Mathematical Statistics
Stockholm University

A General Statistical Framework for Multistage Designs

Maria Grünewald and Ola Hössjer

Research Report 2010:6

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se

Mathematical Statistics
Stockholm University
Research Report **2010:6**,
http://www.math.su.se

A General Statistical Framework for Multistage Designs

Maria Grünewald and Ola Hössjer

Abstract

The efficiency of observational studies may be increased by applying multistage sampling designs. It is however not always transparent how to construct such a design in order to obtain increased efficiency. We here present a general statistical framework for describing and constructing multistage designs. We also provide tools for efficiency and cost-efficiency comparisons, to facilitate the choice of sampling scheme. The comparisons are based on Fisher information matrices and the results are suggested being presented in graphs, where either efficiency or cost adjusted efficiency is plotted against a normalized measure of cost. The former curve resides in the unit square and is analogous to the receiver operating characteristic curve used for testing.

KEY WORDS: Hierarchical multistage model, Multistage sampling, Efficient design, Cost-efficiency, Fisher information

1 Introduction

Likelihood-based methods, such as the maximum likelihood estimator and likelihood ratio test, exhibit good efficiency under rather weak regularity conditions when the underlying model is correctly specified. However, if the cost of collecting the full sample is large, one may collect a subsample at lower total cost, which, by careful choice of the sampling mechanism, only looses little efficiency compared to the full ML estimator. To choose an effective design it is necessary to be able to calculate, and to compare, the effectiveness of different sampling strategies. The aim of this paper is to provide tools for such comparisons, and to present a framework that is general enough be applicable for a broad range of statistical models and sampling schemes.

To this end, we introduce a general hierarchical multistage model with k stages. In this model, the investigator chooses the sampling probabilities at each stage, for each individual, with Stages k and 1 corresponding to full and minimal information. This is a special case of data coarsening (Heitjan & Rubin 1991), where 1) the coarsening variable $J \in \{1, \ldots, k\}$ is observed and 2) the degree of coarsening is hierarchical.

The term two-stage design was introduced by White (1982). Since then two-stage and multistage designs have been explored for different design settings in various areas of research. For instance, Thomas et al. (2004) investigated two-stage case-control designs within a genetic application - testing of association between Single Nucleotide Polymorphism (SNP) markers and disease status. The two-stage design was motivated by the cost of genotyping. In the first stage it is determined what genetic markers should be used in the larger sample. Analytical design optimization was possible here for special cases but generally simulation based optimization is more practical. Asymptotic relative cost-efficiency (ARCE) was used as a measure of performance.

In the example above the sequential design was motivated by differential costs of collecting data on different variables. This situation may arise for example when registry data are available for some variables, while other variables are more costly to measure. It may then be beneficial to formally incorporate costs in the efficiency calculations. Maydrech & Kupper (1978) incorporate cost when providing sample size requirement calculations for cohort and case-control studies. Different costs for exposed/nonexposed or cases/controls are allowed for in the calculations.

Reilly (1996) investigates optimal allocation of available resources for two-stage data. Stage 1 variables are sampled for all individuals, so that complete information is available for these, whereas Stage 2 variables are sampled more

sparsely for a subset of individuals, with sampling probabilities determined by Stage 1 data. Either precision is maximized for a fixed budget or cost is minimized for a fixed precision. Cost functions are used so that cost can differ between sampling in stage one and sampling in stage two. Examples with different data-sets are presented.

The main focus in this paper is to systematically describe the efficiency-cost tradeoff in multistage sampling. To this end we introduce plots of efficiency and cost-adjusted efficiency as functions of average cost. We use the full sampling scheme ($J \equiv k$) as reference, and thus report efficiency as well as average cost in relative terms. A key parameter is the choice of sampling scheme π, defined as the distribution of J, by which the investigator may control the cost-efficiency tradeoff. In particular, we focus on sequential multistage designs, where individuals enter higher stages sequentially based on already collected data. In this way, more data are collected only from individuals that are predicted informative. Mathematically, this corresponds to coarsening at random (CAR, Heitjan & Rubin 1991), although our focus is on design rather than estimation of π. Our framework can be viewed as a generalization of Grünewald & Hössjer (2010), where two-stage retrospective designs are treated.

In Section 2 the multistage sampling model is described. The cost and efficiency of samples are defined in Section 3 and the choice of cost functional is discussed in Section 4. In Section 5 stage-dependent cost functions are incorporated into sequential multistage designs. Strategies for how to compare efficiencies are presented in Section 6. Analytical solutions for the efficiency calculations are not tractable for all statistical models. Therefore, we outline in Section 7 how Monte Carlo methods can be used for computation. A special case of multi-stage designs is the ascertainment problem, where data are not recorded on units that do not have full data. The ascertainment problem is discussed in Section 8. In Section 9 the general theory of multistage sampling is illustrated with a number of examples. Finally, the main conclusions of the paper are discussed in Section 10, and proofs are collected in the appendix.

2 A Multistage Model

Consider a collection of independent and identically distributed (i.i.d.) random variables $Z^1, \ldots, Z^n$ defined on a sample space $\mathcal{Z}$ with common density $f(z;\theta)$ with respect to an underlying measure on $(\mathcal{Z}, \mathcal{B})$, where $\mathcal{B}$ is the Borel

sigma algebra on $\mathcal{Z}$. The p-dimensional parameter vector $\theta = (\theta_1, \ldots, \theta_p)$, belongs to a parameter space Θ. If estimation of θ is of concern, we may use the maximum likelihood estimator

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} L(\theta),$$

where $L(\theta) = \prod_{i=1}^{n} f(z^i; \theta)$ is the likelihood function and z^i the observed value of Z^i. Instead, we may wish to test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \notin \Theta_0$, given some null parameter set $\Theta_0 \subset \Theta$. Then the log likelihood ratio test statistic

$$T_{\mathrm{LR}} = 2\left(\max_{\theta \in \Theta} \log L(\theta) - \max_{\theta \in \Theta_0} \log L(\theta)\right),$$

can be used, with H_0 rejected if T_{LR} exceeds a given threshold.

Sometimes only part of the data are observed. To this end, we introduce a k-stage sampling model, starting with $\mathcal{Z}_k = \mathcal{Z}$, and then defining a sequence of reduced sampling spaces $\mathcal{Z}_{k-1}, \ldots, \mathcal{Z}_1$. The reduction of complexity from Stage $j+1$ to Stage j is achieved by means of the non-invertible transformation $g_j : \mathcal{Z}_{j+1} \to \mathcal{Z}_j$. Let $G_j = g_j \circ g_{j+1} \circ \ldots g_{k-1}$ denote reduction of information (or coarsening) from Stage k down to Stage j. If $Z \sim f(\cdot; \theta)$ is drawn from the full (Stage k) distribution, let

$$Z_j = G_j(Z), \quad j = 1, \ldots, k$$

be the corresponding Stage j random variable. For a graphical representation of the model see Figure 1.

The sampling mechanism is defined using a discrete random variable $J \in \{1, \ldots, k\}$, controlling which stage to sample from. The resulting sampled random variable

$$\tilde{Z} = Z_J$$

is defined on the combined sample space $\tilde{\mathcal{Z}} = \mathcal{Z}_1 \cup \ldots \cup \mathcal{Z}_k$. The property of the sampling mechanism is determined from the joint distribution of Z and J. Write

$$\pi_j(z) = P(J = j | Z = z)$$

for the probability of collecting information on $z \in \mathcal{Z}$ at Stage j, so that $J|Z = z \in \mathrm{Mult}(1, \pi(z))$ has a multinomial distribution, with parameters 1 and $\pi(z) = (\pi_1(z), \ldots, \pi_k(z))$. Also, let

$$\Pi_j(z) = P(J \geq j | Z = z) = \sum_{l=j}^{k} \pi_l(z)$$

be the probability that z is sampled at least up to Stage j and

$$\lambda_j(z) = P(J \geq j | J \geq j-1, Z = z) = \Pi_j(z)/\Pi_{j-1}(z),$$

$j = 2, \ldots, k$ be one minus the discrete hazard rate of J, i.e. the conditional probability of collecting data from Stage j, given that data have been collected from Stage $j-1$ already. For a two-stage design $\Pi_2(z)$ will simplify to $\sum_{l=2}^{2} \pi_l(z) = \pi_2(z)$, or more generally, for a k-stage design, $\Pi_k(z)$ will simplify to $\sum_{l=k}^{k} \pi_l(z) = \pi_k(z)$. Also, $\lambda_2(z)$ will simplify to $\Pi_2(z)/\Pi_1(z) = \Pi_2(z)$ for all k.

Let J^i be the stage that data on individual i, Z^i, are sampled from. Then $(Z^1, J^1), \ldots, (Z^n, J^n)$ is an i.i.d. sequence of random variables. It gives rise to a cost-reduced sample $\tilde{Z}^1, \ldots, \tilde{Z}^n$, where $\tilde{Z}^i = Z^i_{J^i}$. The corresponding ML estimator and LR tests are

$$\hat{\theta}_{\mathrm{ML}}(\pi) = \arg\max_{\theta \in \Theta} L(\theta, \pi), \tag{1}$$

and

$$T_{\mathrm{LR}}(\pi) = 2\left(\max_{\theta \in \Theta} \log L(\theta, \pi) - \max_{\theta \in \Theta_0} \log L(\theta, \pi)\right), \tag{2}$$

where

$$L(\theta, \pi) = \prod_{i=1}^{n} f(\tilde{z}^i; \theta, \pi) \tag{3}$$

is the likelihood function, $\tilde{z}^i$ the observed value of $\tilde{Z}^i$, $\pi = \{\pi(z);\, z \in \mathcal{Z}\}$ the (possibly infinite-dimensional) sampling parameter and $f(\cdot; \theta, \pi)$ the density of $\tilde{Z}$ on $\tilde{\mathcal{Z}}$. It is defined as

$$f(\tilde{z}; \theta, \pi) = f(z_j; \theta) E\left(\pi_j(Z) | G_j(Z) = z_j\right), \quad \text{if } \tilde{z} = z_j \in \mathcal{Z}_j, \tag{4}$$

where $f(z_j; \theta)$ is the density of Z_j on $\mathcal{Z}_j$.

In (1) and (2), we implicitly assume π to be known. In Section 10, we will discuss relaxation of this assumption. The full sample corresponds to $\pi_k(\cdot) \equiv 1$. We denote this sampling scheme by π_{full}, so that $L(\theta) = L(\theta, \pi_{\mathrm{full}})$. At the other extreme, we let $\pi_{\min}$ denote the design $\pi_1(\cdot) \equiv 1$, yielding a data set with minimal possible amount of information.

3 Cost and Efficiency

Let $C_j(z)$ be the cost of sampling $z \in \mathcal{Z}$ at Stage j. We will assume that

$$0 \leq C_1(z) \leq \ldots \leq C_k(z), \quad \forall z \in \mathcal{Z}, \tag{5}$$

so that cost increases when more information on z is gathered. Then, the total average cost (TAC) of the cost-reduced sample is

$$\mathrm{TAC}(\theta, \pi) = nE(C_J(Z)) = n \sum_{j=1}^{k} \int_{\mathcal{Z}} \pi_j(z) C_j(z) f(z; \theta) dz \tag{6}$$

and the relative average cost (RAC) compared to the full sample

$$\mathrm{RAC}(\theta, \pi) = \mathrm{TAC}(\theta, \pi)/\mathrm{TAC}(\theta, \pi_{\mathrm{full}}).$$

Let

$$\psi(\tilde{z}; \theta, \pi) = \frac{\partial \log f(\tilde{z}; \theta, \pi)}{\partial \theta} \tag{7}$$

be the score function, which is a $1 \times p$ vector-valued function defined on $\tilde{\mathcal{Z}}$. The Fisher information of the whole cost-reduced sample $\{\tilde{Z}^i\}_{i=1}^n$ is

$$I(\theta, \pi) = nE\left(\psi(\tilde{Z}; \theta, \pi)^T \psi(\tilde{Z}; \theta, \pi)\right) \tag{8}$$

where ψ^T is the transpose of ψ.

Let $h(I)$ be a scalar function of I satisfying

$$\begin{array}{l} h(tI) = th(I) \text{ for any } t > 0, \\ h(I_1) \leq h(I_2) \text{ if } I_1 \leq I_2, \end{array} \tag{9}$$

where $I_1 \leq I_2$ means that $I_2 - I_1$ is positive semidefinite. The relative efficiency of the cost-reduced sample compared to the full sample is defined as

$$e(\theta, \pi) = h\left(I(\theta, \pi)\right) / h\left(I(\theta, \pi_{\mathrm{full}})\right). \tag{10}$$

The first part of (9) ensures that e has the usual interpretation in terms of relative sample sizes: Asymptotically, when n is large, a sample of size $n/e(\theta, \pi)$ is needed for design π to attain the same accuracy as a sample of size n using the full design π_{full}.

The cost adjusted efficiency

$$\mathrm{CE}(\theta, \pi) = e(\theta, \pi)/\mathrm{RAC}(\theta, \pi)$$

quantifies the relative efficiency of design π at parameter θ compared to a random sample ($\pi = \pi_{\mathrm{full}}$) exhibiting the same total average cost. It has been used in the context of planning genetic association studies by Thomas et al. (2004).

4 Choice of efficiency functional

If estimating $\hat{\theta}_{\mathrm{ML}}(\pi)$ is of interest, the functional $h(I)$ in (9) is typically a function of the asymptotic covariance matrix $V = I^{-1} = (V_{rs})_{r,s=1}^{p}$. Some examples are $\det(V)^{-1/p}$, $\mathrm{tr}(V)^{-1}$ and V_{rr}^{-1}, see for examlpe Melas (2006).

For testing, other functionals can be used. Assume a simple null hypothesis $\Theta_0 = \{\theta_0\}$, and a true parameter value $\theta_0 + a$. Then asymptotically, in the limit of large samples (large I) and local alternatives (small a), $T_{\mathrm{LR}}(\pi)$ in (2) has a noncentral χ^2 distribution with p degrees of freedom and noncentrality parameter $h(I) = a^T I a$, where $I = I(\theta_0, \pi)$. See for instance Serfling (1980). Hence the power of the LR-test is asymptotically a monotone function of $h(I)$. More generally, let H be a distribution on $\mathbb{R}^p$ satisfying $\int a dH(a) = m$ and $\int a^T a dH(a) = \Sigma$. Define the linear functional

$$h(I) = E_H(aIa^T) = \mathrm{tr}\left((m^T m + \Sigma)I\right), \tag{11}$$

so that $h(I) = a^T I a$ when H has a one-point distribution at a. The more general criterion (11) is a robustified version, which allows for uncertainty of a in terms of a prior distribution H.

If $\theta = (\xi, \gamma)$ can be split into structural parameters ξ and nuisance parameters γ, with parameter space $\Theta = \Xi \times \Gamma$, we write

$$I(\theta, \pi) = \begin{pmatrix} I_{\xi\xi}(\theta, \pi) & I_{\xi\gamma}(\theta, \pi) \\ I_{\gamma\xi}(\theta, \pi) & I_{\gamma\gamma}(\theta, \pi) \end{pmatrix}. \tag{12}$$

The estimation-based functionals h are defined as before. For testing, consider a composite null hypothesis $\Theta_0 = \{\xi_0\} \times \Gamma$. Then, an appropriate functional when $\xi_0 + a$ is the true structural parameter, is $h(I) = a I_{\mathrm{profile}} a^T$, where

$$I_{\mathrm{profile}}(\theta, \pi) = I_{\xi\xi} - I_{\xi\gamma} I_{\gamma\gamma}^{-1} I_{\gamma\xi}.$$

is the profile likelihood Fisher information. More generally, $h(I) = \tilde{h}(I_{\mathrm{profile}})$ can be used for linear $\tilde{h}$, as in (11).

More details on efficiency functionals in the context of experimental design can be found in Silvey (1980) and Melas (2006).

5 Sequential Multistage Designs and Stage-Dependent Cost Functions

Consider a fixed observation $Z = Z^i$ and $J = J^i$. It will be helpful to introduce a measurable space $(\Omega, \mathcal{F})$ and think of $Z : \Omega \to \mathcal{Z}$ and $J : \Omega \to \{1, \ldots, k\}$ as random variables measurable with respect to the sigma algebra $\mathcal{F}$. Let $\mathcal{F}_j$ be the σ-algebra on Ω generated by Z_j, so that $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}_k \subset \mathcal{F}$ defines a filtration. We will mainly restrict ourselves to designs such that J is a stopping time with respect to this filtration. That is, we require $\{J = j\} \in \mathcal{F}_j$ for $j = 1, \ldots, k$. Less formally, we will write this as

$$\pi_j(z) = \pi_j(z_j), \quad z \in \mathcal{Z}, j = 1, \ldots, K, \tag{13}$$

which is short for $\pi_j(\cdot)$ being constant on $G_j^{-1}(z_j) = \{z \in \mathcal{Z};\, G_j(z) = z_j\}$. Condition (13) is equivalent to CAR and means that the probability of including information from z up to but not exceeding Stage j should not depend on the outcome of z at stages above j. If (13) was not satisfied we would need information above Stage j in order to decide whether or not to sample Z at this level, which would imply loss of sampled information.

We will refer to all designs satisfying (13) as the class of *sequential multistage designs* $\mathcal{P}$. The name can be motivated as follows: Recall $\Pi_j(z)$, the probability that z is sampled at least up to Stage j. It is easy to see that (13) is equivalent to

$$\Pi_j(z) = \Pi_j(z_{j-1}), \quad z \in \mathcal{Z}, j = 1, \ldots, K, \tag{14}$$

which naturally corresponds to collecting more and more data, starting from Stage 1 and proceeding sequentially up to Stage k. Indeed (14) implies

$$\lambda_j(z) = \Pi_j(z)/\Pi_{j-1}(z) = \Pi_j(z_{j-1})/\Pi_{j-1}(z_{j-2}) = \lambda_j(z_{j-1}), \tag{15}$$

so that the probability of collecting more data only depends on data already present. For $k > 2$ transition between the parameters π (or Π) and λ requires knowledge of Z, so in real data collection it is convenient to use λ, which corresponds directly to the sampling procedure.

With condition (13), (4) simplifies to

$$f(\tilde{z}; \theta, \pi) = \pi_j(z_j) f(z_j; \theta), \quad \text{if } \tilde{z} = z_j \in \mathcal{Z}_j. \tag{16}$$

Inserting (16) into (7) and (8), we get

$$I(\theta,\pi) \;\; = \;\; \textstyle\sum_{j=1}^{k} I_j(\theta,\pi) \tag{17}$$

with

$$I_j(\theta,\pi) = n\int_{\mathcal{Z}_j} \pi_j(z_j)\psi(z_j;\theta)^T\psi(z_j;\theta)f(z_j;\theta)dz_j$$

the part of the total information obtained from individuals sampled at (but not above) Stage j, and

$$\psi(z_j;\theta) = \partial \log f(z_j;\theta)/\partial\theta = E(\psi(Z;\theta)|Z_j = z_j) \tag{18}$$

the score function for Stage j data. Notice in particular that (17) depends linearly on the design distribution π. Alternatively, we may write the Fisher information as

$$I(\theta,\pi) = I(\theta,\pi_{\min}) + \sum_{j=2}^{k} I_{j|j-1}(\theta,\pi), \tag{19}$$

where

$$I_{j|j-1}(\theta,\pi) = n\int_{\mathcal{Z}_{j-1}} \Pi_j(z_{j-1})\mathrm{Cov}(\psi(Z_j;\theta)|Z_{j-1} = z_{j-1})f(z_{j-1};\theta)dz_{j-1}$$

is that part of the total information from Stage j data that are not present at Stage $j-1$. See the appendix for a derivation of (17) and (19).

For cost functions, a simplification similar to (13) is possible. Indeed, most cost functions of practical interest will satisfy

$$C_j(z) = C_j(z_j), \quad z \in \mathcal{Z}, j = 1,\ldots,K, \tag{20}$$

which means that C_j is $\mathcal{F}_j$-measurable. When (20) holds, the cost of gathering information at Stage j does not depend on information from stages above j that is not present. We will refer to (20) as a *stage-dependent* cost function. With (13) and (20), the total average cost (6) becomes

$$\mathrm{TAC}(\theta,\pi) = n\sum_{j=1}^{k}\int_{\mathcal{Z}_j} \pi_j(z_j)C_j(z_j)f(z_j;\theta)dz_j. \tag{21}$$

6 Cost-Efficiency Plots and Optimal Designs

The tradeoff between cost and efficiency at a given parameter θ can be illustrated by varying π and plotting $e(\theta, \pi)$ as a function of $\mathrm{RAC}(\theta, \pi)$, see Grünewald & Hössjer (2010). Some simple but useful properties of such cost-efficiency plots are summarized in the following proposition:

Proposition 1 *Consider a fixed $\theta \in \Theta$ and functional h satisfying (9). Let π, π' be two designs with $\pi \leq \pi'$, i.e. $\Pi_j(z) \leq \Pi'_j(z)$ for all $z \in \mathcal{Z}$ and $j = 1, \ldots, k$. Then*

$$\begin{aligned} &0 \leq RAC(\theta, \pi_{min}) \leq RAC(\theta, \pi) \leq RAC(\theta, \pi') \leq RAC(\theta, \pi_{full}) = 1, \\ &0 \leq e(\theta, \pi_{min}) \leq e(\theta, \pi) \leq e(\theta, \pi') \leq e(\theta, \pi_{full}) = 1, \end{aligned} \tag{22}$$

with equalities $0 = RAC(\theta, \pi_{min})$ and $0 = e(\theta, \pi_{min})$ on the left hand sides of (22) if Stage 1 corresponds to no cost ($C_1(\cdot) \equiv 0$) and no information ($I(\theta, \pi_{min}) = 0$) respectively.

It follows from Proposition 1 that each design π corresponds to a point in the unit square $[0, 1] \times [0, 1]$ of the (RAC, e)-plane, with $(1, 1)$ for the full design and $(0, 0)$ for the minimal design if Stage 1 has zero cost and no information.

An optimal design is one that given θ maximizes $e(\theta, \pi)$ subject to a constraint on $\mathrm{RAC}(\theta, \pi)$. It will typically be a locally optimal design, i.e. depend on θ, which is unknown (in fact, it is the quantity we wish to estimate). In practice, we may use training data to compute a preliminary estimate of θ, which is used as plug-in for the optimal design.

We consider a finite-dimensional subclass

$$\mathcal{Q} = \{\pi \in \mathcal{P};\ \pi(\cdot) = \pi(\cdot; \eta) \text{ for some } \eta\} \tag{23}$$

of all sequential multistage designs $\mathcal{P}$, parameterized by $\eta = (\eta_1, \ldots, \eta_r)$. Keep $\theta \in \Theta$ fixed and let $\mathcal{Q}_R = \{\pi \in \mathcal{Q};\ \mathrm{RAC}(\theta, \pi) \leq R\}$ be the class of sequential multistage designs in $\mathcal{Q}$ with relative average cost not exceeding R. The $\mathcal{Q}$-optimal efficiency function is defined as the non-decreasing function

$$R \to e_{\max}(\theta, R) = \sup_{\pi \in \mathcal{Q}_R} e(\theta, \pi). \tag{24}$$

Write $\pi(\cdot; R)$ for the design attaining the maximum in (24). We will refer to it as a $\mathcal{Q}$-optimal design. When $\mathcal{Q} = \mathcal{P}$, we omit $\mathcal{Q}$ and simply speak of

optimal designs. Define

$$\begin{aligned} R_{\min} &= \min_{\pi\in\mathcal{Q}} \mathrm{RAC}(\theta,\pi), \\ R_{\max} &= \max_{\pi\in\mathcal{Q}} \mathrm{RAC}(\theta,\pi), \\ e_{\min} &= \min_{\pi\in\mathcal{Q}} e(\theta,\pi), \\ e_{\max} &= \max_{\pi\in\mathcal{Q}} e(\theta,\pi). \end{aligned}$$

Proposition 2 *Assume h satisfies (9) and that $\mathcal{Q}$ is convex with $\pi_{full} \in \mathcal{Q}$. Then, given any $\theta \in \Theta$, the optimal efficiency curve (24) satisfies*

$$e_{max}(\theta, R) = \sup_{\pi\in\mathcal{Q}; RAC(\theta,\pi)=R} e(\theta,\pi), \tag{25}$$

for any $R \in (R_{min}, 1]$, and hence the maximal cost efficiency satisfies

$$CE_{max}(\theta, R) := \sup_{\{\pi\in\mathcal{Q}; RAC(\theta,\pi)=R\}} CE(\theta,\pi) = e_{max}(\theta, R)/R.$$

It follows immediately from its definition that $\mathcal{P}$ is convex, so that Proposition 2 applies with $\mathcal{Q} = \mathcal{P}$. It is easy to see that the full design π_{full} is optimal, with $(1, 1)$ the right-hand end point of the optimal efficiency curve. The minimal design $\pi_{\min}$ is optimal as well if we have strict inequalities in (5) and then $(R_{\min}, e_{\min})$ is the left-hand end point of the optimal efficiency curve.

In the following two propositions, some more properties of the $\mathcal{Q}$-optimal efficiency and cost-efficiency curves are stated. We will consider efficiency functionals h with the property

$$h((1-t)I_1 + tI_2) \geq (1-t)h(I_1) + th(I_2), \tag{26}$$

for $0 \leq t \leq 1$ and any positive semidefinite I_1 and I_2. For instance, the linear functional (11) satisfies (26), but not, in general, functionals that operate directly on the covariance matrix V.

Proposition 3 *In addition to the conditions of Proposition 2, suppose that h satisfies (26). Then the $\mathcal{Q}$-optimal efficiency curve (24) is concave for any $\theta \in \Theta$.*

Proposition 4 *Assume the regularity conditions of Proposition 2 hold. Then, the $\mathcal{Q}$-optimal cost-efficiency curve $R \to CE_{max}(\theta, R)$ is non-increasing if either*

$$\begin{aligned} R_{min} &= 0, \\ I(\theta,\pi) &= 0, \forall \pi \in \mathcal{Q}_0, \end{aligned} \tag{27}$$

or if $R_{min} > 0$*,* h *satisfies (26) and*

$$e'_{max}(\theta, R_{min}) \leq e_{max}(\theta, R_{min})/R_{min} \tag{28}$$

where $e'_{max}(\theta, R_{min}) = \partial e_{max}(\theta, R)/\partial R$*.*

7 Computation

For simple models with low-dimensional $\mathcal{Z}_j$, we may compute cost and efficiency directly from (17) and (21). Alternatively, for more complex models, we may use Monte Carlo. Generate an i.i.d. sample $\{Z^i\}_{i=1}^N$ from $f(\cdot;\theta)$ and estimate

$$\begin{aligned} \widehat{\mathrm{TAC}}(\theta,\pi) &= nN^{-1}\textstyle\sum_{i=1}^N \sum_{j=1}^k \pi_j(Z_j^i) C_j(Z_j^i), \\ \widehat{\mathrm{RAC}}(\theta,\pi) &= \widehat{\mathrm{TAC}}(\theta,\pi)/\widehat{\mathrm{TAC}}(\theta,\pi_{\mathrm{full}}), \\ \hat{I}(\theta,\pi) &= nN^{-1}\textstyle\sum_{i=1}^N \sum_{j=1}^k \pi_j(Z_j^i) \psi(Z_j^i;\theta)^T \psi(Z_j^i;\theta), \\ \hat{e}(\theta,\pi) &= h(\hat{I}(\theta,\pi))/h(\hat{I}(\theta,\pi_{\mathrm{full}})), \end{aligned} \tag{29}$$

where $Z_j^i = G_j(Z^i)$. Since $\psi(Z_j^i;\theta)$ is defined by means of an integral when $j < k$ (cf. (18)), we may need to approximate it by a Monte Carlo estimate $\hat{\psi}(Z_j^i;\theta)$. When $\mathcal{Z}_j$ is finite, we put

$$\hat{\psi}(z_j;\theta) = \frac{1}{N_{z_j}} \sum_{i;Z_j^i = z_j} \psi(Z^i;\theta),$$

where $N_{z_j} = |\{i; Z_j^i = z_j\}|$. When $\mathcal{Z}_j$ is continuous, more refined methods based nonparametric regression (e.g. local polynomial kernel regression) can be used.

If $I(\theta,\pi)$ and $e(\theta,\pi)$ are to be computed for several parameter vectors θ, a new sample $\{Z^i\}_{i=1}^N$ has to be generated for each new θ. Alternatively, we may use importance sampling (Hammersley & Handscomb 1964) based on *one* sample $\{Z^i\}_{i=1}^N$ from $f(\cdot;\theta')$, with

$$\begin{aligned} \widehat{\mathrm{TAC}}(\theta,\pi) &= nN^{-1}\textstyle\sum_{i=1}^N \sum_{j=1}^k \pi_j(Z_j^i) C_j(Z_j^i) w(Z^i;\theta), \\ \hat{I}(\theta,\pi) &= nN^{-1}\textstyle\sum_{i=1}^N \sum_{j=1}^k, \pi_j(Z_j^i) \psi(Z_j^i;\theta)^T \psi(Z_j^i;\theta) w(Z^i;\theta), \end{aligned} \tag{30}$$

and weight function $w(z;\theta) = f(z;\theta)/f(z;\theta')$. Each term $\psi(Z_j^i;\theta)$ with $j < k$ may be replaced by an estimate $\hat{\psi}(Z_j^i;\theta)$. For instance, when $\mathcal{Z}_j$ is discrete,

$$\hat{\psi}(z_j) = \sum_{i;Z_j^i=z_j} \psi(Z_j^i;\theta)w(Z_j^i) / \sum_{i;Z_j^i=z_j} w(Z_j^i).$$

The advantage of importance sampling is that that $\{Z^i\}$ can be reused for several θ. On the other hand, the accuracy may be poor if the candidate parameter θ' is far away from θ (Hesterberg 1995).

Approximate $\mathcal{Q}$-optimal designs may be found using (29) or (30) and maximizing $\hat{e}(\theta,\cdot)$ over $\hat{\mathcal{Q}}_R = \{\pi \in \mathcal{Q};\ \widehat{\mathrm{RAC}}(\theta,\pi) \le R\}$.

8 Ascertainment

In Grünewald & Hössjer (2010), ascertainment is viewed as a two-stage problem, where only data from Stage 2 is observed. Various methods of parameter estimation for ascertained data is discussed in Grünewald et al. (2010) and references therein. More generally, ascertainment can be viewed as originating from a multistage design: Suppose that in a k-stage sample, only data from Stage k are observed and used in the analysis. That is, Stage $1,\ldots,k-1$ data are not recorded at all. Such sampling is often referred to as ascertainment when $k = 2$, although we use the word more generally here for $k \ge 2$. A $k = 2$ example of ascertainment is the case-control design, which is frequently used i epidemiology. In the case-control design, selection is typically based on a dichotomous disease status variable, and exposure variables are measured for selected individuals. Often there is not a well defined sampling frame, but patients are instead recruited at clinics.

For ascertainment (13) does not hold in general, and it does thus not fulfill the conditions for coarsening at random. It is well known that the ascertainment procedure must be incorporated into the likelihood in order to avoid inconsistent estimators (Fisher 1934, Rao 1965).

8.1 Unconditional Ascertainment

If it is known which observations that are not ascertained (although their values are unknown), we modify the likelihood (3) to an unconditional ascertainment likelihood

$$L_{\mathrm{asc}}(\theta,\pi) = \prod_{i=1}^{n} f_{\mathrm{asc}}(z^i;\theta,\pi), \tag{31}$$

where

$$f_{\text{asc}}(\tilde{z};\theta,\pi) = \begin{cases} \pi_k(z)f(z;\theta); & \text{if } \tilde{z} \in \mathcal{Z}_k, \\ 1 - P(A|\theta,\pi); & \text{if } \tilde{z} \in \mathcal{Z}_1 \cup \ldots \mathcal{Z}_{k-1}, \end{cases}$$

where

$$A = \{J = k\}$$

denotes the event that data are ascertained and $P(A|\theta,\pi) = P(J = k|\theta)$ is the ascertainment probability.

The total average cost becomes

$$TAC_{\text{asc}}(\theta,\pi) = nP(A|\theta,\pi)E(C_J(Z)|J = k)$$

and the Fisher information (8) changes to

$$I_{\text{asc}}(\theta,\pi) = I_k(\theta,\pi) + nP'(A|\theta,\pi)^T P'(A|\theta,\pi)/(1 - P(A|\theta,\pi)),$$

where I_k is the information from individuals sampled at Stage k and $P'(A|\theta,\pi) = \partial P(A|\theta,\pi)/\partial\theta$.

Efficiency is calculated as $e_{\text{asc}}(\theta,\pi) = h(I_{\text{asc}}(\theta,\pi))/h(I_{\text{asc}}(\theta,\pi_{\text{full}}))$. Since $I_{\text{asc}}(\theta,\pi) \leq I(\theta,\pi)$, it follows from (9) that $e_{\text{asc}}(\theta,\pi) \leq e(\theta,\pi)$.

8.2 Conditional Ascertainment

More commonly, it is not known which observations that are not ascertained, and then also n is unknown. We then condition on ascertainment status and use the likelihood

$$L_{\text{condasc}}(\theta,\pi) \;=\; \textstyle\prod_{i=1}^{n} f_{\text{condasc}}(\tilde{z}^i;\theta,\pi)$$

where

$$f_{\text{condasc}}(\tilde{z};\theta,\pi) = \begin{cases} f_{\tilde{Z}|J=k}(\tilde{z};\theta,\pi) = \pi_k(z)f(z;\theta)/P(A|\theta,\pi), & \text{if } \tilde{z} \in \mathcal{Z}_k, \\ 1, & \text{if } \tilde{z} \in \mathcal{Z}_1 \cup \ldots \cup \mathcal{Z}_{k-1}, \end{cases}$$

so that

$$\begin{aligned} L_{\text{condasc}}(\theta,\pi) &= \textstyle\prod_{i;\, J^i=k} f_{\text{condasc}}(\tilde{z}^i;\theta,\pi) \\ &= \textstyle\prod_{i;\, J^i=k} \pi_k(z^i)f(z^i;\theta)/P(A|\theta,\pi) \\ &\propto \frac{\prod_{i;\, J^i=k} f(z^i;\theta)}{P(A|\theta,\pi)^{n^{asc}}}, \end{aligned}$$

where $n^{asc} = |\{i; 1 \leq i \leq n, J^i = k\}|$. The expected value of n^{asc} is $E(n^{asc}) = nP(A|\theta,\pi)$.

The resulting total average cost

$$\mathrm{TAC}_{\mathrm{condasc}}(\theta,\pi) = nP(A|\theta,\pi)E(C_J(Z)|J=k),$$

is the same as for unconditional ascertainment, but the Fisher information (8) changes to

$$\begin{aligned} I_{\mathrm{condasc}}(\theta,\pi) &= nP(A|\theta,\pi)\mathrm{Cov}\left(\psi(\tilde{Z};\theta,\pi)\right) \\ &= I_k(\theta,\pi) - nP'(A|\theta,\pi)^T P'(A|\theta,\pi)/P(A|\theta,\pi), \end{aligned}$$

where $\mathrm{Cov}(\psi(\tilde{Z}))$ is the $p \times p$ covariance matrix of $\psi(\tilde{Z})$.

Efficiency is calculated as $e_{\mathrm{condasc}}(\theta,\pi) = h(I_{\mathrm{condasc}}(\theta,\pi))/h(I_{\mathrm{condasc}}(\theta,\pi_{\mathrm{full}}))$. Since $I_{\mathrm{condasc}}(\theta,\pi) \leq I_{\mathrm{asc}}(\theta,\pi)$, it follows from (9) that $e_{\mathrm{condasc}}(\theta,\pi) \leq e_{\mathrm{asc}}(\theta,\pi)$.

9 Examples

In this section, we illustrate the general theory by giving a number of examples. The calculations are run in the software R (R Development Core Team 2008).

Example 1 (Two-stage designs and missing data.) If

$$\begin{aligned} C_1(\cdot) &\equiv 0 \\ C_2(\cdot) &\equiv 1, \end{aligned} \tag{32}$$

we get

$$\mathrm{RAC}(\theta,\pi) = P(J=2) = \int_{\mathcal{Z}} \pi_2(z) f(z;\theta) dz =: P(A|\theta,\pi). \tag{33}$$

In the simplest case when no variables are collected in Stage 1, $\mathcal{Z}_1 = \emptyset$, data are either completely missing or completely observed. The likelihood (3) is then very similar to the unconditional likelihood (31). Indeed, the density and score functions of the sampled random variable $\tilde{Z}$ are then

$$f(\tilde{z};\theta,\pi) = \begin{cases} \pi_2(z) f(z;\theta), & \tilde{z} = z, \\ 1 - P(A|\theta,\pi), & \tilde{z} = \emptyset, \end{cases}$$

and

$$\psi(\tilde{z};\theta,\pi) = \begin{cases} \psi(z;\theta), & \tilde{z} = z, \\ -P'(A|\theta,\pi)/(1-P(A|\theta,\pi)), & \tilde{z} = \emptyset, \end{cases}$$

respectively. Inserting the last two equations into (8) we get

$$\begin{aligned} I(\theta,\pi) &= n\Big(\textstyle\int_{\mathcal{Z}} \pi_2(z)\psi(z;\theta)^T\psi(z;\theta)f(z;\theta)dz \\ &+ P'(A|\theta,\pi)^T P'(A|\theta,\pi)/(1-P(A|\theta,\pi))\Big). \end{aligned} \tag{34}$$

When $k = 2$, we can characterize $\pi(\cdot)$ by the single function $\pi_2(\cdot) = 1 - \pi_1(\cdot)$. For a sequential design (13) with $k = 2$, we further have

$$\pi_2(z) = \Pi_2(z_1) = \pi_2(z_1), \tag{35}$$

so that the inclusion probability at Stage 2 is a function of the outcome at Stage 1 only. Viewing all un-sampled data from Stage 2 as missing, we find that (35) corresponds to data missing at random (MAR), which is a special case of CAR (Rubin 1976, Little & Rubin 2002). With completely missing observations, $\mathcal{Z}_1 = \emptyset$, (35) simplifies to

$$\lambda_2(\cdot) = \pi_2(\cdot) \equiv \eta \tag{36}$$

for some $0 \leq \eta \leq 1$. Hence the class of sequential designs is one-dimensional for completely missing data, with the resulting sample resembling a simple random sample (SRS). Condition (36) is referred to as data missing completely at random (MCAR), cf. Little & Rubin (2002). Then (33)-(34) imply that $(RAC(\theta, \pi), e(\theta, \pi)) = (\eta, \eta)$ is located along the diagonal in a cost-efficiency plot. □

Example 2 (Design of 'x-random' experiments.) Let $z = (x, y)$, where x is a set of covariates and y the response. Put $\theta = (\gamma, \xi)$ and

$$f(z; \theta) = P_\gamma(x) P_\xi(y|x) \tag{37}$$

where γ contains nuisance parameters involved in the covariate distribution, and ξ are the regression parameters. Consider a two stage design

$$\begin{array}{rcl} z_1 & = & x, \\ z_2 & = & (x, y), \end{array} \tag{38}$$

so that $f(z_1; \theta) = P_\gamma(x)$. According to (35), a sequential design satisfies

$$\lambda_2(z) = \pi_2(z) = \pi_2(x).$$

Suppose cost function (32) is used, so that

$$\mathrm{RAC}(\theta, \pi) = \int \pi_2(x) P_\gamma(x) dx$$

equals the probability that y is collected for a randomly chosen x, cf. (33). Finding an optimal design $\pi_2(\cdot)$ amounts to deciding for which x to include response information. That is, for which fraction R of the x-variables $\{x^i\}_{i=1}^n$ we should collect responses y^i in order maximize efficiency when estimating

or testing ξ. This is very similar in spirit to optimal design, as described in Silvey (1980) and Melas (2006), although our framework is for random x-variables. Since x is ancillary for estimating ξ, the likelihood function can factorized as

$$L(\theta, \pi) \propto \prod_{i=1}^{n} P_\gamma(x^i) \cdot \prod_{i;J^i=2} P_\xi(y^i|x^i), \tag{39}$$

with the proportionality constant depending on π and data, but not on θ. This implies that only the last term of (39) is important for estimating ξ and $I_{\gamma\xi}(\theta, \pi) = I_{\xi\gamma}(\theta, \pi) = 0$ in (12). Hence, any functional $h(I(\theta, \pi))$ that involves estimation or testing of ξ will be a function of $I_{\xi\xi}(\theta, \pi)$ only. The first term of (39) is the marginal likelihood of covariate data and the second term the prospective likelihood of response data given covariate data. It gives rise to a decomposition (19) of the Fisher information matrix, given by

$$I(\theta, \pi_{\min}) = \begin{pmatrix} 0 & 0 \\ 0 & I_{\gamma\gamma}(\theta, \pi_{\min}) \end{pmatrix}, \quad I_{2|1}(\theta, \pi) = \begin{pmatrix} I_{\xi\xi}(\theta, \pi) & 0 \\ 0 & 0 \end{pmatrix},$$

so that 1) the design π has no effect on estimation of γ and 2) all information about the effect parameters ξ is contained in the prospective likelihood. □

Example 3 (Binary response-selective sampling.) We retain model (37), but consider the two-stage design

$$\begin{array}{rcl} z_1 & = & y, \\ z_2 & = & (x, y), \end{array} \tag{40}$$

so that $f(z_1; \theta) = \int P_\xi(y|x) P_\gamma(x) dx$. According to (35), a sequential multi-stage design must satisfy

$$\pi_2(z) = \pi_2(y).$$

Two important differences of (38) and (40) are that the x-variables are no longer ancillary for estimating ξ, and the first stage sample is informative for estimating ξ in (40). As an effect, the likelihood no longer factorizes as in (39).

For the logistic regression model the response is binary, with

$$P_\xi(y|x) = F(\alpha + \beta x^T)^{\{y=1\}} (1 - F(\alpha + \beta x^T))^{\{y=0\}},$$

where $\xi = (\alpha, \beta)$ consists of one intercept parameter α, a number of slope parameters β and $F(x) = \exp(x)/(1 + \exp(x))$. Putting $\mathcal{P} = \mathcal{Q}$ in (23), any

sequential multistage design is parametrized by $\eta = (\eta_0, \eta_1)$, where $\eta_m = \pi_2(m) = P(J = 2|Y = m)$.

In Figure 2 RAC and efficiency for a logistic regression model with $X \sim \text{Bin}(1, F(\gamma))$ is illustrated. Three different costs are included, $(C_1, C_2) = (0, 1)$, $(1/3, 1)$ and $(1/2, 1)$, representing scenarios where the cost per individual of collecting Y is none, half the cost of collecting X, and the same cost as collecting X, respectively. As x-axis η_0 is chosen, since this measure easily translates into sampling probabilities when planning a study. Other measures can however be used as x-axis, for example the total sampling probability, $P(J = 2|\pi, \theta) = \eta_0 P(Y = 0|\theta) + \eta_1 P(Y = 1|\theta)$, or the relative average cost, RAC. The graphs show an efficiency gain of response selective sampling in the example setting. However, the relative efficiency differs greatly between parameters, and with the values of C_1 and C_2.

□

Example 4 Continuous response selective sampling, with a genetic application.

For continuous y, $\mathcal{P}$ is infinite-dimensional, so we consider $\mathcal{Q} \subset \mathcal{P}$, parameterized by $\eta = (\eta_1, \ldots, \eta_r)$, corresponding to step functions

$$\lambda_2(y) = \pi_2(y) = \sum_{i=1}^{q} \eta_i 1_{\{y \in \mathcal{Y}_i\}}, \tag{41}$$

where $0 \leq \eta_i \leq 1$, $i = 1, \ldots, q$ and $\{\mathcal{Y}_i\}_{i=1}^q$ is a decomposition of the response region into q mutually disjoint subsets. $\eta_{q+1}, \ldots, \eta_r$ are parameters of this decomposition.

Lyon et al. (2007) investigate the association between the C/C genotype of genetic marker rs7566605 and Body Mass Index (BMI) in a number of cohorts. One of these samples (Maywood) was enriched for obese individuals (BMI$\geq$ 30), while four other cohorts (FHS 1, Iceland, KORA S3 and Scandinavia) were not. We investigate if over-sampling of obese individuals would have been cost efficient in these four samples, if analyzing the data as two-stage samples with continuous response variables $Y = \text{BMI}$ and binary covariate $X = 1_{\{\text{genotype}=CC\}}$. We assume that

$$X \sim \text{Bin}(1, F(\gamma)),$$

where F is the logistic distribution function, and a linear regression model

$$\log(Y)|X = x \sim N(\alpha + \beta x, \sigma^2),$$

for the logarithm of BMI. This yields regression parameters $\xi = (\alpha, \beta, \sigma)$. We use BMI=30 as cutoff value, giving $q = 2$ regions in (41), with $\mathcal{Y}_1 = (0, 30)$ and $\mathcal{Y}_2 = [30, \infty)$. Since these regions are fixed, we have $r = q = 2$, $\eta = (\eta_1, \eta_2)$, with $\eta_2 = 1$ kept fixed and $0 < \eta_1 < 1$ varying. Parameter values are calculated as indicated in Table 1. The calculations were adjusted for age and sex in Lyon et al. (2007) while we did not make this adjustment. In Figure 3 the cost adjusted efficiency is illustrated for $(C_1, C_2) = (0, 1)$, $(1/3, 1)$ and $(1/2, 1)$, representing scenarios where the cost per individual of collecting BMI is none, half the cost of genotyping, and the same cost as genotyping, respectively. As can be seen in Figure 3, using $\eta_1 < 1$ (over-sampling obese subjects for genotyping) is only beneficial if the cost of measuring BMI is substantially lower than the cost of genotyping. Also, over-sampling obese subjects was efficient only in three of the four cohorts, while in the Island cohort $\eta_1 = 1$ was most efficient.

□

Example 5 (Sequential inclusion of covariates.) We retain (37), but consider a k-stage design

$$z_j = (x_1, ..., x_{j-1}, y), \quad j = 1, \ldots, k,$$

where z_j contains an increasing number of covariates from x as j increases. Put

$$f(z_j; \theta) = \int P_\xi(y|x) P_\gamma(x|x_1, ..., x_{j-1}) dx \, P_\gamma(x_1, ..., x_{j-1}), \quad j = 1, \ldots, k. \tag{42}$$

A sequential multi-stage design satisfies

$$\pi_j(z) = \pi_j(x_1, ..., x_{j-1}, y), \quad j = 1, \ldots, k.$$

In particular, for $k = 2$ this means $\pi_2(z) = \pi_2(x_1, y)$, i.e. the sampling probability depends on (x_1, y).

As an illustration we consider a three-stage design, as follows: In the first step of the design Y is collected for the whole sample. In the second step a proportion of X_1 is selected, with selection probabilities determined by the value of Y. A cut-off value t is used, letting

$$\lambda_2(y) = \begin{cases} a; & y < t_Y, \\ 1; & y \geq t_Y, \end{cases}$$

and vary the value of a. In the third step X_2 is collected, with selection probabilities determined by the values of Y and X_1 simultaneously. For individuals with X_1 observed selection probabilities are

$$\lambda_3(y, x_1) = \begin{cases} b; \ y < t_Y, x_1 < t_{x_1} \\ 1; \ \text{else}, \end{cases}$$

while for individuals with X_1 is missing, X_2 is not eligible for selection. For this example we further specify dependencies between the variables, as illustrated in Figure 4, and distributions

$$\begin{aligned} X_2 &\sim \text{Bin}(1, \frac{\exp(\alpha_{X_2})}{1+\exp(\alpha_{X_2})}), \\ X_1|\{X_2 = x_2\} &\sim \text{Bin}(1, \frac{\exp(\alpha_{X_1} + \beta_{X_2X_1} \times x_2)}{1+\exp(\alpha_{X_1} + \beta_{X_2X_1} \times x_2))}), \\ Y|\{X_1 = x_1, X_2 = x_2\} &\sim \text{N}(\alpha_Y + \beta_{X_1Y} \times x_1 + \beta_{X_2Y} \times x_2, \sigma_Y^2). \end{aligned}$$

The resulting likelihood contains seven model parameters of potential interest for estimation $\theta = (\alpha_{X_2}, \alpha_{X_1}, \beta_{X_2X_1}, \alpha_Y, \beta_{X_1Y}, \beta_{X_2Y}, \sigma_Y)$. To simplify the presentation we focus on the three effect parameters $\beta_{X_2X_1}$, β_{X_1Y} and β_{X_2Y}, and investigate the efficiency in estimating these parameters. The efficiency can be assessed for each parameter individually, as well as for all three parameters simultaneously. As a summary measure of the efficiency of the three effect parameters

$$h(I(\pi)) = (\sum_{i=3,5,6} \frac{V_{ii}}{V_{ii}(\pi_{\text{full}})})^{-1} \tag{43}$$

is used. This measure assumes equal interest in the three parameters, and disregards the efficiency in estimating the four remaining parameters. Note that when computing $e = h(I(\pi))/h(I(\pi_{\text{full}})$ the denominator is reduced to a constant. To visualize the results of the three-stage design three-dimensional plots are used in Figure 5 with two different plot designs. In the upper graphs a surface is representing the cost efficiency. The same information is visualized in the graphs below, here instead projecting the height of the surface on a two-dimensional grid, letting a color gradient represent the cost efficiency. Cost efficiency is presented both for each effect parameter individually (green) and for the combined efficiency (red) defined through (43). Two cost functions are included in Figure 5, one with cost associated only with sampling X_2 and the second with equal cost of sampling the variables X_2 and X_1, that is (C_1, C_2, C_3) are $(0, 0, 1)$ and $(0, 1, 2)$ respectively.

□

Example 6 (Ascertainment sample.) As an illustration of efficiency of conditional ascertainment, $e_{\text{condasc}}(\theta, \pi)$ is calculated for the logistic regression model introduced in Example 3. The results are presented in Figure 6, together with the two-stage design efficiency, $e(\theta, \pi)$, for comparison. A two-stage design is preferable to an ascertainment design if collecting the Stage 1 data are not associated with extra cost, i.e. $C_1 = 0$, since more information is available in the two-stage data set. In this example the efficiency gain is however most prominent in estimating α, while no such effect is observed in the estimation of β, which is often the parameter of main interest.

□

10 Discussion

Even though multistage designs are often used in observational studies, it is usually not transparent in the design phase how the data selection will affect the efficiency of the study. This paper provides a framework describing the design procedure, in the attempt to facilitate description and discussion of such. We also describe how efficiency, and cost adjusted efficiency, can be calculated using Fisher information matrices adjusted for the selection procedure. We suggest presenting the results in graphs, to assist in choosing between alternative designs. Some examples are presented, using different types of graphs to compare selections schemes.

The relative performance of different selection schemes in the examples varied with costs and parameters values, which illustrates that it may prove unfortunate to rely on rules of thumb in the choice of selection scheme. The results from the examples also suggest it is advisable to consider efficiency at some different sets of parameter values, and choose a design that has acceptable efficiency for most plausible parameter values. An advantage of presenting efficiencies in graphs, rather than calculating only the optimal sampling scheme, is that a graph provides a broader picture of a range of sampling schemes, so that the investigator may decide on a sampling scheme which is more robust to misspecification of the model parameters than the optimal sampling scheme.

A technique that is successfully used in experimental design is to take an iterative approach to design, that is, to first do a small study and then fill in more data where the initial sample suggest is efficient (Montgomery 1984). Similarly pilot studies are sometimes used in observational studies. These are often directed to identify practical issues in data collection (such as phrasing of questionnaires) but can be used for assessment of crude parameter estimates for efficiency calculations.

When investigating the efficiency it is also important to consider the consequences of model assumptions not being valid, such as assuming errors to be normally distributed in an overly simplified model. For example, Allison et al. (1998) argue that selecting only extreme outcomes is not advisable in genetical studies searching for quantitative trait loci (QTL), since extremes are likely to result from rare exposures with strong effects (which are not in the model), rather than from the QTL.

We have implicitly assumed that π is known and only θ is estimated, cf. (1). This is a natural view if the sampling scheme is controlled by the investigator. In many observational studies this is however not the case. Then π is an unknown nuisance parameter, which can be estimated from training data for which the full data set $\{Z^i\}$ is known. Otherwise, when only $\{\tilde{Z}^i\}$ is available, we can maximize $L(\theta, \pi)$ jointly with respect to θ and π. However, for sequential multistage designs, it is not necessary to estimate π. Indeed, because of (16), the joint likelihood can be factorized as

$$L(\theta, \pi) = A(\theta)B(\pi). \tag{44}$$

so that inference on θ can be based solely on $A(\theta)$ if the joint parameter space of θ and π is the product of the individual parameter spaces. See Rubin (1976), Heitjan & Rubin (1991) and Little & Rubin (2002) for more details. In particular, (44) implies that the Fisher information for estimating θ is the same whether we know or estimate π.

We have motivated sequential multi-stage designs as a bottom-up procedure (15) for successively including more information. However, in many applications of data coarsening, such as various types of censoring, grouping, rounding and heaping, it is more appropriate to regard $\tilde{Z} = G_J(Z)$ as a top-down procedure, and focus on the loss of information induced by G_J, see Heitjan (1993). In such application, the CAR condition is not always realistic, although it greatly simplifies analysis, as the nuisance parameter π need not be estimated.

The methodology outlined in this paper is equally valid for other estimators and tests than ML and LR. For instance, let $\hat{\theta}(\pi)$ be a given estimator of θ using design parameter π, with asymptotic covariance matrix $V(\hat{\theta}; \theta, \pi)$. Then, the efficiency of estimating the r^{th} component of θ, compared to the ML-estimator $\hat{\theta}_{\text{ML}}$ of the full data set, is

$$e(\hat{\theta}; \theta, \pi) = V_{rr}(\hat{\theta}_{\text{ML}}; \theta, \pi_{\text{full}}) / V_{rr}(\hat{\theta}; \theta, \pi). \tag{45}$$

This corresponds to our previous definition (10) when $\hat{\theta} = \hat{\theta}_{\text{ML}}$ and $h(I) = V_{rr}^{-1}$. When $\pi = \pi_{\text{full}}$, (45) is the usual definition of efficiency, see for example

Lehmann & Casella (1998). Similarly, $e(T;\theta,\pi)$ for a test statistic T, with design π, is defined by comparing its noncentrality parameter with that of T_{LR} for the full data set (π_{full}).

There may be several reasons for choosing other estimators and tests than ML and LR such as 1) robustness against model misspecification, 2) ease of computation or 3) ease of finding designs that optimize efficiency for the given test/estimator subject to a cost-constraint. For instance, Pepe, Reilly & Fleming (1994), Reilly & Pepe (1995) and Reilly (1996) consider mean-score estimators for the two-stage model (42) and derive explicit expressions for designs π that minimize the variance of this estimator given a bound on the cost.

Appendix A.

Verification of (17). From (16) we deduce

$$\psi(\tilde{z};\theta,\pi) = \psi(z_j;\theta), \quad \text{if } \tilde{z} = z_j \in \mathcal{Z}_j. \tag{A.1}$$

This proves the first equality of (17), and the second follows by writing

$$E\left(\psi(\tilde{Z};\theta)^T\psi(\tilde{Z};\theta)\right) = \sum_{j=1}^{k}\int_{\mathcal{Z}_j}\psi(z_j;\theta)^T\psi(z_j;\theta)f(z_j;\theta,\pi)dz_j$$

which inserted into (8) proves the result. □

Verification of (19). Let $\bar{\psi}(z_1) = \psi(z_1)$ and
$\bar{\psi}(z_j) = \psi(z_j) - E(\psi(Z_j)|Z_{j-1} = z_{j-1}) = \psi(z_j) - \psi(z_{j-1})$ for $j = 2,\ldots,k$. Then, using (A.1), we get

$$\psi(\tilde{z};\theta,\pi) = \sum_{r=1}^{k} 1_{\{j\geq r\}}\bar{\psi}(z_r;\theta), \quad \text{if } \tilde{z} = z_j \in \mathcal{Z}_j.$$

and

$$\begin{aligned} I(\theta,\pi) &= E\left(\psi(\tilde{Z};\theta)^T\psi(\tilde{Z};\theta)\right) \\ &= \textstyle\sum_{r,s=1}^{k} E\left(1_{\{J\geq\max(r,s)\}}\bar{\psi}(Z_r;\theta)^T\bar{\psi}(Z_s;\theta)\right) \\ &= \textstyle\sum_{r=1}^{k} E\left(1_{\{J\geq r\}}\bar{\psi}(Z_r;\theta)^T\bar{\psi}(Z_r;\theta)\right) \\ &= I(\theta,\pi_{\min}) \\ &\quad + \textstyle\sum_{r=2}^{k}\int_{\mathcal{Z}_{r-1}} P(J\geq r|z_{r-1})\mathrm{Cov}(\psi(Z_r)|Z_{r-1}=z_{r-1})f(z_{r-1};\theta)dz_{r-1}, \end{aligned}$$

where, in the second equality the non-diagonal terms $e \neq s$ vanish by conditioning on Z_{s-1} when $r < s$ (and vice versa on Z_{r-1} when $r > s$). In the last equality we conditioned on Z_{r-1} for each $r \geq 2$. This proves (19). □

Proof of Proposition 1. The identities $\mathrm{RAC}(\theta, \pi_{\mathrm{full}}) = e(\theta, \pi_{\mathrm{full}}) = 1$ follow immediately from the definitions of RAC and e. Since $\pi \geq \pi_{\min}$ and $\pi' \leq \pi_{\mathrm{full}}$, it suffices to prove the middle inequalities of (22).

To start with, $\mathrm{RAC}(\theta, \pi) \leq \mathrm{RAC}(\theta, \pi')$ follows from (5) and

$$\mathrm{TAC}(\theta, \pi) = \sum_{j=1}^{k} \int_{\mathcal{Z}_j} (C_j(z) - C_{j-1}(z)) \Pi_j(z) f(z_j; \theta) dz_j,$$

with $C_0(\cdot) \equiv 0$. It remains to prove $e(\theta, \pi) \leq e(\theta, \pi')$. Assume $Z \sim f(\cdot; \theta)$ and that U is independent of Z and has a uniform distribution on $(0, 1)$. Put $J = J(Z, U, \pi) = \max\{j; U \leq \Pi_j(Z)\}$. Then, it is easy to check that $\tilde{Z} = Z_J \sim f(\cdot; \theta, \pi)$. Let I_V denote the Fisher information matrix of a random variable V with distribution depending on theta. It is easy to show that $I_{\tilde{Z}} = I(\theta, \pi) = I_{\tilde{Z},U}$, where the first equality follows directly from the definition of $I(\theta, \pi)$ and the second equality since the conditional distribution of $U|\tilde{Z} = \tilde{z}$ does not depend on θ. Similarly, $I_{\tilde{Z}'} = I(\theta, \pi') = I_{\tilde{Z}',U}$. But $(\tilde{Z}, U) = H(\tilde{Z}', U)$, where H is a non-invertible transformation from $\mathcal{Z} \times [0, 1]$ into itself, defined through

$$H(\tilde{z}', u) = (G_J(G_{J'}^{-1}(\tilde{z}')), u).$$

That H is well-defined follows from (13) and the fact that $\pi \leq \pi'$, which implies $J \leq J'$. Hence $I_{\tilde{Z},U} \leq I_{\tilde{Z}',U}$, which is equivalent to $I(\theta, \pi) \leq I(\theta, \pi')$. Finally, $e(\theta, \pi) \leq e(\theta, \pi')$ follows from (9).

□

Proof of Proposition 2. To prove (25), fix $R \in (R_{\min}, 1]$ and assume, on the contrary, that

$$\sup_{\pi; \mathrm{RAC}(\theta,\pi)=R} e(\theta, \pi) \leq e_{\max}(\theta, R) - \varepsilon \tag{A.2}$$

for some $\varepsilon > 0$. Pick $\pi \in \mathcal{Q}_{\mathcal{R}}$ such that $\mathrm{RAC}(\theta, \pi) < R$ and $e(\theta, \pi) \geq e_{\max}(\theta, R) - \varepsilon$. Let $\pi' = (1-t)\pi + t\pi_{\mathrm{full}}$, where $t = (R - R(\theta, \pi))/(1 - R(\theta, \pi))$. Since $\pi, \pi_{\mathrm{full}} \in \mathcal{Q}$ and $\mathcal{Q}$ is convex, it follows that $\pi' \in \mathcal{Q}$. By linearity of $\mathrm{RAC}(\theta, \cdot)$ (cf. (6), $\mathrm{RAC}(\theta, \pi') = R$. Since $\pi \leq \pi' \leq \pi_{\mathrm{full}}$, it follows from Proposition 1 that $e(\theta, \pi') \geq e(\theta, \pi) \geq e_{\max}(\theta, R) - \varepsilon$, which contradicts (A.2). □

Proof of Proposition 3.
Given $R_{\min} \leq R < R' \leq 1$, pick π and π' such that $\mathrm{RAC}(\theta, \pi) = R$ and $\mathrm{RAC}(\theta, \pi') = R'$. Given $0 < t < 1$, let $\pi'' = (1 - t)\pi + t\pi'$. Then $\pi'' \in \mathcal{Q}$

since $\mathcal{Q}$ is convex. By linearity (cf. (6) and (17)), $\mathrm{RAC}(\theta,\pi'') = (1-t)R + tR'$ and $I(\theta,\pi'') = (1-t)I(\theta,\pi) + tI(\theta,\pi')$. Invoking (26), we obtain $e(\theta,\pi'') \geq (1-t)e(\theta,\pi) + te(\theta,\pi')$. Hence

$$e_{\max}((1-t)R + tR') \geq e(\theta,\pi'') \geq (1-t)e(\theta,\pi) + te(\theta,\pi').$$

Since π and π' can be chosen so that $e(\theta,\pi) \geq e_{\max}(\theta,R) - \varepsilon$ and $e(\theta,\pi') \geq e_{\max}(\theta,R') - \varepsilon$ for arbitrarily small $\varepsilon > 0$, concavity of $e_{\max}(\theta,\cdot)$ follows. □

Proof of Proposition 4.
It suffices to establish

$$e_{\max}(\theta,tR) \geq te_{\max}(\theta,R) \tag{A.3}$$

for $R_{\min} < R \leq 1$ and $R_{\min}/R < t < 1$, since this is equivalent to $CE_{max}(\theta,\cdot)$ being non-increasing on $(R_{\min},1]$. If (27) holds, (A.3) follows as in the proof of Proposition 3. If (26) and (28) hold, we use Proposition 3 and the concavity of $e_{max}(\theta,\cdot)$ to deduce that

$$\begin{aligned} e_{max}(\theta,R) &= e_{max}(\theta,tR) + \int_{tR}^{R} e'_{max}(\theta,r)dr \\ &\leq e_{max}(\theta,tR) + R(1-t)e'_{max}(\theta,tR) \\ &\leq e_{max}(\theta,tR) + R(1-t)e_{max}(\theta,tR)/(tR) \\ &= e_{max}(\theta,tR)/t. \end{aligned}$$

References

Allison, D., Heo, M., Schork, N., Wong, S. & Elston, R. (1998), 'Extreme selection strategies in gene mapping studies of oligogenic quantitative traits do not always increase power.', *Human Hered.* **48**, 97–107.

Fisher, R. (1934), 'The effects of methods of ascertainment upon the estimation of frequencies', *Annals of Eugenics* **6**, 13–25.

Grünewald, M. & Hössjer, O. (2010), 'Efficient ascertainment schemes for maximum likelihood estimation', *Journal of Statistical Planning and Inference* **140**(7), 2078 – 2088.

Grünewald, M., Humphreys, K. & Hössjer, O. (2010), A Stochastic EM type algorithm for parameter estimation in models with continuous outcomes, under complex ascertainment. Working paper -under revision for The International Journal of Biostatistics.

Hammersley, J. M. & Handscomb, D. C. (1964), *Monte Carlo methods*, Methuen, London.

Heitjan, D. (1993), 'Ignorability and coarse data: Some biomedical applications', *Biometrics* (49), 1099–1109.

Heitjan, D. & Rubin, D. (1991), 'Ignorability and coarse data', *The Annals of Statistics* **19**(4), 2244–2253.

Hesterberg, T. (1995), 'Weighted average importance sampling and defensive mixture distributions', *Technometrics* **37**, 185–194.

Ku, H. H. (1966), 'Notes on the use of propagation of error formulas', *Journal of research of the national bureau of standards* **70C**(4).

Lehmann, E. & Casella, G. (1998), *Theory of point estimation*, 2nd edn, Springer, New York.

Little, R. J. A. & Rubin, D. (2002), *Statistical analysis with missing data*, Wiley series in probability and statistics, John Wiley & Sons, Inc., New York; Chichester.

Lyon, H. N., Emilsson, V., Hinney, A., Heid, I. M., Lasky-Su, J., Zhu, X., Thorleifsson, G., Gunnarsdottir, S., Walters, G. B., Thorsteinsdottir, U., Kong, A., Gulcher, J., Nguyen, T. T., Scherag, A., Pfeufer, A., Meitinger, T., Brönner, G., Rief, W., Soto-Quiros, M. E., Avila, L., Klanderman, B., Raby, B. A., Silverman, E. K., Weiss, S. T., Laird, N., Ding, X., Groop, L., Tuomi, T., Isomaa, B., Bengtsson, K., Butler, J. L., Cooper, R. S., Fox, C. S., O'Donnell, C. J., Vollmert, C., Celedón, J. C., Wichmann, H. E., Hebebrand, J., Stefansson, K., Lange, C. & Hirschhorn, J. N. (2007), 'The association of a SNP upstream of INSIG2 with body mass index is reproduced in several but not all cohorts', *PLoS Genet* **3**(4), e61.

Maydrech, E. & Kupper, L. (1978), 'Cost considerations and sample size requirements in cohort and case-control studies', *American journal of epidemiology* **107**, 201–205.

Melas, V. (2006), *Functional Approach to Optimal Experimental Design*, number 184 *in* 'Lecture Notes in Statistics', Springer, United States of America, , Chapter 1.4.

Montgomery, D. C. (1984), *Design and Analysis of Experiments*, second edn, John Wiley & sons, Chapter 1.

Pepe, M., Reilly, M. & Fleming, T. (1994), 'Auxiliary outcome data and the mean score method', *Journal of Statistical Planning and Inference* (42), 137–160.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Rao, C. (1965), *Classical and Contagious Discrete Distributions*, Pergamon Press and Statistical Publishing Society, Calcutta, chapter On discrete distributions arising out of methods of ascertainment., pp. 320–332.

Reilly, M. (1996), 'Optimal sampling strategies for two-stage studies', *American journal of epidemiology* **143**(1), 92–100.

Reilly, M. & Pepe, M. S. (1995), 'A mean score method for missing and auxiliary covariate data in regression models', *Biometrika* **82**(2), 299–314.

Rubin, D. B. (1976), 'Inference and missing data', *Biometrica* **63**(3), 581–592.

Serfling, R. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Silvey, S. (1980), *Optimal Design*, Chapman and Hall, London.

Thomas, D., Xie, R. & Gebregziabher, M. (2004), 'Two-stage sampling designs for gene association studies', *Genetic Epiemiology* **27**, 401–414.

White, J. (1982), 'A two stage design for the study of the relationship between a rare exposure and a rare disease', *American Journal of Epidemiology* **115**, 119–128.

Figures and tables

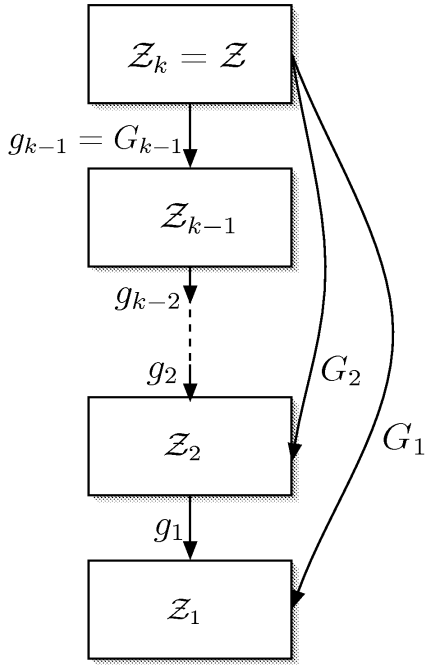

Figure 1: A Multistage Model where sampling spaces are reduced sequentially. Stage k represents the most complex sampling space and Stage 1 represent the most sparse one.

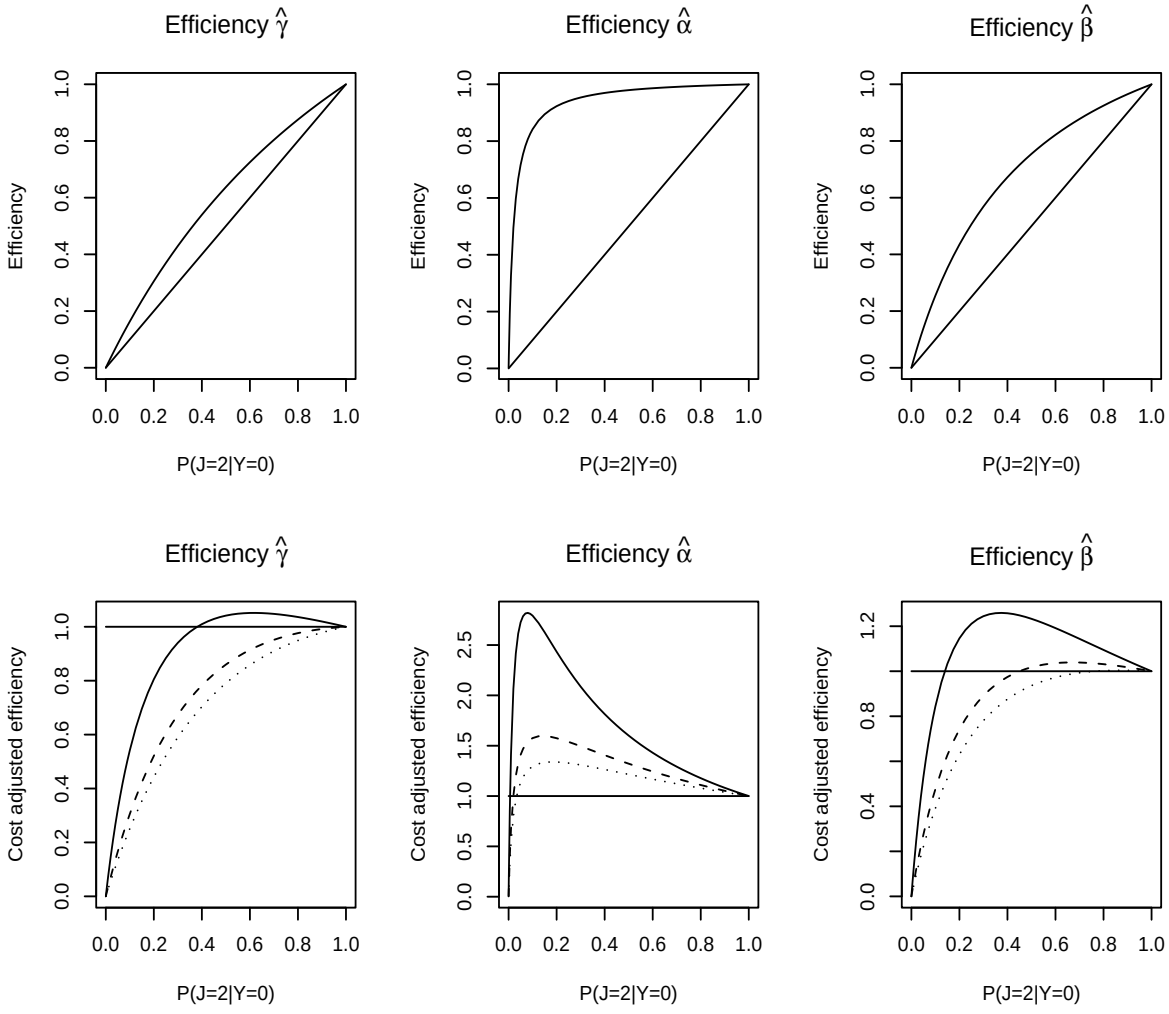

Figure 2: Efficiency and Cost adjusted efficiency for a logistic regression model. For the cost adjusted efficiency solid, dashed and dotted lines represent costs $C_1 = 0, 1/3$ and $1/2$ respectively, for $C_2 = 1$. $\gamma = -1, \alpha = -2, \beta = 2$.

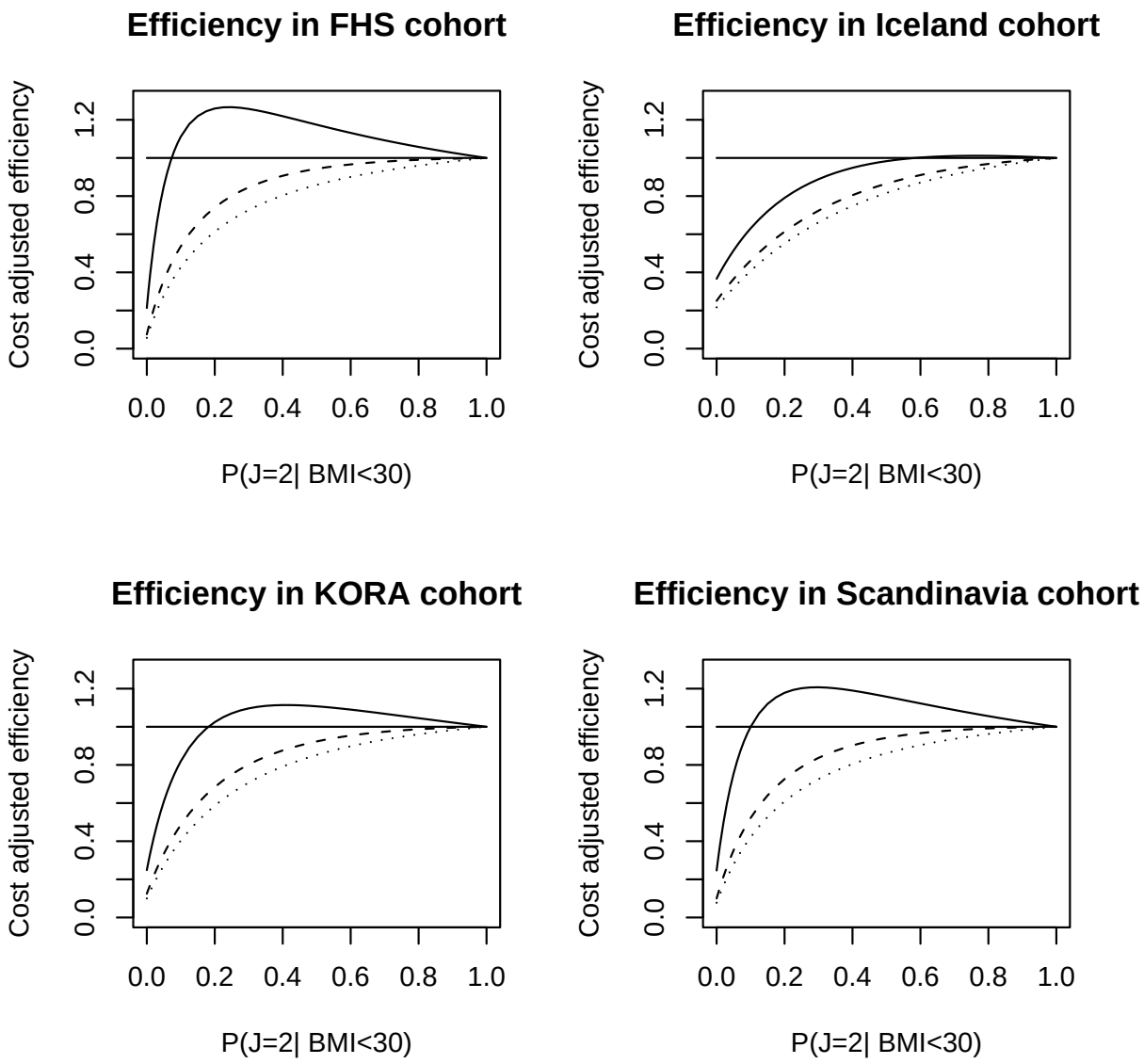

Figure 3: Efficiency of over-sampling obese individuals (BMI$\geq$ 30) in different cohorts in Lyon et al. (2007) when estimating the effect (β) of the C/C genotype of genetic marker rs7566605 on log(BMI) in a linear regression model. Solid, dashed and dotted lines represent $C_1 = 0, 1/3$ and $1/2$ respectively, for $C_2 = 1$.

	FHS	Iceland	KORA	Scandinavia
$\hat{\gamma}$	-2.04	-2.02	-2.12	-2.11
$\hat{\alpha}$	3.22	3.36	3.29	3.28
$\hat{\beta}$	0.0136	0.0237	0.0016	0.0111
$\hat{\sigma}$	0.169	0.232	0.165	0.137

Table 1:
Specification of parameter values in Example 4 that were used in Figure 3. Parameter values were calculated from values presented in Table 1. and Table 2. in Lyon et al. (2007) as follows:

$$\begin{aligned} F(\hat{\gamma}) &= \exp(\hat{\gamma})/(1+\exp(\hat{\gamma})) = n_{C/C}/(n_{C/C}+n_{C/G}+n_{G/G}), \\ \hat{\alpha} &= \log[(\text{Mean BMI}_{C/G} \times n_{C/G} + \text{Mean BMI}_{G/G} \times n_{G/G})/(n_{C/G}+n_{G/G})], \\ \hat{\beta} &= \log(\text{Mean BMI}_{C/C}) - \hat{\alpha}, \\ \hat{\sigma}^2 &= (\text{SD BMI})^2 \times (1/\text{BMI Mean})^2 \end{aligned}$$

Calculation of σ^2 was performed using the propagation of error formulas (Ku 1966).

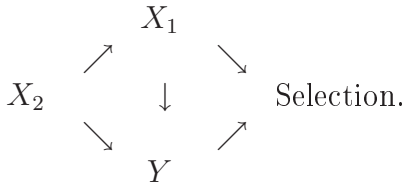

Figure 4: A three-stage design.

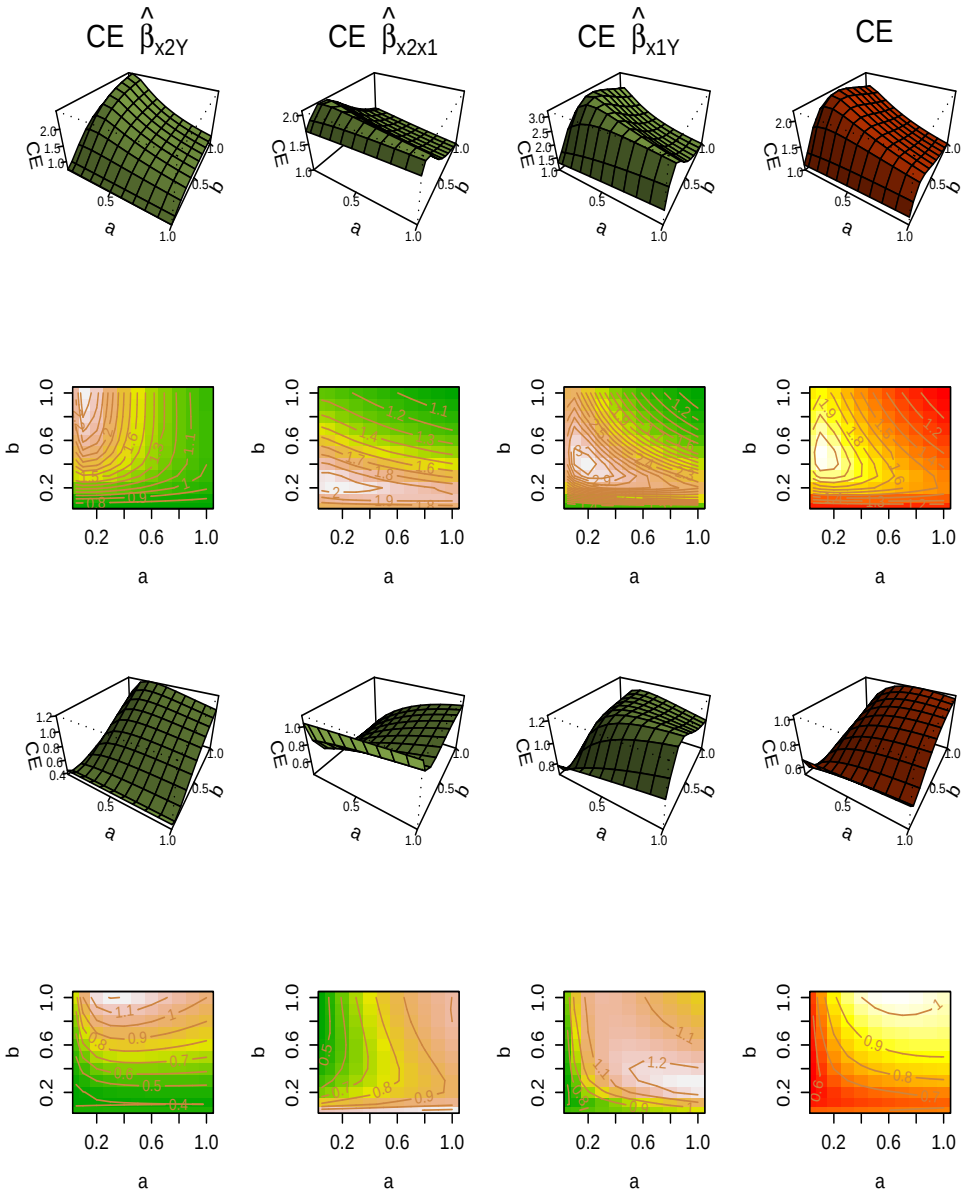

Figure 5: Individual (green) and combined (red) CE for three parameters in a three-stage design, as described in Example 5. Two cost functions are applied, $(C_1, C_2, C_3) = (0, 0, 1)$ (upper rows) and $(C_1, C_2, C_3) = (0, 1, 2)$ (lower rows). $t_Y = 3, t_{x_1} = 0.5$, $\alpha_{X_2} = 0, \alpha_{X_1} = -3, \beta_{X_2X_1} = 2, \alpha_Y = 0, \beta_{X_1Y} = 1, \beta_{X_2Y} = 1, \sigma_Y^2 = 1$. Monte Carlo approximation, based on (29), is used with $N = 10000$.

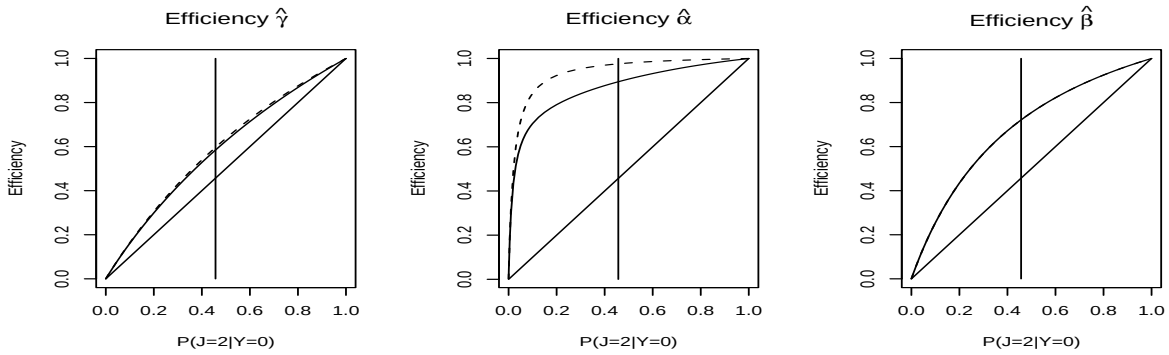

Figure 6: Efficiency for estimation of parameters using two-stage (dashed line) likelihood and conditional ascertainment likelihood (solid line) in logistic regression, cf. Examples 3 and 6, with $\gamma = -1, \alpha = -2, \beta = 2, \eta_1 = P(J = 2|Y = 1) = 1$ and $0 < \eta_0 = P(J = 2|Y = 0) \leq 1$.