mu(Y_1)$ has a finite variance under both $\mu \in \{0, 1\}$.*
(ii) *$\sum_{i=1}^n \mu_i/n \to \gamma \in (0, 1)$, and $f_{1-\mu}(Y_1)/f_\mu(Y_1)$ has a finite variance under one of $\mu \in \{0, 1\}$.*

Then $\mathrm{E}_{\boldsymbol{\mu}}\, \|\hat\mu^S - \hat\mu^{PI}\|^2 = O(1)$.

Proof. Suppose condition (i) holds. Let $K = \sum_{i=1}^n \mu_i$, and suppose, WLOG, that $K \le n/2$. Consider the Bayes model of (1.1). By Bayes Theorem

$$P(M_1 = 1 | Y_1) = \frac{K f_1(Y_1)}{K f_1(Y_1) + (n - K) f_0(Y_1)}.$$

On the other hand

$$
\begin{aligned}
&P(M_1 = 1|Y_{1:n})\\
&\quad = \frac{Kf_1(Y_1)f_{K-1}(Y_{2:n})}{Kf_1(Y_1)f_{K-1}(Y_{2:n}) + (n-K)f_0(Y_1)f_K(Y_{2:n})}\\
&\quad = \frac{Kf_1(Y_1)}{Kf_1(Y_1) + (n-K)f_0(Y_1)}\\
&\qquad \times \Bigl(1 + \frac{(n-K)f_0(Y_1)}{Kf_1(Y_1) + (n-K)f_0(Y_1)}\Bigl(\frac{f_K}{f_{K-1}}(Y_{2:n}) - 1\Bigr)\Bigr)^{-1}\\
&\quad = P(M_1 = 1|Y_1)\Bigl(1 + \gamma\Bigl(\frac{f_K}{f_{K-1}}(Y_{2:n}) - 1\Bigr)\Bigr)^{-1},
\end{aligned}
$$

where, with some abuse of notation $f_k(Y_{2:n})$ is the joint density of $Y_{2:n}$ conditional on $\sum_{j=2}^n \mu_j = k$, and the random variable γ is in $[0,1]$. We prove now that $f_K/f_{K-1}(Y_{2:n})$ converges to 1 in the mean square.

We use Lemma 2.1 (with ψ the square) to compare the testing of $f_K(Y_{2:k})$ vs. $f_{K-1}(Y_{2:k})$ to an easier problem, from which the original problem can be obtained by adding a random permutation. Suppose for simplicity and WLOG that in fact $Y_{2:K}$ are i.i.d. under f_1, while $Y_{K+1:n}$ are i.i.d. under f_0. Then we compare

$$
g_{K-1}(Y_{2:n}) = \prod_{j=2}^{K} f_1(Y_j) \prod_{j=K+1}^{n} f_0(Y_j),
$$

the true distribution, to the mixture

$$
g_K(Y_{2:n}) = g_{K-1}(Y_{2:n}) \frac{1}{n-K} \sum_{j=K+1}^{n} \frac{f_1}{f_0}(Y_j).
$$

However, the likelihood ratio between g_K and g_{K-1} is a sum of $n-K$ terms, each with mean 1 (under g_{K-1}) and finite variance. The ratio between the gs is, therefore, $1 + O_p(n^{-1/2})$ in the mean square. By Lemma 2.1, this applies also to the fs' ratio.

Consider now the second condition. By assumption, K is of the same order as n, and we can assume, WLOG, that the f_1/f_0 has a finite variance under f_0. With this understanding, the above proof holds for the second condition. □

The condition of the theorem is clearly satisfied in the normal shift model: $F_i = N(\mu_i, 1)$, $i = 1,2$. It is satisfied for the normal scale model, $F_i = N(0, \sigma_i^2)$, $i = 1,2$, if K is of the same order as n, or if $\sigma_0^2/2 < \sigma_1^2 < 2\sigma_0^2$.

4. Dense $\boldsymbol{\mu}$'s

We consider now another simple case in which $\boldsymbol{\mu}$ can be ordered $\mu_{(1)}, \dots, \mu_{(n)}$ such that the difference $\mu_{(i+1)} - \mu_{(i)}$ is uniformly small. This will happen if, for example, $\boldsymbol{\mu}$ is in fact a random sample from a distribution with density with respect to Lebesgue measure, which is bounded away from 0 on its support, or more generally, if it is sampled from a distribution with short tails. Denote by $Y_{(1)}, \dots, Y_{(n)}$ and $f_{(1)}, \dots, f_{(n)}$ the Ys and fs ordered according to the μs.

We assume in this section

(B1) For some constants A_n and V_n which are bounded by a slowly converging to infinite sequence:

$$\max_{i,j} |\mu_i - \mu_j| = A_n,$$
$$\mathrm{Var}\biggl(\frac{f_{(j+1)}}{f_{(j)}}(Y_{(j)})\biggr) \le \frac{V_n}{n^2}.$$

Note that condition **(B1)** holds for both the normal shift model and the normal scale model, if $\boldsymbol{\mu}$ behaves like a sample from a distribution with a density as above.

Theorem 4.1. *If Assumption **(B1)** holds then*

$$\sum_{i=1}^{n} |\hat{\mu}_i^{PI} - \hat{\mu}_i^{S}|^2 = O_p(A_n^2 V_n^2/n).$$

Proof. By definition

$$\hat{\mu}_1^S = \frac{\sum_{i=1}^n \mu_i f_i(Y_i)}{\sum_{i=1}^n f_i(Y_i)},$$
$$\hat{\mu}_1^{PI} = \frac{\sum_{i=1}^n \mu_i f_i(Y_1) f_{-i}(Y_{2:n})}{\sum_{i=1}^n f_i(Y_1) f_{-i}(Y_{2:n})},$$

where f_{-i} is the density of $Y_{2:n}$ under $\boldsymbol{\mu}_{-i}$:

$$f_{-i}(y_{2:m}) = \frac{1}{(n-1)!} \sum_{\pi \in \mathcal{P}(2:n)} \prod_{j=2}^{n} f_{\pi(j)}(y_j) \frac{f_1}{f_i}(y_i).$$

The result will follow if we argue that

$$|\mu_1^{PI} - \mu_1^S| \le \max_{i,j} |\mu_i - \mu_j| \biggl(\max_{i,j} \frac{f_{-i}}{f_{-j}}(Y_{2:n}) - 1\biggr) = O_p(A_n V_n/n). \tag{4.1}$$

That is, $\max_i |f_{-i}(Y_{2:n})/f_{-1}(Y_{2:n}) - 1| = O_p(V_n/n)$. In fact we will establish a slightly stronger claim that $\|f_{-i} - f_{-1}\|_{TV} = O_p(V_n/n)$, where $\|\cdot\|_{TV}$ denotes the total variation norm.

We will bound this distance by the distance between two other densities. Let $g_{-1}(y_{2:n}) = \prod_{j=2}^n f_j(y_j)$, the true distribution of $Y_{2:n}$. We define now a similar analog of f_{-i}. Let r_j and $y_{(r_j)}$ be defined by $f_j = f_{(r_j)}$ and $y_{(r_j)} = y_j$, $j = 1, \dots, n$. Suppose, for simplicity, that $r_i < r_1$. Let

$$g_{-i}(y_{2:n}) = g_{-1}(y_{2:n}) \prod_{j=r_i}^{r_1-1} \frac{f_{(j+1)}}{f_{(j)}}(y_{(j)}).$$

The case $r_1 < r_i$ is defined similarly. Note that g_{-i} depends only on $\boldsymbol{\mu}_{-i}$. Moreover, if $\tilde{Y}_{2:n} \sim g_{-j}$, then one can obtain $Y_{2:n} \sim f_{-j}$ by the Markov kernel that takes $\tilde{Y}_{2:n}$ to a random permutation of itself. It follows from Lemma 2.1

$$\begin{aligned}\|f_{-i} - f_{-1}\|_{TV} &\le \|g_{-i} - g_{-1}\|_{TV} \\ &= \mathrm{E}_{\mu_{2:n}} \biggl|\frac{g_{-i}}{g_{-1}}(Y_{2:n}) - 1\biggr| \\ &= \mathrm{E}_{\mu_{2:n}} \biggl|\prod_{j=k}^{r_1-1} \frac{f_{(j+1)}}{f_{(j)}}(Y_{(j)}) - 1\biggr|.\end{aligned}$$

But, by assumption

$$R_k = \prod_{j=k}^{r_1-1} \frac{f_{(j+1)}}{f_{(j)}}(Y_{(j)})$$

is a reversed L_2 martingale, and it follows from Assumption **(B1)** that

$$\max_{k<r_1} |R_k - 1| = O_p(A_n V_n/n).$$

Similar argument applies to i, $r_i > r_1$, yielding

$$\max_i \|f_{-i} - f_{-1}\|_{TV} = O_p(A_n V_n/n).$$

We established (4.1). The theorem follows. □

5. Main Result

We assume:

(G1) For some $C < \infty$: $\max_{i\in\{1,\ldots,n\}} |\mu_i| < C$, and $\max_{i,j\in 1,\ldots,n} \mathrm{E}_{\mu_i}(f_{\mu_j}(Y_1)/f_{\mu_i}(Y_1))^2 < C$. Also, there is $\gamma > 0$ such that $\min_{i,j\in 1,\ldots,n} P_{\mu_i}(f_{\mu_j}(Y_1)/f_{\mu_i}(Y_1) > \gamma) \geq 1/2$.

(G2) The random variables

$$p_j(Y_i) = \frac{f_j(Y_i)}{\sum_{k=1}^n f_k(Y_i)}, \quad i, j = 1, \ldots, n,$$

are bounded in expectation by

$$\mathrm{E}\sum_{i=1}^n \sum_{j=1}^n \bigl(p_j(Y_i)\bigr)^2 < C,$$

$$\sum_{i=1}^n \mathrm{E}\,\frac{1}{n \min_j p_j(Y_i)} < Cn,$$

$$\mathrm{E}\sum_{i=1}^n \frac{\sum_{j=1}^n \bigl(p_j(Y_i)\bigr)^2}{n \min_j p_j(Y_i)} < C.$$

Both assumptions describe a situation where the μs do not "separate". They cannot be too far one from another, geometrically or statistically (Assumption **(G1)**), and they are dense in the sense that each Y can be explained by many of the μs (Assumption **(G2)**). The conditions hold for the normal shift model if $\boldsymbol{\mu}_n$ are uniformly bounded: Suppose the common variance is 1 and $|\mu_j| < A_n$. Then

$$\begin{aligned}
\mathrm{E}\sum_{j=1}^n \left(\frac{f_j(Y_1)}{\sum_{k=1}^n f_k(Y_1)}\right)^2 &= \mathrm{E}\,\frac{\sum_{j=1}^n f_j^2(Y_1)}{(\sum_{k=1}^n f_k(Y_1))^2} \\
&\leq \mathrm{E}\,\frac{ne^{-Y_1^2+2A_n|Y_1|-A_n^2}}{(ne^{-(Y_1^2-2A_n|Y_1|+A_n^2)/2})^2} \\
&= \frac{1}{n}\mathrm{E}\,e^{4A_n|Y_1|} \\
&= \frac{1}{n}\Bigl(e^{8A_n^2+4A_n\mu_1} + e^{8A_n^2-4A_n\mu_1}\Bigr) \leq \frac{2}{n}e^{12A_n^2}
\end{aligned}$$

and the first part of **(G2)** hold. The other parts follow a similar calculations.

Theorem 5.1. *Assume that **(G1)** and **(G2)** hold. Then*

(i) $$\mathrm{E}\,\|\Delta^S_{\boldsymbol{\mu}} - \Delta^{PI}_{\boldsymbol{\mu}}\|^2 = O(1),$$

(ii) $$r^S_n - r^{PI}_n = O(1).$$

Corollary 5.2. *Suppose $\mathcal{F} = \{N(\mu, 1) : \ |\mu| < c\}$ for some $c < \infty$, then the conclusions of the theorem follow.*

Proof. It was mentioned already in the introduction that when we are restricted to permutation invariant procedure we can consider the Bayesian model under which $(\pi, Y_{1:n})$, π a random permutation, have a distribution given by (1.1). Fix now $i \in \{1, \ldots, n\}$. Under this model we want to compare

$$\mu^S_i = E(\mu_{\pi(i)}|Y_i), \quad i = 1, \ldots, n$$

to

$$\mu^{PI}_i = E(\mu_{\pi(i)}|Y_{1:n}), \quad i = 1, \ldots, n.$$

More explicitly:

$$\begin{aligned}\mu^S_i &= \frac{\sum_{j=1}^n \mu_j f_j(Y_i)}{\sum_{j=1}^n f_j(Y_i)} \\ &= \sum_{j=1}^n \mu_j p_j(Y_i), \quad i = 1, \ldots, n, \\ \mu^{PI}_i &= \frac{\sum_{j=1}^n \mu_j f_j(Y_i) f_{-j}(Y_{-i})}{\sum_{j=1}^n f_j(Y_i) f_{-j}(Y_{-i})} \\ &= \sum_{j=1}^n \mu_j p_j(Y_i) W_j(Y_{-i}, Y_i), \quad i = 1, \ldots, n,\end{aligned} \tag{5.1}$$

where for all $i, j = 1, \ldots, n$, $f_j(Y_i)$ was defined in Section 2, and

$$\begin{aligned} p_j(Y_i) &= \frac{f_j(Y_i)}{\sum_{k=1}^n f_k(Y_i)}, \\ W_j(Y_{-i}, Y_i) &= \frac{f_{-j}(Y_{-i})}{\sum_{k=1}^n p_k(Y_i) f_{-k}(Y_{-i})}. \end{aligned}$$

Note that $\sum_{k=1}^n p_k(Y_i) = 1$, and $W_j(Y_{-i}, Y_i)$ is the likelihood ratio between two (conditional on Y_i) densities of Y_{-i}, say g_{j0} and g_1. Consider two other densities (again, conditional on Y_i):

$$\begin{aligned} \tilde{g}_{j0}(Y_{-i}|Y_i) &= f_i(Y_j) \prod_{m \neq i,j} f_m(Y_m), \\ \tilde{g}_{j1}(Y_{-i}|Y_i) &= \tilde{g}_{j0}(Y_{-i}|Y_i)\Big(\sum_{k \neq i,j} p_k(Y_i)\frac{f_j}{f_k}(Y_k) + p_i(Y_i)\frac{f_j}{f_i}(Y_j) + p_j(Y_i)\Big). \end{aligned}$$

Note that $g_{j0} = \tilde{g}_{j0} \circ \mathbb{K}$ and $g_1 = \tilde{g}_{j1} \circ \mathbb{K}$, where $\mathbb{K}$ is the Markov kernel that takes Y_{-i} to a random permutation of itself. It follows from Lemma 2.1 that

$$\begin{aligned} \mathrm{E}\big(|W_j(Y_{-i}, Y_i) - 1|^2 \big| Y_i\big) &\leq \mathrm{E}_{\tilde{g}_{j1}}\Big(\frac{\tilde{g}_{j0}}{\tilde{g}_{j1}} - 1\Big)^2 \\ &= \mathrm{E}_{\tilde{g}_{j0}}\Big(\frac{\tilde{g}_{j0}}{\tilde{g}_{j1}} - 2 + \frac{\tilde{g}_{j1}}{\tilde{g}_{j0}}\Big). \end{aligned} \tag{5.2}$$

This expectation does not depend on i except for the value of Y_i. Hence, to simplify notation, we take WLOG $i = j$. Denote

$$\begin{aligned} L &= \frac{\tilde g_{j1}}{\tilde g_{j0}} = p_j(Y_j) + \sum_{k\neq i} p_k(Y_j)\frac{f_j}{f_k}(Y_k), \\ V &= \frac{n}{4}\gamma \min_k p_k(Y_j), \end{aligned}$$

where γ is as in **(G1)**. Then by (5.2)

$$\begin{aligned} \mathrm{E}\big(|W_j(Y_{-j},Y_j)-1|^2\big|Y_j\big) &\le \mathrm{E}_{\tilde g_{j0}}\Big(\frac{1}{L}-2+L\Big) \\ &= \mathrm{E}_{\tilde g_{j0}}\frac{(L-1)^2}{L} \\ &\le \frac{1}{V}\mathrm{E}_{\tilde g_{j0}}(L-1)^2\mathbf{I}(L>V) + \mathrm{E}_{\tilde g_{j0}}\frac{\mathbf{I}(L\le V)}{L} \\ &\le \mathrm{E}_{\tilde g_{j0}}\frac{\mathbf{I}(L\le V)}{L} + \frac{1}{V}\sum_{k=1}^n p_k^2(Y_j), \end{aligned} \tag{5.3}$$

by **G1**. Bound

$$L \ge \gamma \min_k p_k(Y_j)\sum_{k=1}^n \mathbf{I}\Big(\frac{f_j}{f_k}(Y_k)>\gamma\Big) \ge \gamma\min_k p_k(Y_j)(1+U),$$

where $U \sim B(n-1,1/2)$ (the 1 is for the ith summand). Hence

$$\begin{aligned} \mathrm{E}_{\tilde g_{j0}}\frac{\mathbf{I}(L\le V)}{L} &\le \frac{1}{\gamma\min_k p_k(Y_j)}\sum_{k=0}^{\lceil n/4\rceil}\frac{1}{k+1}\binom{n-1}{k}2^{-n+1} \\ &= \frac{1}{\gamma n\min_k p_k(Y_j)}\sum_{k=0}^{\lceil n/4\rceil}\binom{n}{k+1}2^{-n+1} \\ &= O(e^{-n})\frac{1}{\gamma n\min_k p_k(Y_j)} \end{aligned} \tag{5.4}$$

by large deviation.

From **(G1)**, **(G2)**, (5.1), (5.3), and (5.4):

$$\begin{aligned} \mathrm{E}\,\mathrm{E}\big((\mu_i^S-\mu_i^{PI})^2\big|Y_i\big) &= \mathrm{E}\,\mathrm{E}\left(\left(\sum_{j=1}^n \mu_j p_j(Y_i)\big(W_j(Y_{-i},Y_i)-1\big)\right)^2\middle|Y_i\right) \\ &\le \max_j|\mu_j|^2\mathrm{E}\left(\left(\sum_{j=1}^n p_j(Y_i)\mathrm{E}\big(W_j(Y_{-i},Y_i)-1\big)^2\right)\middle|Y_i\right) \\ &\le \kappa C^3/n, \end{aligned}$$

for some κ large enough. Claim (i) of the theorem follows. Claim (ii) follows (i) by Proposition 1.1. □

References

[1] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

[2] Brown, L. D. and Greenshtein, E. (2009). Non parametric empirical Bayes and compound decision approaches to estimation of a high dimensional vector of normal means. *Ann. Statist.* **37** 1685–1704.

[3] Copas, J. B. (1969). Compound decisions and empirical Bayes (with discussion). *J. Roy. Statist. Soc. Ser. B* **31** 397–425.

[4] Diaconis, P. and Freedman, D. (1980). Finite exchangeable sequences. *Ann. Probab.* **8** 745–764.

[5] Hannan, J. F. and Robbins, H. (1955). Asymptotic solutions of the compound decision problem for two completely specified distributions. *Ann. Math. Statist.* **26** 37–51.

[6] Lehmann, E. L. (1986). *Testing Statistical Hypothesis*, 2nd ed. Wiley, New York.

[7] Robbins, H. (1951). Asymptotically subminimax solutions of compound decision problems. In *Proc. Third Berkeley Symp.* 157–164.

[8] Robbins, H. (1983). Some thoughts on empirical Bayes estimation. *Ann. Statist.* **11** 713–723.

[9] Samuel, E. (1965). On simple rules for the compound decision problem. *J. Roy. Statist. Soc. Ser. B* **27** 238–244.

[10] Zhang, C. H. (2003). Compound decision theory and empirical Bayes methods. (Invited paper.) *Ann. Statist.* **31** 379–390.

[11] Wenuha, J. and Zhang, C. H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* To appear.