



Combining estimators of a common parameter across samples

Eric Slud^{a,b}, Ilia Vonta^c and Abram Kagan^a

^aMathematics Department, University of Maryland College Park, College Park, MD, USA; ^bCenter for Statistical Research and Methodology, US Census Bureau, Washington DC, USA; ^cDept. of Mathematics, National Technical University of Athens, Athens, Greece

ABSTRACT

In many settings, multiple data collections and analyses on the same topic are summarised separately through statistical estimators of parameters and variances, and yet there are scientific reasons for sharing some statistical parameters across these different studies. This paper summarises what is known from large-sample theory about when estimators of a common structural parameter from several independent samples can be combined functionally, or more specifically linearly, to obtain an asymptotically efficient estimator from the combined sample. The main idea is that such combination can be done when the separate-sample nuisance parameters, if any exist, vary freely and independently of one another. The issues are illustrated using data from a multi-centre lung cancer clinical trial. Examples are presented to show that separate estimators cannot always be combined in this way, and that the functionally combined separate estimators may have low or 0 efficiency compared to the unified analysis that could be performed by pooling the datasets.

ARTICLE HISTORY

Received 21 March 2018
Accepted 28 September 2018

KEYWORDS

Efficient estimator; estimating equation; Fisher information; information bound; meta-analysis; regular estimator

1. Introduction

In many fields of social and biomedical science, multiple studies estimating the same parameter are conducted and summarised separately. The parameter of interest is often a measure of effectiveness, that is, of difference in response between groups with and without a certain treatment. Effectiveness may be quantified by the positive effect of a biomedical treatment or of an intervention in a social-science context such as education or criminal justice. The most common version of the problem, addressed by *Meta-Analysis* (Hartung, Knapp, & Sinha, 2008), arises when many independent and uncoordinated studies are not individually large enough to make a definitive statement about positivity of the single (usually scalar) parameter θ which has scientific or policy interest. Another setting in which analyses are combined is the less common one of large similar studies (say, done in different geographic regions) with a shared parameter, where there may or may not be shared nuisance parameters. Whether large or small, few or many, the separate studies might differ in the precise characteristics of the population investigated, the criteria of inclusion in each study, or in the measurements collected either because of the choice of auxiliary variables or because of the definitions and research methods used.

The primary objective here, as in the numerous examples of meta-analysis in Hartung et al. (2008), is to combine the results of many moderate-sample

studies to distill a consensus parameter estimate, or a corresponding test of significance of specific component parameter(s) representing treatment effectiveness. Since it is not always possible to gain simultaneous access to the raw unit-level data of previously published studies, researchers seeking a definitive test or estimator of treatment effectiveness often attempt to combine the study results through a function of their separate summary statistics rather than through a re-analysis of all pooled study subject-level data based on a unified model.

The active field of statistical meta-analysis is occupied with methods of simultaneously modelling disparate studies or the estimated parameters from those studies in such a way that they can be combined (Hartung et al., 2008). Much of this effort has gone into models that account for differences in study methodology through random effects. Random-effect, empirical-Bayes, and hierarchical Bayes methods have all proved useful in this effort to combine study results. Efforts to ‘borrow strength’ across distinct experimental entities are ubiquitous in the random-effect and Bayes community (Carlin & Louis, 2008; Efron, 1996), but arise also under the heading of Small Area Estimation in the survey world (Rao & Molina, 2015). The validity of analyses depending on models to combine different experiments with shared parameters and random effects can always be questioned, but sometimes such analyses turn out to be surprisingly robust (Slud

& DeMissie, 2011). We discuss and illustrate some of the relevant modelling issues in Section 2 below.

This paper studies the problem of combining studies using a frequentist large-sample approach based on the standard tools of Fisher information and large-sample asymptotics. We begin by giving notations, statements and explanations of relevant results.

Suppose that k large independent samples $\mathbf{X}_j = \{\mathbf{x}_{ij}\}_{i=1}^{n_j}$ of independent vectors of data $\mathbf{x}_{ij} \sim f_j(\cdot, \beta, \lambda_j)$ are observed, for $j = 1, \dots, k$, with the goal of estimating the common parameter $\beta \in U \subset \mathbb{R}^p$ efficiently. The densities f_j of individual data-vectors \mathbf{x}_{ij} are assumed known except for the parameters $(\beta, \lambda_j) \in U \times \Lambda_j \subset \mathbb{R}^{p+q_j}$, where the *nuisance parameters* λ_j may not be present for all j , but k and the parameter dimensions $p, \{q_j\}_{j=1}^k$ are fixed while the sample-sizes n_j all tend to ∞ . Assume that the separate samples \mathbf{X}_j are summarised only through estimators $\tilde{\beta}_j$ of β , along with estimators \tilde{V}_j/n_j of their variances, and that it is desired to estimate β as efficiently as possible from these statistics. The vectors $\tilde{\beta}_j$ might then be treated as independent data with approximate means β and with known variance matrices \tilde{V}_j/n_j . We refer to $(\tilde{\beta}_j, \tilde{V}_j)$ as *separate-sample* estimators and to estimators within a unified model of the k -sample data considered together as *combined-sample* estimators.

When β is scalar, the best unbiased linear-combination estimator $\hat{\beta} = \sum_{j=1}^k w_j \tilde{\beta}_j$ with respect to mean-squared error is well known to have $w_j = (n_j/\tilde{V}_j) / \sum_{l=1}^k (n_l/\tilde{V}_l)$. This estimator $\hat{\beta}$ has often been viewed in Meta-Analysis (Hartung et al., 2008, Chapter 4) as arising from the model

$$\tilde{\beta}_j = \beta + e_j, \quad e_j \stackrel{\text{indep.}}{\sim} \mathcal{N}(0, \Sigma_j), \quad \Sigma_j = \tilde{V}_j/n_j \quad (1)$$

where the variances Σ_j are treated as known. This model is treated as the source of the meta-analytic estimator (3) below also in the p -vector setting.

Olkin and Sampson (1998) considered a balanced ANOVA model

$$\mathbf{x}_{ij} = \mu_j(1, \dots, 1)^{\text{tr}} + (b_1, \dots, b_{p-1}, 0)^{\text{tr}} + \epsilon_{ij} \quad (2)$$

which fits into our setting, where $\beta = (b_1, \dots, b_{p-1}, 0)^{\text{tr}}$, the components ϵ_{aij} of the p -vectors ϵ_{ij} are uncorrelated and all have mean 0 and (known or unknown) variance σ^2 , and $\tilde{\beta}_j = \bar{\mathbf{x}}_j - \bar{\mathbf{x}}_{p,j}(1, \dots, 1)^{\text{tr}}$ are the least-squares estimators of the β vectors in terms of the j th-sample data $\{\mathbf{x}_{ij}\}_{i=1}^{n_j}$. Olkin and Sampson (1998) prove – in a setting extending (2) to unbalanced ANOVA – that the best linear estimator (3) below in model (1) is identical to the least squares estimator of β within the combined unit-level model (2). When the errors ϵ_{aij} are normal, the Gauss-Markov Theorem implies that the linear-combination estimator (3) is Maximum Likelihood and therefore efficient within (2).

Two recent articles, Lin and Zeng (2010) and Liu, Liu, and Xie (2015), have treated in different ways the problem of efficiently combining separate-study estimators of a shared parameter when there are nuisance parameters. Lin and Zeng (2010) discusses the parametric combination of separate-study estimators when only the separate-study estimators $\tilde{\beta}_j$ and variance estimators \tilde{V}_j are available. They prove the result given as (V) below, in the case where the same parameter β is estimated by maximum likelihood in each study and all nuisance parameters vary freely and unconstrained between studies. They do not cover the case where nuisance parameters are infinite-dimensional or where the separate studies may estimate only projections of β , a situation that occurs naturally in meta-analysis, as illustrated in the example of Section 2 below. These extensions are covered in the present paper, in Section 3 and Appendices 1 and 3. The second issue, that separate studies may estimate only functions of a common parameter, is treated fully by Liu et al. (2015) in a more general setting where nuisance parameters may constrain each other across studies. The paper of Liu et al. (2015), using the idea of *Confidence distributions*, establishes under general regularity conditions the form of the optimal combination of the separate-study estimators of *all* of the parameters, not just the shared structural parameters but the nuisance parameters too. That paper is therefore less directly relevant to meta-analysis in practice, since it is very uncommon for investigators in separate studies to report the estimated nuisance parameters as well as joint variance estimates of all structural and nuisance parameters, as was noted by Lin and Zeng (2010, 1st paragraph of Section 2.2).

In the rest of this paper, we allow the possibility that in the j th sample, the estimand may be not the full p -vector parameter β but rather a projection $\Pi_j \beta$ to a known subspace of the p -dimensional parameter space, with ‘structural zeroes’ in place of the other coordinates $(\mathbf{I} - \Pi_j)\beta$. In case the range of the projection Π_j has dimension $< p$, the j th sample estimator $\tilde{\beta}_j$ is assumed to be either an efficient estimator or one derived by solving an estimating equation of $\Pi_j \beta$, with asymptotic variance $V_j(\beta)/n_j$, where $V_j(\beta)$ is consistently estimated by \tilde{V}_j . Both $V_j(\beta)$ and \tilde{V}_j are assumed to have range-spaces the same as that of Π_j , and respective generalised-inverses $V_j(\beta)^-, \tilde{V}_j^-$ which are inverses of $V_j(\beta), \tilde{V}_j$ on $\text{range}(\Pi_j)$ and are the $\mathbf{0}$ operator on the orthogonal-complement space $\text{range}(\mathbf{I} - \Pi_j)$. Appendix 1 proves under general large-sample regularity conditions, that if $\tilde{\beta}_j$ are asymptotically efficient estimators of $\Pi_j \beta$ from the separate samples, then the overall p -vector estimator

$$\hat{\beta} = \left[\sum_{j=1}^k n_j \tilde{V}_j^- \right]^{-1} \sum_{j=1}^k n_j \tilde{V}_j^- \tilde{\beta}_j \quad (3)$$

is an efficient estimator of β from the combined sample $\mathbf{X} \equiv \{\mathbf{x}_{ij} : 1 \leq i \leq n_j, 1 \leq j \leq k\}$, i.e., has minimal asymptotic variance, if there exists a regular efficient estimator depending (smoothly) on $(\tilde{\beta}_j, \tilde{V}_j : j = 1, \dots, k)$ alone. Alternatively, this efficiency is proved to hold in Section 3 under the restriction that the nuisance parameters λ_j range freely without any constraints connecting them for different j . In a further reformulation in Section 3, it is shown that when the nuisance parameters λ_j range freely without any constraints connecting them for different j , and the estimators $\tilde{\beta}_j$ are obtained by solving M-estimating equations

$$\sum_{i=1}^{n_j} \Psi_j(\mathbf{x}_{ij}, \beta, \lambda_j) = \mathbf{0},$$

$$\Psi_j(\mathbf{x}, \beta, \lambda_j) = \nabla_{\beta, \lambda_j} M(\mathbf{x}, \beta, \lambda_j) \quad (4)$$

the estimator $\hat{\beta}$ is efficient in the sense of having the same asymptotic variance as the best estimating-equation estimator linearly combining the estimating functions $\Psi_j(\mathbf{x}, \beta, \lambda_j)$ with matrix coefficients. Through examples in Section 4, we show that when the parameters λ_j are related by constraints, the estimator (3) is generally not efficient and may have efficiency 0, which means intuitively that its large-sample variance has larger order of magnitude than the best possible estimator based on the combined sample.

Before going on to theoretical developments, we illustrate these ideas in the next section using data from a real clinical trial.

2. Data example from a multi-centre clinical trial

Large randomised clinical trials are often conducted simultaneously at different medical centres. They are generally governed by the same clinical protocols — including formal entry criteria, randomisation methods, baseline and cross-sectional measurements to be collected, and study endpoints — but can differ in many of the same ways as completely separate studies: the patient populations from which study participants are recruited, slight differences in the way entry criteria and medical procedures are applied by medical personnel at the different centres, and the patient management strategies of individual physicians associated with the different centres. However, since the study design is shared by all the centres, such large clinical trials can be excellent test-beds for methods purporting to estimate shared parameters by combining analyses done in separate smaller studies. We describe the separate and combined analyses for just such a study, the Eastern Cooperative Oncology Group (ECOG) EST 1582 clinical trial of two different combination chemotherapies for treatment of small cell lung cancer, which has previously been analysed by Gray (1994) and studied

by meta-analysis in Slud and DeMissie (2011). The standard therapy in this trial was CAV, a combination of cyclophosphamide, adriamycin and vincristine, and the experimental treatment regimen (CAV-HEM) alternated cycles of CAV with hexamethylmelamine, etoposide and methotrexate. Allocation to these two treatment arms was randomised and equal.

The covariates in the study were binary indicators: **Trt** for experimental treatment, **bone** for bone metastasis, **liver** for liver metastasis, **wtloss** for weight-loss prior to study entry, and a measure **Perf** of performance status at baseline. These covariates entered significantly into an overall Bayesian proportional-hazards analysis by Gray (1994), who found that both a coefficient for **Trt** and one for an interaction term **Trt-by-bone** were significant. Here, as in Slud and DeMissie (2011), building on the earlier MS thesis work of DeMissie (2009), we analyse the same data separately by centre taking the **Trt-by-bone** interaction into account.

The data for subject i in study-centre j consist of Y_{ij} equal to the logarithm of survival time T_{ij} (or log of time until censoring, for the 10 out of 570 patients who were lost to follow-up at a time before death), Δ_{ij} indicating Y_{ij} as a failure time rather than censoring, and \mathbf{z}_{ij} the vector of covariates consisting of the constant 1, **Trt**, **bone**, **liver**, **Perf**, **wtloss** in *Model 1*, augmented in *Model 2* by **Trt*bbonCtr**, where **bbonCtr** is a recoded covariate obtained by subtracting from **bone** its study-centre mean. The main **Trt** effect in *Model 2* is still the **Trt** coefficient due to the centering in **bbonCtr**. The 18 study-centre clusters were obtained from the 26 original hospitals, as in DeMissie (2009), after merging some smaller ones by clustering for similarity of covariate means, so that the smallest number of individuals in any centre became 17. The underlying models we consider here are

$$\log T_{ij} = \sum_{r=1}^p z_{ij,r} b_r + \sigma e_{ij} \quad (5a)$$

or

$$\log T_{ij} = \sum_{r=1}^p z_{ij,r} b_r + u_j + \sigma e_{ij} \quad (5b)$$

where \mathbf{z}_{ij} are as defined above, $u_j \sim \mathcal{N}(0, \sigma_u^2)$ are independent random cluster-effects included only in model (5b), and e_{ij} are independent random errors distributed as extreme-value, i.e., as the logarithm of a unit exponential variable. These models are called *Model 1* (a or b) when the parameter β common to all study-centres is the scalar coefficient of **Trt**. *Model 2* (a or b) differs only by including the **Trt*bbonCtr** covariate, with β the 2-vector of coefficients of **Trt** and **Trt*bbonCtr**.

Models 1a and 2a are first fitted for each study-centre, and their separately estimated β coefficients

Table 1. Estimated parameters and SE's from two models, with coefficients β_1 for **Trt** and β_2 for **Trt*bonCtr**, in separate study centres. '*' denotes structural 0.

j	n_j	Model 1		Model 2			
		$\tilde{\beta}_j$	$\tilde{V}_j^{1/2}$	$\tilde{\beta}_{j,1}$	$\tilde{V}_{j,11}^{1/2}$	$\tilde{\beta}_{j,2}$	$\tilde{V}_{j,22}^{1/2}$
1	21	0.307	0.166	0.383	0.154	1.083	0.420
2	17	0.266	0.205	0.266	0.205	*	*
3	18	0.682	0.205	0.687	0.207	-0.187	0.427
4	27	0.400	0.343	0.406	0.340	0.243	0.840
5	46	-0.159	0.183	-0.147	0.177	0.367	0.452
6	31	0.407	0.294	0.410	0.295	-0.186	0.620
7	17	-0.246	0.293	-0.246	0.293	*	*
8	59	-0.035	0.154	-0.031	0.150	0.288	0.313
9	56	0.415	0.181	0.424	0.177	-0.304	0.413
10	31	0.646	0.296	0.645	0.296	-0.012	0.571
11	22	0.591	0.250	0.603	0.216	1.737	0.623
12	39	0.353	0.286	0.342	0.281	-0.313	0.623
13	27	0.181	0.202	0.188	0.187	-0.437	0.362
14	53	-0.269	0.211	-0.296	0.203	0.671	0.365
15	17	0.352	0.655	0.526	0.626	-1.365	1.198
16	42	0.026	0.160	0.033	0.162	-0.216	0.386
17	23	-0.107	0.400	-0.122	0.388	0.896	0.710
18	24	-0.219	0.312	-0.221	0.299	0.543	0.489

and standard errors, obtained from the R function *survreg*, are exhibited in Table 1. (Note that in each of Centres 2 and 7, the **Trt*bone** values are all 0, so that in these centres **Trt*bonCtr** is an affine function of **Trt** and β_2 is a structural 0.) We then construct the estimators (3) in these two models, scalar for Model 1a and 2-vector for Model 2a: these are the estimators that would be produced in a meta-analysis, if Model 1a and 2a estimation results (including 2×2 estimated variance matrices for estimates $\tilde{\beta}_j$ in 2a) were separately published from studies at distinct centres. For Model 1a, the meta-analytic estimates (3) of β and standard error are $\hat{\beta} = 0.176$, $\hat{V}_*^{1/2} = 0.053$. For Models 2a, the meta-analytic coefficient estimates are $\hat{\beta} = (0.176, 0.191)$, with respective SE's 0.051, 0.117. In all of these analyses, the **Trt** coefficient β_1 is highly significant, indicating that the experimental treatment prolonged life as was found by Gray (1994) and DeMissie (2009), but the separate centres' models seem to yield conflicting information.

Models 1a and 2a are fixed effects models for many small-sample datasets. Fitting them separately in each centre reflects an assumed lack of connection across centres j for the coefficients of covariates $z_{ij,r}$ not involving treatment. In Model 1a, the standard errors of the treatment coefficient are roughly 0.2 in each centre, and the standardised coefficient ranges across centres from -1.3 to $+3.3$, significant (at $\alpha = .05$, two-sided) in only 4 centres. In Model 2a, the standardised coefficient for **Trt** ranges from -1.5 to $+3.2$, with 8 significant, and that for **Trt*bonCtr** ranges from -1.2 to 2.8 , with only 2 significant. A unified model 1a, assumed to hold in all centres with common β , yields $\hat{\beta} = 0.272$ with $SE = 0.069$, and the unified model 2a with shared 2-vector β yields $\hat{\beta}_1 = 0.268$ with $SE = 0.068$. However, since the β_1 estimates vary considerably across centre in separately fitted models, the unified

model should accommodate the differences through a random-effects model like (5b), with independent random **Trt** effects across cluster. Random treatment effects were previously considered both by Gray (1994) and DeMissie (2009). The unified model (5b), fitted by SAS *proc nlmixed* to allow random **Trt** effects by centre, results in the following estimates and standard errors:

$$\begin{aligned} \text{Model 1b: } \quad & \hat{\beta} = 0.292, \quad SE(\hat{\beta}) = 0.102 \\ \text{Model 2b: } \quad & \hat{\beta} = (0.286, 0.306), \quad SE's = (0.100, 0.145) \end{aligned}$$

The unified models 1b and 2b agree in their estimate of the **Trt** coefficient, disagree only slightly from the unified models 1a and 2a, but the unified analysis disagrees markedly from the meta-analysis. The unified estimate of the **Trt*bonCtr** coefficient appears significant both in the unified model and in the meta-analysis. The **Trt*bonCtr** coefficient can be seen to be very noisily fitted in the individual clusters, and perhaps should also be treated with a random effect in the unified model.

The important point of this example is that a properly specified combined-sample analysis can be expected, under the kind of mixed-effect linear model described here, to give substantially the same results as a meta-analysis under a model with adequately detailed interactions and random effects (Slud & DeMissie, 2011). In this setting, as in most real meta-analyses, the separate samples (here, the analyses at individual centres) are too small to identify interactions such as treatment-by-covariate interactions which are clearly significant in unified-model analysis. The operation of meta-analysis, which functionally combines the separate-sample coefficient estimates, shows through goodness-of-fit assessments the necessity of including interaction terms (such as treatment-by-covariate) and random effects where separate-sample coefficients vary considerably.

3. Results from large-sample theory

Throughout the rest of this paper, we assume standard regularity and nondegeneracy conditions about parameters (β, λ_j) (as in Bickel & Doksum, 2007, Theorem 6.2.2, and Van der Vaart, 1998, Theorem 5.39) which for finite-dimensional λ_j imply the following. First, joint maximum likelihood (ML) estimators $(\tilde{\beta}_j, \tilde{\lambda}_j)$ for (β, λ_j) exist and are consistent and locally uniquely determined as solutions of the score or likelihood equations in the j th sample, and

$$\sqrt{n_j} \begin{pmatrix} \tilde{\beta}_j - \beta \\ \tilde{\lambda}_j - \lambda_j \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, W_j) \quad \text{as } n_j \rightarrow \infty \quad (6)$$

The variance W_j is the inverse of the $(p + q_j) \times (p + q_j)$ Fisher information matrix $\mathcal{I}^{(j)}(\beta, \lambda_j)$ for the j th sample. The upper-left $p \times p$ block of W_j is denoted by $V_j(\beta)$ or

by $\{\mathcal{I}^{(j)}(\beta)\}^{-1}$, and is the smallest possible asymptotic variance (in the sense of matrix ordering if $p > 1 : K \leq L$ if and only if $L - K$ is nonnegative definite) within the class of all ‘regular’ estimators (Bickel, Klaassen, Ritov, & Wellner, 1998, pp. 17–21, or Van der Vaart, 1998, p. 115), which includes (again subject to regularity conditions) all estimators defined as solutions of estimating equations

$$\sum_{i=1}^{n_j} \Psi_j(\mathbf{x}_{ij}, \beta, \lambda_j) = 0 \quad (7)$$

where the functions $\Psi_j(\cdot)$ are known and nonrandom $(p + q_j)$ -vector estimating-function summands whose form depends on study j but not on subject-index i . Any *efficient* regular estimator $\tilde{\beta}_j$ defined from \mathbf{X}_j , i.e., one for which the asymptotic variance of $\sqrt{n_j}(\tilde{\beta}_j - \beta)$ is no larger than $V_j(\beta)$, differs from the ML estimator of β by a remainder of order smaller than $n_j^{-1/2}$ in probability. (This assertion follows from the Hájek-LeCam convolution theorem, Van der Vaart, 1998, p. 115.) This notion of *efficiency* can be extended also to the case where the nuisance parameters λ_j are infinite-dimensional, a notion of ‘first-order optimality’ defined as *semiparametric* or asymptotic efficiency of regular estimators (Van der Vaart, 1998, Section 25.3). Although the information bounds $\mathcal{I}^{(j)}(\beta)$ do depend on the nuisance parameters λ_j , we suppress that dependence to keep the notation as simple as possible.

Suppose that the values of an efficient estimator $\tilde{\beta}_j$ and a consistent estimator \tilde{V}_j of $V_j(\beta)$ are reported from separate analysis of the j th sample in order to obtain confidence intervals for components or linear combinations of the coordinates of β . The main question of interest in this paper is: under what circumstances will there exist a (smooth) function $\hat{\beta} = g(\{\tilde{\beta}_j, \tilde{V}_j\}_{j=1}^k)$ of the summary statistics $(\tilde{\beta}_j, \tilde{V}_j)$ such that $\hat{\beta}$ is efficient? In the case where the separate-sample estimators $\tilde{\beta}_j$ are assumed efficient, this means that the asymptotic variance $V_*(\beta)$ of $\sqrt{n}(\hat{\beta} - \beta)$ is the inverse of the per-observation Fisher information $\mathcal{I}^*(\beta)$ for β based on the combined sample of size $n = \sum_{j=1}^k n_j$. In the case where the estimators $\tilde{\beta}_j$ are obtained from estimating equations as in (4), *efficiency* means that $\hat{\beta}$ has asymptotic variance no larger than the best combined-sample estimating equation estimator obtained from a matrix-weighted linear combination of the estimating functions $\sum_{i=1}^{n_j} \Psi_j(\mathbf{x}_{ij}, \beta, \lambda_j)$.

In the next part of this Section, leading up to paragraphs (I)–(VI), we develop notions of single- and combined-sample Fisher information about the parameter β . Recall that the per-observation Fisher information about a parameter θ (here, (β, λ_j) or $(\beta, \underline{\lambda})$ for a vector $\underline{\lambda}$ combining all of the free parameters in $(\lambda_1, \dots, \lambda_k)$) is a matrix expectation defined as

$$\mathcal{I}(\theta) \equiv n^{-1} E(\mathbf{S}_\theta \mathbf{S}_\theta^{\text{tr}}) \equiv n^{-1} E(\mathbf{S}_\theta^{\otimes 2})$$

where \mathbf{S}_θ is a *score statistic* obtained as the column of partial derivatives of the log-likelihood of the data-sample of size n , with respect to the parameters. (Here and below, we also use the convenient notation $\mathbf{v}^{\otimes 2} \equiv \mathbf{v}\mathbf{v}^{\text{tr}}$ for any vector \mathbf{v} .) In this section, we express statistical properties of estimators in terms of the linear algebra of information matrices.

To begin, we review concepts and develop formulas related to separate-sample Fisher information about β . Denote the score statistic for data \mathbf{X}_j with respect to β by $\mathbf{S}_{\beta,j}$ and with respect to λ_j by $\mathbf{S}_{\lambda_j,j}$. The information matrices per observation for the separate-sample parameters are

$$\mathcal{I}^{(j)}(\beta, \lambda_j) = \frac{1}{n_j} E \left\{ \begin{pmatrix} \mathbf{S}_{\beta,j} \\ \mathbf{S}_{\lambda_j,j} \end{pmatrix}^{\otimes 2} \right\} = \begin{pmatrix} I_{11}^{(j)} & I_{12}^{(j)} \\ I_{21}^{(j)} & I_{22}^{(j)} \end{pmatrix} \quad (8)$$

In the block-decomposition on the right-hand side of (8), $I_{11}^{(j)}$ is $p \times p$.

The information $\mathcal{I}^{(j)}(\beta)$ about β in the j th sample is defined as the inverse of the upper-left $p \times p$ block of $\{\mathcal{I}^{(j)}(\beta, \lambda_j)\}^{-1}$ and is given by $\mathcal{I}^{(j)}(\beta) = I_{11}^{(j)} - I_{12}^{(j)} (I_{22}^{(j)})^{-1} I_{21}^{(j)}$. This linear-algebra fact about inverses of block-decomposed matrices can be found in virtually every book about regression (cf. Draper & Smith, 1981, Appendix 2A), and can be interpreted as the expression of the residual variance of $\mathbf{S}_{\beta,j}$ after regression on $\mathbf{S}_{\lambda_j,j}$. This interpretation is developed in terms of associated Linear Regression theory by Draper and Smith (1981, Section 2.6), or more abstractly in terms of projections by Rao (1973, Sections 4.a.1–2 and 4.a.6).

Since we allow the possibility that some of the coordinates of the parameter vectors λ_j are shared or constrained across different samples \mathbf{X}_j , let $\underline{\lambda}$ denote a parameter vector, of dimension $d \leq \sum_{j=1}^k q_j$, consisting of all free parameters among $\{\lambda_j\}_j$, so that all λ_j vectors are smooth functions $\lambda_j \equiv g_j(\underline{\lambda})$ of $\underline{\lambda}$. Then we can write all of the densities

$$f_j(\mathbf{x}, \beta, \lambda_j) = f_j(\mathbf{x}, \beta, g_j(\underline{\lambda})) \equiv f_j^*(\mathbf{x}, \beta, \underline{\lambda})$$

where f_j^* is smooth in all components of its parameter arguments. Denote the score statistic for the sample \mathbf{X}_j with respect to the parameter vector $\underline{\lambda}$ as $\mathbf{S}_{\underline{\lambda},j}$. Then, in terms of the $q_j \times d$ Jacobian matrix $J_{g_j}(\underline{\lambda}) = (\nabla_{\underline{\lambda}} g_j^{\text{tr}}(\underline{\lambda}))^{\text{tr}}$ of the q_j -vector valued function $g_j(\underline{\lambda})$ with respect to $\underline{\lambda}$,

$$\begin{aligned} \mathbf{S}_{\underline{\lambda},j} &\equiv \nabla_{\underline{\lambda}} \log f_j(\mathbf{X}_j, \beta, g_j(\underline{\lambda})) \\ &= (J_{g_j}(\underline{\lambda}))^{\text{tr}} \nabla_{\lambda_j} \log f_j(\mathbf{X}_j, \beta, \lambda_j) = (J_{g_j}(\underline{\lambda}))^{\text{tr}} \mathbf{S}_{\lambda_j,j} \end{aligned}$$

By independence of the samples \mathbf{X}_j , and therefore of the score-statistics $(\mathbf{S}_{\beta,j}, \mathbf{S}_{\lambda_j,j}, \mathbf{S}_{\underline{\lambda},j})$ for different j , the per-observation Fisher information for the combined

parameter $(\beta, \underline{\lambda})$ in the combined sample is the symmetric $(p + d) \times (p + d)$ matrix

$$\mathcal{I}^*(\beta, \underline{\lambda}) = \frac{1}{n} \sum_{j=1}^k E \left\{ \begin{pmatrix} \mathbf{S}_{\beta,j} \\ \mathbf{S}_{\underline{\lambda},j} \end{pmatrix}^{\otimes 2} \right\} = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} \quad (9)$$

where

$$I_{11} = \sum_{j=1}^k \frac{n_j}{n} I_{11}^{(j)}, \quad I_{12} = \sum_{j=1}^k \frac{n_j}{n} I_{12}^{(j)} J_{g_j}(\underline{\lambda}) \quad (10)$$

The corresponding complete-sample information $\mathcal{I}^*(\beta)$ is then expressed either as

$$\mathcal{I}^*(\beta) = I_{11} - I_{12} I_{22}^{-1} I_{21} \quad (11)$$

in case the I_{22} block is invertible, or more generally via the regression-residual interpretation analogous to that of $I^{(j)}(\beta)$ above, as

$$\mathcal{I}^*(\beta) = \frac{1}{n} \inf_K \sum_{j=1}^k E \left[\left(\mathbf{S}_{\beta,j} - K \mathbf{S}_{\underline{\lambda},j} \right)^{\otimes 2} \right] \quad (12)$$

where K in the infimum is an arbitrary $p \times d$ matrix, and \inf is understood in the sense of nonnegative-definite matrix ordering.

We now present a series of propositions relating combined-sample information and variance to those of the separate samples.

(I) Kagan and Rao (2003, Lemma 2) establish for finite-dimensional λ_j the *superadditivity of information*

$$\mathcal{I}^*(\beta) \geq \sum_{j=1}^k \frac{n_j}{n} \mathcal{I}^{(j)}(\beta) \quad (13)$$

Since inverse information is equal to the smallest attainable asymptotic variance for ('regular') estimators under the conditions assumed here, the inequality (13) can be interpreted to say that the best possible variance $V_*(\beta)$ is at most the asymptotic variance $\{\sum_{j=1}^k (n_j/n) [V_j(\beta)]^{-1}\}^{-1}$ of the right-hand side of (3).

(II) The last statement can be recast as in Janicki (2009, Theorem 6.1.2), to say that *additivity of information*, or equality in (13), holds if and only if $\hat{\beta}$ in (3) has asymptotic variance $\{\mathcal{I}^*(\beta)\}^{-1}$. In a setting without nuisance parameters λ_j , Janicki (2009, Theorem 6.1.2) shows for $\hat{\beta}_j$ defined by estimating equations, that an optimal combined estimator $\hat{\beta}$ is obtained either as the weighted linear combination (3) of estimators, or as solution to the linear combination of j th-sample estimating equations

$$\sum_{j=1}^k A_j(\beta) \sum_{i=1}^{n_j} \Psi_j(\mathbf{x}_{ij}, \beta) = \mathbf{0} \quad (14)$$

where Ψ_j are any estimating functions such that $E(\Psi_j(\mathbf{x}_{1j}, \beta)) = \mathbf{0}$ and $E(-\nabla_{\beta} \Psi_j^{\text{tr}}(\mathbf{x}_{1j}, \beta))$ is nonsingular, and the $p \times p$ matrices $A_j(\beta)$ are defined by

$$A_j(\beta) = \{-E[\nabla_{\beta} \Psi_j^{\text{tr}}(\mathbf{x}_{1j}, \beta)]\}^{\text{tr}} \\ \times \{E[\Psi_j(\mathbf{x}_{1j}, \beta) \Psi_j(\mathbf{x}_{1j}, \beta)^{\text{tr}}]\}^{-1}$$

In these expressions, functions are evaluated at the same β governing the data \mathbf{x}_{ij} .

(III) For an efficient regular combined-sample estimator of the form $\hat{\beta} = g(\beta_1, \dots, \tilde{\beta}_k, \tilde{V}_1, \dots, \tilde{V}_k)$ with g continuously differentiable, Taylor linearisation of g in terms of its $\tilde{\beta}_j$ arguments and their Jacobians J_{x_j} (the 'Delta Method') shows that

$$\sqrt{n}(\hat{\beta} - \beta) \stackrel{P}{\approx} \sum_{j=1}^k (n/n_j)^{1/2} J_{x_j}(\beta, \dots, \beta; V_1, \dots, V_k) \\ \times \sqrt{n_j}(\hat{\beta}_j - \beta) \quad (15)$$

i.e., when n_j are all of order n , shows that $\sqrt{n}(\hat{\beta} - \beta)$ can be represented as a linear combination of the normalised centred ML estimators $\sqrt{n_j}(\hat{\beta}_j - \beta)$, up to a remainder converging to 0 in probability. Since (3) (with variances $V_j^{-1} = V_j^-$ now all assumed to exist) is shown in Appendix 1 to be the unique optimal matrix-linear-combination estimator in the sense of minimal asymptotic variance, up to $o_P(1)$ remainders, it follows that $\hat{\beta}$ is equal to the linear combination (3) plus a remainder of smaller order than $\sum_{j=1}^k n_j^{-1/2}$ in probability.

(IV) There is one more case where the additivity of information (equality in (13)) is obvious, namely the case where all $I_{12}^{(j)}$ are 0. This case is called *adaptive* because, by (11) or (12), the j th sample information $\mathcal{I}^{(j)}(\beta) = I_{11}^{(j)}$ is exactly the same as if λ_j were known in advance. But then $I_{11} = \sum_{j=1}^k (n_j/n) \mathcal{I}^{(j)}(\beta)$ by (10), while (12) with $K = \mathbf{0}$ implies $\mathcal{I}^*(\beta) \leq I_{11}$. Then equality must hold in (13).

(V) We now come to the main result of this paper, which says that when the nuisance parameters λ_j in the separate samples are distinct and unrelated, then (13) becomes an equality and an efficient estimator of β of the form (3) exists. *In the multi-sample context described above, under standard regularity conditions, if all $\lambda_j \in \Lambda_j$ vary freely, unconstrained by one another or by β , so that $\underline{\lambda} = (\lambda_1^{\text{tr}}, \lambda_2^{\text{tr}}, \dots, \lambda_k^{\text{tr}})^{\text{tr}} \in \Lambda_1 \times \dots \times \Lambda_k$ and $d = \sum_{j=1}^k q_j$, then equality holds in (13).* The restriction to unconstrained nuisance parameters is essential in this statement, as will be shown by example in Section 4.

There are a few different proofs of (V) in different cases. When the separate-sample estimators $\tilde{\beta}$ of the same shared parameter vector β are Maximum Likelihood estimators in the presence of the nuisance parameters λ_j , Lin and Zeng (2010) observe that the

maximum profile likelihood estimator for β (i.e., the maximiser of the log-likelihood partially maximised over nuisance parameters) in each of the separate studies and in the combined study is efficient and that the log profile likelihood for the combined study is simply the sum of the log profile likelihoods of the separate studies. Thus the combined-study information is the sum of the separate-study informations. That is the complete proof in the MLE case. Since any separate-study efficient regular estimators $\tilde{\beta}_j$ of β are asymptotically equivalent in probability to the corresponding MLEs or profile-likelihood maximisers, according to the parametric Hájek-LeCam convolution theorem, this proof establishes the same result for the combination of any efficient regular separate-study estimators. Further technicalities would have been needed to carry this proof idea forward to the case of freely varying infinite-dimensional nuisance parameters λ_j . Lin and Zeng (2010) did not do that, but our Appendix 3 does. In another direction of generalisation, when the separate-sample estimators β_j estimate not β but projections $\Pi_j \beta$ (with common null-space $\mathbf{0}$), the profile-likelihood proof sketched above yields the same result when coupled with the verification in Appendix 1 that the combined-sample MLE has $\mathcal{I}^*(\beta) = \sum_{j=1}^k (n_j/n) \mathcal{I}^{(j)}(\beta)$.

(VI) The same result in (V) – the optimality of estimator (3) – also holds when the separate-study estimators $\tilde{\beta}_j, \tilde{\lambda}_j$ are obtained through the solution of M-estimating equations (4), when the notion of *optimality* is suitably clarified. Here the estimating functions $\Psi_j = \nabla_{\beta, \lambda_j} M_j$ with values in \mathbb{R}^{p+q_j} are assumed to satisfy standard smoothness with respect to parameters and other regularity conditions, including that $E(\Psi_j(\mathbf{x}_{j1}, \beta, \lambda_j)) = \mathbf{0}$, where expectations are taken and the integrand functions evaluated at the same true parameters β, λ_j , and where

$$C_j \equiv E\left(-\nabla_{\beta, \lambda_j} \Psi_j^{\text{tr}}(\mathbf{x}_{j1}, \beta, \lambda_j)\right)^{\text{tr}},$$

$$B_j \equiv E\left(\Psi_j(\mathbf{x}_{j1}, \beta, \lambda_j) \Psi_j(\mathbf{x}_{j1}, \beta, \lambda_j)^{\text{tr}}\right)$$

are nonsingular $(p + q_j) \times (p + q_j)$ matrices. The solution of (4) is locally unique, in the vicinity of the true (β, λ_j) , with probability converging to 1 for large n , by general estimating equation theory (Van der Vaart, 1998, Chapter 5 or Janicki, 2009, Chapter 2), and estimators $(\tilde{\beta}_j, \tilde{\lambda}_j)$ are regular asymptotically linear with asymptotic distributions

$$\sqrt{n_j} \begin{pmatrix} \tilde{\beta}_j - \beta \\ \tilde{\lambda}_j - \lambda_j \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, W_j), \quad \mathcal{I}^{(j)}(\beta, \lambda_j) \equiv W_j^{-1}$$

where $W_j = C_j^{-1} B_j (C_j^{\text{tr}})^{-1}$. Beyond this point, we maintain the same notations as throughout Section 3, namely that $V_j(\beta)$ is the upper-left $p \times p$ block of W_j

and

$$\mathcal{I}^{(j)}(\beta, \lambda_j) = W_j^{-1} = \begin{pmatrix} I_{11}^{(j)} & I_{12}^{(j)} \\ I_{21}^{(j)} & I_{22}^{(j)} \end{pmatrix},$$

$$\mathcal{I}^{(j)}(\beta) = (V_j(\beta))^{-1} = I_{11}^{(j)} - I_{12}^{(j)} (I_{22}^{(j)})^{-1} I_{21}^{(j)}$$

although these inverse variance matrices are no longer Fisher information.

From now on, assume as in paragraph (V) that the parameters λ_j range freely and are not constrained by one another, so that the combined-sample unknown parameters are $(\beta, \underline{\lambda}) = (\beta, \lambda_1, \dots, \lambda_k) \in \mathbb{R}^{p+d}$. The first step in extending (V) is to extend the optimality result in (II) by viewing the separate-study estimating equations all as estimating equations for the combined parameters $(\beta, \underline{\lambda})$. Let

$$\begin{aligned} \Psi_j^0(x, \beta, \underline{\lambda}) &= R_j \Psi_j(x, \beta, \lambda_j) \\ &= \left(\nabla_{\beta}^{\text{tr}} M_j(x, \beta, \lambda_j) \mid \mathbf{0}_{q_1}^{\text{tr}} \mid \dots \mid \mathbf{0}_{q_{j-1}}^{\text{tr}} \mid \right. \\ &\quad \left. \nabla_{\lambda_j}^{\text{tr}} M_j(x, \beta, \lambda_j) \mid \mathbf{0}_{q_{j+1}}^{\text{tr}} \mid \dots \mid \mathbf{0}_{q_k}^{\text{tr}} \right)^{\text{tr}} \end{aligned}$$

where R_j is the $(p + d) \times (p + q_j)$ matrix with the block decomposition containing 0's everywhere except for the identity matrix in the upper-left $p \times p$ block and another $q_j \times q_j$ identity matrix in the submatrix with consecutive row-indices from $p + \sum_{a=1}^{j-1} q_a + 1$ through $p + \sum_{a=1}^j q_a$ and consecutive column-indices from $p+1$ through $p + q_j$.

Now consider the class of all combined-sample estimating equations defined by

$$\Psi(\mathbf{X}, \beta, \underline{\lambda}) = \sum_{j=1}^k A_j(\beta, \underline{\lambda}) \sum_{i=1}^{n_j} \Psi_j^0(\mathbf{x}_{ji}, \beta, \lambda_j) = \mathbf{0} \quad (16)$$

where the $(p + d) \times (p + d)$ matrices $A_j(\beta, \underline{\lambda})$ are continuously differentiable functions of their arguments and where

$$\begin{aligned} C^* &\equiv \frac{1}{n} E(-\nabla_{\beta, \underline{\lambda}} \Psi^{\text{tr}}(\mathbf{X}, \beta, \underline{\lambda}))^{\text{tr}} \\ &= \frac{1}{n} \sum_{j=1}^k n_j A_j(\beta, \underline{\lambda}) R_j C_j R_j^{\text{tr}} \text{ is nonsingular} \quad (17) \end{aligned}$$

The same arguments as in standard estimating-equation theory show that for large n , the solution of (16) is, with probability approaching 1, locally unique for $(\beta, \underline{\lambda})$ in a non-shrinking neighbourhood of the true values, and defines \sqrt{n} -consistent asymptotically normal estimators $(\hat{\beta}, \hat{\underline{\lambda}})$ with asymptotic variance

$(C^*)^{-1} B^* (C^{*\text{tr}})^{-1}$, where

$$\begin{aligned} B^* &= \frac{1}{n} E \left(\Psi(\mathbf{X}, \beta, \underline{\lambda}) \Psi(\mathbf{X}, \beta, \underline{\lambda})^{\text{tr}} \right) \\ &= \sum_{j=1}^k \frac{n_j}{n} A_j(\beta, \underline{\lambda}) R_j B_j R_j^{\text{tr}} A_j(\beta, \underline{\lambda})^{\text{tr}} \end{aligned} \quad (18)$$

Then a slight reworking of the proof of Janicki (2009, Theorem 6.1.2), using the fact that the generalised inverse of $R_j B_j C_j^{-1} R_j^{\text{tr}}$ is $R_j C_j B_j^{-1} R_j^{\text{tr}}$, shows that the combined estimating equation (16) with smallest asymptotic variance in the sense of positive-definite ordering of matrices is achieved when $A_j(\beta, \underline{\lambda}) = R_j C_j B_j^{-1} R_j^{\text{tr}}$, and we fix this choice for $A_j(\beta, \underline{\lambda})$ in equations (16)–(18) from now on.

Denote by $(\hat{\beta}, \hat{\underline{\lambda}})$ the solution of the estimating equation (16), which takes the form

$$\Psi(\mathbf{X}, \beta, \underline{\lambda}) = \sum_{j=1}^k R_j C_j B_j^{-1} \sum_{i=1}^{n_j} \Psi_j(X_{ji}, \beta, \lambda_j) = \mathbf{0} \quad (16')$$

with equations (17) and (18) giving

$$\begin{aligned} B^* &= C^* = \sum_{j=1}^k \frac{n_j}{n} R_j C_j B_j^{-1} C_j^{\text{tr}} R_j^{\text{tr}} \\ &= \sum_{j=1}^k \frac{n_j}{n} R_j \mathcal{I}^{(j)}(\beta, \lambda_j) R_j^{\text{tr}} \end{aligned} \quad (19)$$

Now we return to our objective of comparing the combined-sample information-analogue $\mathcal{I}^*(\beta)$ for β , which is the inverse of the combined-sample asymptotic variance for its optimal estimating equation-estimator $\hat{\beta}$, with the information-analogue from equation (3). The asymptotic variance matrix for $(\hat{\beta}, \hat{\underline{\lambda}})$ is $(C^*)^{-1}$, so the combined-sample information-analogue is $\mathcal{I}^*(\beta, \underline{\lambda}) \equiv C^*$. To find $\mathcal{I}^*(\beta)$ as in equations (9) and (11), we block-decompose

$$\mathcal{I}^*(\beta, \underline{\lambda}) = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}$$

where the upper-left block I_{11} is $p \times p$ and the lower-right is $d \times d$.

The upper-left $p \times p$ block of C^* in (19) is obtained by definition of R_j as the weighted sum of upper-left blocks of $\mathcal{I}^{(j)}(\beta, \lambda_j)$, or $I_{11}^{(j)}$, that is,

$$I_{11} = \sum_{j=1}^k \frac{n_j}{n} I_{11}^{(j)}$$

Again by definition of R_j , the $p \times d$ block I_{12} is obtained from (19) as

$$I_{12} = \left(\frac{n_1}{n} I_{12}^{(1)} \mid \frac{n_2}{n} I_{12}^{(2)} \mid \dots \mid \frac{n_k}{n} I_{12}^{(k)} \right)$$

while the $d \times d$ block I_{22} is block-diagonal when decomposed in successive blocks of sizes q_1, q_2, \dots, q_k ,

with j th diagonal block given by $(n_j/n) I_{22}^{(j)}$, for $j = 1, \dots, k$. Therefore,

$$\begin{aligned} \mathcal{I}^*(\beta) &= I_{11} - I_{12} I_{22}^{-1} I_{21} \\ &= \sum_{j=1}^k \frac{n_j}{n} I_{11}^{(j)} - \sum_{j=1}^k \frac{n_j}{n} I_{12}^{(j)} (I_{22}^{(j)})^{-1} I_{21}^{(j)} \end{aligned} \quad (20)$$

Since the inverse asymptotic variance of $\tilde{\beta}_j$ is $\mathcal{I}^{(j)}(\beta) = I_{11}^{(j)} - I_{12}^{(j)} (I_{22}^{(j)})^{-1} I_{21}^{(j)}$, we have proved the desired result, that the asymptotic variance of the inverse-variance weighted linear combination (3) is the same as the asymptotic variance for the estimator derived from the optimal weighted estimating function (16) in this section.

Remark on the Proof of (VI). The proof relies on the M-estimating function property of Ψ_j both in achieving the specific formula in (17) for C^* and because that is the setting where the nuisance-parameter block of the combined-sample information analogue is block-diagonal. The application of (VI) is to meta-analyses where only the structural parameter estimates and their estimated variances are reported in the separate studies. However, the fairly dramatic conclusion is that in the setting of M-estimating functions with unconstrained λ_j 's, even if all estimates and variances for nuisance parameters were also available for combined analysis, the asymptotic variances of the combined estimates of structural parameters would be no better.

4. Examples where information additivity fails

It remains to clarify that the asymptotic optimality of weighted linear combinations of separate-sample estimators in the combined sample cannot persist generally when the nuisance parameters in the two samples are *coupled*, i.e., partially shared or related through common constraints. The matrix weights in the combined estimator (3) are optimal in the sense of minimising variance, and they lead to a combined estimator with asymptotic information $\mathcal{I}^*(\beta)$. So when the lower bound on information is not attained at $\mathcal{I}^*(\beta)$ — and by (13) is smaller with respect to the positive-definite ordering — linear combinations cannot be optimal. An example of a coupled parameterisation is the two-sample location-scale model where λ_1 is finite-dimensional and for a common location parameter β and a positive unknown scalar σ ,

$$\begin{aligned} f_1(x, \beta, \lambda_1) &= f_0(x - \beta, \lambda_1), \\ f_2(y, \beta, \lambda_2) &\equiv \frac{1}{\sigma} f_0\{(y - \beta)/\sigma, \lambda_1\} \end{aligned} \quad (21)$$

Thus in this example, $\lambda_2 \equiv (\sigma, \lambda_1)$, and we replace λ_1 by λ in the notations below. In the case of finite-dimensional λ , straightforward calculation shows that

the per-observation combined-sample Fisher information matrix has the form

$$\mathcal{I}^*(\beta, \sigma, \lambda) = \begin{pmatrix} I_{\beta\beta} & I_{\beta\sigma} & I_{\beta\lambda} \\ I_{\sigma\beta} & I_{\sigma\sigma} & I_{\sigma\lambda} \\ I_{\lambda\beta} & I_{\lambda\sigma} & I_{\lambda\lambda} \end{pmatrix} \quad (22)$$

with the expressions for $I_{\sigma\beta} = I_{\beta\sigma}$, $I_{\beta\lambda} = (I_{\lambda\beta})^{\text{tr}}$, $I_{\sigma\lambda} = (I_{\lambda\sigma})^{\text{tr}}$ given in terms of $c = n_1/n$ in Appendix 2.

The goal of this example is to show that generally the equality in (13) does not hold, by showing that the reciprocal of the (β, β) or upper-left element of $\{\mathcal{I}^*(\beta, \sigma, \lambda)\}^{-1}$ is *not* equal to the sum of the reciprocals of the upper-left elements of the inverses respectively of

$$\begin{aligned} & \frac{n_1}{n} \mathcal{I}^{(1)}(\beta, \lambda) \\ & \equiv c \begin{pmatrix} \left(c + \frac{1-c}{\sigma^2}\right)^{-1} I_{\beta\beta} & \left(c + \frac{1-c}{\sigma}\right)^{-1} I_{\beta\lambda} \\ \left(c + \frac{1-c}{\sigma}\right)^{-1} I_{\beta\lambda} & I_{\lambda\lambda} \end{pmatrix} \end{aligned} \quad (23)$$

and of

$$\begin{aligned} & \frac{n_2}{n} \mathcal{I}^{(2)}(\beta, \sigma, \lambda) \\ & \equiv \begin{pmatrix} \left(\frac{c\sigma^2}{1-c} + 1\right)^{-1} I_{\beta\beta} & I_{\beta\sigma} & \left(\frac{c\sigma}{1-c} + 1\right)^{-1} I_{\beta\lambda} \\ I_{\sigma\beta} & I_{\sigma\sigma} & I_{\sigma\lambda} \\ \left(\frac{c\sigma}{1-c} + 1\right)^{-1} I_{\beta\lambda} & I_{\lambda\sigma} & (1-c)I_{\lambda\lambda} \end{pmatrix} \end{aligned} \quad (24)$$

These two displayed information matrices are the sample-1 and 2 total-information matrices divided by the total sample-size n .

We supply below a numerical comparison of the single-sample and combined-sample information bounds for estimating β . Before doing so, we indicate a class of densities within this framework where separate-sample informations *do* add to give the combined-sample information.

Suppose in the location-scale two-sample setting just described that each density $f_0(\cdot, \lambda)$ is symmetric (i.e., an even function) and that its derivative with respect to λ is also even. Then it is easy to check that f'_0 is odd, and inspection of the integral formulas in Appendix 2 shows that $I_{\beta\sigma} = 0$ and $I_{\lambda\beta} = \mathbf{0}$. In that case, estimation of β is adaptive as in (IV) (just as efficient without as with knowledge of the nuisance parameters), with separate-sample per-observation information numbers for β respectively $\mathcal{I}^{(1)}(\beta) = \int (f'_0)^2 / f_0$ and $\mathcal{I}^{(2)}(\beta) = \sigma^{-2} \int (f'_0)^2 / f_0$, and combined-sample information $\mathcal{I}^*(\beta) = (c + ((1-c)/\sigma^2) \int (f'_0)^2 / f_0$. The

t distributions form a special case of this example:

$$f_0(x, \lambda) = (1 + x^2)^{-\lambda-1} / h(\lambda),$$

$$h(\lambda) = \sqrt{\pi} \Gamma(\lambda + 1/2) / \Gamma(\lambda + 1)$$

This paragraph shows that within the two-sample location-scale setting, we must look within *skewed* density classes to find examples where strict inequality holds in (13).

So we turn to two-sample location family examples (21) in which f_0 is chosen to be asymmetric. One such density family is the *skew-normal* due to Azzalini (1985)

$$f_0(x, \lambda) = 2 \phi(x) \Phi(\lambda x)$$

In this family, in the special case where the true value of $\lambda = 0$, integration shows that the combined-sample per-observation information matrix is

$$\begin{pmatrix} c + (1-c)/\sigma^2 & 0 & \sqrt{2/\pi} (c + (1-c)/\sigma) \\ 0 & 2(1-c)/\sigma^2 & 0 \\ \sqrt{2/\pi} (c + (1-c)/\sigma) & 0 & 2/\pi \end{pmatrix}$$

which is nonsingular if and only if $\sigma \neq 1$, and in that case the combined-sample information for β is $c(1-c)(\sigma-1)^2/\sigma^2$. However, this example is remarkable in that the single-sample information matrices are

$$\begin{aligned} & c \begin{pmatrix} 1 & \sqrt{2/\pi} \\ \sqrt{2/\pi} & 2/\pi \end{pmatrix} \\ & \text{and } \frac{1-c}{\sigma^2} \begin{pmatrix} 1 & 0 & 2\sigma/\sqrt{2\pi} \\ 0 & 2 & 0 \\ 2\sigma/\sqrt{2\pi} & 0 & 2\sigma^2/\pi \end{pmatrix} \end{aligned}$$

and are both singular, and the single-sample information numbers for β in the presence of nuisance parameters tend to 0 as λ decreases to 0.

Non-zero values of the parameter λ further illustrate the phenomenon of non-additivity of separate-sample information bounds for estimating the common location parameter β when nuisance parameters in the two samples are related. Table 2 shows several numerically calculated values (using the function *integrate* in R, R Development Core Team, 2017) of separate-sample and combined information for estimating a location parameter, all divided by the total sample size n . The table refers to the skew-normal two-sample location-scale problem, where λ is the skew-parameter, σ the scale parameter, and $c = 0.5$ the proportion of observations in sample 1. The point of these examples is that the last two rows do not sum to the combined-sample information $\mathcal{I}^*(\beta)$, although this is nearly true when $\sigma = 1$. In the case previously discussed, with $\lambda = 0$, both of the separate-sample information entries $c^{2-j} (1-c)^{j-1} \mathcal{I}^{(j)}(\beta)$ were 0, while the combined-sample information was $\mathcal{I}^*(\beta) = c(1-c)(1-1/\sigma)^2$.

Table 2. Information on β in combined and separate skew-normal samples. In all cases, $c = 0.5$.

λ	0.25	0.5	1.0	.05	0.1	0.25	0.5	1.0	2.0
σ	1	1	1	2	2	2	2	2	2
$\mathcal{I}^*(\beta)$.0382	.1378	.4162	.0638	.0676	.0929	.1718	.3868	.8053
$0.5 \mathcal{I}^{(1)}(\beta)$.0377	.1307	.3572	.0016	.0063	.0377	.1307	.3572	.7345
$0.5 \mathcal{I}^{(2)}(\beta)$.0002	.0034	.0389	.0000	.0000	.0001	.0008	.0097	.0563

Examples like these have practical consequences. In paragraph (V) above, suppose that $\gamma \equiv (\beta_1, \dots, \beta_r) \in \mathbb{R}^r$ is a subvector of the parameter β , where $r < p = \dim(\beta)$. Even if additivity of information (i.e., equality in (13)) holds for two-sample inference about β , it does **not** necessarily hold also for inference about γ based on the same data. As an example, consider the same two-sample skew-normal location-scale problem just presented, but now let the common parameter of interest be $\tilde{\beta} = (\beta, \lambda_1)$, let the sample-1 nuisance parameter $\tilde{\lambda}_1$ be null (i.e., absent), and replace λ_2 by $\tilde{\lambda}_2 \equiv \sigma$. Apart from the change in notation, let the single-sample densities be the same as before. Then the nuisance-parameter varies separately and freely over the two samples, and paragraph (V) implies that information additivity *does* hold for $\tilde{\beta}$. But we have seen in the discussion and Table 2 above that inequality (13) is *strict* for the sub-vector β .

5. Discussion: meta-analysis and shared parameters

It is common in statistical applications to model separately collected datasets with shared parameters. Meta-analysis is one approach, within biomedical or social science, to the pooling of information across separate centres or data-collections. Parameters like treatment effects — those that are important for practical decisions — are most often shared across sub-samples, but common nuisance parameters may also naturally arise when cross-classifying variables are adjusted away using models. We have seen in this paper that there is no obstacle to the efficient weighted linear combination of separate-sample ML or M-estimators when nuisance parameters are either absent or vary freely without constraints across samples, but that otherwise, efficient functional combination may be impossible. The clear message for statistical practice is that whenever possible, separate studies which might be combined should report the estimator along with the estimated variance of a parameter vector which includes both β and whichever components of λ_j might be shared across the models for separate samples.

Acknowledgments

The authors gratefully acknowledge the Eastern Cooperative Oncology Group as the source for the ECOG EST 1582 dataset, and the suggestion of a referee to expand our treatment of (V) to estimating equations.

Disclosure statement

This paper is released to inform interested parties of research and to encourage discussion. Any views expressed are the authors' and not necessarily those of the U.S. Census Bureau.

Notes on contributors

Eric Slud is Professor in the Statistics Program within the Mathematics Department of the University of Maryland, College Park, and is Area Chief for Mathematical Statistics in the Center for Statistical Research and Methodology of the US Census Bureau.

Ilia Vonta is an Associate Professor at the Department of Mathematics of the School of Applied Mathematical and Physical Sciences of the National Technical University of Athens, Athens, Greece and an Associate Professor-Tutor of the Hellenic Open University.

Abram Kagan is Professor in the Statistics Program within the Mathematics Department of the University of Maryland, College Park.

References

- Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12, 171–178.
- Bickel, P., & Doksum, K. (2007). *Mathematical statistics* (2nd ed., Vol. I). Upper Saddle River, NJ: Pearson Prentice Hall. updated printing.
- Bickel, P., Klaassen, C., Ritov, Y., & Wellner, J. (1998). *Efficient and adaptive estimation for semiparametric models*. Berlin: Springer.
- Carlin, B., & Louis, T. (2008). *Bayesian methods for data analysis* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.
- DeMissie, M. (2009). *Investigating center effects in a multicenter clinical trial study using a parametric proportional hazards meta-analysis model* (MS Thesis). Univ. of Maryland Statistics Program, DRUM Digital Repository. Retrieved from <https://drum.lib.umd.edu/handle/1903/9588>
- Draper, N., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: Wiley.
- Efron, B. (1996). Empirical Bayes methods for combining likelihoods. *Journal of the American Statistical Association*, 91, 538–550.
- Gray, R. (1994). A Bayesian analysis of institutional effects in a multicenter cancer clinical trial. *Biometrics*, 50, 244–253.
- Hartung, J., Knapp, G., & Sinha, B. (2008). *Statistical meta-analysis with applications*. Hoboken, NJ: Wiley.
- Janicki, R. (2009). *Statistical inference based on estimating functions in exact and misspecified models* (Ph.D. thesis). Univ. of Maryland Statistics Program, DRUM Digital Repository. Retrieved from <https://drum.lib.umd.edu/handle/1903/9690>

- Kagan, A., & Rao, C. R. (2003). Some properties and applications of the efficient Fisher score. *Journal of Statistical Planning & Inference*, 116, 343–352.
- Lin, D. Y., & Zeng, D. (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, 97, 321–332.
- Liu, D., Liu, R., & Xie, M. (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: Efficiency and robustness. *Journal of the American Statistical Association*, 110, 326–340.
- Olkin, I., & Sampson, A. (1998). Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics*, 54, 347–352.
- R Development Core Team (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org>, ISBN 3-900051-07-0.
- Rao, C. R. (1973). *Linear statistical inference* (2nd ed.). New York: Wiley.
- Rao, J. N. K., & Molina, I. (2015). *Small area estimation* (2nd ed.). Hoboken, NJ: Wiley.
- Slud, E., & DeMissie, M. (2011). Validity of regression meta-analyses versus pooled analyses of mixed-effect linear models. *Mathematics in Engineering, Science and Aerospace*, 2(4), 251–266.
- Tsiatis, A. (2006). *Semiparametric theory and missing data*. New York: Springer.
- Van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge: Cambridge University Press.

Appendices

Appendix 1. Optimality of the linear combination in equation (3)

We are interested in *regular* estimators of $\beta \in \mathbb{R}^p$ based on separate-sample efficient estimators $\tilde{\beta}_j$ of the projected parameters $\Pi_j\beta$, where Π_j are projections onto known subspaces of \mathbb{R}^p . The most important and well-studied case is where the same parameter is estimated in all samples, and $\Pi_j = \mathbf{I}$, but it turns out to be almost as easy to consider the general case where $\{\Pi_j\}_{j=1}^k$ determines β , or equivalently, where $\cap_{j=1}^k \text{null}(\Pi_j) = \emptyset$. If $\text{range}(\Pi_j)$ is of dimension less than p , then the projection $\Pi_j\beta$ is understood as having ‘structural zeroes’ in place of the parameter components $(\mathbf{I} - \Pi_j)\beta$, such as would occur if the β entries were regression coefficients for the (independent, randomly generated) rows of a $n_j \times p$ design matrix D_j , and if in the j th sample the rows of D_j were structurally constrained to be in the range space of Π_j , so that $D_j(\mathbf{I} - \Pi_j) = \mathbf{0}$ is almost surely the zero matrix. Further assume (without loss of generality) that the estimators $\tilde{\beta}_j$ are constrained to fall in the range space of Π_j , so that $(\mathbf{I} - \Pi_j)\tilde{\beta}_j = \mathbf{0}$, and that the asymptotic variance matrix $V_j(\beta)$ of $\sqrt{n_j}(\tilde{\beta}_j - \Pi_j\beta)$ exists and is consistently estimated by \tilde{V}_j . Without loss of generality $(\mathbf{I} - \Pi_j)V_j(\beta) = (\mathbf{I} - \Pi_j)\tilde{V}_j = \mathbf{0}$, and $V_j(\beta)$ is an invertible linear operator from $\text{range}(\Pi_j)$ to itself, with inverse which we denote by \mathcal{I}_j^V . Then $V_j^- \equiv \mathcal{I}_j^V \circ \Pi_j$ is a *generalised inverse* of $V_j = V_j(\beta)$ (Rao, 1973, p. 24), which means that

$$V_j(\beta) V_j^- V_j(\beta) = V_j(\beta) \quad \text{and} \quad V_j^- V_j(\beta) V_j^- = V_j^-$$

More specifically, we have by definition of V_j^- ,

$$V_j(\beta) V_j^- = V_j^- V_j(\beta) = \Pi_j \quad (\text{A1})$$

Similar notations \tilde{V}_j^- and generalised-inverse properties also hold for the estimated variance matrices \tilde{V}_j . The projection operators Π_j need not all have the same range, but we assume that they have no common nontrivial null-space.

We restrict to estimators of the linear form $\sum_{j=1}^k A_j \tilde{\beta}_j$, as in paragraph (III) of Section 3, where A_j are $p \times p$ matrices that depend continuously on $\tilde{\beta}_j$ and the variance-estimators \tilde{V}_j , and there is no loss of generality in imposing the structural-zero constraint $A_j = A_j \Pi_j$. It follows from regularity that for large n , the matrices A_j must satisfy the constraint

$$\sum_{j=1}^k A_j - \mathbf{I} \xrightarrow{P} \mathbf{0} \quad (\text{A2})$$

where \mathbf{I} denotes the $p \times p$ identity matrix. This holds because regularity (Bickel et al., 1998, pp. 17–21, or Van der Vaart, 1998, p. 115) of estimators $\tilde{\beta}_j$ and of $\sum_{j=1}^k A_j \tilde{\beta}_j$ implies that $\sqrt{n_j}(\tilde{\beta}_j - \Pi_j\beta)$ and

$$\begin{aligned} \sqrt{n} \left(\sum_{j=1}^k A_j \tilde{\beta}_j - \beta \right) &= \sum_{j=1}^k \left(\frac{n}{n_j} \right)^{1/2} A_j \sqrt{n_j} (\tilde{\beta}_j - \Pi_j\beta) \\ &+ \sqrt{n} \left(\sum_{j=1}^k A_j - \mathbf{I} \right) \beta \end{aligned}$$

each converge in distribution to the same respective limits whenever β is replaced by any element $\beta + c/\sqrt{n}$ in a neighbourhood of extent $O(1/\sqrt{n})$ about β .

We now prove that among estimators $\sum_{j=1}^k A_j \tilde{\beta}_j$ satisfying (A2) and $A_j = A_j \Pi_j$, the smallest asymptotic variance matrix with respect to positive-definite ordering is attained only when A_j is asymptotically equivalent to (differs by $o_P(1)$ from)

$$A_j^o \equiv \left\{ \sum_{l=1}^k n_l V_l^- \right\}^{-1} n_j V_j^- \quad (\text{A3})$$

If A_j^* denotes the large-sample in-probability limit of A_j , for $j = 1, \dots, k$, then $A_j^* = A_j^* \Pi_j$ and the constraint (A2) immediately implies that $\sum_{j=1}^k A_j^* \Pi_j = \mathbf{I}$. Then the asymptotic variance matrix of the estimator $n^{1/2}(\sum_{j=1}^k A_j \tilde{\beta}_j - \beta)$ is, by independence of the samples \mathbf{X}_j , the limit of

$$\begin{aligned} n \sum_{j=1}^k A_j^* \{V_j(\beta)/n_j\} A_j^{*\text{tr}} &= n \sum_{j=1}^k A_j^o \{V_j(\beta)/n_j\} A_j^{o\text{tr}} \\ &+ n \sum_{j=1}^k (A_j^* - A_j^o) \{V_j(\beta)/n_j\} (A_j^* - A_j^o)^{\text{tr}} \quad (\text{A4}) \end{aligned}$$

The equality in (A4) follows immediately from the constraint $\sum_j A_j^* \Pi_j = \mathbf{I}$, noting by (A1) and (A3) that

$$\begin{aligned} \sum_{j=1}^k A_j^o \{V_j(\beta)/n_j\} A_j^{o\text{tr}} &= \left[\sum_{l=1}^k n_l V_l^- \right]^{-1} \sum_{j=1}^k \Pi_j A_j^{*\text{tr}} \\ &= \left[\sum_{l=1}^k n_l V_l^- \right]^{-1} \end{aligned}$$

which is equal to $\sum_{j=1}^k A_j^o \{V_j(\beta)/n_j\} A_j^{o\text{tr}}$. The unique variance-minimising property of $A_j^* = A_j^o$ follows from the fact that the last matrix on line (A4) is nonnegative definite

and is $\mathbf{0}$ only when all $A_j^* = A_j^0$. The desired result has been proved.

Appendix 2. Information matrices in Section 4

Section 4 presents a two-sample location-scale problem with respective densities

$$f_1(x, \beta, \lambda) = f_0(x - \beta, \lambda),$$

$$f_2(y, \beta, \sigma, \lambda) \equiv \frac{1}{\sigma} f_0(\{y - \beta\}/\sigma, \lambda)$$

and with sample sizes n_1 and $n_2 = n - n_1$ with $c \equiv n_1/n$. The combined-sample information matrix (22) for the parameters (β, σ, λ) is given as

$$\begin{aligned} & c E(\{\nabla_{\beta, \sigma, \lambda} \log f_1(x_{11}, \beta, \lambda)\}^{\otimes 2}) \\ & + (1 - c) E(\{\nabla_{\beta, \sigma, \lambda} \log f_2(x_{12}, \beta, \sigma, \lambda)\}^{\otimes 2}) \\ & \equiv c \begin{pmatrix} \{\mathcal{I}^{(1)}(\beta, \lambda)\}_{11} & 0 & \{\mathcal{I}^{(1)}(\beta, \lambda)\}_{12} \\ 0 & 0 & 0 \\ \{\mathcal{I}^{(1)}(\beta, \lambda)\}_{12} & 0 & \{\mathcal{I}^{(1)}(\beta, \lambda)\}_{22} \end{pmatrix} \\ & + (1 - c) I^{(2)}(\beta, \lambda) \end{aligned}$$

since $\nabla_{\sigma} \log f_1(x_{11}, \beta, \lambda)$ is 0. Then straightforward integration yields the entries of $\mathcal{I}^{(1)}(\beta, \lambda) = E(\{\nabla_{\beta, \lambda} \log f_1(x_{11}, \beta, \lambda)\}^{\otimes 2})$ as

$$I_{11}^{(1)} = \int \frac{(f_0')^2}{f_0} dx, \quad I_{12}^{(1)} = - \int \frac{f_0'}{f_0} (\nabla_{\lambda} f_0) dx,$$

$$I_{22}^{(1)} = \int \frac{(\nabla_{\lambda} f_0)^2}{f_0} dx$$

where f_0', f_0'' denote derivatives of $f_0(x, \lambda)$ with respect to the first argument x , and all functions f_0 and derivatives are evaluated at (x, λ) and integrated with respect to the x variable on $(-\infty, \infty)$.

Similarly one calculates directly the entries of the symmetric 3×3 matrix $\mathcal{I}^{(2)}(\beta, \lambda) = E(\{\nabla_{\beta, \sigma, \lambda} \log f_2(x_{12}, \beta, \sigma, \lambda)\}^{\otimes 2})$ as

$$I_{11}^{(2)} = \frac{1}{\sigma^2} I_{11}^{(1)}, \quad I_{12}^{(2)} = \int \frac{x(f_0')^2}{\sigma^2 f_0}, \quad I_{13}^{(2)} = \frac{1}{\sigma} I_{12}^{(1)}$$

$$I_{22}^{(2)} = \frac{1}{\sigma^2} \left(\int \frac{x^2(f_0')^2}{f_0} dx - 1 \right),$$

$$I_{23}^{(2)} = - \int (\nabla_{\lambda} f_0) \frac{x f_0'}{\sigma f_0} dx, \quad I_{33}^{(2)} = I_{22}^{(1)}$$

From these integral expressions we derive the entries of (22) as

$$I_{\beta\beta} = \left(c + \frac{1-c}{\sigma^2} \right) I_{11}^{(1)}, \quad I_{\beta\sigma} = (1-c) I_{12}^{(2)},$$

$$I_{\sigma\sigma} = (1-c) I_{22}^{(2)},$$

$$I_{\lambda\lambda} = I_{22}^{(1)}, \quad I_{\beta\lambda} = \left(c + \frac{1-c}{\sigma} \right) I_{12}^{(1)}, \quad I_{\sigma\lambda} = (1-c) I_{23}^{(2)}$$

along with the separate-sample information expressions (23) and (24).

Appendix 3. Infinite Dimensional Nuisance Parameters

This Appendix provides an extension to infinite-dimensional nuisance parameters, of the result of paragraph (V) of Section 3 on the efficient functional combination of separate-sample MLEs to provide an efficient combined-sample estimator of a structural (finite-dimensional) parameter β shared across samples. The result states that such an efficient combination exists when the separate-sample nuisance parameters vary freely and independently of one another.

We maintain the notation of Section 3. The nuisance-parameter spaces Λ_j may now be infinite-dimensional, $q_j \leq \infty$. Assume that for all q_j dimensional vectors $\lambda_j \in \Lambda_j$ and any q_j dimensional vectors w_j with at most finitely many components nonzero, there exist $t_j > 0$ such that $\lambda_j + t_j w_j \in \Lambda_j$. We also assume standard regularity and nondegeneracy conditions about finite-dimensional submodels with parameters (β, λ_j) (as in Bickel & Doksum, 2007, Theorem 6.2.2, and Van der Vaart, 1998, Theorem 5.39). In particular, for each finite-dimensional affine subset $M_j \subset \Lambda_j$, (a set of the form $\{\mathbf{u} + \mathbf{v} : \mathbf{v} \in \mathbf{V}_j\}$ where \mathbf{V}_j is a vector space) and $(\beta, \lambda_j) \in U \times M_j$, these assumed conditions imply that the joint ML estimators

$$(\hat{\beta}_j, \hat{\lambda}_j) \equiv \operatorname{argmax} \left\{ \sum_{i=1}^{n_j} \log f_j(\mathbf{x}_{ij}, \beta, \lambda_j) : (\beta, \lambda_j) \in U \times M_j \right\}$$

for (β, λ_j) exist and are consistent and locally uniquely determined as solutions of the score or likelihood equations in the j th sample, and are asymptotically normal in the following sense. Let $\lambda_j \in M_j$ and

$$M_j^0 \equiv \{\mathbf{v}_j - \lambda_j : \mathbf{v}_j \in M_j\}$$

This is a vector space whose dimension $d_j \geq 1$ we have assumed to be finite. Let $(a_l^{(j)} : l = 1, \dots, d_j)$ denote an orthonormal basis of M_j , and define the operator $H_j : M_j^0 \mapsto \mathbb{R}^{d_j}$ by the rule

$$H_j \left(\sum_{l=1}^{d_j} c_l a_l^{(j)} \right) \equiv (c_1, \dots, c_{d_j}) \in \mathbb{R}^{d_j}$$

Then H_j is a vector space isomorphism, and

$$\sqrt{n_j} \begin{pmatrix} \hat{\beta}_j - \beta \\ H_j(\hat{\lambda}_j - \lambda_j) \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \Sigma_j) \quad \text{as } n_j \rightarrow \infty \quad (\text{C1})$$

The variance Σ_j is the inverse of the $(p + d_j) \times (p + d_j)$ Fisher Information matrix $\mathcal{I}^{(j)}(\beta, \lambda_j; M_j)$ for the j th sample and depends on the finite-dimensional parameter space M_j (the *submodel*). The upper-left $p \times p$ block of Σ_j is denoted by $\{\mathcal{I}^{(j)}(\beta; M_j)\}^{-1}$. The infimum in the sense of positive-definite matrix ordering ($K \leq L$ if and only if $L - K$ is nonnegative definite)

$$\mathcal{I}^{(j)}(\beta) \equiv \inf \left\{ \mathcal{I}^{(j)}(\beta; M_j) : M_j \text{ finite-dimensional} \subset \Lambda_j \right\}$$

can be shown to exist (Bickel et al., 1998, pp. 17-21, or Van der Vaart, 1998, p. 115).

As discussed in Section 3, another expression for $\mathcal{I}^{(j)}(\beta; M_j)$ is

$$\mathcal{I}^{(j)}(\beta; M_j) = \frac{1}{n_j} \inf_{\mathbf{v}, K} E \left\{ \sum_{i=1}^{n_j} \left(\nabla_{\beta} \log f_j(\mathbf{x}_{ij}, \beta, \lambda_j) - K \frac{d}{dt} \log f_j(\mathbf{x}_{ij}, \beta, \lambda_j + t\mathbf{v}) \Big|_{t=0} \right)^{\otimes 2} \right\}$$

where \mathbf{v} ranges over all vectors in the d_j dimensional space M_j^o , the $p \times d_j$ matrix K is arbitrary, and we adopt the notation $\mathbf{w}^{\otimes 2} = \mathbf{w} \mathbf{w}^{\text{tr}}$ for any vector \mathbf{w} , and denote by ∇_β the gradient operator. The *inf* in the displayed expression is actually achieved, and leads to the expression

$$\mathcal{I}^{(j)}(\beta) = \frac{1}{n_j} \min_{\mathbf{w}, K} E \left\{ \sum_{i=1}^{n_j} \left(\nabla_\beta \log f_j(\mathbf{x}_{ij}, \beta, \lambda_j) - K \frac{d}{dt} \log f_j(\mathbf{x}_{ij}, \beta, \lambda_j + t\mathbf{w}) \Big|_{t=0} \right)^{\otimes 2} \right\}$$

where \mathbf{w} now ranges over all (sufficiently small) vectors in $\{\mathbf{u} - \lambda_j : \mathbf{u} \in \Lambda_j\}$ differing from 0 in only m coordinates, for some finite m , and K ranges over all $p \times m$ real matrices.

Any efficient regular estimator $\tilde{\beta}_j$ defined from \mathbf{X}_j , i.e., one for which the asymptotic variance of $\sqrt{n_j}(\tilde{\beta}_j - \beta)$ is no larger than $\{\mathcal{I}_j(\beta)\}^{-1}$, differs from any other such estimator by a remainder of order smaller than $n_j^{-1/2}$ in probability (Hájek-LeCam convolution theorem, Van der Vaart, 1998, p. 115). This notion of *semiparametric efficiency* applies under general conditions to the case discussed here where the nuisance parameters λ_j are infinite-dimensional. Semiparametric efficient regular estimators are discussed in Van der Vaart (1998, Section 25.3). Although the information bounds $\mathcal{I}^{(j)}(\beta)$ depend on the nuisance parameters λ_j , we suppress that dependence to keep the notation as simple as possible.

We next define the notion of *combined-sample* Fisher information about β , based on all data (\mathbf{X}_j , $j = 1, \dots, k$). Allowing the possibility that some of the coordinates of the parameter vectors λ_j are shared or constrained across different samples \mathbf{X}_j , let $\underline{\lambda}$ denote a maximal parameter vector consisting of all free parameters among $\{\lambda_j\}_j$, so that all λ_j vectors are well-behaved functions of $\underline{\lambda}$. Then all of the densities can be written $f_j(\mathbf{x}, \beta, \lambda_j) \equiv f_j^*(\mathbf{x}, \beta, \underline{\lambda})$, where f_j^* is smooth in all components of its parameter arguments. By independence of the samples \mathbf{X}_j , and by analogy with the displayed formula above for the separate-sample per-observation Fisher information matrices $\mathcal{I}^{(j)}(\beta)$, there now exists a variationally defined combined-sample information matrix for β ,

$$\mathcal{I}^*(\beta) = \frac{1}{n} \inf_{\mathbf{v}, K} \sum_{j=1}^k \sum_{i=1}^{n_j} E \left\{ \left\{ \nabla_\beta \log f_j^*(\mathbf{x}_{ij}, \beta, \underline{\lambda}) - K \frac{d}{dt} \log f_j^*(\mathbf{x}_{ij}, \beta, \underline{\lambda} + t\mathbf{v}) \Big|_{t=0} \right\}^{\otimes 2} \right\} \quad (\text{C2})$$

where for some finite $m \geq 1$, the vector \mathbf{v} ranges over all subvectors of dimension the same as $\underline{\lambda}$ which differ from $\mathbf{0}$ in at most m places; K ranges over all $p \times m$ real matrices; and the *inf* is in the sense of nonnegative-definite matrix ordering. By these variational considerations, the inequality (13) on *super-additivity of information*, presented in (I) of Section 3 in the finite-dimensional setting, continues to hold in the setting allowing infinite-dimensional λ_j .

We now come to the main result of this Appendix, that when the nuisance parameters λ_j in the separate samples are unrelated, then (13) becomes an equality. The proof arguments are in the spirit of Bickel et al. (1998), Van der Vaart (1998, Chapter 25), and Tsiatis (2006). It is a general phenomenon that independent samples with freely varying nuisance parameters allow combined optimal estimators by simple linear combination of separate-sample optimal estimators.

Proposition A.1: *Under the setting and assumptions above, suppose also that the possibly infinite-dimensional nuisance parameters $\lambda_j \in \Lambda_j$ vary freely, unconstrained by β or by each other. Assume further that as the sample-size n increases, all of the ratios n_j/n have limits $c_j \geq 0$. Then the combined-sample semiparametric per-observation information bound is $\mathcal{I}^*(\beta) = \sum_{j=1}^k c_j \mathcal{I}^{(j)}(\beta)$.*

Proof: We define several Hilbert subspaces of the space $L_p^2(\Omega, \sigma(\{\mathbf{X}_j\}_{j=1}^k), P)$ of p -vector square-integrable random variables defined measurably from the combined samples \mathbf{X}_j , where P is the probability measure corresponding to the true parameter values $\beta, \{\lambda_j\}_{j=1}^k$. First, for $j = 1, \dots, k$, define the closed linear spaces \mathcal{H}_j of (assumed square-integrable) random p -vectors with coordinates spanned by directional derivatives $(d/dt) \log f_j(\mathbf{x}_{ij}, \beta + t\mathbf{b}, \lambda_j + t\mathbf{v})$ evaluated at $t = 0$, where $\mathbf{b} \in \mathbb{R}^p$ and \mathbf{v} is any vector with at most finitely many entries nonzero such that $\lambda_j + t\mathbf{v} \in \Lambda_j$ for all sufficiently small t . Denote by \mathcal{H} the span of all the spaces \mathcal{H}_j , $j = 1, \dots, k$. Define $\mathcal{L}_j \subset \mathcal{H}_j$ to be the closed linear span of the subset of these vectors of directional derivatives for which $\mathbf{b} = \mathbf{0}$, and

$$\mathcal{B} = \left\{ A \nabla_\beta \log f_j(\mathbf{x}_{ij}, \beta, \lambda_j) : p \times p \text{ matrices } A, \right. \\ \left. 1 \leq j \leq k, 1 \leq i \leq n_j \right\}$$

Inner products on the spaces \mathcal{H}_j of p -vectors are: $\langle \xi, \eta \rangle = E(\xi^{\text{tr}} \eta)$. Since all elements of \mathcal{H} have mean 0, the inner product is the sum of componentwise covariances.

Let $\Pi\{\cdot | \mathcal{M}\}$ denote the linear projection within \mathcal{H} onto the closed linear space \mathcal{M} . The separate-sample information bounds $\mathcal{I}^{(j)}(\beta)$ can generally (Bickel et al., 1998; Tsiatis, 2006; Van der Vaart, 1998) be interpreted as the variances of the ‘efficient influence functions’, the projection of the respective p -vector scores $\nabla_\beta \log f_j(\mathbf{x}_{ij}, \beta, \lambda_j)$ onto the orthogonal complement of \mathcal{L}_j within \mathcal{H}_j , denoted \mathcal{L}_j^\perp , or (equivalently, because each sample is independent identically distributed)

$$\mathcal{I}^{(j)}(\beta) = n_j^{-1} E \left(\Pi \left\{ \nabla_\beta \sum_{i=1}^{n_j} \log f_j(\mathbf{x}_{ij}, \beta, \lambda_j) \Big| \mathcal{L}_j^\perp \right\} \right)^{\otimes 2} \\ = n_j^{-1} E \left(\nabla_\beta \sum_{i=1}^{n_j} \log f_j(\mathbf{x}_{ij}, \beta, \lambda_j) - \Pi \left\{ \nabla_\beta \sum_{i=1}^{n_j} \log f_j(\mathbf{x}_{ij}, \beta, \lambda_j) \Big| \mathcal{L}_j \right\} \right)^{\otimes 2}$$

The Hilbert space \mathcal{H} is called the *tangent space* of the *finite-dimensional submodels*. Because of the assumption that all of the parameters λ_j vary freely and unconstrained, \mathcal{H} can be expressed as the direct sum

$$\mathcal{H} = \mathcal{B} \oplus \mathcal{L}_1 \oplus \mathcal{L}_2 \oplus \dots \oplus \mathcal{L}_k$$

Since the subspaces \mathcal{L}_j are mutually orthogonal, being formed from independent random variables for different $j = 1, \dots, k$, for each $i = 1, \dots, n_j$,

$$\Pi\{\nabla_\beta \log f_j(\mathbf{x}_{ij}, \beta, \lambda_j) | \mathcal{L}_j^\perp\} = \\ \Pi\{\nabla_\beta \log f_j(\mathbf{x}_{ij}, \beta, \lambda_j) | (\mathcal{L}_1 \oplus \mathcal{L}_2 \oplus \dots \oplus \mathcal{L}_k)^\perp\}$$

It follows immediately that the projection of $\sum_{j=1}^k \sum_{i=1}^{n_j} \nabla_{\beta} \log f_j(\mathbf{x}_{ij}, \beta, \lambda_j)$ in \mathcal{H} orthogonal to $\mathcal{L}_1 \oplus \cdots \oplus \mathcal{L}_k$, which is the efficient influence function for the combined-sample data problem, is also precisely the same as

$$\sum_{j=1}^k \Pi \left\{ \sum_{i=1}^{n_j} \nabla_{\beta} \log f_j(\mathbf{x}_{ij}, \beta, \lambda_j) \mid \mathcal{L}_j^{\perp} \right\}$$

The norm-squared or variance of this projection is, by mutual independence of \mathbf{X}_j , equal to a sum of variance-covariance

matrices of projections:

$$\begin{aligned} & \sum_{j=1}^k E \left(\Pi \left\{ \sum_{i=1}^{n_j} \nabla_{\beta} \log f_j(\mathbf{x}_{ij}, \beta, \lambda_j) \mid \mathcal{L}_j^{\perp} \right\}^{\otimes 2} \right) \\ &= \sum_{j=1}^k n_j \mathcal{I}^{(j)}(\beta) \end{aligned}$$

Thus $\mathcal{I}^*(\beta)$ is asymptotically equal to $\sum_{j=1}^k c_j \mathcal{I}^{(j)}(\beta)$, completing the proof. \blacksquare