ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 23 — 1991

Point Estimation and Confidence Set Estimation in a Parallelism Model: an Empirical Bayes Approach

Christian P. ROBERT, A. K. Md. Ehsanes SALEH *

ABSTRACT. – When several simple regression models are assumed to have similar slopes, empirical Bayes methods can efficiently process this vague information by estimating the hyperparameters of a conjugate prior. The shrinkage estimators we obtain are shown to be minimax and, furthermore, dominate usual confidence regions in terms of coverage probability.

Estimateurs et régions de confiance pour un modèle de régression en parallèle : une résolution bayésienne empirique

RÉSUMÉ. – Lorsque plusieurs équations de régression simple sont supposées avoir une pente similaire, la meilleure méthode tirant profit de cette information est l'approche bayésienne empirique, qui imite l'approche bayésienne en estimant les paramètres de la loi *a priori*. Les estimateurs ainsi obtenus sont minimax et, de plus, les régions de confiance qui s'en déduisent dominent les régions usuelles (au sens où elles ont une plus grande probabilité de couverture).

---

* C. ROBERT : LSTA, Université Paris-VI, 4, place Jussieu, F-75252 Paris Cedex 05; A. K. Md. E. SALEH: Carleton University, Dept. of Math. & Stat., Ottawa, Ontario, Canada K1S5B6. The first author is grateful to E. George for his numerous and constructive comments which contributed to a better presentation of the paper. We also thank the referees which greatly help to develop this paper from an earlier version in order to adapt it to a wider audience. This research was supported by Natural Sciences and Engineering Research Canada Grant No. A-3088 and l'Institut National de la Statistique et des Études Économiques (Paris, France). It was performed while the first author was visiting Carleton University.

1 Introduction

One of the main justifications of the Bayesian approach is to provide an efficient method to deal with prior information in a statistical estimation problem under the assumption that this prior information can be represented as a probability distribution. However, this ideal setting rarely occurs in practice and it is often the case that this information is not precise enough (or not quantitative enough) to be modeled as a prior distribution. In these situations of *uncertain prior information* (also called *diffuse information*), the statistician has to make use of some clues about the parameters of the model that lead him (or her) to suspect the existence of a functional relation between these parameters is rather likely. For instance, for a linear model and the simultaneous estimation of several regression equations, there may be a linear relation existing between the coefficients of the exogeneous variables.

In a situation of uncertain prior information, the empirical Bayes method is able to make use of the vague prior information in order to produce estimators which take advantage of this information, while being *robust* with respect to a possible misspecification of this prior information. This approach is located at the interface between frequentist (*i. e.* classical) and Bayesian domains, and was introduced by ROBBINS [1955, 1964] in a non-parametric setting. Since the prior information is not precise enough to provide a prior probability distribution on the parameters β of the model, the empirical Bayes approach starts with a mathematically tractable prior depending on unspecified *hyperparameters*, $\pi(\beta \mid \lambda)$, and on the information, and *estimates* these hyperparameters λ by a classical method (moments, maximum likelihood, etc.) through the *marginal* distribution of the observations. Once an estimation $\hat{\lambda}$ is obtained, the empirical Bayes approach proceeds similarly to a usual Bayesian method, but with the *estimated prior distribution*, $\pi(\beta \mid \hat{\lambda})$. For a detailed coverage of empirical Bayes techniques, *see* MARITZ and LWIN [1989].

Even though the prior may be chosen in relation with the prior information, this method is obviously not very dependable from a Bayesian point of view. This is why the *validation* of empirical Bayes methods rather pertains to the frequentist criteria. For a given loss function, one has to show that the *risk* of the produced estimator is at least as small as the risk of the classical estimator (e. g., the LSE, in the case of the linear model). Therefore, once the prior is selected, we will be concerned in this paper with the frequentist properties of our estimator and will not try to justify the chosen prior from a Bayesian point of view. Note that there are strong theoretical arguments in favor of the use of Bayesian methods (considered merely as *tools*) in a frequentist decision theoretic framework (*see* BERGER [1985]). Actually, when faced with this problem, a Bayesian would rather model the uncertainty on the prior information by building up another distribution on the hyperpriors λ, thus leading to the so-called *hierarchical Bayes approach*. It has also been argued that this approach is superior to the empirical Bayes approach (*see* Berger [1985] or Berger and Robert [1990]), but if often leads to a difficult evaluation of the frequentist properties, even though the computational aspects have been partially alleviated recently (*see* Gelfand and Smith [1990]).

A frequentist alternative to the empirical Bayes approach would be to use a *restricted least squares estimator* (RLSE), since it takes into account the relation between the regression equations. However, this estimator is not *robust w.r.t. a misspecification of the relation.* Namely, the risk of the RLSE is usually unbounded when the relation is not satisfied. And it often occurs, particularly in econometric models, that the assumed relation cannot be taken for granted. (Some theoretical economic property may be inapplicable for the considered case, an unobserved change of model may have occurred for the last equation, etc.). The empirical Bayes estimator has therefore the advantage to be more "flexible", since it appears as a combination of the least squares estimator (LSE) and the RLSE, the weights being determined by the likelihood of the alleged relation. This combination allows for better frequentist properties, e. g. *uniform domination of the LSE*, and good performances when the relation is actually satisfied.

In this paper, we illustrate the problem of uncertain prior information and its empirical Bayes resolution through the example of a *parallelism model*, namely the case of several *simple* regression equations where the slopes are likely to be quite similar. This type of model appears in Econometrics (*see* Section 2) and in simple analysis of covariance models (ANCOVA), in particular in some bioassays problems (*see* Sen [1971]). Moreover, the simplicity of this restrictive model allows for a better exposure of the methods we present here, even through they can be used in more complicated (and more realistic) settings. In particular, we show that our empirical Bayes estimators have actually good frequentist properties under squared error loss, namely that they are *minimax.* These results are related with Ghosh, Saleh and Sen [1989]. They studied the estimation of regression coefficients when a subset of coefficients are suspected to be null. In fact, by using a suitable linear transformation, the parallelism model can be expressed in this form. However, apart from the fact that working with the original model rather than with a transformed model is always easier, the results we get are more general as they are also applicable to the whole vector of parameters as well as to the *intercepts.* Furthermore, by deriving minimaxity results from a previous sufficient minimaxity condition (Fraisse, Robert and Roy [1987]), we show the potentialities of this condition for more intricated hypotheses. Moreover, we also establish the domination of the truncated estimator (*see* Section 5).

Another related reference is Hui and Berger [1983], who discuss the empirical Bayes estimation of several simple regression equations under an exchangeability assumption; they also consider different variances for the regressions equations. In this respect, their model is more general than ours. However, they do not consider the risk properties of the deduced estimators, because they present this approach as an alternative to an untractable hierarchical Bayes modeling. More generally, the empirical Bayes approach is also related to the *Stein effect*, *i.e.* the domination of the LSE by shrinkage estimators, as shown in EFRON and MORRIS [1975] and MORRIS [1983], because it exposes the underlying reasons for this seemingly paradoxical phenomenon (*see* also BERGER [1985]).

In Section 6, we propose some additional results about *confidence set* estimation for the parallelism problem, in order to give a complete estimation procedure. Following CASELLA and HWANG [1987], we give arguments in favor of *recentered confidence sets* for the estimation of the intercepts.

Let us stress once more that the restricted model we consider here is mainly intended to *illustrate* the techniques of the empirical Bayes approach, but that the empirical Bayes scope goes far beyond this type of model, as shown in HUI and BERGER [1983], BERGER and CHEN [1987] or GHOSH *et al.* [1989]. For instance, the restriction to equal variances (*homocedasticity*) for the different equations can be removed, while keeping the minimaxity results. Moreover, in CELLIER, FOURDRINIER and ROBERT [1990] and KUBOKAWA, ROBERT and SALEH [1991], we established that the normal framework is not necessary and can be replaced by a *spherical symmetry* assumption, at no loss for the minimaxity conditions, even if the distributions of the errors are only partially known.

2 Model

The model we are considering consists of p simple regression equations

$$y_i = \theta_i + \beta_i t_i \qquad (1 \leqslant i \leqslant p). \tag{1}$$

For each equation. we consider n_i independent observations $(1 \leqslant i \leqslant p)$; so we have

$$\mathbf{y}_i = \theta_i \mathbf{1}_i + \beta_i \mathbf{t}_i + \varepsilon_i \tag{2}$$

where

$$\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^t,$$
$$\mathbf{1}_i = (1, \ldots, 1)^t \in \mathbb{R}^{n_i}, \qquad \mathbf{t}_i = (t_{i1}, t_{i2}, \ldots, t_{in_i})^t$$

and $\varepsilon_i \sim N_{n_i}(0, \sigma^2 \mathbf{I}_{n_i})$. Furthermore, we assume that the vectors ε_i are mutually independent. We are mainly interested in the estimation of the vector of the *intercepts*, $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$, given that the components of the *slopes* vector $\beta = (\beta_1, \ldots, \beta_p)^t$ are likely to be equal. An alternative study for this *parallelism problem* and some additional references are given in SALEH and SEN [1986].

A justification for this model is provided by the following example: the p equations (1) represent the production costs for p factories, $\mathbf{y}_i$ being the costs and $\mathbf{t}_i$ the outputs (production volume). Assuming that all these enterprises produce the same object, cars say, and that they are at the same technological level, it is quite reasonable to suppose that the slopes β_i are the same. However, in a study about the profitabilities of these enterprises, one would be mainly concerned with the level of the constant cost (or *entry cost*), θ_i, which can be really different from each factory, due to some structural constraints (distance from raw materials plants, cost of buildings, etc.). The model is also applicable in bioassay problems (*see* SEN [1971]).

We give below the formulae for the least squares and restricted least squares estimators. These formulae are obviously well-known, but are of use in the sequel. Let us denote by $\gamma=(\theta^t, \beta^t)^t$, the whole vector of parameters. Then, for the model (2), the *least squares estimator* of γ is $\tilde{\gamma}=(\tilde{\theta}^t, \tilde{\beta}^t)^t$, given by $(1\leqslant i\leqslant p)$

(3) $$\tilde{\theta}_i=\bar{y}_i-\tilde{\beta}_i\bar{t}_i, \qquad \tilde{\beta}_i=\frac{\sum_{j=1}^{n_i}(t_{ij}-\bar{t}_i)\,y_{ij}}{\sum_{j=1}^{n_i}(t_{ij}-\bar{t}_i)^2},$$

where $\bar{y}_i=\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ and $\bar{t}_i=\frac{1}{n_i}\sum_{j=1}^{n_i} t_{ij}$. If we write $\mathbf{X}_i=(\mathbf{1}_i\, t_i)$, we have

(4) $$(\mathbf{X}_i^t\mathbf{X}_i)^{-1}=\begin{pmatrix}\dfrac{t_i^t\,t_i}{d_i n_i} & -\dfrac{\bar{t}_i}{d_i}\\ -\dfrac{\bar{t}_i}{d_i} & \dfrac{1}{d_i}\end{pmatrix},$$

with $d_i=\sum_{j=1}^{n_i}(t_{ij}-\bar{t}_i)^2$, and the variance covariance matrix of $\tilde{\gamma}$ is given by $\sigma^2\boldsymbol{\Sigma}$, with

(5) $$\boldsymbol{\Sigma}=\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{21}\\ \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{22}\end{pmatrix}$$

and

$$\boldsymbol{\Sigma}_{11}=\operatorname{diag}\left(\frac{\mathbf{t}_i^t\mathbf{t}_i}{n_i d_i}\right),$$

$$\boldsymbol{\Sigma}_{12}=\boldsymbol{\Sigma}_{21}=\operatorname{diag}\left(\frac{-\bar{t}_i}{d_i}\right), \qquad \boldsymbol{\Sigma}_{22}=\operatorname{diag}\left(\frac{1}{d_i}\right).$$

In the case of the restricted model, (2) becomes

(6) $$\mathbf{y}_i=\theta_i\mathbf{1}_i+\beta_0\mathbf{t}_i+\varepsilon_i \qquad (1\leqslant i\leqslant p)$$

and the *restricted least squares estimator* of γ is $\hat{\gamma}=(\hat{\theta}, \beta \mathbf{1}_p)$ with

$$(7) \qquad \hat{\beta}_0=\frac{\sum_{i=1}^{p} d_i \tilde{\beta}_i}{\sum_{i=1}^{p} d_i}, \qquad \hat{\theta}_i=\bar{y}_i-\hat{\beta}_0 \bar{t}_i.$$

The RLSE can be expressed as a linear transformation of the LSE. If we denote $\mathbf{q}=(d_1, d_2, \ldots, d_p)^t$ and $d=\sum_{i=1}^{p} d_i$, it is easy to see that

$$(8) \qquad \hat{\gamma}=\begin{pmatrix} \mathbf{I} & \mathbf{D} \\ \mathbf{O} & \mathbf{P}_2 \end{pmatrix}\tilde{\gamma}$$

where

$$\mathbf{P}_2=\frac{1}{d}\mathbf{1}\,\mathbf{q}^t \quad \text{and} \quad \mathbf{D}=\operatorname{diag}(\bar{t}_i)(\mathbf{I}-\mathbf{P}_2).$$

3 Empirical Bayes Estimation

By sufficiency considerations, we can reduce the model to the observation of the statistics $\tilde{\gamma} \sim N_{2p}(\gamma, \sigma^2 \Sigma)$ and $s^2 \sim \sigma^2 \chi^2_{n-2p}/(n-2p+2)$, where

$$s^2=\sum_{i=1}^{p} \|\mathbf{y}_i-\tilde{\theta}_i \mathbf{1}_i-\tilde{\beta}_i \mathbf{t}_i\|^2/(n-2p+2),$$

and $n=\sum_{i=1}^{p} n_i$. This estimator of σ^2 is actually *the best scale equivariant estimator* under the loss

$$(9) \qquad L(\sigma^2, d)=\left(1-\frac{d}{\sigma^2}\right)^2$$

and thus dominates the maximum likelihood and the unbiased estimators of σ^2.

Since the available prior information says that the slopes β_i are likely to be equal, we can select a prior distribution on the parameters, $\pi(\beta, \sigma^2)$, which reflects this piece of information. Furthermore, the *evaluation* of the estimator being frequentist, the choice of this prior distribution does not need to be thoroughly justified on a subjective basis, as long as the resulting estimator has good frequentist properties. In the normal case, a possible choice for γ is to take

$$\gamma \sim \mathcal{N}_{2p}(\mathrm{v}, \tau^2 \mathbf{V}),$$

where $\mathbf{V}$ is a $2p \times 2p$ matrix. This prior is called *conjugate* because the posterior distribution is in the same family, namely

$$\gamma \mid \tilde{\gamma}, \sigma \sim \mathcal{N}_{2p}([\sigma^2 \mathbf{V}^{-1} + \tau^2 \mathbf{\Sigma}^{-1}]^{-1} [\sigma^2 \mathbf{V}^{-1} \mathrm{v} + \tau^2 \mathbf{\Sigma}^{-1} \tilde{\gamma}], \\ [\tau^{-2} \mathbf{V}^{-1} + \sigma^{-2} \mathbf{\Sigma}^{-1}]^{-1}).$$

Conjugate priors are intensively used in empirical Bayes analysis, mainly because they lead to analytically tractable posterior distributions, even though their use is not necessarily justified from a Bayesian point of view (*see* DE GROOT [1970] or BERGER [1985]). For the parallelism model, a reasonable choice for v is $\mathrm{v}^t = (\mathrm{v}_1^t, \mathrm{v}_2 \mathbf{1}^t)$, with $\mathrm{v}_1 \in \mathbb{R}^p$. Actually, since the coefficients β_i $(1 \leqslant i \leqslant p)$ are likely to be identical, or at least similar, we can formalize this information by assuming that these coefficients have the same prior mean.

Two usual choices for the variance-covariance matrix $\mathbf{V}$ are $\mathbf{V} = \mathbf{I}_{2p}$ and $\mathbf{V} = \mathbf{\Sigma}$, because they lead to tractable expressions. The first choice corresponds to *ridge regression* (*see* GHOSH *et al.* [1989]) and the second choice is commonly called *g-prior* by ZELLNER [1971, 1986] (*see* also BLATTBERG and GEORGE [1991]). We will consider this choice of $\mathbf{V}$, as it leads to easier computations and good properties of the resulting estimators. A drawback is that this prior somehow depends on the observed data. However, as the entire model is conditional on $\mathbf{t}_i$ $(1 \leqslant i \leqslant p)$, this criticism may be easily sidepassed (for instance, the coefficients β_i themselves are defined with respect to the $\mathbf{t}_i$'s). Another common attack against *g*-priors is that they are not intuitively sound. The fact is that they do not correspond to exchangeable priors for the coefficients (θ_i, β_i) but for the means, $\theta_i + \beta_i t_i$. One may still feel legitimate to use such priors, if exchangeable priors on the means are more justified. For instance, they are invariant under rescaling and translation of the regressors. At last, recall that we consider the prior distribution mainly as a *tool* that produces a potentially interesting estimator: the fact that this prior may be subjectively unsound has no relevance whatsoever from a frequentist point of view.

Therefore, we will use in the sequel the prior distribution

$$\gamma \sim \mathcal{N}_{2p}(\mathrm{v}, \tau^2 \mathbf{\Sigma}),$$

which gives the following posterior distribution conditionally on σ^2,

(10) $$\gamma \mid \tilde{\gamma}, \sigma^2 \sim \mathcal{N}_{2p}\left(\frac{\sigma^2 \mathrm{v} + \tau^2 \tilde{\gamma}}{\sigma^2 + \tau^2}, \frac{\tau^2 \sigma^2}{\sigma^2 + \tau^2} \mathbf{\Sigma}\right).$$

Conditionally on σ^2, the associated posterior mean can be written

(11) $$\delta^\pi(\tilde{\gamma}) = \frac{\sigma^2 \mathrm{v} + \tau^2 \tilde{\gamma}}{\sigma^2 + \tau^2} = \mathrm{v} + \left(1 - \frac{\sigma^2}{\sigma^2 + \tau^2}\right)(\tilde{\gamma} - \mathrm{v}).$$

Since no prior information is available about the variance factor σ^2, a possible choice, from a Bayesian point of view, is to use the so-called *non-informative prior*, $\pi(\sigma^2)=1/\sigma^2$. We can also consider σ^2 as another hyperparameter and estimate it from the marginal distribution, namely from $(n-2p+2)s^2 \sim \sigma^2 \chi^2_{n-2p}$, which leads to s^2 for the equivariant loss (9). If we use the non-informative prior with this loss, the Bayes estimator of σ^2 is also s^2.

Similarly, the empirical Bayes approach for the *estimation* of the hyperparameters ν and $\tau^2$is based on the *marginal distribution* of $\tilde{\gamma}$,

$$\tilde{\gamma} \mid \sigma^2 \sim \mathcal{N}_{2p}(\nu, (\sigma^2+\tau^2)\Sigma).$$

As

(12) $$\mathbf{E}^m[\hat{\beta}]=\nu_2, \qquad \mathbf{E}^m[\hat{\theta}]=\nu_1,$$

where $\mathbf{E}^m[.]$ denotes the expectation with respect to the marginal distribution, we will take $\hat{\beta}$, $\hat{\theta}$ as empirical Bayes estimators of ν_2, ν_1. Furthermore, the variance-covariance matrix of $(\tilde{\beta}-\hat{\beta}\,\mathbf{1})$ is

(13) $$\mathbf{V}(\tilde{\beta}-\hat{\beta}\,\mathbf{1})=(\sigma^2+\tau^2)(\mathbf{I}_p-\mathbf{P}_2)\boldsymbol{\Sigma}_{22}(\mathbf{I}_p-\mathbf{P}_2)^t$$

and

(14) $$(\tilde{\beta}-\hat{\beta}\,\mathbf{1})\boldsymbol{\Sigma}_{22}^{-1}(\tilde{\beta}-\hat{\beta}\,\mathbf{1}) \sim (\sigma^2+\tau^2)\chi^2_{p-1}.$$

Therefore,

$$\mathbf{E}^m\left[\frac{1}{(\tilde{\beta}-\hat{\beta}\,\mathbf{1})\boldsymbol{\Sigma}_{22}^{-1}(\tilde{\beta}-\hat{\beta}\,\mathbf{1})}\right]=\frac{1}{(p-3)(\sigma^2+\tau^2)}$$

and an unbiased estimator of $(\sigma^2+\tau^2)^{-1}$ is

(15) $$\frac{p-3}{(\tilde{\beta}-\hat{\beta}\,\mathbf{1})^t\boldsymbol{\Sigma}_{22}^{-1}(\tilde{\beta}-\hat{\beta}\,\mathbf{1})}.$$

Since we derived estimators of the hyperparameters and of σ^2 through (14), (15) and s^2, we can propose the following empirical Bayes estimator

(16) $$\varphi^{\mathrm{EB}}(\mathbf{y})=\hat{\gamma}+\left(1-\frac{(p-3)s^2}{(\tilde{\beta}-\hat{\beta}\,\mathbf{1})^t\boldsymbol{\Sigma}_{22}^{-1}(\tilde{\beta}-\hat{\beta}\,\mathbf{1})}\right)(\tilde{\gamma}-\hat{\gamma}),$$

which is the estimated version of (11). φ^{EB} is thus a convex combination of the LSE and the RLSE, with weights depending logically on the test statistics,

(17) $$\frac{(\tilde{\beta}-\hat{\beta}\,\mathbf{1})^t\boldsymbol{\Sigma}_{22}^{-1}(\tilde{\beta}-\hat{\beta}\,\mathbf{1})}{s^2},$$

which is an F-statistics for the test on the equality of the slopes. This estimator generalizes Stein-type estimators (*see* Cellier *et al.* [1989] or Fraisse *et al.* [1987]). Actually, if we define, for $\mathbf{z}\in\mathbb{R}^{2p}$, the norm $\|\mathbf{z}\|_*^2=z^t\boldsymbol{\Sigma}^{-1}z$, it can be established, by straightforward algebra, that

$$(\tilde{\beta}-\hat{\beta}\,\mathbf{1})\boldsymbol{\Sigma}_{22}^{-1}(\tilde{\beta}-\hat{\beta}\,\mathbf{1})=\|\tilde{\gamma}-\hat{\gamma}\|_*^2$$

and (16) can be written as

$$\varphi^{EB}(\mathbf{y})=\hat{\gamma}+\left(1-\frac{(p-3)\,s^2}{\|\tilde{\gamma}-\hat{\gamma}\|_*^2}\right)(\tilde{\gamma}-\hat{\gamma}).$$

A deduced empirical Bayes estimator for the intercepts, $\theta_i\,(1\leqslant i\leqslant p)$, is then

$$\varphi^{EB}(\mathbf{y})=\hat{\theta}+\left(1-\frac{(p-3)\,s^2}{(\tilde{\beta}-\hat{\beta}\,\mathbf{1})\,\Sigma_{22}^{-1}\,(\tilde{\beta}-\hat{\beta}\,\mathbf{1})}\right)(\tilde{\theta}-\hat{\theta}) \tag{18}$$

Therefore, it is possible to use the assumption on the slopes of the model to improve the estimation of the intercepts, as we will establish rigorously in the next section.

Remark 1: The smaller (17) is, the closer (16) should be from the RLSE. However, when the statistics (17) goes to 0, the estimator has a strange behavior which will be corrected in Section 5, but does not impair minimaxity results.

Remark 2: An estimator similar to (16) has been obtained by RAO [1983], following a non-Bayesian approach.

Remark 3: As in every empirical Bayes study, the choice of the estimators of the hyper-parameters is partially arbitrary, since there is no Bayesian guideline to show which estimator is best. For instance, σ^2 and $(\sigma^2+\tau^2)^{-1}$ could be estimated by *maximum likelihood* and this would lead to the constant

$$\frac{n-2p+2}{n-2p}\,p,$$

instead of $(p-3)$ in the estimators (15) and (16). However, we will see in the next section that the constant $(p-3)$ is optimal from a Bayesian point of view.

4 Minimaxity

The estimators (16) and (18) have now to be evaluated under a frequentist criterion. We use quadratic risks associated with symmetric non-negative matrices $\mathbf{Q}$,

$$R(\varphi,\gamma)=\mathbf{E}_{\gamma,\sigma}[(\varphi(\mathbf{y})-\gamma)^t\,\mathbf{Q}\,(\varphi(\mathbf{y})-\gamma)]. \tag{19}$$

The empirical Bayes estimator φ^{EB} *dominates* the LSE $\tilde{\gamma}$ if it satisfies, for every (γ, σ^2),

$$R(\varphi^{EB}, \gamma) \leqslant R(\tilde{\gamma}, \gamma) = \mathrm{tr}(\mathbf{\Sigma Q}).$$

In this case, since the least square estimator $\tilde{\gamma}$ is *minimax* [*i.e.* $\mathrm{tr}(\mathbf{\Sigma Q}) \leqslant \sup_\gamma R(\delta, \gamma)$ for every estimator δ], the empirical Bayes is minimax too. (Note that, since $\tilde{\gamma}$ has a constant risk, any other minimax estimator dominates $\tilde{\gamma}$.)

Minimaxity is often an important property in empirical Bayes studies, since it validates these estimators from a frequentist point of view and also shows the robustness of the estimator w.r.t. a misspecification of the prior, partially chosen for computational reasons. As the papers quoted in the introduction have shown, it is usually the case that empirical Bayes estimators behave well for this criterion. (Hui and Berger [1983] do not consider minimaxity because they work from a Bayesian perspective and only use empirical Bayes techniques for their computational convenience.)

When the empirical Bayes estimators can be expressed in a simple form, it is often possible to apply general sufficient minimaxity conditions (*see*, e. g., Judge and Bock [1978], Stein [1981], or Berger [1985]). For instance, in the parallelism model, for the estimation of γ as well as the estimation of θ, one may derive sufficient minimaxity conditions from Fraisse *et al.* [1987], as shown below. In fact, the same derivation could have been done for Ghosh *et al.* [1989].

4.1. Preliminaries

Although we are mainly interested in the estimation of the intercepts, θ, we still consider the whole estimator (16) because the minimaxity of (18) then appears as a corollary. Note that, by (8), (16) can be written as

(20) $$\varphi^{EB}(\mathbf{y}) = \left(\mathbf{I}_{2p} - \frac{(p-3)\, s^2}{\|\tilde{\gamma} - \hat{\gamma}\|_*^2}\, \mathbf{C}\right)\tilde{\gamma}$$

where

$$\mathbf{C} = \begin{pmatrix} \mathbf{0} & -\mathbf{D} \\ \mathbf{0} & \mathbf{I} - \mathbf{P}_2 \end{pmatrix}.$$

Furthermore,

$$\|\tilde{\gamma} - \hat{\gamma}\|_*^2 = \tilde{\gamma}^t \mathbf{B} \tilde{\gamma},$$

with

$$\mathbf{B} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{I} - \mathbf{P}_2)^t \mathbf{\Sigma}_{22}^{-1} (\mathbf{I} - \mathbf{P}_2) \end{pmatrix}$$

Therefore, the estimator (20) belongs to the general class,

(21) $$(\mathbf{I}_{2p} - h(\tilde{\gamma}^t \mathbf{B} \tilde{\gamma}/s^2)\, \mathbf{C})\, \tilde{\gamma},$$

considered in JUDGE and BOCK [1978]. Their minimaxity results under (19) have been generalized in many directions. In particular, FRAISSE *et al.* [1987] have examined larger families of matrices **B** and **C**, including *non-definite matrices*, which are of interest in this case (Proposition 3).

As we are focussing on the estimation of the intercepts, we will consider with particular care matrices of the form

$$\mathbf{Q}=\begin{pmatrix}\mathbf{Q}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix},$$

where $\mathbf{Q}_1$ is $p\times p$, since they only penalize the estimation of θ. As mentioned above, to work on with estimators of γ (rather than with the intercepts θ) makes the derivation from FRAISSE *et al.* [1987] easier.

4.2. Bayesian Optimality of the Constant c

Before considering the minimaxity of the estimators (16) and (18), we derive from FRAISSE *et al.* the optimality of the empirical Bayes estimator (16) among the estimators

$$\varphi_c(\mathbf{y})=\left(\mathbf{I}_{2p}-c\,\frac{s^2}{\tilde{\gamma}^t\mathbf{B}\tilde{\gamma}}\,\mathbf{C}\right)\tilde{\gamma}, \tag{22}$$

independently of the loss matrix $\mathbf{Q}$; in fact, for $c^*=p-3$, φ_{c^*} minimizes the Bayes risk in the class (22). Similar results have been obtained by MORRIS [1983], GHOSH *et al.* [1989] and GHOSH [1989]. They are quite interesting, as they generalize JAMES and STEIN's result [1961], *i.e.* the fact that an optimal constant exists for the *frequentist risk*, to a larger class of models and estimators when one considers the *Bayes risk*.

The following lemma, derived from FRAISSE *et al.* [1987] (Proposition 2), gives an *unbiased estimator* of the risk of a shrinkage estimator of the class (22).

LEMMA 1 : The frequentist risk of φ_c,

$$\mathrm{R}(\varphi_c;\,\gamma,\,\sigma^2)=\frac{1}{\sigma^2}\mathbf{E}_{\gamma,\sigma}[(\varphi_c(\mathbf{y})-\gamma)^t\,\mathbf{Q}\,(\varphi_c(\mathbf{y})-\gamma)],$$

can be expressed as

$$\mathrm{R}(\varphi_0;\,\gamma,\,\sigma^2) \tag{23}$$
$$-2\,\mathbf{E}_{\gamma,\sigma}\left[c\,\frac{s^2}{\tilde{\gamma}^2\mathbf{V}\tilde{\gamma}}\left(\operatorname{tr}(\Sigma\mathbf{Q}\mathbf{C})-2\,\frac{\tilde{\gamma}^t\mathbf{B}\Sigma\mathbf{Q}\mathbf{C}\tilde{\gamma}}{\gamma^t\mathbf{B}\tilde{\gamma}}\right)\right]\frac{n-2p+2}{n-2p}$$
$$+\mathbf{E}_{\gamma,\sigma}\left[c^2\,\frac{s^2\,\tilde{\gamma}^t\,\mathbf{C}^t\,\mathbf{Q}\mathbf{C}\tilde{\gamma}}{(\tilde{\gamma}^t\mathbf{B}\tilde{\gamma})^2}\right]\frac{n-2p+2}{n-2p}.$$

Taking the expectation of (23) w.r.t. the prior distribution (10) on γ, we get the following result (*see* Appendix 1 for the proof).

PROPOSITION 2 : For any prior distribution on σ^2 and the prior distribution (10) on γ, the estimator minimizing the Bayes risk among the class (22) is φ_{c^*}.

4.3. Risk Domination of the LSE

If we come back to the frequentist perspective under (19), we have the following sufficient minimaxity condition (FRAISSE *et al.* [1987], Proposition 8). As $\mathbf{B}$ is not necessarily invertible, the condition relies on a *generalized inverse* of $\mathbf{B}$, *i.e.* any matrix $\mathbf{B}^-$ satisfying $\mathbf{B}\mathbf{B}^-\mathbf{B}=\mathbf{B}$.

PROPOSITION 3: An estimator of the class (21) will dominate uniformly the least squares estimator for the risk associated with $\mathbf{Q}$, if, for every $t \geqslant 0$,

(*a*) $th(t)$ is non-decreasing;

(24) (*b*) $$0 \leqslant th(t) \leqslant 2\frac{tr(\Sigma\mathbf{Q}\mathbf{C}) - 2\,ch_{\max}(\Sigma\mathbf{Q}\mathbf{C})}{ch_{\max}(\mathbf{C}^t\mathbf{Q}\mathbf{C}\mathbf{B}^-)},$$

where $ch_{\max}$ denotes the maximum eigenvalue.

Remark 4: Two generalized inverses of $\mathbf{B}$ are

$$\mathbf{B}_1^- = \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \quad \text{and} \quad \mathbf{B}_2^- = \begin{pmatrix} 0 & 0 \\ 0 & (\mathbf{I}-\mathbf{P}_2)\Sigma_{22}(\mathbf{I}-\mathbf{P}_2)^t \end{pmatrix},$$

as $\mathbf{I}-\mathbf{P}_2$ is idempotent.

For the estimation of the intercepts, we have the following consequence (*see* Appendix 2 for the proof).

COROLLARY 4: The estimator

$$\varphi(\mathbf{y}) = \hat{\theta} + [1 - h(\|\tilde{\gamma} - \hat{\gamma}\|_*^2/s^2)](\tilde{\theta} - \hat{\theta})$$

dominates the LSE $\tilde{\theta}$ for the risk associated with $\mathbf{Q}_1$ if

(*a*) $th(t)$ is non-decreasing;

(*b*) $$0 \leqslant th(t) \leqslant 2\left(\left[\frac{tr(-\Sigma_{12}\mathbf{Q}_1\mathbf{D})}{ch_{\max}(-\Sigma_{12}\mathbf{Q}_1\mathbf{D})}\right] - 2\right).$$

For instance, if $\mathbf{Q}_1 = \text{diag}(d_i \bar{t}_i^2)$, the above bound in (*b*) is $2(p-3)$ and the empirical Bayes estimator (18) is minimax. For other choices of $\mathbf{Q}_1$ ($\mathbf{I}_p$, Σ_{11}^{-1}, etc.), the minimaxity of (18) will depend on the relative magnitude of the regressors t_i. For instance, if $\mathbf{Q}_1 = \Sigma_{11}^{-1}$, the quantity

$$\sum_{i=1}^{p} \frac{n_i \bar{t}_i^2}{\mathbf{t}_i^t \mathbf{t}_i} \Big/ \max_{1 \leqslant i \leqslant p} \left(\frac{n_i \bar{t}_i^2}{\mathbf{t}_i^t \mathbf{t}_i}\right)$$

must be larger than 2. Therefore, there is a *homogeneity requirement* on the regressors of the different equations for the exchangeability assumption on the mean to be transfered efficiently to the estimation of the coefficients, *i.e.* to insure minimaxity of the derived empirical Bayes estimator (*see* also Remark 6 below and the remarks on the *g-priors* at the beginning of Section 3).

Remark 5: Proposition 3 can also be applied to the estimation of the slopes only, by using an appropriate loss matrix. For instance, if the loss on β is associated with Σ_{22}^{-1}, the bound in (24) is $2(p-3)$ and we obtain LINDLEY and SMITH's result [1972].

Remark 6: BROWN [1975] gave a necessary condition for the existence of minimax estimators in the class (21),

$$tr(\Sigma\mathbf{QC}) - 2\, ch_{\max}(\Sigma\mathbf{QC}) > 0.$$

Therefore, the class (21) is of no interest if $p-3 \leqslant 0$ or if the variables $\mathbf{t}_i$ are already centered (as $\mathbf{C}=\mathbf{0}$). This second result is quite logical, as the information about β cannot be of any help in the estimation of θ in this case. The first condition is more surprising: it was known that $p-3>0$ is a necessary condition for the use of a James-Stein estimator shrinking toward $\hat{\beta}\mathbf{1}$ (EFRON and MORRIS [1972]); it appears here that it is also a necessary condition if we want to estimate the other parameters.

Remark 7: As we mentioned at the end of the introduction, the normality restriction could be replaced by a weaker spherical symmetry assumption, at no loss for the minimaxity results, as shown in CELLIER *et al.* [1989] and in KUBOKAWA *et al.* [1991], where we considered a common mean problem for two multiple regression equations with the errors distributions partially unknown (GEORGE [1991]). The drawback with this semi-parametric extension is that empirical Bayes techniques do not apply and the shape of the estimators is therefore harder to derive. This is why a normal modeling can be useful, at least as a first stage.

5 Improvement by Truncation

When the expression (17) goes to 0, the estimators (16) and (18) have a counter-intuitive behavior, in the sense that they go away from the RLSE while the test statistic increasingly supports the null hypothesis. This drawback is not dramatic in the sense that (16) and (18) are still dominating the corresponding LSE. However, if $\tilde{\beta}$ and $\hat{\beta}\mathbf{1}$ are very close, it does not seem quite proper to advocate the use of an obviously wrong estimator! This bad property in the neighborhood of $\tilde{\beta}=\hat{\beta}\mathbf{1}$ can be suppressed by *truncation* of the shrinkage factor. The *positive-part version* of the estimator (22) is

(25) $$\varphi_c^+(\mathbf{y}) = \hat{\gamma} + \left(1 - \frac{cs^2}{\|\tilde{\gamma}-\hat{\gamma}\|_*^2}\right)^+ (\tilde{\gamma}-\hat{\gamma})$$

where $m^+=\max(0,m)$. The positive-part of (18) can be derived in the same way. Therefore, these estimators are equal to the RLSE when $\|\tilde{\gamma}-\hat{\gamma}\|_*^2<cs^2$, *i.e.* when a test of the equality of the slopes accepts the null hypothesis. From an empirical Bayes point of view, the truncation is also justified by the fact that $\frac{\tau^2}{\sigma^2+\tau^2}$ should always be estimated by a term smaller than 1. (Note that the maximum likelihood approach automatically produces a truncated estimator.)

The following result shows why truncation is also justified from a frequentist point of view (*see* Appendix 3 for the proof).

PROPOSITION 5: Under the quadratic loss associated with Σ^{-1}, the positive-part estimators φ_c^+ are minimax for

$$0\leqslant c\leqslant 2(p-3).$$

Furthermore, they also dominate φ_c under this loss.

A similar result also holds if we consider the estimation of the intercepts under the loss Σ_{11}^{-1}, as $\|\tilde{\theta}-\hat{\theta}\|_{\Sigma_{11}}^2=\|\tilde{\gamma}-\hat{\gamma}\|_*^2$. For other losses, one has then to use a *vectorial truncation*, as advocated in JUDGE and BOCK [1978]. In all cases, the positive-part estimator behaves much better than the original estimator (22) and improves significantly the estimation of γ in terms of the risk. For instance, in the neighborhood of the restricted model, the risk reduction w.r.t., the LSE can be as large as p (BERGER [1985]).

6 Confidence Set Estimators

So far, we have considered empirical Bayes estimators that perform well from the frequentist perspective (minimaxity) and the Bayesian perspective (Proposition 2). However, from a practical point of view, it is also important to propose some *confidence sets* and, while one could still use the usual sets,

$$C_0(\mathbf{y})=\{\theta;\ (\theta-\tilde{\theta})^t\Sigma_{11}^{-1}(\theta-\tilde{\theta})\leqslant\omega^2 s^2\}, \tag{26}$$

it is more logical to propose sets directly associated with the empirical Bayes estimators, in order to make further use of the prior information. For instance, we can propose the *recentered confidence sets*

$$C^{EB}(\mathbf{y})=\{\theta;\ (\theta-\psi^{EB}(\mathbf{y}))^t\Sigma_{11}^{-1}(\theta-\psi^{EB}(\mathbf{y}))\leqslant\omega^2 s^2\},$$

i.e. ellipsoids with the same radii than the usual sets but centered at the empirical Bayes estimator. For a review, see ROBERT and SALEH [1989]. We study in this section the performance of these recentered set estimators, first for the whole parameter γ and then for the intercepts, θ. Obviously, they can also be used for testing purposes.

6.1. Estimation of γ

A natural comparison criterion for set estimators with the same volume is the *coverage probability*,

$$P_{\gamma, \sigma^2}(C^{EB}(y) \ni \theta).$$

When σ^2 is known, HWANG and CASELLA [1982, 1984] established that confidence sets recentered at positive-part James-Stein estimators have a uniformly larger coverage probability to contain the true value of the parameter than the usual confidence sets if the shrinkage constant is small enough. CASELLA and HWANG [1987] exploit this result in the reverse way: they proposed recentered sets with *variable radii* which have at least the same coverage probability than $C_0(y)$ and derived the radii from an empirical Bayes approach. Unfortunately, they were not able to derive an analytical proof of the domination of C_0 and we will not consider these sets in this paper.

When the variance factor σ^2 is unknown, there is no existing analytical domination result for the usual mean estimation problem in the literature. Still, ROBERT and CASELLA [1990] have obtained such a result for the multivariate t-distributions and HWANG and ULLAH [1989] give asymptotic and numerical motivations for the extension of the domination result to the unknown variance case. These references could then induce one to feel confident with the use of the estimator φ_{p-3}^{+} [as defined in (25)] for both point estimation and confidence set estimation.

Yet, the lack of analytical proof is bothering theoretically and we will consider now an alternative estimator for which a domination proof is available. The drawback is the restriction on the variance factor $\sigma^2 \geqslant \varepsilon$, even though such bounds may often be obtained in practice. Before stating the result (established in Appendix 4), let us define the two following functions

$$G^{\varepsilon}(a, \omega) = \left(\frac{\omega/2 + \sqrt{(\omega/2)^2 + a}}{\sqrt{a}}\right)^{2p-3} e^{-\omega\sqrt{a/2\varepsilon}}$$

and

$$H^{\varepsilon}(a, \omega) = \left(\frac{\omega + \sqrt{\omega^2 + a}}{\sqrt{a}}\right)^{2p-2} e^{-\omega\sqrt{a/\varepsilon}} \left(\frac{\sqrt{\omega^2 + 4a} - \omega}{2\sqrt{a}}\right).$$

PROPOSITION 6: If $a(s^2)$ is the function defined as the minimum of

$$\rho = 2\frac{tr(\mathbf{\Sigma QC}) - 2\, ch_{\max}(\mathbf{\Sigma QC})}{ch_{\max}(\mathbf{C\Sigma CB}^{-})}$$

and of the two solutions of

$$\text{(27)} \qquad G^{\varepsilon}(as^2, \omega s) = 1 \quad \text{and} \quad H^{\varepsilon}(as^2, \omega s) = 1,$$

the estimator

$$\text{(28)} \qquad \varphi(\mathbf{y}) = \hat{\gamma} + \left(1 - a(s^2)\frac{s^2}{\|\tilde{\gamma} - \hat{\gamma}\|_*^2}\right)^+ (\tilde{\gamma} - \hat{\gamma})$$

is minimax for the risk associated with $\mathbf{Q}$ and the confidence set

$$C^* = \{\gamma;\ \|\varphi(\mathbf{y}) - \gamma\|_*^2 \leqslant \omega^2 s^2\}$$

has a uniformly larger coverage probability than the usual confidence set

$$C^0 = \{\gamma;\ \|\gamma - \tilde{\gamma}\|_*^2 \leqslant \omega^2 s^2\}.$$

This result is derived from Hwang and Casella [1984] by conditioning on s^2, in a manner similar to Kim [1987]. We obtain in addition the minimaxity of the estimator φ. Simulations have also shown that the bounds (27) are certainly not optimal and that one may work confidently with larger bounds, as shown numerically in Hwang and Ullah [1989]. Some values for the bounds (27) are given in Hwang and Casella [1984] and Robert and Casella [1990].

6.2. Estimation of the Intercepts

As we are mainly interested in the estimation of the intercepts $\theta_i\,(1 \leqslant i \leqslant p)$, it would be more interesting to consider a confidence set for θ alone rather than for the whole vector γ. A natural choice is to take the recentered set

$$\text{(29)} \qquad C_a(\mathbf{y}) = \left\{\theta;\ \left\|\theta - \hat{\theta} - \left(1 - \frac{as^2}{\|\tilde{\gamma} - \hat{\gamma}\|_*^2}\right)^+ (\tilde{\theta} - \hat{\theta})\right\|_{\Sigma_{11}}^2 \leqslant \omega^2 s^2\right\},$$

where $\|\mathbf{z}\|_{\Sigma_{11}}^2 = \mathbf{z}^t \Sigma_{11}^{-1}\mathbf{z}$, as opposed to the usual set C_0 defined in (26). We know from Corollary 4 that the empirical Bayes estimator of θ is minimax if a is small enough and previous arguments apply to advise the use of the recentered confidence set (29).

From a theoretical point of view, the analytical domination of C_0 by C_a is not easier to establish than in Section 6.1. The case σ^2 known can be derived from Casella and Hwang [1987], since $\|\tilde{\gamma} - \hat{\gamma}\|_*^2 = \|\tilde{\theta} - \hat{\theta}\|_{\Sigma_{11}}^2$, but the case where σ^2 is unknown remains unsolved. We conclude this section with some heuristics arguments supporting the domination of C_0 by C_a. In the event they may be formalized, this could lead to a general proof for recentered confidence sets. The coverage probability of C_a can be written as

$$\text{(30)} \qquad P_\theta\left(\left\|\tilde{\gamma} - \theta - h\left(\frac{\|\tilde{\gamma} - \hat{\gamma}\|_{\Sigma_{11}}^2}{s^2}\right)(\tilde{\gamma} - \hat{\gamma})\right\|_{\Sigma_{11}}^2 \leqslant \omega^2 s^2\right),$$

as opposed to

$$P_\theta(\|\tilde{\gamma}-\theta\|^2_{\Sigma_{11}} \leqslant \omega^2 s^2) = 1-\alpha,$$

for the usual set. Therefore, we are adding to the usual vector, $\tilde{\gamma}-\theta$, another vector ξ such that

(i) $$\|\xi\|^2 \leqslant a s^2$$

(ii) $$E[(\tilde{\gamma}-\theta)^t \Sigma_{11}^{-1} \xi] \leqslant 0$$

[in fact, the scalar product in (ii) can even be bounded away from 0]. When a is small enough, the property (ii) should increase the probability to be close of θ. [Note also that properties (i) and (ii) are satisfied in the case of the positive-part James-Stein estimator.]

APPENDIX 1

Optimality of the Constant

Given Lemma 1, the Bayes risk, $r(\varphi_c, \pi)$, can be expressed as

$$r(\varphi_c, \pi) + \frac{n-2p+2}{n-2p} \times \left\{ -2c\,\mathbf{E}^m\left[\left(tr(\mathbf{\Sigma QC}) - 2\frac{\tilde{\gamma}^t \mathbf{B\Sigma C}\tilde{\gamma}}{\tilde{\gamma}^t \mathbf{B}\tilde{\gamma}}\right)\frac{s^2}{\tilde{\gamma}^t \mathbf{B}\tilde{\gamma}}\right] + c^2\,\mathbf{E}^m\left[s^2 \frac{\tilde{\gamma}^t \mathbf{C}^t \mathbf{QC}\tilde{\gamma}}{(\tilde{\gamma}^t \mathbf{B}\tilde{\gamma})^2}\right]\right\}$$

and the optimal choice of c is

$$c^* = \mathbf{E}^m\left[(tr(\mathbf{\Sigma QC}) - 2\tilde{\gamma}^t \mathbf{B\Sigma QC}\tilde{\gamma}/\tilde{\gamma}^t \mathbf{B}\tilde{\gamma})\frac{s^2}{\tilde{\gamma}^t \mathbf{B}\tilde{\gamma}}\right] \Big/ \mathbf{E}^m\left[s^2 \frac{\tilde{\gamma}^t \mathbf{C}^t \mathbf{QC}\tilde{\gamma}}{(\tilde{\gamma}^t \mathbf{B}\tilde{\gamma})^2}\right] \quad (31)$$

We have then just to show that the right hand side of (31) is equal to $(p-3)$.

Following the same type of arguments as in GHOSH *et al.* [1989], as the ratios $\frac{\tilde{\gamma}^t \mathbf{B\Sigma Q}\tilde{\gamma}}{\tilde{\gamma}^t \mathbf{B}\tilde{\gamma}}$ and $\frac{\tilde{\gamma}^t \mathbf{C}^t \mathbf{QC}\tilde{\gamma}}{\tilde{\gamma}^t \mathbf{B}\tilde{\gamma}}$ are *ancillary* statistics, we have

$$\begin{aligned}
&\mathbf{E}^m\left[\left(tr(\mathbf{\Sigma QC}) - 2\frac{\tilde{\gamma}^t \mathbf{B\Sigma QC}\tilde{\gamma}}{\tilde{\gamma}^t \mathbf{B}\tilde{\gamma}}\right)\frac{1}{\tilde{\gamma}^t \mathbf{B}\tilde{\gamma}}\right] \\
&= \left(tr(\mathbf{\Sigma QC}) - 2\,\mathbf{E}^m\left[\frac{\tilde{\gamma}^t \mathbf{B\Sigma QC}\tilde{\gamma}}{\tilde{\gamma}^t \mathbf{B}\tilde{\gamma}}\right]\right)\mathbf{E}^m\left[\frac{1}{\tilde{\gamma}^t \mathbf{B}\tilde{\gamma}}\right] \\
&= \left(tr(\mathbf{\Sigma QC}) - 2\frac{\mathbf{E}^m[\tilde{\gamma}^t \mathbf{B\Sigma QC}\tilde{\gamma}]}{\mathbf{E}^m[\tilde{\gamma}^t \mathbf{B}\tilde{\gamma}]}\right)\mathbf{E}^m\left[\frac{1}{\tilde{\gamma}^t \mathbf{B}\tilde{\gamma}}\right]
\end{aligned}$$

and

$$\mathbf{E}^m\left[\frac{\tilde{\gamma}^t \mathbf{C}^t \mathbf{QC}\tilde{\gamma}}{(\tilde{\gamma}^t \mathbf{B}\tilde{\gamma})^2}\right] = \mathbf{E}^m[\tilde{\gamma}^t \mathbf{C}^t \mathbf{QC}\tilde{\gamma}]\,(\mathbf{E}^m[\tilde{\gamma}^t \mathbf{B}\tilde{\gamma}])^{-1}\,\mathbf{E}^m\left[\frac{1}{\tilde{\gamma}^t \mathbf{B}\tilde{\gamma}}\right].$$

Note that $\mathbf{\Sigma QC}$ can be expressed as

$$\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{U(I-P)} \end{pmatrix}.$$

Then,

$$\begin{aligned}
\mathbf{E}^m(\tilde{\gamma}^t \mathbf{B\Sigma QC}\tilde{\gamma}) &= tr(\mathbf{\Sigma}_{22}(\mathbf{I}-\mathbf{P})^t \mathbf{\Sigma}_{22}^{-1}(\mathbf{I}-\mathbf{P})\mathbf{U}(\mathbf{I}-\mathbf{P}))\\
&= tr((\mathbf{I}-\mathbf{P})\mathbf{\Sigma}_{22}(\mathbf{I}-\mathbf{P})^t \mathbf{\Sigma}_{22}^{-1}(\mathbf{I}-\mathbf{P})\mathbf{U})\\
&= tr\left(\left(\mathbf{I}-\frac{1}{d}\mathbf{1}\mathbf{1}^t\mathbf{\Sigma}_{22}^{-1}\right)(\mathbf{I}-\mathbf{P})\mathbf{U}\right)\\
&= tr((\mathbf{I}-\mathbf{P})(\mathbf{I}-\mathbf{P})\mathbf{U})\\
&= tr(\mathbf{\Sigma QC}).
\end{aligned}$$

In the same way,

$$\begin{aligned}
\mathbf{E}^m[\tilde{\gamma}^t \mathbf{C}^t\mathbf{QC}\tilde{\gamma}] &= tr(\mathbf{\Sigma C}^t\mathbf{QC})\\
&= tr((\mathbf{I}-\mathbf{P})\mathbf{\Sigma}_{22}(\mathbf{I}-\mathbf{P})^t\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{12}\mathbf{Q}_1\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{12})\\
&= tr((\mathbf{I}-\mathbf{P})\mathbf{\Sigma}_{12}\mathbf{Q}_1\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{12})\\
&= tr(\mathbf{\Sigma QC}),
\end{aligned}$$

as $\operatorname{diag}(\bar{t}_i)=\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{12}$ and $\mathbf{U}=\mathbf{\Sigma}_{12}\mathbf{Q}_1\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{12}(\mathbf{I}-\mathbf{P})$. Furthermore,

$$\mathbf{E}^m[\tilde{\gamma}^t\mathbf{B}\tilde{\gamma}]=p-1.$$

Therefore,

$$c^*=\left(tr(\mathbf{\Sigma QC})-2\frac{tr(\mathbf{\Sigma QC})}{p-1}\right)\Big/[tr(\mathbf{\Sigma QC})/(p-1)]=p-3.$$

Minimaxity of the Empirical Bayes Estimator

We just have to express the bound in (24). It is easy to see that

$$\boldsymbol{\Sigma}\mathbf{QC}=\begin{pmatrix}\mathbf{0} & -\boldsymbol{\Sigma}_{11}\mathbf{Q}_1\mathbf{D}\\ \mathbf{0} & -\boldsymbol{\Sigma}_{12}\mathbf{Q}_1\mathbf{D}\end{pmatrix}$$

and

$$\mathbf{C}^t\mathbf{QCB}=\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}^t\mathbf{Q}_1\mathbf{D}\boldsymbol{\Sigma}_{22}\end{pmatrix}.$$

Further,

$$-\boldsymbol{\Sigma}_{12}\mathbf{Q}_1\mathbf{D}=\operatorname{diag}\left(\frac{\bar{t}_i}{d_i}\right)\mathbf{Q}_1\operatorname{diag}(\bar{t}_i)(\mathbf{I}-\mathbf{P})=\mathbf{U}(\mathbf{I}-\mathbf{P})$$

and

$$\begin{aligned}\mathbf{D}^t\mathbf{Q}_1\mathbf{D}\boldsymbol{\Sigma}_{22}&=(\mathbf{I}-\mathbf{P})^t\operatorname{diag}(\bar{t}_i)\mathbf{Q}_1\operatorname{diag}(\bar{t}_i)(\mathbf{I}-\mathbf{P})\operatorname{diag}\left(\frac{1}{d_i}\right)\\ &=(\mathbf{I}-\mathbf{P})^t\operatorname{diag}(\bar{t}_i)\mathbf{Q}_1\operatorname{diag}(\bar{t}_i)\operatorname{diag}\left(\frac{1}{d_i}\right)(\mathbf{I}-\mathbf{P})^t\\ &=(\mathbf{I}-\mathbf{P})^t\mathbf{U}^t(\mathbf{P})^t.\end{aligned}$$

Thus

$$\begin{aligned}ch_{\max}(\boldsymbol{\Sigma}\mathbf{QC})&=ch_{\max}(-\boldsymbol{\Sigma}_{12}\mathbf{Q}_1\mathbf{D})\\ &=ch_{\max}(\mathbf{U}(\mathbf{I}-\mathbf{P}))\\ &=ch_{\max}((\mathbf{I}-\mathbf{P})^t\mathbf{U}^t(\mathbf{I}-\mathbf{P})^t)\\ &=ch_{\max}(\mathbf{C}^t\mathbf{QCB}^-).\end{aligned}$$

Domination by the Truncated Estimator

This result can be derived from BARANCHICK [1970] if we can express the model in the canonical form. Note that $\hat{\gamma}$ is the orthogonal projection of $\tilde{\gamma}$ on the subspace of $\mathbf{R}^{2p}$

$$A = \{(\theta, \beta);\ \beta_i = \beta_1,\ 2 \leqslant i \leqslant p\},$$

for the scalar product defined by $\mathbf{\Sigma}$, *i.e.* $\hat{\gamma} = P_A \tilde{\gamma}$. Therefore,

$$(32)\quad \|\hat{\gamma} + [1 - h(\|\tilde{\gamma} - \hat{\gamma}\|_I^2/s^2)](\tilde{\gamma} - \hat{\gamma}) - \gamma\|_*^2 \\ = \|\hat{\gamma} - P_A \gamma\|_*^2 + \|(1 - h(\|\tilde{\gamma} - \hat{\gamma}\|_*^2/s^2))(\tilde{\gamma} - \hat{\gamma}) - (\gamma - P_A \gamma)\|_*^2$$

In the comparison of the positive-point estimator φ_c^* with the original estimators φ_c, the first term of (32) vanishes and the second term can be expressed under the canonical form, thus allowing to use BARANCHICK [1970].

Improvement of the Confidence Sets

The principle of the proof is to work conditionally on s^2. We have

$$\begin{aligned} P(C^*) &= E^{s^2}[P(C^* \mid s^2)] \\ &= E^{s^2}[P(\| \psi(y) - \gamma \|_*^2 \leqslant \omega^2 s^2 \mid s^2)] \end{aligned}$$

and

$$\begin{aligned} &P(\| \psi(y) - \gamma \|_*^2 \leqslant \omega^2 s^2 \mid s^2) \\ &= P\left(\left\| \hat{\gamma} + \left(1 - \frac{a(s^2)s^2}{\|\tilde{\gamma} - \hat{\gamma}\|_*^2}\right)^+ (\tilde{\gamma} - \hat{\gamma}) - \gamma \right\|_*^2 \leqslant \omega^2 s^2 \mid s^2 \right) \\ &= P\left(\left\| \frac{\hat{\gamma}}{\sigma} - \left(1 - \frac{a(s^2)s^2/\sigma^2}{\|\tilde{\gamma} - \hat{\gamma}\|_*^2/\sigma^2}\right)^+ \frac{(\tilde{\gamma} - \hat{\gamma})}{\sigma} - \frac{\gamma}{\sigma} \right\|_*^2 \leqslant \omega^2 \frac{s^2}{\sigma^2} \middle| s^3 \right) \\ &= P\left(\left\| \hat{v} + \left(1 - \frac{a(s^2)s^2/\sigma^2}{\|\tilde{v} - \hat{v}\|_*^2}\right)^+ (\tilde{v} - \hat{v}) - v \right\|_*^2 \leqslant \omega^2 \frac{s^2}{\sigma^2} \middle| s^2 \right) \end{aligned}$$

with obvious notations. From CASELLA and HWANG [1987], it follows that this probability will be superior to $P\left(\|\tilde{v} - v\|_*^2 \leqslant \omega^2 \frac{s^2}{\sigma^2} \middle| s^2 \right)$ for every v if

$$G_{2p}(a(s^2)s^2/\sigma^2, \omega s/\sigma) \geqslant 1 \quad \text{and} \quad H_{2p}(a(s^2)s^2/\sigma^2, \omega s/\sigma) \geqslant 1.$$

These conditions are equivalent to

$$\left(\frac{\omega s + \sqrt{\omega^2 s^2 + a(s^2)s^2}}{\sqrt{a(s^2)s^2}} \right)^{2p-2} e^{-\omega s^2 \sqrt{a(s^2)}/2\sigma^2} \geqslant 1 \tag{33}$$

and

$$\left(\frac{\omega s + \sqrt{\omega^2 s^2 + a(s^2)s^2}}{\sqrt{a(s^2)s^2}} \right)^{2p-2} e^{-\psi s \sqrt{a(s^2)}/\sigma^2} \left(\frac{\sqrt{w^2 + 4a(s^2)} - w}{2\sqrt{a(s^2)}} \right) \geqslant 1. \tag{34}$$

Now (33) and (34) will be satisfied if (27) is satisfied as

$$e^{-\omega s^2} \sqrt{a(s^2)}/2\sigma^2 \geqslant e^{-\omega s^2 \sqrt{a(s^2)/2\varepsilon}}$$

and

$$e^{-\omega s \sqrt{a(s^2)}/\sigma^2} \geqslant e^{-\omega s \sqrt{a(s^2)}/\varepsilon}.$$

Thus, for every s^2,

$$P(\| \gamma - \psi(y) \|_*^2 \leqslant \omega^2 s^2 \mid s^2) \geqslant P(\| \gamma - \tilde{\gamma} \|_*^2 \leqslant \omega^2 s^2 \mid s^2)$$

and we deduce the domination result for the confidence set.

To establish the minimaxity of the estimator (28), it is sufficient to show that $a(s^2)$ is a decreasing function (Proposition 8, FRAISSE *et al.* [1987]). And this will be proved if we establish that the solutions of (27) are decreasing.

The solution of $G^{\varepsilon}(as^2, \omega s) = 1$ satisfies

$$\left[\left(\frac{\omega}{2\sqrt{a}}\right)+\sqrt{\left(\frac{\omega}{2\sqrt{a}}\right)^2+1}\,\right]^{2p-3} = e^{\omega s^2 \sqrt{a/2\varepsilon}} \tag{35}$$

The right hand side of (35) is an increasing function of a while the left hand side is decreasing. Therefore, if a was increasing in s^2, the *rhs* would increase while the *lhs* would decrease; this is impossible. The same kind of argument works for the solution of $H^{\varepsilon}(as^2, \omega s) = 1$, as the function $(u+\sqrt{u^2+1})^{2p-2}(\sqrt{u^2+4}-u)$ is increasing for every $p \geqslant 2$.

• References

BARANCHICK, A. (1970). – "A family of Minimax Estimators of the Mean of a Multivariate Normal Distribution", *Ann. Math. Statist.*, 42, pp. 642-645.

BERGER , J. (1985). – *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York.

BERGER, J. and CHEN, S. Y. (1987). – "Minimaxity of Empirical Bayes Estimators Derived from Subjective Hyperpriors", In *Advances in Multivariate Statistical Analysis*, A. K. GUPTA, ed., D. REIDEL Publishing Company, pp. 1-12.

BERGER, J. and ROBERT, C. (1991). – "Subjective Hierarchical Bayes Estimation of a Multivariate Normal Mean: on the Frequentist Interface", *Ann. Statist.*, 18, pp. 617-651.

BLATTBERG, R. and GEORGE, E. I. (1990). – "Shrinkage Estimation of Price and Promotional Elasticities: Seemingly Unrelated Equations", *J. Amer. Statist. Assoc.*, 86, pp. 304-315.

BROWN, L. (1975). – "Estimation with Incompletely Specified Loss Functions", *J. Amer. Statist. Assoc.*, 70, pp. 417-425.

CASELLA, G. and HWANG, J. (1987). – "Employing Vague Prior Information in the Construction of Confidence Sets", *J. Multivariate Analysis*, 21, pp. 79-104.

CELLIER, D., FOURDRINIER, D. and ROBERT, C. (1989). – "Robust Shrinkage Estimators of the Location Parameter for Elliptically Symmetrical Distributions", *J. Multivariate Anal.*, 29, pp. 39-52.

DEGROOT, M. (1970). – *Optimal Statistical Decisions*. McGraw-Hill, New York.

EFRON, B. and MORRIS, C. (1972). – "Limiting the Risk of Bayes and Empirical Bayes Estimators – Part II: The Empirical Bayes Case". *J. Amer. Statist. Assoc.*, 67, pp. 130-139.

FRAISSE, A. M., ROBERT, C. and ROY, M. (1987). – "Estimateurs à rétrécisseur matriciel différentiable pour un coût quadratique général", *Annales d'Economie et de Statistique*, 8, p. 161-175.

GELFAND, A. et SMITH, A. F. (1990). – "Sampling Based Approaches to Calculating Marginal Densities", *J. Amer. Statist. Assoc.*, 85, pp. 398-409.

George, E. (1991). – "Shrinkage domination in the multivariate common mean problem", *Ann. Statist.*, 19, pp. 952-960.

Ghosh, M. (1989). – "Hierarchical and Empirical Bayes Multivariate Estimation", Tech. Report, Univ. of Florida, Gainesville.

Ghosh, M., Saleh, A. and Sen, P. (1989). – "Empirical Bayes Subset Estimation in Regression Models", *Statist. Decis.*, 7, pp. 15-35.

Hui, S. and Berger, J. (1983). – "Empirical Bayes Estimation of Rates in Longitudinal Studies", *J. Amer. Statist. Assoc.*, 78, pp. 753-760.

Hwang, J. and Casella, G. (1982). – "Minimax Confidence Sets for the Mean of a Multivariate Normal Distribution", *Ann. Statist.*, 10, (3), pp. 868-881.

Hwang, J. and Casella, G. (1984). – "Improved Set Estimators for a Multivariate Normal Mean", *Statist. Decis.*, Supplement Issue, 1, pp. 3-16.

Hwang, J. and Ullah, A. (1989). – "Confidence Sets Recentered at James-Stein Estimators, A Surprise Concerning the Unknown Variance Case", Tech. Report, Statistics Center, Cornell University.

James, W. and Stein, C. (1961). – "Estimation with Quadratic Loss", *Proc. 4th Berkeley Symp.*, 1, pp. 362-379.

Judge, G. and Bock, M. E. (1978). – *The Statistical Implications of Pre-test and Stein-rule estimators in Econometrics,* North-Holland, Amsterdam.

Kim, P. (19987). – "Recentered Confidence Sets for the Mean of a Multivariate Distribution when the Scale Parameter is Unknown", Ph. D. Thesis, Department of Mathematics, University of California at San Diego.

Kubokawa, T., Robert, C. and Saleh, A. (1991). – "Robust Estimation of Common Regression Coefficients under Spherical Symmetry", *Ann. Instit. Statist. Math.*, (to appear).

Lindley, D. and Smith, A. (1972). – "Bayes Estimate for the Linear Model" (with discussion), *J. Roy. Statist. Soc.*, (Ser. B), 34, pp. 1-41.

Maritz, J. and Lwin, T. (1989). – *Empirical Bayes Methods*, 2nd edition, Chapman and Hall, London.

Morris, C. (1983). – "Parametric Empirical Bayes Inference: Theory and Applications" (with discussion), *J. Amer. Statist. Assoc.*, 78, pp. 47-65.

Rao, C. R. (1983). – "Prediction of Future Observations in Polynomial Growth Curve Models", (1), Tech. Rep. #83-05, University of Pittsburgh.

Robert, C. and Casella, G. (1990). – "Improved Confidence Sets in Spherically Symmetric Distributions", *J. Multivariate Anal.*, 32, (1), pp. 84-94

Robert, C. and Saleh, A. (1989). – "Recentered Confidence Sets: a Review", Tech. Report, L.S.T.A., Université Paris-VI.

Robbins, H. (1955). – "An Empirical Bayes Approach to Statistics", In *Proc. Third Berkeley Symp. Math. Statist. Probab.*, 1, University of California Press, Berkeley.

Robbins, H. (1964). – "The Empirical Bayes Approach to Statistical Decision Problems", *Ann. Math. Statist.*, 35, pp. 1-20.

Saleh, A. and Sen, P. (1986). – "On Shrinkage Least Squares Estimation in a Parallelism Problem", *Comm. Statist. Theor. Meth.*, 15, pp. 1451-1466.

Sen, P. (1971). – "Robust Statistical Procedures in Problem in Linear Regression with Special References to Quantitative Bioassays", *Rev. Int. Statist. Inst.*, 39, pp. 21-38.

Stein, C. (1981). – "Estimation of the mean of a multivariate normal distribution", *Ann. Statist.*, 9, pp. 1135-1151.

Zellner, A. (1971). – *An Introduction to Bayesian Inference in Econometrics*, Wiley, NY.

Zellner, A. (1986). – "On Assessing Prior Distributions and Bayesian Regression Analysis with *g*-Prior Distributions", *Bayesian Inference and Decision Techniques*, North-Holland, Amsterdam.