STATISTICS IN MEDICINE, VOL. 16, 285–319 (1997)

TOWARD A CURSE OF DIMENSIONALITY APPROPRIATE (CODA) ASYMPTOTIC THEORY FOR SEMI-PARAMETRIC MODELS

JAMES M. ROBINS

Harvard School of Public Health, Boston, MA 02115, U.S.A.

AND

YA'ACOV RITOV

Hebrew University, Jerusalem, Israel

SUMMARY

We argue, that due to the curse of dimensionality, there are major difficulties with any pure or smoothed likelihood-based method of inference in designed studies with randomly missing data when missingness depends on a high-dimensional vector of variables. We study in detail a semi-parametric superpopulation version of continuously stratified random sampling. We show that all estimators of the population mean that are uniformly consistent or that achieve an algebraic rate of convergence, no matter how slow, require the use of the selection (randomization) probabilities. We argue that, in contrast to likelihood methods which ignore these probabilities, inverse selection probability weighted estimators continue to perform well achieving uniform $n^{\frac{1}{2}}$-rates of convergence. We propose a curse of dimensionality appropriate (CODA) asymptotic theory for inference in non- and semi-parametric models in an attempt to formalize our arguments. We discuss whether our results constitute a fatal blow to the likelihood principle and study the attitude toward these that a committed subjective Bayesian would adopt. Finally, we apply our CODA theory to analyse the effect of the 'curse of dimensionality' in several interesting semi-parametric models, including a model for a two-armed randomized trial with randomization probabilities depending on a vector of continuous pre-treatment covariates X. We provide substantive settings under which a subjective Bayesian would ignore the randomization probabilities in analysing the trial data. We then show that any statistician who ignores the randomization probabilities is unable to construct nominal 95 per cent confidence intervals for the true treatment effect that have both: (i) an expected length which goes to zero with increasing sample size; and (ii) a guaranteed expected actual coverage rate of at least 95 per cent over the ensemble of trials analysed by the statistician during his or her lifetime. However, we derive a new interval estimator, depending on the Randomization probabilities, that satisfies (i) and (ii).

1. INTRODUCTION

In the analysis of data obtained by stratified random sampling, 'likelihoodist' and Bayesian statisticians often claim that inference concerning the population mean should be the same regardless of whether the stratum-specific randomization (selection) probabilities are or are not known to the data analyst. This paper was motivated by a desire to study this claim in the following infinite-dimensional (semi-parametric) superpopulation version of continuously stratified random sampling.

CCC 0029–5981/97/030285–35
© 1997 by John Wiley & Sons, Ltd.

We draw n independent and identically distributed realizations of a random vector (Y,X,R). Y and R are dichotomous $\{0,1\}$ variables, X is a k-dimensional continuous random vector with support in the unit cube $[0,1]^k$ in R^k. X and R are always observed. However, Y is only observed when $R = 1$ and remains unobserved when $R = 0$. The known selection (randomization) probability $\pi(X) = \Pr[R = 1 \mid X, Y]$ is bounded away from 0 so all subjects have a positive probability of having Y observed. The goal is to estimate the mean of Y, $E(Y)$, from the observed data. In this model, the statistic $\{(R_i,X_i); i = 1,\ldots,n\}$ is ancillary for $E(Y)$ when X has a known distribution.

Through a careful study of this model, this paper attempts to formalize arguments given in references 1–5 that, in semi-parametric models afflicted by the curse of dimensionality, it is often not possible to construct estimators for a finite dimensional parameter (for example, $E(Y)$) that perform well in moderate samples unless prior knowledge is available concerning the marginal distribution of an ancillary statistic, such as the conditional distribution $P_r(R = 1 \mid X)$ of the randomization indicator R; when such prior knowledge is available, this knowledge can be used to construct locally efficient semi-parametric estimators.

In the model described above, X_i is observed for all n subjects but Y_i is observed only on a random subsample with randomization (selection) probabilities depending on X_i. In the sample survey literature, our problem would be referred to as a semi-parametric superpopulation version of continuously stratified random sampling since: (i) our parameter of interest is $\psi_0 \equiv E(Y)$, the mean of Y in a hypothetical superpopulation, rather than $\sum_i Y_i/n$, the mean of Y in our n study subjects; (ii) the stratification variables X are continuous rather than discrete; and (iii) the model is infinite dimensional (semi-parametric) since the law $p(X) = E(Y \mid X)$ is completely unspecified. In this model Y is missing at random in the sense of Rubin[6,7] since, by design, $\pi(X) = P(R = 1 \mid Y,X)$.

In Section 4, we conclude that, due to the curse of dimensionality, when X is of high dimension, say $k = 5$, and the sample size is moderate (say, $n = 10^5$), it is essential for the analyst to use the known randomization probabilities $\pi(X)$ in order to construct accurate estimates of $\psi_0 \equiv E(Y)$. Hansen *et al.*[32] reached a similiar conclusion. According to the standard asymptotic theory of estimation in semi-parametric models as described by Bickel *et al.*[8], the semi-parametric variance bound for estimation of the mean of Y is the same whether or not the randomization probabilities are known to the data analyst. In this paper we adopt the point of view that an appropriate asymptotic theory is, by definition, one whose results provide guidance to the moderate sample performance of estimation procedures. We therefore conclude that the standard asymptotic theory for semi-parametric models is not a curse of dimensionality appropriate (CODA) theory. We therefore propose a new CODA asymptotic theory that should provide more appropriate guidance to moderate sample performance. The new theory applies to any non- or semi-parametric model – not only to our continuously stratified random sampling model. Our new CODA theory reduces to the standard theory when the curse of dimensionality does not affect the estimation problem. The CODA theory correctly indicates that in our continuously stratified random sampling model, estimators of $E(Y)$ with good moderate sample performance will not exist in the absence of knowledge of $\pi(X)$. In contrast, the theory implies that if X were discrete with only a moderate number of levels, good moderate sample performance would not require any knowledge of $\pi(X)$.

We apply our CODA theory to several semi-parametric models, and we obtain interesting results. We argue that if, in our continuously stratified random sampling model, (i) $E(Y \mid X)$ follows a logistic regression on X, that is, $\text{logit}E(Y \mid X) = \psi_0 X$, and (ii) data on a surrogate variable V correlated with Y is available on all study subjects,[9] the surrogate data could be used to construct estimators of ψ_0 that are both (a) more efficient than the complete case estimator (logistic regression restricted to subjects for whom Y was observed), and (b) perform well in moderate size samples only if knowledge is available concerning the randomization probabilities $\pi(X)$. In contrast, if

X is discrete with only a moderate number of levels, data on V can be used to construct more efficient estimators of ψ_0 even with $\pi(X)$ completely unknown.[10]

We then consider a survival analysis setting in which Y is the logarithm of failure time and $Y = \psi_0 X + \varepsilon$ with ε independent of X. It is known that when Y is subject to independent right-censoring, estimators of ψ_0 can be constructed that perform well in moderate size samples even if (a) X has 5 continuous components, and (b) the dependence of the censoring distribution on X is completely unspecified (See Tsiatis;[11] Ritov;[12] and Buckley and James[13]). In contrast, when (a) and (b) are true, no reasonably performing estimator of ψ_0 has ever been proposed in the less restrictive censored median regression model that assumes only that the median of ε does not depend on X. We use CODA theory to argue that this is not due to a lack of imagination but rather that no such estimator exists. In contrast, if we knew that the censoring distribution was independent of X or followed a parametric or semi-parametric model (for example, a Cox proportional hazards model) for the dependence of censoring on X, good estimators of ψ_0 can be constructed.[5,14]

Finally, we consider a two-arm placebo-controlled randomized trial with data $(Y_i, R_i, X_i), i = 1, \ldots, n$, where Y_i is a continuous outcome variable, R_i is a dichotomous treatment arm indicator, X_i is a vector of pre-treatment continuous covariates such as age, height, weight, etc., and the known randomization probabilities $\pi(X) = P_r[R = 1 \mid X]$ depend in a possibly complicated way on the covariates X. For simplicity, we assume the effect of the treatment R_i on the outcome Y_i is additive, that is, $Y_i = \psi_0 R_i + \varepsilon_i$, where ε_i is the outcome that would be observed if, possibly contrary to fact, subject i received placebo ($R_i = 0$). A number of likelihoodists and Bayesians have argued that knowledge of the randomization probabilities should be ignored when making inferences on the treatment effect ψ. However, we show that, due to the curse of dimensionality, any estimate of ψ that fails to use knowledge of $\pi(X)$ can perform poorly in moderate samples. In contrast, the propensity score estimator of Rosenbaum,[15] which is an inefficient member of the class of G-estimators proposed by Robins,[16] depends on the propensity score[33] $\pi(X)$ and will perform well in moderate samples. Robins[16] describes how one can construct locally semi-parametric efficient G-estimators of ψ_0. Indeed, in Section 8, we argue that data analysts who ignore the randomization probabilities are unable to construct nominal 95 per cent confidence intervals for the treatment effect that have both (i) expected length which goes to zero with increasing sample size and (ii) a guaranteed expected actual coverage rate of at least 95 per cent over the ensemble of trials analysed by the statistician during his or her lifetime. We derive a new interval estimator, depending on $\pi(X)$ that satisfies (i) and (ii).

We now provide a rather detailed introductory overview of the paper, the purpose of which is to informally describe our main ideas before delving into more technical material. In Section 2, we study the continuously stratified random sampling model and show that if we could specify a parametric model for $E(Y \mid X)$, then the maximum likelihood estimator (MLE) of $E(Y)$ is the same whether or not the selection probabilities $\pi(X)$ are known. In Section 3, we note that the method of maximum likelihood is an example of an inferential method that obeys the 'strict likelihood principle (SLP)' that states that 'if 2 experiments give proportional likelihoods for parameter θ, then inference concerning θ should not depend on which experiment generated the data.' We then argue that any method of inference that obeys the SLP will result in inferences concerning $E(Y)$ that do not depend on whether randomization probabilities are known.

In Section 4 we argue that, in practice, it will not be possible to correctly specify a parametric model for $E(Y \mid X)$, and adopt the semi-parametric model in which $E(Y \mid X)$ is completely unrestricted. However, the randomization probabilities $\pi(X)$ will be known, since they are under the control of the investigator. We believe this semi-parametric model is appropriate, since historically the most common argument for random sampling has been that it allows valid inference for $E(Y)$

without requiring any assumptions concerning the joint distribution of (Y,X). It is well known that when $\pi(X)$ is known, the Horvitz–Thompson[17] estimator $\hat{\psi}_{HT} = n^{-1}\sum_i R_i Y_i/\pi(X_i)$ is unbiased and $n^{\frac{1}{2}}$-consistent for $E(Y)$, even when $E(Y \mid X)$ is completely unrestricted. Rosenbaum and Rubin in a series of papers also stressed the robustness properties of estimators that use knowledge of the randomization probabilities $\pi(X)$ which they refer to as the propensity score. See references 7, 33 and 15 for examples of their work. The estimator $\hat{\psi}_{HT}$ will also perform well in moderate size samples (in the sense that, whatever be $\pi(X)$ and $E(Y \mid X)$, the estimator will be approximately normal and centred on $E(Y)$). However, as noted in Section 3, the HT estimator depends on the known selection probabilities $\pi(X)$ and thus violates the SLP. When k is of high dimension, say, $k = 5$, it is difficult to imagine how, in the moderate size samples found in practice, one would go about constructing an estimator $\tilde{\psi}$, as a competitor to $\hat{\psi}_{HT}$, that did not depend on the known $\pi(X)$. One candidate would be $\tilde{\psi} = n^{-1}\sum_i \tilde{E}(Y \mid X_i)$ where $\tilde{E}(Y \mid X_i)$ is a multivariate kernel regression estimator of $E(Y \mid X_i) = E(Y \mid X_i, R_i = 1)$ among subjects with $R_i = 1$. However, unless the bandwidth is chosen to be very small, $\tilde{\psi}$ may have substantial finite sample bias if $E(Y \mid X)$ and $\pi(X)$ are both complicated functions of X. If the bandwidth is chosen to be small enough to avoid substantial bias, the finite sample variance of $\tilde{\psi}$ will be excessive. In the jargon of epidemiologists, X is a high-dimensional confounding variable, and prior knowledge of either the association of X with the outcome Y (that is, of $E(Y \mid X)$) or of the association of X with the selection indicator, R (that is, of $\pi(X)$) must be known in order to effectively control selection bias (that is, confounding) due to X in estimation of $E(Y)$. So, when $E(Y \mid X)$ is unknown, we conclude from such informal finite sample considerations that, due to the curse of dimensionality, $E(Y)$ can be well estimated for all $\pi(X)$ and $E(Y \mid X)$ only if the estimator violates the SLP and depends on the known $\pi(X)$. The difficulty with the SLP is that it (inappropriately in our view) proscribes the data analyst from using knowledge of the population association of X with R (that is, $\pi(X)$) to control selection bias (confounding) due to X.

The above 'informal' finite sample considerations are, as is often the case, hard both to formalize and to generalize in a finite sample setting (although see Section 8.1) but are easy to do so in an asymptotic setting. Specifically, in Section 4, we prove our main result; the Horvitz–Thompson estimator $\hat{\psi}_{HT}$ that uses knowledge of the randomization probabilities is uniformly asymptotically normal and unbiased (as defined in Section 8) and thus uniformly $n^{\frac{1}{2}}$-consistent for $\psi_0 \equiv E(Y)$; in contrast, there is no estimator $\hat{\psi}$ of ψ that satisfies the SLP and thus does not depend on $\pi(X)$ that (i) will be uniformly consistent for ψ_0, or (ii) will achieve a (pointwise) algebraic rate of convergence (much less an $n^{\frac{1}{2}}$-rate) whatever be $\pi(X)$ and $E(Y \mid X)$ generating the data. Part (ii) states that there exists a law indexed by $\pi(X)$ and $E(Y \mid X)$such that if the data were generated under this law, $n^{\alpha}(\hat{\psi} - \psi_0)$ will fail to converge to zero in probability for all $\alpha > 0$.

Throughout the paper, we often stress uniform (over the possible laws $E(Y \mid X)$ and $\pi(X)$ that might be generating the data) properties of our estimators, such as uniform consistency. This reflects the fact that in practice we are faced with a particular sample size n. If an estimate $\hat{\psi}$ is not uniformly consistent, then, for any fixed sample size, including the actual sample size n, there will be the laws $E(Y \mid X)$ and $\pi(X)$, such that, if the data were generated under these laws, then the difference between $\hat{\psi}$ and ψ_0 will not be small. On the other hand, if $\hat{\psi}$ is uniformly asymptotically normal and unbiased for ψ_0 (as defined in Section 8), then there will exist a minimal sample size n at which the estimator will be approximately normally distributed and centred on ψ_0 regardless of the laws $\pi(X)$ and $E(Y \mid X)$ generating the data. Technically it is this minimal sample size n that characterizes the size of the 'moderate size' sample at which $\hat{\psi}$ is guaranteed to perform well.

In Section 5, we consider the attitude a committed subjective Bayesian would take toward these results. We first show that any Bayesian whose prior belief concerning the function $E(Y \mid X)$

were *a priori* independent of her beliefs concerning $\pi(X)$ would perform inference consistent with the SLP and ignore knowledge of $\pi(X)$ and thus fail to be uniformly consistent. We then provide substantive settings in which a subjective Bayesian would indeed have independent priors. We then argue that the criterion that an estimator be uniformly consistent whatever be $\pi(X)$ and $E(Y \mid X)$ is essentially a minimax criterion and thus, in general, of no interest to a subjective Bayesian; a subjective Bayesian has no worries about doing poorly at certain laws that have essentially zero subjective prior probability if she can do extremely well at laws with substantial prior probability. That is, a subjective Bayesian has no particular desire to be robust by being approximately minimax.

We then show that when $\pi(X)$ is known, there are Bayes estimators of $E(Y)$ based on dependent priors that are $n^{\frac{1}{2}}$-consistent. In fact, the minimax estimator of $E(Y)$ under squared error loss is approximated by a Bayes estimator based on dependent priors. This does not conflict with our previous claim, since Bayes estimators based on dependent priors depend on $\pi(X)$ and hence do not satisfy the SLP.[16,18] Further, we discuss how a 'robust' Bayesian might use a mixture prior consisting of a mixture of an independent prior and a dependent minimax prior that can afford him/her both the benefits of their subjective beliefs encoded in the independence prior, as well as robustness to those beliefs being wrong, as encoded in an approximately minimax prior. Indeed, the Bayes estimator based on this mixture prior will be $n^{\frac{1}{2}}$-consistent for all $\pi(X)$ and $E(Y \mid X)$. Rubin[7] had previously proposed that 'robust' Bayesians ignore their subjective beliefs and use dependent priors when $\pi(X)$ is known. Rubin refers to $\pi(X)$ as the propensity score and suggests incorporating the propensity score in the prior for $E(Y \mid X)$ – thus creating a form of dependent prior. Godambe and Thompson[34] had earlier discussed related ideas. In a separate report, Ritov and Robins[19] show how one can use Rubin's idea to actually construct $n^{\frac{1}{2}}$-consistent Bayes estimators ψ_0 that are more efficient than the HT estimator.

The continuously stratified random sampling model $\mathcal{M}(p)$ in which $p(X) \equiv E(Y \mid X)$ is unrestricted and $\pi(X)$ is known and the model $\mathcal{M}(p, \pi)$ in which both $E(Y \mid X)$ and $\pi(X)$ are unrestricted and unknown except for $\pi(X)$ being bounded away from zero are examples of infinite dimensional models. Bickel *et al.*[8] describe an asymptotic theory for inference in infinite-dimensional semi- and non-parametric models. A key concept in this theory is the semi-parametric variance bound for a parameter ψ_0. The bound is equal to the supremum of the Cramer–Rao variance bounds for ψ_0 over all regular fully parametric submodels. The semi-parametric variance bound for ψ_0 is a lower bound for the asymptotic variance of any regular $n^{\frac{1}{2}}$-consistent estimator of ψ_0. An estimator is regular if its convergence to its limiting distribution is uniform in shrinking $n^{-\frac{1}{2}}$ neighbourhoods.[8] In Section 4, we show that the semi-parametric variance bound for $\psi_0 = E(Y)$ is finite and the same in both model $\mathcal{M}(p, \pi)$ and $\mathcal{M}(p)$. However, as discussed above, we know that, for all $\alpha > 0$, there is no n^{α}-consistent (much less an $n^{\frac{1}{2}}$-consistent) estimator of ψ_0 in model $\mathcal{M}(p, \pi)$ when X has continuous components, since knowledge of $\pi(X)$ is not available in this model. Thus the lower bound is far from being sharp. Ritov and Bickel[20] have previously exhibited a number of examples of this same phenomenon. In contrast, there are $n^{\frac{1}{2}}$-consistent estimators of ψ_0 in model $\mathcal{M}(p)$, for example, $\hat{\psi}_{\mathrm{HT}}$. As described earlier, this observation led us to attempt to develop a new CODA asymptotic theory for semi- and non-parametric models.

The standard semi-parametric variance bound for model $\mathcal{M}(p, \pi)$ is an attainable lower bound for the asymptotic variance of a regular estimator of ψ when X is continuous only under the additional assumptions that $\pi(X)$ and $p(X) = E(Y \mid X)$ are locally smooth in X, for example, have bounded higher order derivatives with respect to X. The CODA semi-parametric variance bound is the attainable asymptotic variance of a regular estimator of ψ when there are no smoothness assumptions imposed. In high dimensional problems with moderate sample sizes, local smoothness assumptions, even if true, are of little use in practice, since essentially no two subjects will

have X-vectors close enough to one another to allow the 'borrowing of information' necessary for smoothing. In this paper, we take the point of view that an appropriate asymptotic theory is, by definition, one whose results provide guidance to moderate sample performance. Thus an appropriate asymptotics for high dimensional problems is an asymptotics that does not impose smoothness assumptions.

In Section 7 we develop a CODA algorithm that, in any semi-parametric model with finite dimensional parameter of interest ψ_0, produces a class of estimators that we believe contains (up to asymptotic equivalence) all $n^{\frac{1}{2}}$-uniformly consistent regular asymptotically linear estimators of ψ_0 under CODA asymptotics (and thus all estimators expected to have good moderate sample properties). The CODA semi-parametric variance bound for ψ_0 is defined to be the infimum of the asymptotic variances of the estimators in this class. As an example, as expected, in model $\mathcal{M}(p,\pi)$, the CODA variance bound is infinite, while in model $\mathcal{M}(p)$, the CODA variance bound is the finite standard semi-parametric variance bound. We conjecture that if, as in model $\mathcal{M}(p,\pi)$, the CODA variance bound is infinite, there will exist no uniformly consistent estimators of ψ_0.

When the CODA variance bound is finite as in model $\mathcal{M}(p)$, it will in general be possible to construct a regular estimator that attains the bound globally (that is, at all laws allowed by the model); however, it will not in general be possible to construct an estimator that attains the bound uniformly and globally (that is, uniformly over all laws allowed by the model). Note that a regular estimator need not converge uniformly to its limiting distribution (since, by definition, that would require uniform convergence in fixed, rather than shrinking, neighbourhoods). An essentially equivalent point has previously been made by Klaassen,[21] Bickel and Klaassen,[22] and Donoho[23] in their discussion of the symmetric location model. However, as argued in Section 8, under regularity conditions, it should generally be possible to obtain uniform locally efficient estimators, that is, estimators that attain the bound uniformly at a smooth submodel (for example, a parametric submodel for $p(X)$) and remain uniformly asymptotically normal and unbiased for ψ_0 off the submodel. Furthermore, the 'Wald' interval formed by taking a uniform locally efficient estimator ± 1.96 empirically estimated standard errors (for example, obtained by bootstrapping) will be a valid 95 per cent asymptotic confidence interval since an essentially necessary and sufficient condition for the validity of the large sample Wald confidence interval is that the estimator be uniformly asymptotically normal and unbiased.

We describe how to construct such locally efficient estimators. In particular, we show how to construct a locally efficient estimator in model $\mathcal{M}(p)$ that can be much more efficient than the Horvitz–Thompson estimator. We conclude that inferences concerning ψ_0 should be based on locally efficient estimators and the associated Wald confidence intervals when possible.

Until Section 9, we consider only unconditional inference. However, in a number of our models, there are ancillary statistics and the conditionality principle suggests that inference be performed conditionally on these ancillary statistics. We show how the use of locally efficient estimators allows us to both 'nearly' satisfy the conditionality principle in model $\mathcal{M}(p)$ and yet not sacrifice our claim that we can produce uniformly $n^{\frac{1}{2}}$-consistent estimators.

2. PARAMETRIC MODELS

For convenience to simplify the problem, until Section 8 we shall assume the marginal distribution of $X = (X_1, \ldots, X_k)$ is known to be uniform with density equal to 1 on $[0,1]^k$. Denote by $p(X)$ the probability that $Y = 1$ given X. Thus, $p(X) = E(Y \mid X)$ and $E(Y) = E[p(X)] \equiv \int p(x)\,\mathrm{d}x$ with $\mathrm{d}x = \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_k$.

In this Section, we shall study this problem first under the assumption that we can specify parametric models for $p(x)$ and $\pi(x)$ before proceeding to study the models of real interest to us,

non- and semi-parametric models. Hence, we assume $p(x) = p(x;\theta_0)$ and $\pi(x) = \pi(x;\gamma_0)$ where $p(x;\theta)$ and $\pi(x;\gamma)$ are known functions taking values in $[0,1]$ (for example, $p(x;\theta) = \text{expit}(\theta' x)$ and $\pi(x;\gamma) = \text{expit}(\gamma' x)$, with $\text{expit}(b) = e^b/\{1+e^b\}$). θ takes values in a finite-dimensional subset θ^*, γ takes values in a finite-dimensional subset γ^* and θ and γ are variation-independent, that is, (θ,γ) takes values in $\theta^* \times \gamma^*$. We can represent the observed data as $(R_i, X_i, R_i Y_i), i = 1,\ldots,n$.

Since X is marginally uniform, the likelihood function is

$$L = L_1(\theta) L_2(\gamma) \tag{1}$$

$$L_1(\theta) = \prod_{i=1}^{n} \left[p(X_i;\theta)^{Y_i} (1 - p(X_i;\theta))^{(1-Y_i)} \right]^{R_i}. \tag{2}$$

$$L_2(\gamma) = \prod_{i=1}^{n} \pi(X_i;\gamma)^{R_i} \{1 - \pi(X_i;\gamma)\}^{1-R_i} \quad . \tag{3}$$

$L_1(\theta) = f(Z \mid W;\theta)$ is the conditional density of $Z = \{Y_i R_i; I = 1,\ldots,n\}$ given $W = \{(R_i, X_i); i = 1,\ldots,n\}$. $L_2(\gamma) = f(W;\gamma)$ is the marginal density of W. Note, in (2), $\left\{ p(X_i;\theta)^{Y_i} \right\}^{R_i}$ only depends on Y_i through the product $Y_i R_i$. When the true value γ_0 of γ is known, W is ancillary for θ since then $f(W;\gamma_0)$ is a known density. When γ_0 is unknown, W is said to be S-ancillary for θ since the marginal distribution of W depends on a variation-independent parameter γ.[24] The likelihood may factor as in (1) and yet there exists no ancillary or S-ancillary statistic for θ. See Section 7 for an example. Even in this setting, following Robins et al.,[25] we shall refer to $L_2(\gamma)$ as an 'ancillary process' for θ.

Let $\psi(\theta) \equiv \int p(x;\theta)\,dx$ so $\psi_0 \equiv \psi(\theta_0)$ is the parameter of interest $E(Y)$. Notice that since the likelihood factors as in (1), the maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$ of θ_0 and $\hat{\psi}_{\text{MLE}} \equiv \psi\left(\hat{\theta}_{\text{MLE}}\right)$ of ψ_0 do not depend on whether the true value γ_0 of γ is known or unknown. The same holds true for associated likelihood-based confidence intervals for ψ based either on the Wald interval or on inverting the likelihood ratio or score test. We shall have to refer to any method, such as maximum likelihood, that leads to the same inference, whether γ_0 is known or unknown, to be a strict factorization-based method.

Definition: We define any method of inference for $\psi(\theta_0)$ to be a strict factorization-based (SFB) method if, whenever the likelihood factors as in (1) and θ and γ are variation-independent, inference on $\psi(\theta_0)$ is the same whether the true value γ_0 of γ is known or unknown.

In addition to maximum likelihood, inference based on the profile likelihood for θ (that is, the maximized relative likelihood)[26] or Bayesian inference with independent priors (that is, $f(\theta,\gamma) = f(\theta) f(\gamma)$) are SFB methods. However, Bayesian inference based on dependent priors is not an SFB method,[6,18] since the posterior distribution of θ with γ unknown

$$f(\theta \mid data) = L_1(\theta) \int L_2(\gamma) f(\theta,\gamma)\,d\gamma \Big/ \left\{ \int L_1(\theta) \left[\int L_2(\gamma) f(\theta,\gamma)\,d\gamma \right] d\theta \right\} \tag{4}$$

differs from the posterior distribution of θ when γ is known to be γ_0

$$\begin{aligned} f(\theta \mid data;\gamma_0) &= L_1(\theta) L_2(\gamma_0) f(\theta \mid \gamma_0) \Big/ \left\{ \int L_1(\theta) L_2(\gamma_0) f(\theta \mid \gamma_0)\,d\theta \right\} \\ &= L_1(\theta) f(\theta \mid \gamma_0) \Big/ \left\{ \int L_1(\theta) f(\theta \mid \gamma_0)\,d\theta \right\}. \end{aligned} \tag{5}$$

In particular, the posterior distribution of θ when γ_0 is known depends on the known value. Note, as mentioned above, when θ and γ are *a priori* independent, equations (4) and (5) are the same and equal $L_1(\theta) f(\theta) / \int L_1(\theta) f(\theta) \mathrm{d}\theta$ (which does not depend on γ_0). A Bayesian with dependent priors uses a likelihood-based method of inference; however, it is not an SFB method. We now see that the claim that 'for data obtained by stratified random sampling, knowledge of the randomization probabilities is irrelevant for inference concerning the population mean' essentially amounts to the claim that we should use an SFB method of inference. In particular, Bayesians who make this claim must have independent priors on θ and γ[6].

We next show that the decision to use an SFB method of inference follows as a logical consequence of two more 'fundamental' principles: the 'strict likelihood principle (SLP)' and the 'non-informative nuisance parameter principle (NNPP)' described by Berger and Wolpert (reference 18, p. 42).

> Strict likelihood principle. If two experiments depending on the same parameter θ give proportional likelihoods for θ, then inference about θ should be the same regardless of the experiment generating the data.
>
> Non-informative nuisance parameter principle (NNPP) (Berger and Wolpert[18]). Suppose a model depends on parameters (θ, γ). Suppose further that if γ were precisely known, inference about θ would be the same for all $\gamma \in \gamma^*$. Then we should make the same inference about θ even when γ is unknown.

Theorem 1. The SLP and NNPP imply that one use a SFB method of inference.

Proof. Suppose the likelihood function factors as in (1) with (θ, γ) variation independent. If it were known that $\gamma = \gamma^\dagger$, then $L_2(\gamma)$ would be a known constant $L_2(\gamma^\dagger)$ and the likelihood for any θ would be $L^\dagger(\theta) \equiv L(\theta, \gamma^\dagger) = L_1(\theta) L_2(\gamma^\dagger)$. (Recall that when viewing $L(\theta, \gamma^\dagger)$ as a likelihood function, the data are regarded as fixed constants.) If γ were known to equal γ^Δ, then the likelihood for θ would be $L^\Delta(\theta) = L_1(\theta) L_2(\gamma^\Delta)$. Thus the likelihoods for θ would be proportional with constant of proportionality $L_2(\gamma^\Delta) / L_2(\gamma^\dagger)$. Thus, letting γ^Δ and $\gamma^\dagger$ index the two experiments, by the SLP, inference concerning θ must be the same whether γ was known to be $\gamma^\dagger$ or γ^Δ. $\gamma^\dagger$ and γ^Δ were arbitrary, so inference about θ must be the same when γ is known, whatever be γ. Hence, by the NNPP, inference about θ must be the same even if γ were unknown; that is, the method of inference is SFB.

Remark. A Bayesian with dependent priors will violate the SLP since, although the likelihoods $L^\dagger(\theta)$ and $L^\Delta(\theta)$ will be proportional, the priors for θ, $f(\theta \mid \gamma = \gamma^\dagger)$ and $f(\theta \mid \gamma = \gamma^\Delta)$ would differ depending on whether the experiment was that indexed by $\gamma^\dagger$ or that indexed by γ^Δ. Thus, we might refer to the SLP as the non-Bayesian likelihood principle. The correct likelihood principle for a Bayesian is:

Bayesian likelihood principle. If two experiments depending on a parameter θ give proportional likelihoods for θ and if knowledge of which of the two experiments was performed provides no independent evidence concerning θ, then inference about θ should be the same regardless of the experiment generating the data.

3. THE HORVITZ–THOMPSON ESTIMATOR

When the randomization probabilities $\pi(X_i; \gamma_0)$ are known, the Horvitz–Thompson (HT) estimator of $E(Y)$ is

$$\hat{\psi}_{\mathrm{HT}} \equiv n^{-1} \sum\nolimits_i R_i Y_i / \pi(X_i; \gamma_0). \tag{6}$$

Clearly the HT estimator is not SFB since it is not computable when γ_0 is unknown. Thus it also violates the SLP. Yet the HT estimator is commonly used to analyse data obtained by stratified random sampling. If as we assume, the $\pi(X_i;\gamma_0)$ are bounded away from 0, that is

$$\pi(X_i;\gamma_0) > \sigma > 0 \text{ with probability } 1 \tag{7}$$

so that every subject has a probability greater than a prespecified constant σ (say, 0.05) of being selected to have Y observed, then the HT estimator is unbiased, and, by the central limit theorem, uniformly asymptotically normal and unbiased for $E(Y)$.

Proof: Note the $R_iY_i/\pi(X_i)$ are independent identically distributed random variables with finite variance. Equation. (7) guarantees that the variance is finite. Further, $E[R_iY_i/\pi(X_i)] = E\left[E(R_i \mid Y_i,X_i)\,Y_i/\pi(X_i)\right] = E[Y_i]$, since $E(R_i \mid Y_i,X_i) = \pi(X_i)$ by assumption.

From standard likelihood theory, it follows that the asymptotic variance of $\hat{\psi}_{\mathrm{HT}}$ is greater than or equal to that of $\hat{\psi}_{\mathrm{MLE}}$. Hence, if we can specify a parametric model $p(x;\theta)$, it seems reasonable to adopt the SLP and use SFB methods. After all, $\hat{\psi}_{\mathrm{MLE}}$ is asymptotically as efficient as any other estimator and the performance of $\hat{\psi}_{\mathrm{MLE}}$ in the moderate sized samples occurring in practice would usually be well approximated by its asymptotic properties, when the dimension of θ is not too large.

4. SEMI-PARAMETRIC MODELS

In practice, it will not be possible to correctly specify a parametric model for $E(Y \mid X) \equiv p(X)$ and we would wish to adopt non-parametric models for $p(x)$. If we use random sampling, the randomization probabilities $\pi(x)$ will be known, since they are under the control of the investigator. It seems natural to consider a semi-parametric model in which $\pi(x)$ is a known and satisfies (7) and $p(x)$ is an unknown and unrestricted measurable function of x.

To do so, let the infinite dimensional abstract parameter θ index the measurable (that is, integrable) functions of X (w.r.t. Lesbesgue measure on $[0,1]^k$) with range in $(0,1)$ and let Θ be the set of all such functions. That is, each value of θ denotes a particular measurable function which we write as $p(x;\theta)$. We let θ_0 denote the measurable function $p(x;\theta_0) \equiv p(x)$ where $p(x) \equiv E(Y \mid X = x)$ generates our data Y given $X = x$. Note each θ is not a number or vector; it is a function. However, $\psi(\theta) = \int p(x;\theta)\,dx$ is a number and $\psi(\theta_0) = E(Y)$. Note that each $p(x;\theta)$ must be measurable (integrable) for $\psi(\theta)$ to be well-defined. Similarly, let γ index the measurable functions of X with range on $[\sigma,1]$ (for a prespecified constant σ, say 0.05), and denote the function associated with γ by $\pi(x;\gamma)$. Let Γ be the set of all such functions. Let γ_0 denote the function $\pi(x;\gamma_0) \equiv \pi(x)$ where $\pi(x) \equiv E\left[R \mid X = x\right]$. Thus, since the randomization probabilities are known, $E(Y \mid X)$ is unknown and the data are missing at random, it is natural to consider the semi-parametric model $\mathcal{M}(p)$ with parameter set θ defined as follows.

By definition, any model is just a set of allowable probability distributions for the data. Given that the data are missing at random and that the marginal law of X is known to be uniform, an ordered pair (θ,γ) specifies a possible law for (R,Y,X) with $p(x;\theta)$ indexing $E(Y \mid X = x)$ and $\pi(x;\gamma)$ indexing $E\left[R \mid Y,X = x\right]$. Our model $\mathcal{M}(p)$ is the particular set of laws indexed by $\Theta \times \gamma_0$.

Remark. If readers feel uncertain as to the meaning of the sets Θ and Γ of measurable functions $p(x;\theta)$ and $\pi(x;\gamma)$, it is important to emphasize that every result obtained in this paper remains true if we replace the sets Θ and Γ by their subsets Θ_c and Γ_c consisting of functions $p(x;\theta)$ and $\pi(x;\gamma)$ that are *continuous* in X on $[0,1]$ with the possible exception of a finite set of points.

Our proof that $n^{\frac{1}{2}}(\hat{\psi}_{\mathrm{HT}}-\psi_0)$ is uniformly asymptotically normal and unbiased with finite variance remains valid under semi-parametric model $\mathcal{M}(p)$ so that $\hat{\psi}_{\mathrm{HT}}$ is uniformly $n^{\frac{1}{2}}$-consistent for ψ_0. Thus we have already proved theorem 2.

Theorem 2. (*i*) For each $\gamma_0 \in \Gamma$, $\hat{\psi}_{\mathrm{HT}}$ is $n^{\frac{1}{2}}$-consistent for $\psi_0 \equiv E(Y)$ under model $\mathcal{M}(p)$. (ii) Furthermore, $\hat{\psi}_{\mathrm{HT}}$ is asymptotically normal and unbiased (and thus $n^{\frac{1}{2}}$-consistent) for ψ_0 uniformly for $(\theta_0, \gamma_0) \in \Theta \times \Gamma$.

Remark. Theorem 2 means that $n^{\frac{1}{2}} \mid \hat{\psi}_{\mathrm{HT}}(\gamma_0) - \psi(\theta_0) \mid$ is uniformly bounded in probability for $(\theta_0, \gamma_0) \in \Theta \times \Gamma$ where we have stressed the dependence of $\hat{\psi}_{\mathrm{HT}}$ on γ_0 by writing $\hat{\psi}_{\mathrm{HT}}(\gamma_0)$. Thus for all $\delta > 0$, $\varepsilon > 0$, no matter how small,

$$\sup_{(\theta_0,\gamma_0)\in\Theta\times\Gamma} P_{r_{\theta_0,\gamma_0}}\left\{n^{\frac{1}{2}-\varepsilon} \mid \hat{\psi}_{\mathrm{HT}}(\gamma_0) - \psi(\theta_0) \mid > \delta\right\} \to 0 \quad \text{as} \quad n \to \infty \tag{8}$$

where $P_{r_{\theta_0,\gamma_0}}$ refers to probabilities when the data are generated under the law (θ_0, γ_0).

5. NON-EXISTENCE OF AN $n^{\frac{1}{2}}$-CONSISTENT SFB ESTIMATOR

When $p(x)$ is parametrically modelled, the SFB estimator $\hat{\psi}_{\mathrm{MLE}}$ is uniformly $n^{\frac{1}{2}}$-consistent for ψ_0 over the finite dimensional parameter space $\theta^* \times \gamma^*$. In contrast, we now show there is no SFB estimator in the semi-parametric model $\mathcal{M}(p)$ which achieves an algebraic rate of convergence (much less an $n^{\frac{1}{2}}$-rate) for $E(Y)$ for all γ_0. To do so, it will be useful to consider model $\mathcal{M}(p, \pi)$ in which both $p(x)$ and $\pi(x)$ are unknown and unrestricted except for being measurable and for $\pi(x)$ satisfying (7), that is, model $\mathcal{M}(p, \pi)$ is indexed by the set $\Theta \times \Gamma$. In this big semi-parametric model, the likelihood function $L(\theta, \gamma)$ is still well-defined and given by (1), except now (i) θ indexes the measurable functions $p(x)$, and (ii) γ indexes the measurable functions $\pi(x)$ satisfying (7). Model $\mathcal{M}(p)$ is the submodel of model $\mathcal{M}(p, \pi)$ in which $\pi(x)$ is known and equal to $\pi(x; \gamma_0)$. In the Appendix, we sketch a proof of the following theorem.

Theorem 3. Even when X is univariate, (i) there is no estimator of $\psi_0 = E(Y)$ in model $\mathcal{M}(p, \pi)$ that achieves a (pointwise) rate of convergence better than $(\log\log n)^2 \log n$ (much less an n^{α} rate for any $\alpha > 0$); (ii) further, there is no estimator of ψ_0 that is uniformly consistent; (iii) on the other hand, if $p(x)$ is Riemann integrable (for example, $p(x)$ is continuous in x except at possibly a finite number of points), then there exists a (pointwise) consistent estimator of ψ_0.

Remark 1. Theorem 3(i) states that, given any sequence a_n such that $a_n/[(\log\log n)^2 \log n] \to \infty$ (for example, $a_n = n^{1/1000}$) and any estimator $\hat{\psi}$, there will exist $(\theta_0, \gamma_0) \in \Theta \times \Gamma$ such that $a_n \mid \hat{\psi} - \psi_0 \mid$ does not converge to zero in (θ_0, γ_0)-probability. In fact, we show in the Appendix that at any point $(\theta, \gamma) \in (\Theta \times \Gamma)$, there exists a (θ_0, γ_0) within a distance ε of (θ, γ) (in the sense that, for all x, $\mid \pi(x; \gamma) - \pi(x; \gamma_0) \mid < \varepsilon$ and $\mid p(x; \theta) - p(x; \theta_0) \mid < \varepsilon$) for any ε, no matter how small, such that $\hat{\psi}$ does not converge to ψ_0 at rate a_n under (θ_0, γ_0). This shows that the failure to attain $n^{\frac{1}{2}}$-consistent rates in model $\mathcal{M}(p, \pi)$ is, in this sense, the rule, rather than the exception.

Theorem 3(ii) implies that for any estimator $\hat{\psi}$ there exists $\delta > 0$ such that

$$\sup_{(\theta_0,\gamma_0)\in\Theta\times\Gamma} P_{r_{\theta_0,\gamma_0}}\left\{\mid \hat{\psi} - \psi_0 \mid > \delta\right\} \not\to 0 \quad \text{as} \quad n \to \infty. \tag{9}$$

This result holds even when we restrict consideration to laws within an arbitrarily small distance ε of one another. However, Theorem 3(iii) states that if $p(x)$ is Riemann integrable, there exists $\hat{\psi}$ such that for all $(\theta_0, \gamma_0) \in \Theta \times \Gamma$ and all $\delta > 0$

$$P_{r_{\theta_0,\gamma_0}}\left\{|\hat{\psi} - \psi_0| > \delta\right\} \to 0 \text{ as } n \to \infty. \tag{10}$$

However, as argued in the introduction, since our concern is with the performance of an estimator at a fixed sample size n, uniform rather than pointwise consistency is more relevant. If an estimator is pointwise but not uniformly consistent then, for each sample size n, there will exist laws (θ_0, γ_0) under which $\hat{\psi} - \psi_0$ is not small with high probability.

Remark 2. Relationship to the asymptotics used in the sample survey literature. Heretofore, we have assumed that the law (θ_0, γ_0) generating the data was the same at each sample size n. In the sample survey literature, a common asymptotic theory allows the law (θ_n, γ_n) generating the data at sample size n to vary with n. If (θ_n, γ_n) varies with n in model $\mathscr{M}(p, \pi)$, then (even if $p(x) \equiv p(x; \theta_n)$ is Riemann integrable for each n), no (pointwise) consistent estimator exists. That is, for any estimator $\hat{\psi}$ there exists a sequence of laws $\{(\theta_n, \gamma_n)\}, n = 1, \ldots,$ and a $\delta > 0$ such that

$$P_{r_{\theta_n,\gamma_n}}\left\{|\hat{\psi} - \psi_n| > \delta\right\} \not\to 0 \text{ as } n \to \infty. \tag{11}$$

This follows because (9) and (11) are logically equivalent. In fact, as shown in the Appendix, equation (11) remains true even if we restrict attention to laws (θ_n, γ_n) for which the mean ψ_n of Y is the same fixed constant for each n.

Corollary 3. Even if the dimension of X is 1, (i) for each SFB estimator $\hat{\psi}$, there exists $\gamma_0 \in \Gamma$ such that $\hat{\psi}$ does not achieve a rate of convergence better than $(\log\log n)^2 \log n$ (much less an n^α rate for $\alpha > 0$) in model $\mathscr{M}(p)$. In particular, Theorem 2 is false for any SFB estimator $\hat{\psi}$; the particular γ_0's for which $\hat{\psi}$ fails to achieve a better rate of convergence will vary depending on the particular estimator $\hat{\psi}$. (ii) Furthermore, for any SFB estimator $\hat{\psi}$, equations (9)–(11) remain true. Hence no SFB estimator $\hat{\psi}$ can converge to ψ_0 uniformly over all (θ_0, γ_0) in $\Theta \times \Gamma$.

Proof of Corollary 3: Since, by definition of an SFB estimator, $\hat{\psi}$ must be the same whether γ is completely unknown or known to be γ_0, $\hat{\psi}$ can depend on the data but not on parameter γ_0 generating the data. Hence, $\hat{\psi}$ could be used as an estimator of ψ_0 in model $\mathscr{M}(p, \pi)$. Since $\Theta \times \Gamma$ is the set of probability distributions allowed by the models $\mathscr{M}(p)$ as γ_0 varies in Γ, it follows that if Corollary 3(i) were false, $\hat{\psi}$ would achieve a rate of convergence to ψ_0 better than $(\log\log n)^2 \log n$ at each probability distribution in model $\mathscr{M}(p, \pi) = \Theta \times \Gamma$, contradicting Theorem 3. Corollary 3(ii) follows by a similar argument.

Remark 3. Sample survey asymptotics: If, as in the sample survey literature, we allow (θ_n, γ_n) generating the data to vary with sample size n, the model $\mathscr{M}(p)$ is indexed by sequences $\theta_1 \times \gamma_1, \theta_2 \times \gamma_2, \ldots, \theta_n \times \gamma_n, \ldots$ where $\gamma_1, \gamma_2, \ldots, \gamma_n, \ldots$ is a sequence of known randomization of probabilities and each unknown θ_n lies in Θ. Then Corollary 3(ii) implies, by equation (11), that for each SFB estimator $\hat{\psi}$ of the mean ψ_n of Y there exists a sequence $\gamma_1, \gamma_2, \ldots, \gamma_n$ of randomization probabilities under which $\hat{\psi}$ is inconsistent for ψ_n in this sample survey version of model $\mathscr{M}(p)$.

Theorem 3 might seem surprising to readers familiar with the theory of semi-parametric efficiency bounds since, as we show below, the semi-parametric variance bound for regular estimators of ψ_0 in model $\mathscr{M}(p, \pi)$ is finite. If a semi-parametric model is sufficiently smooth, it will be possible to construct $n^{\frac{1}{2}}$-consistent estimators of any parameter ψ_0 with finite variance bound.

However, our model $\mathcal{M}(p,\pi)$ allows quite unsmooth distributions, since $p(x;\theta)$ and $\pi(x;\gamma)$ need only be measurable. As a consequence, as stated in Theorem 3, construction of an $n^{\frac{1}{2}}$-consistent estimator of ψ_0 is not possible. Ritov and Bickel[20] have previously discussed other examples of semi- and non-parametric models in which the variance bound for a parameter ψ_0 was finite but (even pointwise) $n^{\frac{1}{2}}$-consistent estimators of ψ_0 did not exist due to lack of smoothness.

Construction of uniformly $n^{\frac{1}{2}}$-consistent estimators of ψ_0 would be possible in submodels $\mathcal{M}(p_\sigma,\pi) = \Theta_\sigma \times \Gamma$, $\mathcal{M}(p,\pi_\sigma) = \Theta \times \Gamma_\sigma$ and $\mathcal{M}(p_\sigma,\pi_\sigma) = \Theta_\sigma \times \Gamma_\sigma$ where Θ_σ and Γ_σ are the "smooth" subsets of Θ and Γ of the measurable functions $p(x;\theta)$ and $\pi(x;\gamma)$ with, say, $(k+1)$ bounded derivatives with respect to x. Specifically, consider the candidate estimators $\hat{\psi}_{a1} = \int \hat{p}(x)\,dx, \hat{\psi}_{a2} = n^{-1}\sum_i \hat{p}(X_i)$, and $\hat{\psi}_b = n^{-1}\sum_i R_i Y_i/\hat{\pi}(X_i)$ where $\hat{p}(X)$ and $\hat{\pi}(X)$ are multivariate kernel regression estimators of $p(x) = E(Y \mid x)$ (based on subjects with $R = 1$) and $\pi(X) = E(R \mid X)$. With bandwidths appropriately chosen, $\hat{\psi}_{a1}$ and $\hat{\psi}_{a2}$ will be $n^{\frac{1}{2}}$-consistent for ψ_0 in models $\mathcal{M}(p_\sigma,\pi)$ and $\mathcal{M}(p_\sigma,\pi_\sigma)$ since, by $p(x)$ smooth in x, $\hat{p}(x)$ will converge to $p(x)$ at a sufficient rate to imply an $n^{\frac{1}{2}}$ convergence rate for $\hat{\psi}_{a1}$ and $\hat{\psi}_{a2}$. Similarly, by $\pi(x)$ smooth in x, $\hat{\psi}_b$ will be $n^{\frac{1}{2}}$-consistent in models $\mathcal{M}(p,\pi_\sigma)$ and $\mathcal{M}(p_\sigma,\pi_\sigma)$. See Section 8.2 for additional details.

However, none of the above estimators will converge at an algebraic rate at all $(\theta_0,\gamma_0) \in \Theta \times \Gamma$ since, at distributions where $p(x)$ and $\pi(x)$ are both very non-smooth, $\hat{p}(x)$ and $\hat{\pi}(x)$ will fail to converge to $p(x)$ and $\pi(x)$ at a suitable rate.

5.1. *Proof that the Semi-parametric Variance Bound for ψ_0 in Model $\mathcal{M}(p,\pi)$ is Finite*

Consider model $\mathcal{M}(p)$. According to Theorem 2, $\hat{\psi}_{\mathrm{HT}}$ is an $n^{\frac{1}{2}}$-consistent estimator of ψ_0 with finite variance. Further, $\hat{\psi}_{\mathrm{HT}}$ is a regular estimator. Since the asymptotic variance of a regular $n^{\frac{1}{2}}$-consistent estimator of ψ_0 must exceed the semi-parametric variance bound,[8] it follows that the bound is finite in model $\mathcal{M}(p)$. However, by the factorization (1), the MLE of ψ_0 and thus of the Cramer–Rao variance bound for ψ_0 are the same in any parametric submodel of model $\mathcal{M}(p,\pi)$ (with variation-independent parameters) and do not depend on whether γ is known or unknown. It follows that the semi-parametric variance bound in model $\mathcal{M}(p,\pi)$ must equal that of model $\mathcal{M}(p)$ and thus be finite, since the supremum of Cramer-Rao variance bounds over parametric submodels with variation-independent parameters equals the supremum over all parametric submodels.

5.2. *Philosophical and Practical implications*

We have observed that in the non-parametric model $\mathcal{M}(p,\pi)$ with continuous X, although the semi-parametric variance bound for $\psi_0 = E(Y)$ is finite, no estimator of $\hat{\psi}$ will be $n^{\frac{1}{2}}$-consistent for ψ_0 at all distributions (θ,γ) allowed by the model unless we impose additional local smoothness assumptions on $p(x)$ and/or $\pi(x)$. We believe that the implicit attitude of the literature on semi-parametric models has been to view the imposition of smoothness assumptions on $p(x)$ or $\pi(x)$ as 'mild regularity' conditions (since it is often true that one or both will be smooth), and, thus, to view theorems (such as our Theorem 3 or those of Ritov and Bickel[20]) which demonstrate the non-existence of $n^{\frac{1}{2}}$-consistent estimators in the absence of smoothness as mathematical curiosities. We, however, take the opposite point of view in substantive settings in which X has many continuous components (say 5 or greater) and the number of observations is moderate (say less than 10^5). We argue as follows.

As discussed in the introduction, due to the curse of dimensionality, it is impossible to construct an estimator of $\hat{\psi}$ that performs well in the moderate sized samples occurring in practice (in the sense that under all probability laws allowed by our model, the estimator $\hat{\psi}$ will be approximately

normal and centred on ψ_0). This reflects the fact that $\hat{\psi}$ will not be centred on ψ_0 unless we can obtain an accurate estimate of $p(X) \equiv E(Y \mid X)$ or $\pi(X) \equiv E(R \mid X)$. However, it is not possible to obtain accurate estimates, because estimation of unrestricted conditional expectations given X requires effective smoothing in at least five dimensions; but, due to the curse of dimensionality, essentially no subjects will have X-values close enough to one another to allow the 'borrowing of information' necessary for effective smoothing. This inability to obtain an estimator that performs well in moderate samples will be true even if we admit that $p(x)$ and $\pi(x)$ both have many continuous derivatives – that is, even if we impose model $\mathcal{M}(p_\sigma, \pi_\sigma)$ of Section 5. We take the point of view that an appropriate asymptotic theory is, by definition, one whose results provide guidance to moderate sample performance. Thus, an asymptotic theory that adopts smoothness as 'mild regularity' conditions (that is, replaces model $\mathcal{M}(p, \pi)$ by model $\mathcal{M}(p_\sigma, \pi_\sigma)$ and thus declares that regular $n^{\frac{1}{2}}$-consistent estimators of ψ_0 can be constructed) is not a curse of dimensionality appropriate (CODA) theory. However, an asymptotic theory that assumes no smoothness conditions on $p(x)$ and $\pi(x)$ is CODA. To put it another way, we are suggesting that even when $p(x)$ and $\pi(x)$ are known to be smooth, but, due to the curse of dimensionality, we are unable to effectively smooth the functions without bias in the moderate sample sizes occurring in practice, we should employ an asymptotic theory that does not impose smoothness; that is, for asymptotic calculations, we only assume model $\mathcal{M}(p, \pi)$ even if we know the true law generating the data lies in model $\mathcal{M}(p_\sigma, \pi_\sigma)$.

In this spirit, we define the CODA variance bound for ψ_0 in model $\mathcal{M}(p_\sigma, \pi_\sigma)$ whenever $\pi(x)$ and $p(x)$ cannot be effectively estimated by smoothing to be the minimum asymptotic variance obtainable by any $n^{\frac{1}{2}}$-consistent regular asymptotically linear estimator of ψ_0 in model $\mathcal{M}(p, \pi)$ (the model which imposes no smoothness). This bound is infinite, since in model $\mathcal{M}(p, \pi)$, no $n^{\frac{1}{2}}$-consistent estimator exists and thus differs from the ordinary semi-parametric variance bound.

Remark. It may be preferable to define the bound at a particular distribution (θ, γ) to be the minimum asymptotic variance obtained by a regular asymptotic linear estimator that is $n^{\frac{1}{2}}$-consistent for each law within a distance ε of (θ, γ) with distance as defined in the remark following Theorem 3. This makes it clear that the problem with model $\mathcal{M}(p, \pi)$ is local to each distribution(θ, γ) and not just something that occurs at the boundaries. Under this redefinition, the bound would still be infinite.

Now if $\pi(x) \equiv \pi(x; \gamma_0)$ is known, the CODA variance bound for ψ_0 is finite since in model $\mathcal{M}(p)$, $\hat{\psi}_{\mathrm{HT}}$ is a regular asymptotically linear $n^{\frac{1}{2}}$-consistent estimator with finite asymptotic variance even in the absence of smoothness assumptions on $p(x)$. Thus, although models $\mathcal{M}(p, \pi)$ and $\mathcal{M}(p)$ have identical semi-parametric variance bounds in the ordinary theory, they have quite different CODA variance bounds. In Section 7, we consider a general CODA asymptotic theory for arbitrary semi- and non-parametric models.

6. BAYESIAN PERSPECTIVE

Because of the importance and influence of the Bayesian perspective on both the philosophical foundations of statistics and its practice, we examine here the attitude to our result that a committed subjective Bayesian would adopt. We first provide some technical results concerning the properties of Bayes estimators. We have seen that in model $\mathcal{M}(p)$, any Bayes estimator $\hat{\psi}_{\mathrm{IB}}$ with independent priors on θ and γ would be a strict factorization based (SFB) estimator and thus, for some $\gamma_0 \in \Gamma$, would fail to converge to ψ_0 at an $n^{\frac{1}{2}}$-rate at some law (θ_0, γ_0) allowed by model $\mathcal{M}(p)$. This raises the question of whether there exists a Bayes estimator $\hat{\psi}_{\mathrm{DB}}$ based on (θ, γ) having dependent priors that, for each $\gamma_0 \in \Gamma$, is $n^{\frac{1}{2}}$-consistent on model $\mathcal{M}(p)$. We now show

that the answer is yes. Indeed, given γ_0, let $\hat{\psi}_{MM}(\gamma_0)$ be the minimax estimator of ψ_0 with respect to the squared error loss function $\left\{n^{\frac{1}{2}}(\psi - \psi_0)\right\}^2$. $\hat{\psi}_{MM}(\gamma_0)$ must be $n^{\frac{1}{2}}$-consistent (otherwise it would not be minimax, since $\hat{\psi}_{HT}(\gamma_0)$ is $n^{\frac{1}{2}}$-consistent). However, in our model $\mathcal{M}(p)$, it can be shown that the minimax estimator can be approximated by Bayes estimators, and thus there must exist an approximately minimax prior, $p_{0,\gamma_0}(\theta)$ say. Let $p_0(\theta,\gamma)$ be any prior for (θ,γ) such that the conditional density θ given γ_0 is $p_{0,\gamma_0}(\theta)$, that is, $p_0(\theta \mid \gamma_0) = p_{0,\gamma_0}(\theta)$. Then the Bayes estimator using $p_0(\theta,\gamma)$ is $n^{\frac{1}{2}}$-consistent on model $\mathcal{M}(p)$ for each γ_0.

With this technical background, we next examine conditions under which a committed subjective Bayesian would use independent priors and when the Bayesian would use the approximately 'minimax' dependent prior $p_0(\theta,\gamma)$. Suppose that the randomization probabilities $\pi(x)$ were to be picked by a person A. Suppose our Bayesian B believes that person A may have greater knowledge of θ, that is, of $p(x)$ than she (perhaps because A is a subject matter expert) and, prior to data generation, B queries A concerning A's knowledge of θ and then B appropriately updates his priors concerning θ to $p_B(\theta)$, say. Then B would be expected to have independent priors on θ and γ, that is, $p_B(\theta \mid \gamma_0) = p_B(\theta)$. For even if B believes that A's choice of $\pi(x)$ (that is, of γ_0) depended on A's beliefs and/or knowledge of θ, this would not affect B's beliefs about θ, since B has already fully considered B's beliefs concerning θ in updating his prior to $p_B(\theta)$. (In contrast, if B had not queried A about A's beliefs concerning θ, B would typically have dependent priors.) Let the SFB estimator $\hat{\psi}_{IB}$ be B's Bayes estimate of ψ_0 and suppose B learns that the value of γ_0 selected by A is precisely one of those values for which $\hat{\psi}_{IB}$ fails to be $n^{\frac{1}{2}}$-consistent at a subset $\Theta^\dagger$ (depending on γ_0 and $\hat{\psi}_{IB}$) of Θ. Then B knows that if the true θ is in $\Theta^\dagger$, his estimate $\hat{\psi}_{IB}$ will fail to be $n^{\frac{1}{2}}$-consistent and typically will converge to a value that differs from ψ_0 (that is, B knows he will be inconsistent). However, to B, as a committed subjective Bayesian, this will be of no matter, since $\hat{\psi}_{IB}$ is his optimal estimate and B must have placed essentially zero prior subjective probability on the set $\Theta^\dagger$.That is, the desire to be $n^{\frac{1}{2}}$-consistent (or even consistent) for ψ_0 on all Θ is a 'minimax' criterion and is of no interest to a Bayesian who has subjective probability essentially zero that θ lies in the subset $\Theta^\dagger$.

However, now suppose that just before announcing his Bayes estimate, $\hat{\psi}_{IB}$, B realizes that A might well be lying to him and that: (i) A might know θ exactly, (ii) A might be purposely misleading B about θ; and (iii) A's choice of γ_0 was made so that B's estimate $\hat{\psi}_{IB}$ would do poorly. That is, B now thinks that A is his adversary in a game. To protect himself, even a Bayesian might use the approximately 'minimax' dependent prior $p_0(\theta,\gamma)$ and report the $n^{\frac{1}{2}}$-consistent minimax estimator. More generally, if our Bayesian B thought that, with subjective probability α, A was being straight with him and with subjective probability $1-\alpha$, A was lying and trying to beat him in a game, then to try to protect himself, our Bayesian might use as a prior a mixture of his independent prior and $p_0(\theta,\gamma)$ with mixing parameter α. This is somewhat heuristic, since Bayesian rules of rationality may not apply when there is some subjective probability that one is playing against an adversary. As discussed in the Remark in Section 7.1.1, the above remarks about the attitude that a subjective Bayesian would take apply equally to a two-armed randomized trial with randomization probabilities depending on a vector of continuous pre-treatment covariates X.

We have seen that the minimax estimator is approximately Bayes and is $n^{\frac{1}{2}}$-consistent for ψ_0 on model $\mathcal{M}(p)$ whatever be γ_0. However, minimax estimators are notoriously difficult to compute and may have undesirable performance at many points in the parameter space. This raises the question whether there are other, more easily computable $n^{\frac{1}{2}}$-consistent Bayes estimators that have reasonable performance. The answer to this question is yes. Specifically, Ritov and Robins[19] construct a Bayes estimator $\hat{\psi}_{HTB}$, based on dependent priors, that is asymptotically

equivalent to the Horvitz–Thompson estimator, that is, $n^{\frac{1}{2}}\left(\hat{\psi}_{\mathrm{HTB}}-\hat{\psi}_{\mathrm{HT}}\right)$ converges to zero in probability on $\mathscr{M}(p)$ for each γ_0. $\hat{\psi}_{\mathrm{HTB}}$ is a 'smoothed' version of the propensity score estimator proposed by Rubin.[7] Rubin[7] originally proposed the idea of constructing Bayes estimators based on dependent priors that are guaranteed to be what he called 'randomization valid' (which is essentially equivalent to our criterion that they be $n^{\frac{1}{2}}$-consistent for ψ_0 on $\mathscr{M}(p)$ for each γ_0).

From the above discussion, one might have the misleading impression that robust Bayes estimators are only needed when there is an adversary selecting $\pi(x)$. The following somewhat simplified example shows the general importance of using robust estimators (that is, $n^{\frac{1}{2}}$-consistent estimators). Consider a study in which Brazil is divided into 10,000 geographic regions, each with population 25,000. Within each of the 10,000 regions, 500 inhabitants are randomly selected from the census roles and arranged on a list in random order. In each region on five consecutive nights, a single interviewer is dispatched from the various study centres at 5 p.m. and visits consecutive subjects on the list until 9 p.m. Hence, X is discrete with 10,000 levels $x, x = 1,\ldots,10{,}000$ and is marginally uniform in the sense that, for each x, the probability $X = x$ is $1/10{,}000$ for each of the 10,000 values of x.

Suppose the following complication arose, unforeseen by the study designer A: due to the remoteness of the poorer regions from the study centres and inclement weather, much of the interviewer time is consumed by travel and therefore, as a consequence, the number of interviews completed in 5 days (and thus the sampling fraction) in the remote poor regions was less than 50 per cent of that of affluent regions. Further suppose the outcome Y is a measure of affluence so that, due to this 'complication', $p(x)$, the mean affluence in region x, and $\hat{\pi}(x)$, the fraction of subjects actually interviewed in region x, were highly positively correlated. Indeed, due to inclement weather, $\hat{\pi}(x)$ was 0 for a non-negligible fraction of the regions.

Now suppose that neither the study designer A nor our Bayesian B had strong knowledge or beliefs concerning the relative distribution of affluence among the 10,000 regions. This would be the case if both A and B had available to them only code numbers 1–10,000 denoting the regions but did not know which code number was associated with which region. Further suppose that neither A nor B had the insight nor the imagination to entertain the possibility that there would be a high correlation between $\hat{\pi}(x)$ and $p(x)$ due to some complication. Then, after consulting A and correctly assuming that A is not his adversary, B would indeed have independent priors for $p(x)$ and $\pi(x)$, and his prior for $\{p(1),\ldots,p(10{,}000)\}$ would be exchangeable. Typically, by appeal to DeFinetti's Theorem, one models exchangeable priors using a hierarchical model which assumes $p(1),\ldots,p(10{,}000)$ constitute independent identically distributed draws of p from a distribution $f(p;\delta)$ depending on a parameter δ with prior distribution $f(\delta)$. Let $\hat{p}(x)$ be the posterior mean of $p(x)$ under this prior. Under such a prior, the Bayes estimator of $E(Y)$, $\hat{\psi}_{\mathrm{IB}} = \sum_{x=1}^{10{,}000}\hat{p}(x)$, will in general be badly biased when $\hat{\pi}(x)$ and $p(x)$ are correlated, since, for each region j, $\hat{p}(j)$ will borrow information from (that is, be influenced by) all regions k, $k \neq j$, when properly $\hat{p}(j)$ should only be strongly influenced by regions k for which either $\hat{\pi}(j)$ and $\hat{\pi}(k)$ or $p(j)$ and $p(k)$ are close.

If our Bayesian B had had the imagination to see that the study design might lead to important correlations between $\hat{\pi}(x)$ and $p(x)$, his beliefs about $p(x)$ would have been influenced by knowledge of $\hat{\pi}(x)$ and he or she would not have entertained the independent priors just described.

This example suggests that, even when a Bayesian cannot imagine any mechanism that would result in $\hat{\pi}(x)$ and $p(x)$ being highly associated, he or she should have enough humility to recognize that, due to lack of information and imagination, (i) there may nonetheless be a correlation (positive or negative) between $\hat{\pi}(x)$ and $p(x)$ and attempt to use the data to empirically test such an association.

In this example, the sampling probabilities were not set by design. Therefore, in order to relate this example more closely to our earlier discussion, we might let $\pi(x)$ be the expected selection probabilities in region x which would depend on such fixed factors as the distance of region x from the study centre and the quality of the roads. The observed fraction $\hat{\pi}(x)$ would also be affected by 'independent random' influences, such as the local (region-specific) weather and the particular driver used. Finally, assume that knowledge of $\pi(x)$ can be obtained by consulting an expert who has both (i) knowledge of how road quality and distance jointly determine $\pi(x)$ and (ii) region x-specific data on road quality and distance. Thus the HT estimator or a robust 'dependent' Bayes estimator that uses knowledge of $\pi(x)$ will be uniformly asymptotically normal and unbiased and thus perform well unconditionally (that is, in hypothetical repetitions with $\pi(x)$ fixed but with the observed selection probabilities $\hat{\pi}(x)$ random). Obviously, no estimator can be guaranteed to perform well conditional on the observed sampling fractions $\hat{\pi}(x)$ since a non-negligible fraction of regions x have $\hat{\pi}(x) = 0$ (see Section 9). Further, this example demonstrates how difficult and often artificial the construction of an appropriate statistical model may be when the sampling probabilities are not fixed by design.

7. A CODA ASYMPTOTICS

In Section 7.1 we review the standard asymptotic theory of inference in semi-parametric models. In Section 7.2, we propose a general algorithm for CODA asymptotic inference in semi-parametric models. This algorithm (some readers may prefer pseudo-algorithm, scheme, etc.) applies to any semi-parametric model – not only to models in which the likelihood factors as in (1). In Section 7.3, we provide an informal discussion of how to interpret the output of the CODA algorithm. We then offer some conjectures as to formal mathematical properties of the output of the algorithm. We believe the conjectures to be true for any semi-parametric model but have only proved it for specific models. In Section 7.4, we apply the CODA algorithm to a number of interesting semi-parametric models.

7.1. Standard Semi-parametric Theory

7.1.1. Semi-parametric models

Let $\mathcal{L}(\psi, \rho)$ be the likelihood function with respect to a dominating measure for a single subject in a semi-parametric model indexed by an unknown finite dimensional parameter ψ taking values in $\Psi \subset R^q$ and an unknown nuisance parameter ρ taking values in an infinite dimensional set $\mathcal{R}$ with ρ and ψ locally variation-independent (that is, for any given (ψ, ρ) there is some ψ-neighbourhood in Ψ and ρ-neighbourhood in $\mathcal{R}$ so that any value of ψ and any value of ρ in these neighbourhoods together generate a distribution allowed by our model). We have suppressed the dependence on $\mathcal{L}(\psi, \rho)$ on the observed random variables O. Throughout, we assume that we have obtained n independent and identically distributed realization $O_1, \ldots, O_n$ of O.

Example. Let model $\mathcal{M}(p, \pi, F_X)$ be our continuously stratified random sampling model $\mathcal{M}(p, \pi)$, except now with the marginal law of X and the randomization probabilities both unrestricted and unknown. Then $O = (RY, R, X)$ and there are three unknown functions: $p(x)$, the conditional probability of $Y = 1$ given X; $\pi(x)$, the conditional probability of observing Y (that is, $R = 1$); and the marginal distribution F_X of X. The parameter of interest is still ψ, the expectation of Y. Write $\theta = (\psi, v)$ where v indexes the function $q(x; v) = p(x; \theta) - \psi$. The function $q(x; v)$ is the difference between the conditional mean $p(x; \theta)$ of Y given X and the unconditional mean ψ of Y so $\theta = (\psi, v)$. Let μ be the nuisance parameter corresponding to the law of X. Then $\rho =$

(ν, μ, γ), $\mathscr{L}(\psi, \rho) = \mathscr{L}_1(\psi, \nu, \mu)\mathscr{L}_2(\gamma), \mathscr{L}_1(\psi, \nu, \mu) = \Pr\left[Y \mid X; \psi, \nu\right]^R f(X; \mu)$, and $\mathscr{L}_2(\gamma) = \pi(X; \gamma)^R \{1 - \pi(X; \gamma)\}^{1-R}$. (The parameters μ and ν are restricted by $0 = \int q(x; \nu)\, \mathrm{d}F(x; \mu)$ since $Y - \psi$ has unconditional mean zero under law (ψ, ρ). However, the three sets $(\mu, \nu), \psi$ and γ are mutually locally variation-independent, so that ψ and ρ are locally variation-independent.) In this model, $\mathscr{L}_2(\gamma)$ is an 'ancillary process' for ψ because of the likelihood factorization[25]. However, (R, X) are not S-ancillary for ψ.

Example. As a second example, let $\mathscr{M}(p, F_X)$ be model $\mathscr{M}(p, \pi, F_X)$ but with γ_0 (that is, $\pi(x)$) known, so $\rho = (\nu, \mu)$ and $\mathscr{L}(\psi, \rho) = \mathscr{L}(\psi, \rho, \gamma_0) = \mathscr{L}_1(\psi, \nu, \mu)\mathscr{L}_2(\gamma_0)$.

Example. As a third example, suppose the marginal law of X is known but γ_0 is unknown. This model is model $\mathscr{M}(p, \pi)$ of Section 2, $\rho = (\nu, \gamma)$ since μ_0 is known, and the likelihood is $\mathscr{L}(\psi, \rho) \equiv \mathscr{L}(\psi, \rho, \mu_0) = \mathscr{L}_1(\psi, \nu, \mu_0)\mathscr{L}_2(\gamma)$.

Example. As a final example, suppose γ_0 is also known; then this model is $\mathscr{M}(p)$ of Section 2, and $\rho = \nu$, $\mathscr{L}(\psi, \rho) \equiv \mathscr{L}(\psi, \rho, \mu_0, \gamma_0) = \mathscr{L}_1(\psi, \nu, \mu_0)\mathscr{L}_2(\gamma_0)$.

Remark. Note that the model $M(p, F_X)$ will also serve as an appropriate superpopulation model for analysing a dichotomous outcome Y in treatment arm $R = 1$ of a two-arm randomized trial with known randomization probabilities $\Pr\left[R = 1 \mid X\right]$ depending on a vector of covariates X. Further, the outcome in the other treatment arm can be analysed under the same model except with treatment arm $R = 0$ replacing treatment arm $R = 1$. Finally, an estimate of the average treatment effect can be obtained by taking the difference between the estimates $\hat{\psi}$ of the treatment arm-specific means ψ. Thus, all the results in this paper obtained for model $M(p, F_X)$ also obtain for a two-armed randomized trial with the average treatment effect as the parameter of interest. See the Remark in Section 8.1 for details.

7.1.2. Parametric Submodels

A parametric submodel of a semi-parametric model imposes the additional restriction that the parameter ρ is a 1-1 function $\tau(\eta)$ of a finite-dimensional Euclidean parameter η and thus lives in a finite-dimensional subspace of $\mathscr{R}$. Given a parametric submodel, we can write the nuisance score $S_\eta(\psi, \rho) \equiv \partial \log \mathscr{L}(\psi, \tau(\eta))/\partial\eta$ as a function of ρ, since $\tau(\eta)$ is 1 - 1 and thus invertible. Let $A = a(O)$ be any fixed vector-valued function $a(\cdot)$ of the data O of the same dimension q as ψ that, under law (ψ, ρ), (i) has mean zero and (ii) is uncorrelated with the nuisance score $S_\eta(\psi, \rho)$ for all parametric submodels. That is, $E_{\psi,\rho}(A) = 0$ and $E_{\psi,\rho}\left[AS_\eta(\psi, \rho)\right] = 0$ for all $S_\eta(\psi, \rho)$ where $E_{\psi,\rho}$ indicates expectations computed under the law (ψ, ρ). Let $\Lambda^\perp \equiv \{A\}$ be the set of all such A. This set of random variables is called the orthogonal complement to the nuisance tangent space at law (ψ, ρ) in the semi-parametric literature, since it is a collection of random variables uncorrelated with (that is, orthogonal to) any nuisance score. Since whether a particular random variable A is in $\Lambda^\perp$ depends on the law generating the data, that is, on (ψ, ρ), we write $\Lambda^\perp(\psi, \rho) = \{A(\psi, \rho)\}$ with $A(\psi, \rho) \equiv a(O; \psi, \rho)$ to emphasize the dependence on the law (ψ, ρ).

Example. Model $\mathscr{M}(p, F_X)$ with $\rho = (\nu, \mu)$. Henceforth, let $g(x)$ be an arbitrary function of x and b an arbitrary constant. Propositions (8.1) and (8.2) of Robins et al.[3] imply that

$$\begin{aligned} &\Lambda^\perp(\psi, \rho) \equiv \Lambda^\perp(\psi) = \{bA_1(g, \psi, \gamma_0)\} \\ &\text{with } A_1(g, \psi, \gamma_0) \equiv A_1(g, \psi) \equiv \\ &[R/\pi(X; \gamma_0)]^{-1}(Y - \psi) - [R - \pi(X; \gamma_0)][\pi(X; \gamma_0)]^{-1} g(X). \end{aligned} \tag{12}$$

The reader who is unfamiliar with how to calculate $\Lambda^{\perp}(\psi,\rho)$ using methods described in references 3 or 8 need not be concerned, since we will always provide the results of the necessary calculations.

We observe that, up to a constant b, $\Lambda^{\perp}(\psi)$ is indexed by arbitrary functions $g(X)$. More precisely, $\Lambda^{\perp}(\psi)$ is formed by subtracting the mean ψ from Y, multiplying by the indicator variable R for complete data, dividing by the known conditional probability of observing complete data given the data Y and X, that is, $\pi(X;\gamma_0) = \Pr(R = 1 \mid Y,X)$, and then subtracting R minus its conditional mean $\pi(X;\gamma_0)$, dividing by $\pi(X;\gamma_0)$ and multiplying by an arbitrary function $g(X)$ of X. Note $\Lambda^{\perp}(\psi)$ does not depend on the unknown $\rho = (\mu,\nu)$ but it does depend on the known γ_0.

7.1.3. Unbiased estimating functions

When $A(\psi,\rho) = A(\psi)$ depends only on ψ and not on ρ, $E_{\psi,\rho}[A(\psi)] = 0$. Hence, $A(\psi)$ is an unbiased estimating function. It follows from the general theory of unbiased estimating functions that, under regularity conditions, there exists (i) an $n^{\frac{1}{2}}$-consistent estimator $\hat{\psi}(A)$ solving $\sum_{i=1}^{n} A_i(\psi) = 0$ with $A_i(\psi) \equiv a(O_i;\psi)$, and (ii) $\hat{\psi}(A)$ is a regular asymptotically linear (RAL) estimator of ψ with influence function $\Upsilon^{-1}(\psi,\rho)A(\psi)$ when (ψ,ρ) generates the data, where $\Upsilon(\psi,\rho) \equiv \Upsilon(A,\psi,\rho) = -\partial E_{\psi,\rho}[A(\psi^*)]/\partial\psi^*$ evaluated at $\psi^* = \psi$. The above constitutes a slight abuse of notation. Technically $\hat{\psi}(A)$ and $\Upsilon(A,\psi,\rho)$ are functions of a $(.;.)$ not A.

Remark. By definition, an estimator $\hat{\psi}$ of ψ is asymptotically linear with influence function $B(\psi,\rho) \equiv b(O;\psi,\rho)$ if, under the law (ψ,ρ), $n^{\frac{1}{2}}(\hat{\psi} - \psi) = n^{-\frac{1}{2}}\sum_i B_i(\psi,\rho) + o_p(1)$ with $E_{\psi,\rho}[B(\psi,\rho)] = 0$ and $E\left[B(\psi,\rho)'B(\psi,\rho)\right] < \infty$. If $\hat{\psi}$ is asymptotically linear it is $n^{\frac{1}{2}}$-consistent since, by the Central Limit Theorem and Slutzky's Theorem, $n^{\frac{1}{2}}\left(\hat{\psi} - \psi\right)$ is asymptotically normal with mean zero and variance $E\left[B(\psi,\rho)^{\otimes 2}\right]$. Conversely, Bickel *et al.*[8] show that the influence function of any RAL estimator lies in the set $\left\{[\Upsilon(\psi,\rho)]^{-1}A(\psi)\right\}$ so $\left\{\hat{\psi}(A); A(\psi) \in \Lambda^{\perp}(\psi)\right\}$ is, up to asymptotic equivalence, the set of all RAL estimators of ψ, where two RAL estimators are asymptotically equivalent if they have the same influence function.

Example. In model $\mathcal{M}(p,F_X)$, the solution to $0 = \sum_i A_{1i}(g,\psi)$ is $\hat{\psi}(g) = \sum_i \pi(X_i;\gamma_0)^{-1}\{R_iY_i - (R_i - \pi(X_i;\gamma_0))g(X_i)\} / \left\{\sum_i \pi(X_i;\gamma_0)^{-1}R_i\right\}$. The HT estimator of ψ_0 is asymptotically equivalent to $\hat{\psi}(g)$ with $g(X) \equiv \psi$.

Example. In model $\mathcal{M}(p)$, $\rho = \nu$, γ_0 and μ_0 are known, and

$$\Lambda^{\perp}(\psi,\rho) \equiv \Lambda^{\perp}(\psi) = \{bA_2(g,g^*,\psi,\gamma_0,\mu_0)\}$$

where for arbitrary functions $g(x)$ and $g^*(x)$

$$A_2(g,g^*,\psi,\gamma_0,\mu_0) = \left[A_1(g,\psi,\gamma_0) - g^*(X) + \int g^*(x)f(x;\mu_0)\,dx\right].$$

Thus knowledge of the marginal distribution of X in model $\mathcal{M}(p)$ implies there are more unbiased estimating functions and thus more RAL estimators than in model $\mathcal{M}(p,F_X)$.

7.1.4. Nuisance parameters

In many models, $A(\psi,\rho)$ will depend on a function of ρ, that is, $A(\psi,\rho) = A(\psi,r)$ where $r \equiv r(\rho)$ is a function of ρ. We assume $r \equiv r(\rho)$ is minimal in the sense that $A(\psi,\rho_1) = A(\psi,\rho_2) \Leftrightarrow r(\rho_1) = r(\rho_2)$ so that r is uniquely defined up to a $1-1$ transformation.

Example. Model $\mathcal{M}(p,\pi,F_X)$ with $\rho=(\mu,v,\gamma)$. Results in Ref. 3 imply that

$$\Lambda^{\perp}(\psi,\rho)=\Lambda^{\perp}(\psi,r)=\{bA_1(q(v),\psi,\gamma)\}, \tag{13}$$

where $q(X,v)\equiv p(X;\theta)-\psi$ is the difference between the conditional mean of Y given X and the unconditional mean of Y.

Remark. Often, the form of $\Lambda^{\perp}(\psi,\rho)$ does not depend on whether additional smoothness restrictions are imposed and/or whether the data are discrete or continuous. For example, $\Lambda^{\perp}(\psi,\rho)$ is given by (13) in the more restricted smooth model $\mathcal{M}(p_\sigma,\pi_\sigma,F_X)$ where $\mathcal{M}(p_\sigma,\pi_\sigma,F_X)$ is model $\mathcal{M}(p_\sigma,\pi_\sigma)$ of Section 5, except the law of X is unrestricted. That is, $\mathcal{M}(p_\sigma,\pi_\sigma,F_X)$ differs from model $\mathcal{M}(p,\pi,F_X)$ only in that $\pi(x)$ and $p(x)=E(Y\mid x)$ are now restricted to being smooth functions. Similarly, if X were discrete rather than being multivariate with continuous components, $\Lambda^{\perp}(\psi,\rho)$ for model $\mathcal{M}(p,\pi,F_X)$ still equals (13).

As in the case in which $\Lambda^{\perp}(\psi,\rho)=\Lambda^{\perp}(\psi)$, the influence function of any RAL estimator when $\Lambda^{\perp}(\psi,\rho)=\Lambda^{\perp}(\psi,r)$ lies in

$$\Lambda^{\perp *}(\psi,\rho)\equiv\left\{A^*(A,\psi,\rho)\equiv\Upsilon^{-1}(A,\psi,\rho)A(\psi,r)\right\}$$

where $\Upsilon^{-1}(A,\psi,\rho)\equiv-\partial E_{\psi\rho}[A(\psi^*,r)]/\partial\psi^*$ evaluated at $\psi^*=\psi$. Conversely, under sufficient smoothness conditions, for all $A^*(A,\psi,\rho)$ in $\Lambda^{\perp *}(\psi,\rho)$ there will exist an RAL estimator with that influence function. When such an estimator exists, it can essentially be obtained by solving $0=\sum_i A_i(\psi,\hat r(\psi))$ where $\hat r(\psi)$ is an appropriate smooth estimator of r that may depend on ψ. (In addition, when such an estimator exists, replacing r by its estimate $\hat r(\psi)$ does not affect the asymptotic distribution, that is, the influence function, of the estimator.) However, in the absence of smoothness, no RAL estimator with this influence function may exist. For example, according to Theorem 3, when X is continuous and no smoothness assumptions are imposed, no $n^{\frac{1}{2}}$-consistent estimator of ψ_0 exists in model $\mathcal{M}(p,\pi)$ and thus in model $\mathcal{M}(p,\pi,F_X)$ (since model $\mathcal{M}(p,\pi,F_X)$ is even less restrictive). Thus, since all RAL estimators are $n^{\frac{1}{2}}$-consistent, no RAL estimator exists in model $\mathcal{M}(p,\pi,F_X)$. A goal of the CODA algorithm will be to determine which $A^*(A,\psi,r)$ are influence functions of RAL estimators in the absence of smoothness assumptions. In model $\mathcal{M}(p,\pi,F_X)$ when X is discrete, with only a moderate number of levels, we would expect an RAL estimator to exist even without additional smoothness or modelling assumptions, since for subjects without Y observed, there will be many other subjects with the same value of X from whom information concerning the law of Y given X can be borrowed.

Example. In model $\mathcal{M}(p,\pi,F_X)$ when X is discrete, $\hat\psi$ solves $0=\sum_i A_i(\psi,\hat r(\psi))=\sum_i\{1/\pi(X_i;\hat\gamma)\}\left[R_i(Y_i-\psi)-\{R_i-\pi(X_i;\hat\gamma)\}q(X_i;\hat v(\psi))\right]$ where $q(X_i;\hat v(\psi))=\hat E(Y\mid X_i)-\psi$ and $\hat E(Y\mid X_i)$ is the sample average of Y among subjects with $R=1$ and $X=X_i$; and $\pi(X_i;\hat\gamma)$ is the sample average of R among subjects with $X=X_i$. Solving for $\hat\psi$, we obtain

$$\hat\psi=n^{-1}\sum_i\pi(X_i;\hat\gamma)^{-1}\left\{R_iY_i-[R_i-\pi(X_i;\hat\gamma)]\hat E(Y\mid X_i)\right\}. \tag{14}$$

In model $\mathcal{M}(p_\sigma,\pi_s,F_X)$, with X one-dimensional and continuous, we would use kernel or histogram estimators $\hat E(Y\mid X_i)$ and $\pi(X_i;\hat\gamma)$, provided we believe that, at the sample size available, we can effectively estimate the conditional expectations by smoothing. Further discussion of this example is provided in Section 8.2.

7.1.5. Efficient influence functions and efficient scores

There will exist an element $\mathrm{EIF}(\psi,\rho)$ of $\Lambda^{\perp *}(\psi,\rho)$ called the efficient influence function (EIF) that has minimal variance under the law (ψ,ρ), in the positive definite sense.[8] Since $\Lambda^{\perp *}(\psi,\rho)$ contains the influence functions of all RAL estimators, no RAL estimator can have asymptotic variance less than $I^{-1}(\psi,\rho) \equiv E_{\psi,\rho}\left[\mathrm{EIF}(\psi,\rho)^{\otimes 2}\right]$. Bickel *et al.*[8] show that (i) no $n^{\frac{1}{2}}$-consistent regular estimator can have asymptotic variance less than $I^{-1}(\psi,\rho)$, (ii) $I^{-1}(\psi,\rho)$, which we call the semi-parametric variance bound, is the supremum of the Cramer–Rao variance bounds for ψ over all parametric submodels $\mathscr{L}(\psi,\rho(\eta))$, and (iii) any regular estimator that attains the bound must be asymptotically linear. Of course, as we have seen, there may be no estimator whose asymptotic variance actually attains the bound.

Furthermore, if $\mathscr{L}(\psi,\rho) = \mathscr{L}(\psi,\nu,\mu)\,\mathscr{L}(\gamma)$, the efficient influence function is the same whether γ is or is not known, so the EIF (ψ,ρ) is the same in models $\mathscr{M}(p,\pi,F_X)$ and $\mathscr{M}(p,F_X)$ and equals $A^*\{A_1(q(v),\psi,\gamma),\psi,\rho\} = A_1(q(v),\psi,\gamma)$. Similarly the efficient influence function in models $\mathscr{M}(p,\pi)$ and $\mathscr{M}(p)$ is the same and equals $A^*\{A_2[q(v),q(v),\psi,\gamma,\mu_0],\psi,\rho\} = A_2[q(v),q(v),\psi,\gamma,\mu_0]$ but differs from the EIF in model $\mathscr{M}(p,\pi,F_X)$ and $\mathscr{M}(p,F_X)$ since the likelihood as a function of μ and ψ does not factor.

Further, there exists a unique element $\mathrm{ES}(\psi,\rho)$ of $\Lambda^{\perp}(\psi,\rho)$ known as the efficient score such that the variance $E_{\psi,\rho}\left[\mathrm{ES}(\psi,\rho)^{\otimes 2}\right]$ of $\mathrm{ES}(\psi,\rho)$ is the inverse $I(\psi,\rho)$ of the semiparametric variance bound. In fact, $EIF(\psi,\rho) = E_{\psi,\rho}\left[\mathrm{ES}(\psi,\rho)^{\otimes 2}\right]^{-1}\mathrm{ES}(\psi,\rho)$ so that EIF and ES are proportional.

Note in model $M(p,F_X)$, $A_1(q(v),\psi,\gamma_0)$ and thus $\mathrm{ES}(\psi,\rho)$ depends on v, since the particular choice of g such that $A_1(g,\psi,\gamma_0)$ is proportional to $\mathrm{ES}(\psi,\rho)$ depends on the parameter v generating the data. Define $r_e(\rho)$ to be the function of ρ on which $\mathrm{ES}(\psi,\rho)$ depends. Note that $r = r(\rho)$ as defined in Section 7.1.4 will also always be a function of r_e. For example, in model $M(p,F_X)$, $r = \emptyset$, but $r_e = v$. In model $M(p,\pi,F_X)$, $r = r_e = (v,\gamma)$.

This concludes a review of the standard theory of asymptotic inference in semi-parametric models. We now present the CODA algorithm.

7.2. The CODA Algorithm

CODA algorithm step 1: given a semi-parametric model $\mathscr{L}(\psi,\rho)$, calculate $\Lambda^{\perp}(\psi,\rho) = \Lambda^{\perp}(\psi,r)$, $r = r(\rho)$.

This is accomplished using methods described in Chapters 3 and 4 of reference 8 or for missing data models in reference 3. Again, the reader of this paper need not know how to carry out step 1.

CODA algorithm step 2: decompose r into components (h,s) where h, taking values in a set $\mathscr{H}$, represents the components of r that cannot be well-estimated in moderate size samples due to a combination of lack of smoothness and/ or the curse of dimensionality and s, taking values in $\mathscr{S}$, represents the components that can be. This step is the subjective step in the algorithm as the following example shows.

Example: Consider model $\mathscr{M}(p,\pi,F_X)$ with $r = (v,\gamma)$ and $\Lambda^{\perp}(\psi,r)$ as in (13). If $X = (X_1,\ldots,X_k)$ is high-dimensional $(k = 5)$ with continuous components, we choose $h = (v,\gamma), s = \emptyset$. If each X_k were dichotomous with $k = 5$, we would choose $h = \emptyset, s = (v,\gamma)$. However, even if the X_k were dichotomous but k were 50, we would choose $h = (v,\gamma)$ and $s = \emptyset$. Similarly, if we were in model $\mathscr{M}(p_\sigma,\pi_\sigma,F_X)$ so $\pi(x)$ and $p(x) = E(Y \mid x)$ were smooth, nonetheless, if $k = 5$, we would choose $h = (v,\gamma)$ and $s = \emptyset$. However, if k were one-dimensional, we would choose

$h = \emptyset$ and $s = (v, \gamma)$. Similarly, if we were in model $\mathcal{M}(p, \pi_\sigma, F_X)$ in which $\pi(x)$ is smooth but $p(x)$ need not be, and $k = 1$, we would choose $h = v$ and $s = \gamma$. In model $\mathcal{M}(p_\sigma, \pi, F_X)$ in which $p(x)$ is smooth but $\pi(x)$ need not be and $k = 1$, we would choose $h = \gamma$ and $s = v$. For borderline cases, such as X_k dichotomous but $k = 9$ (so there are 512 levels of X), it is subjective and debatable whether one can successfully borrow information if, say, the sample size is 30,000–40,000.

CODA algorithm Step 3 (a): define

$$\Lambda^{\perp}_{\text{CODA}}(\psi, r) = \left\{ A(\psi, r) \equiv A(\psi, h, s) \in \Lambda^{\perp}(\psi, r); E_{\psi,h,s}\left[A(\psi, h^*, s)\right] = 0 \text{ for all } h^* \in \mathcal{H} \right\}.$$

Remark: $\Lambda^{\perp}_{\text{CODA}}(\psi, r)$ is the subspace of the orthogonal complement $\Lambda^{\perp}(\psi, r)$ of the nuisance tangent space that has mean zero under law (ψ, h, s) even if we misspecify h by the arbitrary value h^*.

CODA algorithm step 3 (b): define the CODA set of influence functions to be

$$\Lambda^{\perp *}_{\text{CODA}}(\psi, \rho) = \left\{ A^*(A, \psi, \rho) \in \Lambda^{\perp *}(\psi, \rho); A \equiv A(\psi, r) \in \Lambda^{\perp}_{\text{CODA}}(\psi, r) \right\}.$$

CODA algorithm step 3 (c): if $\Lambda^{\perp *}_{\text{CODA}}(\psi, \rho)$ is non-empty, then define (i) the CODA efficient influence function, $\text{EIF}_{\text{CODA}}(\psi, \rho)$, to be the unique minimizer (in the positive definite sense) of $E_{(\psi,\rho)}\left[A^*(A, \psi, \rho)^{\otimes 2}\right], A^*(A, \psi, \rho) \in \Lambda^{\perp *}_{\text{CODA}}(\psi, \rho)$, and (ii) the CODA efficient score $\text{ES}_{\text{CODA}}(\psi, \rho)$ to be the unique element of $\Lambda^{\perp *}_{\text{CODA}}(\psi, r)$ with variance equal to the inverse of the variance of $\text{EIF}_{\text{CODA}}(\psi, \rho)$.

Technical remark: Uniqueness of the minimizer follows by a Hilbert space projection argument, since $\Lambda^{\perp}_{\text{CODA}}(\psi, \rho)$ is a closed linear subspace of the Hilbert space $\Lambda^{\perp}(\psi, r)$.

CODA algorithm Step 3 (d): define the CODA variance bound $I^{-1}_{\text{CODA}}(\psi, \rho)$ to be the variance $E_{\psi,\rho}[\text{EIF}_{\text{CODA}}(\psi, \rho)^{\otimes 2}]$ of $\text{EIF}_{\text{CODA}}(\psi, \rho)$ if $\Lambda^{\perp *}_{\text{CODA}}(\psi, \rho)$ is non-empty, and to be infinite if $\Lambda^{\perp *}_{\text{CODA}}(\psi, \rho)$ is empty.

7.3. Interpretation of CODA Algorithm Output

As yet, we have not succeeded in proving any general theorems concerning the interpretation of the output of the CODA algorithm. An eventual goal is to be able to formalize and prove the following informal conjectures (i)–(iv) under appropriate regularity conditions.

(i) If the set $\mathcal{H}$ is suitably large and unsmooth, then no uniformly consistent estimator of ψ exists when the CODA variance bound is infinite.

Technical remark: By $\mathcal{H}$ large and unsmooth, we mean that there is no norm $\| \ \|$ that can be placed on $\mathcal{H}$ such that both (a) $A(\psi, h, s)$ is L_2-continuous in h and (b) there exists an estimator $\hat{h}$ of h that is either uniformly consistent over $\Psi \times \mathcal{R}$ or converges to h at some rate for all $(\psi, \rho) \in \Psi \times \mathcal{R}$. Nonetheless, even when $\mathcal{H}$ is large and unsmooth, it will generally be possible to find a norm for which $A(\psi, h, s)$ is continuous in h and h can be consistently estimated, that is there exists $\hat{h}$ such that $\|\hat{h} - h\|$ converges to zero in probability as $n \to \infty$ (although $\hat{h}$ will not converge uniformly to h).

(ii) The influence function of any $n^{\frac{1}{2}}$-consistent RAL estimator of ψ will lie in $\Lambda^{\perp *}_{\text{CODA}}(\psi, \rho)$ and will have asymptotic variance greater than or equal to $I^{-1}_{\text{CODA}}(\psi, \rho)$.

(iii) For any $A^*(A, \psi, \rho) \subset \Lambda^{\perp *}_{\text{CODA}}(\psi, \rho)$, if $\mathcal{S}$ is a sufficiently smooth class, there will exist an RAL estimator $\hat{\psi}$ with influence function $A^*(A, \psi, \rho)$. Specifically, given $A(\psi, h, s) \in \Lambda^{\perp}_{\text{CODA}}(\psi, r)$ and any choice of $h^* \in \mathcal{H}$, $\hat{\psi}$ solving $0 = \sum_i A_i(\psi, h^*, \hat{s}(\psi))$ will be such an estimator, where

$\hat{s}(\psi)$ is an appropriately smoothed estimator of s that converges to s under (ψ,ρ) at a suitable rate.

Technical remark: Newey et al.[27] show that $\hat{\psi}$ solving $0 = \sum_i A_i(\psi,\hat{h}(\psi),\hat{s}(\psi))$ where $\|\hat{h}(\psi) - h\|$ converges to zero in probability would generally be a RAL estimator of ψ under less restrictive smoothness conditions on $\mathscr{S}$ than $\hat{\psi}$ solving $0 = \sum_i A_i(\psi,h^*,\hat{s}(\psi))$.

(iv) Let $r_{ce} \equiv r_{ce}(\rho)$ be the minimal function of ρ on which the CODA efficient score $ES_{CODA}(\psi,\rho) = ES_{CODA}(\psi,r_{ce})$ depends. Further let $r_{ce} = (h_{ce},s_{ce})$ where $s_{ce} \in \mathscr{S}_{ce}$ are the components of r_{ce} that can be effectively smoothed in moderate samples and $h_{ce} \in \mathscr{H}_{ce}$ are the components that cannot be. Then, if the CODA variance bound is finite and $\mathscr{S}_{ce}$ is a sufficiently smooth class, there will exist a solution $\hat{\psi}_{eff}$ to $\sum_i ES_{CODA,i}(\psi,\hat{h}_{ce}(\psi),\hat{s}_{ce}(\psi)) = 0$ that will be RAL with the influence function $EIF(\psi,\rho)$ for all (ψ,ρ). Here $\hat{h}_{ce}(\psi)$ is consistent for h_{ce} under a suitable norm and $\hat{s}_{ce}(\psi)$ converges to s_{ce} at a suitable rate.

Remark: The informal conjectures (ii)–(iv) all refer to RAL estimators. As discussed in the Introduction, in order to guarantee good performance at a particular (hopefully moderate) sample size for all laws (ψ,ρ), we need to be concerned with the uniform properties of $\hat{\psi}$ over all $\Psi \times \mathscr{R}$; however, RAL estimators need not have good uniform performance. Section 8 considers additional informal conjectures about the interpretation of the output of the CODA algorithms in terms of uniform performance of estimators.

7.4. *Applications of CODA Algorithm:*

7.4.1. *Example 1: Continuously stratified random sampling*

In model $\mathscr{M}(p,F_X)$, $\Lambda^{\perp}(\psi,\rho) = \Lambda^{\perp}(\psi)$ given in equation (12). Hence, $r = h = s = \emptyset$. Thus, CODA and standard asymptotics are identical in the sense that $\Lambda^{\perp}(\psi) = \Lambda^{\perp}_{CODA}(\psi), \Lambda^{\perp *}(\psi,\rho) = \Lambda^{\perp *}_{CODA}(\psi,\rho)$, and $I^{-1}_{CODA}(\psi,\rho) = I^{-1}(\psi,\rho)$.

In contrast, in model $\mathscr{M}(p,\pi,F_X)$ with $r = h = (v,\gamma)$, it follows from equation (13), upon setting $b = 1$, and defining $\varepsilon(\psi) = Y - \psi$ so $q(X;v) = E[\varepsilon(\psi) \mid X;v]$, that

$$
\begin{aligned}
&E_{\psi,\rho}[A(\psi,h^*)] \equiv E_{\psi,\rho}[A_1\{q(v^*),\psi,\gamma^*\}] = \\
&E_{\psi,\rho}\left[\varepsilon(\psi) + \{R - \pi(X;\gamma^*)\}\pi(X;\gamma^*)^{-1}\{\varepsilon(\psi) - E[\varepsilon(\psi) \mid X;v^*]\}\right] = \\
&E_{\psi,\rho}\left[\pi(X;\gamma^*)^{-1}\{\pi(X;\gamma) - \pi(X;\gamma^*)\}\{E[\varepsilon(\psi) \mid X;v] - E[\varepsilon(\psi) \mid X;v^*]\}\right]
\end{aligned}
\tag{15}
$$

which is not equal to zero for all h^*. Thus, $\Lambda^{\perp *}_{CODA}(\psi,\rho)$ is empty and $I^{-1}_{CODA}(\psi,\rho)$ is infinite.

Analogous calculations in models $\mathscr{M}(p)$ and $\mathscr{M}(p,\pi)$ show that (i) in model $\mathscr{M}(p)$ CODA asymptotics are standard asymptotics, and (ii) $I^{-1}(\psi,\rho)$ is infinite in model $\mathscr{M}(p,\pi)$. These results are in accord with the results obtained in Section 5, indicating that the informal conjectures described above are appropriate for these models. Suppose, now, in model $\mathscr{M}(p,\pi,F_X)$, X is discrete with only a few levels, and thus we choose $h = \emptyset$ and $s = r$; we immediately obtain that CODA and standard asymptotics are the same. Next, assume X is univariate and continuous and consider model $\mathscr{M}(p_\sigma,\pi,F_X)$ which assumes $E(Y \mid X)$ but not $\pi(X)$ is smooth. Then, $h = \gamma$ and $s = v$, and $E_{\psi,\rho}[A(\psi,h^*,s)]$ is given by equation (15) except with $v^* = v$ and is thus equal to 0. So CODA and standard asymptotics are the same. Similarly, consider the model $\mathscr{M}(p,\pi_\sigma,X_F)$ with X univariate, in which case we would choose $h = v$ and $s = \gamma$. Then $E_{\psi,\rho}[A(\psi,h^*,s)]$ is given by equation (15) with $\gamma^* = \gamma$ and is thus equal to 0. Hence CODA and standard asymptotics are identical. This model is discussed further in Section 8.

7.4.2. Example 2: Surrogate marker data in regression:

Following Pepe[9] and Rotnitzky and Robins,[10] consider the model $\mathcal{M}(p_{\text{sur}}, \pi, F_X)$ which assumes that: (i) $E(Y \mid X) = \text{expit}(X\psi)$ where $\text{expit}(X\psi) = \{1 + \exp[-\psi_1 - \psi_2' X]\}^{-1}$ is the logistic function; (ii) the available data are independent identically distributed random vectors $(R_i, R_i Y_i, X_i, V_i), i = 1, \ldots, n$ where V is a surrogate variable highly correlated with Y and available on all study subjects; and (iii) $E[R \mid Y, X, V] = E(R \mid X) \equiv \pi(X)$ so missingness in Y depends only on X. Our goal is the estimation of ψ. When data on V are unavailable, the most efficient estimator of ψ is the logistic regression estimator $\hat{\psi}_{\text{cc}}$ of Y on X among the complete cases, that is, subjects with $R = 1$. We will show that, when data on V are available and $\pi(X)$ is known or can be estimated effectively, $\hat{\psi}_{\text{cc}}$ can be quite inefficient. However, if $\pi(X)$ is completely unknown, then, due to the curse of dimensionality, $\hat{\psi}_{\text{cc}}$ is the most efficient estimator of ψ_0 that performs well in moderate size samples even in the presence of a surrogate V highly correlated with Y.

Redefine ν to be the infinite dimensional parameter corresponding to the law of V given (Y, X). With $\rho = (\nu, \mu, \gamma)$ the subject-specific likelihood in model $\mathcal{M}(p_{\text{sur}}, \pi, F_X)$ is $\mathcal{L}(\psi, \rho) = \mathcal{L}_1(\psi, \nu, \mu)\mathcal{L}_2(\gamma)$ where

$$\mathcal{L}_2(\gamma) = \pi(X;\gamma)^R \{1 - \pi(X;\gamma)\}^{1-R} \text{ and } \mathcal{L}_1(\psi, \nu, \mu) \equiv$$
$$f(X;\mu)\{f[V \mid Y, X; \nu] f[Y \mid X; \psi]\}^R \left\{ \sum_{y=0}^{1} f(V \mid y, X; \nu) f(y \mid X; \psi) \right\}^{1-R}.$$

Define $\varepsilon(\psi) = Y - \text{expit}(X\psi)$ so $E[\varepsilon(\psi) \mid X] = 0$.

Before analysing this model, we first analyse the more restrictive model $\mathcal{M}(p_{\text{sur}}, F_X)$ in which γ is known to be γ_0 and $\rho = (\mu, \nu)$ are unrestricted. Robins *et al.*[3] proved that $\Lambda^{\perp}(\psi, \rho) = \Lambda^{\perp}(\psi) = \{A(g_1, g_2, \psi)\}$ where $A(g_1, g_2, \psi) = \pi(X;\gamma_0)^{-1}[R\,\varepsilon(\psi) g_1(X) - \{R - \pi(X;\gamma_0)\} g_2(X, V)]$ and $g_1(X)$ and $g_2(X, V)$ are arbitrary vector-valued functions (of the dimension of ψ) chosen by the data analyst. Since $\Lambda^{\perp}(\psi, \rho)$ does not depend on ρ in this model $\mathcal{M}(p_{\text{sur}}, F_X)$, it follows $h = \emptyset$, and thus CODA and standard asymptotics are identical. Rotnitzky and Robins[10] prove that $\text{EIF}(\psi, \rho) = \Upsilon(\text{ES}, \psi, \rho)^{-1} \text{ES}(\psi, \rho)$ where $\text{ES}(\psi, \rho) \equiv A(g_{1\text{eff}}, g_{2\text{eff}}, \psi)$ with $g_{1\text{eff}}(X) \equiv g_{1\text{eff}}(X, \rho)$ as given in reference 10 and $g_{2\text{eff}}(X, V) \equiv g_{2\text{eff}}(X, V; \rho) = g_{1\text{eff}}(X; \rho) E[\varepsilon(\psi) \mid X, V; \nu]$. In contrast, the most efficient estimator $\hat{\psi}_{\text{cc}}$ if data on V were unavailable is the logistic regression estimator of Y on X among complete cases which has influence function $\Upsilon(A_{\text{cc}}, \psi, \rho)^{-1} A_{\text{cc}}(\psi)$ with $A_{\text{cc}}(\psi) \equiv A(g_{1,\text{cc}}, g_{2,\text{cc}}, \psi), g_{2,\text{cc}}(X, V) = 0, g_{1,\text{cc}}(X) = \pi(X;\gamma_0)$. When data on V are available, $\hat{\psi}_{\text{cc}}$ will be quite inefficient with asymptotic variance much greater than $I^{-1}(\psi, \rho)$.

Turning now to model $\mathcal{M}(p_{\text{sur}}, \pi, F_X)$ in which the randomization probabilities are unknown and $\rho = (\nu, u, \gamma)$ is unrestricted, $\Lambda^{\perp}(\psi, \rho) = \Lambda^{\perp}(\psi, r) \equiv A(g_1, g_2, \psi, r)$ with

$$A(g_1, g_2, \psi, r) = \pi(X;\gamma)^{-1}[R\,\varepsilon(\psi) g_1(X) - \{R - \pi(X;\gamma)\}\{g_2(X, V) - E[g_2(X, V) \mid X; \psi, \nu]\}]$$

and $r = (\gamma, \nu)$. Now since $\mathcal{L}(\psi, \rho) = \mathcal{L}_1(\psi, \nu, \mu)\mathcal{L}_2(\gamma)$, $\text{EIF}(\psi, \rho)$ is the same in models $\mathcal{M}(p_{\text{sur}}, \pi, F_X)$ and $\mathcal{M}(p_{\text{sur}}, F_X)$, since the additional restriction in $\mathcal{M}(p_{\text{sur}}, F_X)$ is on the parameter of the ancillary process $\mathcal{L}_2(\gamma)$. If, in model $\mathcal{M}(p_{\text{sur}}, \pi, F_X)$, (i) $h = \emptyset, s = r$ as would be the case with X discrete with only a moderate number of levels, (ii) $h = \gamma, s = \nu$ as would be the case with X univariate and continuous with $f(V \mid X, Y; \nu)$ assumed smooth, or (iii) $h = \nu, s = \gamma$ as would be the case if X were univariate and continuous and $\pi(X;\gamma)$ were assumed to be smooth,

then $E_{\psi,\rho}[A(g_1,g_2,\psi,h^*,s)] = 0$ and CODA and standard asymptotics are the same. However, if $h = r, s = \emptyset$ as would be the case when X is highly multivariate and continuous

$$E_{\psi,\rho}\left[A(g_1,g_2,\psi,h^*)\right] = E_{\psi,\rho}\left[\pi(X;\gamma^*)^{-1}\left[\begin{array}{l}\pi(X;\gamma)E\left[\varepsilon(\psi)\mid X\right]g_1(X) - \{\pi(X;\gamma) - \pi(X;\gamma^*)\}\times \\ \quad\{E\left[g_2(X,V)\mid X;\psi,v\right] - E\left[g_2(X,V)\mid X;\psi,v^*\right]\}\end{array}\right]\right]. \tag{16}$$

Since $E\left[\varepsilon(\psi)\mid X\right] = 0$, equation (16) is 0 for all v^*,γ^* if and only if $g_2(X,V) = g_2(X)$ does not depend on V. Hence, $\Lambda^{\perp}_{\mathrm{CODA}}(\psi,\rho) = \Lambda^{\perp}_{\mathrm{CODA}}(\psi,\gamma) = \left\{\pi(X;\gamma)^{-1}R\ \varepsilon(\psi)g_1(X)\right\}$ which does not depend on the data through V. We conclude there is no RAL $n^{\frac{1}{2}}$-consistent estimator whose influence function depends on the data through V. Indeed, it is easy to show that $\hat{\psi}_{\mathrm{cc}}$'s influence function is $\mathrm{EIF}_{\mathrm{CODA}}(\psi,\rho)$. That is, due to the curse of dimensionality, $\hat{\psi}_{\mathrm{cc}}$ is the most efficient estimator of ψ that performs well in moderate size samples.

7.4.3. Example 3. Randomly censored median regression:

We observe the minimum Y^* of the logarithm Y of failure time and the logarithm Q of censoring time, as well as the failure indicator $\tau = I[Y^* = Y]$ and covariates X. We assume that Q and Y are independent given X and that

$$\varepsilon(\psi) = Y - \psi X \text{ has conditional median 0 given } X. \tag{17}$$

To simplify the problem, we assume that for some σ

$$\mathrm{pr}\left[\tau = 1 \mid X\right] > \sigma > 0 \text{ with probability 1.} \tag{18}$$

See Robins[5] for an approach when (18) fails. Our goal is estimation of the median regression parameters ψ. We will show that, due to the curse of dimensionality, there will be no estimator of ψ that performs well in moderate size samples when X is multivariate and continuous. However, if, as in the accelerated failure time model, $\varepsilon(\psi)$ is assumed to be independent of X, then estimators with good moderate sample performance exist.

Letting γ index the law of Q given X, v index the law of $\varepsilon(\psi)$ given X, and μ the law of X, we can calculate that

$$\mathscr{L}(\psi,\rho) = \mathscr{L}_1(\psi,v,\mu)\mathscr{L}_2(\gamma), \mathscr{L}_2(\gamma) = f(Y^*\mid X;\gamma)^{1-\tau}\left\{\int_{Y^*}^{\infty} f(Q\mid X;\gamma)\,\mathrm{d}Q\right\}^{\tau}$$

$$\mathscr{L}_1(\psi,v,\mu) = f(X;\mu)f(\varepsilon(\psi)\mid X;v)^{\tau}\left\{\int_{Y^*-\psi X}^{\infty} f(\varepsilon\mid X;v)\,\mathrm{d}\varepsilon\right\}^{1-\tau}$$

restricted only by $E\left[I(\varepsilon(\psi) < 0) - 1/2 \mid X;v\right] = 0$ by assumption (17). Robins[5] proves that $\Lambda^{\perp}(\psi,\rho) = \{A(g,\psi,\rho)\}$, $g(X)$ an arbitrary function of X picked by the investigator. Here $A(g,\psi,\rho) \equiv D(g,\psi) - \int_{-\infty}^{\infty}\mathrm{d}\mathscr{M}(u;\gamma)\{D(g,\psi) - E_{\psi,\rho}\left[D(g,\psi)\mid Y^* > u,X\right]\}$ where $D(g,\psi) \equiv g(X)\{I(\varepsilon(\psi) < 0) - 1/2\}$ and $\mathrm{d}\mathscr{M}(u;\gamma) = \mathrm{d}N_Q(u) - \lambda_Q\left[u\mid X;\gamma\right]I(Y^* > u)\,\mathrm{d}u, N_Q(u) = I[Y^* \le u, \tau = 0], \lambda_Q\left[u\mid \bullet\right]$ is the hazard of Q given $\bullet$. Write $E_{\psi,\rho}\left[D(g,\psi)\mid Y^* > u,X\right] = z(X,u,v)$ since it depends on the unknown parameters only through v. Let $z = z(v)$ be a parameter indexing the functions $z(X,u,v)$. Then $A(g,\psi,\rho) = A(g,\psi,r)$ with $r = (\gamma,z)$. Now if X is of high dimension, neither γ nor z could be well estimated by smoothing, so we take $h = r$ and $s = \emptyset$. We then

have

$$E_{\psi,\rho}\left[A(g,\psi,h^*)\right] =$$

$$E_{\psi,\rho}\left\{ \int_{-\infty}^{\infty} \mathrm{d}u \left\{\lambda_Q\left[u \mid X;\gamma\right] - \lambda_Q\left[u \mid X;\gamma^*\right]\right\} \{z(X,u,v) - z(X,u,v^*)\} \operatorname{pr}\left[Y^* > u \mid X;\psi,\gamma,v\right] \right\} \tag{19}$$

which is not equal to 0 for all $h^* = (v^*,\gamma^*)$. Hence the CODA variance bound $I^{-1}_{\mathrm{CODA}}(\psi,\rho)$ is infinite. On the other hand, if X were univariate and $\lambda_Q\left[u \mid X;\gamma\right]$ were assumed smooth as in Ying *et al.*,[14] or it were assumed that $\lambda_Q\left[u \mid X;\gamma\right]$ does not depend on X or followed a lower dimensional model (for example, a Cox proportional hazards model), then we would choose $h = z$ and $s = \gamma$. We would then replace γ^* by γ in (19). Hence (19) would equal to 0 for all h^* and the CODA variance bound would equal the finite standard variance bound.

If, as in the accelerated failure time model, we impose the additional assumption that $\varepsilon(\psi)$ were independent of X under (ψ,ρ), then, even if X were high-dimensional and continuous, we show below that $z(X,u,v)$ can be well estimated in moderate samples without smoothing. Thus we would choose $h = \gamma$ and $s = z$ and would replace $z(X,u,v^*)$ by $z(X,u,v)$ in (19) to obtain that (19) equals 0. Hence, it follows $\hat{\psi}$ which solves $0 = \sum_i A_i(g,\psi,\gamma^*,\hat{z}(\psi))$ will, under this additional independence restriction, be a regular asymptotically linear estimator that performs well in moderate size samples, where $\hat{z}(\psi)$ is the $n^{\frac{1}{2}}$-consistent estimator of $z(X,u,v)$ given by $\hat{z}(\psi) = g(X)\sum_{i=1}^{n} J_i / \sum_{i=1}^{n} J_i^{\dagger}$ with $J^{\dagger} \equiv I\left[\varepsilon(\psi) > u - \psi X\right]$ and $J = J^{\dagger}\left\{I\left[\varepsilon(\psi) < 0\right] - 1/2\right\}$.

7.4.4. Example 4. A Randomized Trial:

Consider a two-arm randomized trial with $O = (Y,R,X)$, Y continuous, R a dichotomous treatment arm indicator, X a vector of continuous covariates, and the randomization probabilities $\pi(X;\gamma) = \Pr\left[R = 1 \mid X;\gamma\right]$ indexed by γ. We assume that $Y = \psi R + \varepsilon$ with ε independent of R given X due to randomization, so ψ is an additive treatment effect. Then, in model $M(p_{\mathrm{rand}},\pi,F_X)$with the randomization probabilities unknown, $\rho = (v,\gamma,\mu)$ and $\mathscr{L}(\psi,\rho) = \mathscr{L}_1(\psi,v)\mathscr{L}_2(\gamma,\mu)$ where $\mathscr{L}_1(\psi,v) = f\left[\varepsilon(\psi) \mid X;v\right]$ with $\varepsilon(\psi) \equiv Y - \psi R$, v indexes the law of $\varepsilon(\psi)$ given X, and $\mathscr{L}_2(\gamma,\mu) = \pi(X;\gamma)^R\left[1 - \pi(X;\gamma)\right]^{1-R} f(X;\mu)$.

In the model $M(p_{\mathrm{rand}},F_X)$ with randomization probabilities known (that is, the true value of γ_0 of γ known), $\rho = (v,\mu)$, $\mathscr{L}(\psi,\rho) = \mathscr{L}_1(\psi,v)\mathscr{L}_2(\gamma_0,\mu)$ where $\mathscr{L}_1$ and $\mathscr{L}_2$ are as above; $\Lambda^{\perp}(\psi,\rho) = \Lambda^{\perp}(\psi) = \{bg[\varepsilon(\psi),X][R - \pi(X;\gamma_0)]\}$ where $g(\cdot,\cdot)$ is an arbitrary function and b is an arbitrary constant.[16] Thus $\Lambda^{\perp}_{\mathrm{CODA}}(\psi) = \Lambda^{\perp}(\psi)$. Further the efficient score $\mathrm{ES}(\psi,\rho) = g_{\mathrm{eff}}(\varepsilon(\psi),X;v)[R - \pi(X;\gamma_0)]$ with $g_{\mathrm{eff}}(u,X;v) = \partial \log f\left[u \mid X;v\right]/\partial u$. The solution $\hat{\psi}$ to $0 = \sum_i g\left\{\varepsilon_i(\psi),X_i\right\}[R_i - \pi(X_i;\gamma_0)]$ will be RAL where $g(\cdot,\cdot)$ is chosen by the analyst.[16] For example, if $g(u,X) = u - v_1^* X$, v_1^* chosen by the investigator, then $\hat{\psi} = \sum_i \left(Y_i - v_1^* X_i\right)[R_i - \pi(X_i;\gamma_0)] / \sum_i R_i[R_i - \pi(X_i;\gamma_0)]$ will be RAL. Further, $\hat{\psi}$ will obtain the semi-parametric variance bound if, in truth, $\varepsilon(\psi) \mid X \sim \mathrm{N}\left(v_1^* X, v_2^{*2}\right)$ since then $g(u,X)$ is proportional to $g_{\mathrm{eff}}(u,X;v)$. One can obtain a global RAL estimator by adaptively estimating the derivative of the log density of $\varepsilon(\psi)$ given X[28]

In contrast, in model $M(p_{\mathrm{rand}},\pi,F_X)$ with γ unknown, we have $\Lambda^{\perp}(\psi,\rho) = \Lambda^{\perp}(\psi,v,\gamma) = \left\{b\left[g(\varepsilon(\psi),X) - E\left\{g(\varepsilon(\psi),X) \mid X;v\right\}\right][R - \pi(X;\gamma)]\right\}$. Hence with $h = (v,\gamma)$ and $s = \emptyset$ (that is neither $\pi(X;\gamma)$ nor $f(\varepsilon \mid X;v)$ can be effectively smoothed in moderate samples), one can calculate that $\Lambda^{\perp *}_{\mathrm{CODA}}(\psi,v,\gamma)$ is empty and the CODA variance bound is infinite. Indeed, Ritov and Bickel[20] show that, as expected by our informal conjecture (i) of Section 7.3, no uniformly

consistent estimator of ψ exists. This implies, by arguing as in Corollary 3 of Section 5, that no SFB estimator can converge to the true ψ uniformly over all (ψ,ρ) in $\Psi \times \mathcal{R}$, $\rho \equiv (\nu,\gamma,\mu)$ even when γ is known (since an SFB estimator does not use this knowledge).

8. UNIFORMITY, ASYMPTOTIC CONFIDENCE INTERVALS, AND LOCAL AND GLOBAL CODA EFFICIENCY:

In model $\mathcal{M}(p,F_X)$, $\hat{\psi}_{\mathrm{HT}}$ is uniformly asymptotically linear, but it is not CODA efficient in the sense that its asymptotic variance exceeds the CODA variance bound. This raises the question of whether CODA efficient estimators exist in model $\mathcal{M}(p,F_X)$. To satisfactorily answer this question, we require some additional concepts. Armed with these concepts, we study the issue of efficient estimation in model $\mathcal{M}(p,F_X)$ and model $\mathcal{M}(p,\pi,F_X)$. We then attempt to generalize our results to arbitrary semi-parametric models.

8.1. Concepts of Uniform Convergence

The following definitions will refer to an arbitrary semi-parametric model $\mathcal{L}(\psi,\rho)$. In the following, we abbreviate $\sup_{(\psi,\rho)\in\Psi\times\mathcal{R}}$ by $\sup_{(\psi,\rho)}$.

Definition. An estimator $\hat{\psi}_n$ (with n indexing sample size) is uniformly regular Gaussian (URG) with uniform asymptotic variance $\sigma^2(\psi,\rho)$ if, for each t,

$$\sup_{(\psi,\rho)} \left| \Pr_{(\psi\rho)}\left[n^{\frac{1}{2}}\left(\hat{\psi}_n - \psi\right) < t\right] - \phi\left(t;\sigma^2(\psi,\rho)\right)\right| \to 0 \text{ as } n \to \infty \tag{20}$$

where $\phi\left(t;\sigma^2\right)$ is the cumulative distribution function of a normal random variable with mean zero and variance σ^2. If $\hat{\psi}_n$ is a uniformly asymptotic linear estimator of ψ (that is, the $o_p(1)$ term in the definition of an asymptotically linear estimator is uniformly $o_p(1)$ over all laws in $\Psi \times \mathcal{R}$), then $\hat{\psi}_n$ is URG. However, $\hat{\psi}_n$, a regular asymptotic linear (RAL) estimator, does not imply $\hat{\psi}_n$ is URG.

Definition. The estimator $\hat{\psi}_n$ is uniformly asymptotically normal and unbiased (UANU) for ψ if there exists a sequence $\sigma_n^2(\psi,\rho)$ such that the z-statistic $n^{\frac{1}{2}}\left(\hat{\psi}_n - \psi\right)/\sigma_n(\psi,\rho)$ converges uniformly to a N(0,1) random variable, that is

$$\sup_{(\psi,\rho)} \left| \Pr_{(\psi,\rho)}\left[n^{\frac{1}{2}}\left(\hat{\psi}_n - \psi\right)/\sigma_n(\psi,\rho) < t\right] - \phi(t;1)\right| \to 0 \text{ as } n \to \infty\ . \tag{21}$$

$\hat{\psi}_n$ URG implies $\hat{\psi}_n$ UANU but the converse is false. However, if $\hat{\psi}_n$ is UANU and $\sigma_n(\psi,\rho)$ converges uniformly to $\sigma(\psi\rho)$ that is

$$\sup_{(\psi,\rho)} \left|\, \sigma_n(\psi\rho) - \sigma(\psi,\rho)\,\right| \to 0 \text{ as } n \to \infty \tag{22}$$

the $\hat{\psi}_n$ is URG. Furthermore, if $\hat{\psi}_n$ is UANU and there exists an estimator $\hat{\sigma}_n$ of $\sigma_n(\psi,\rho)$ such that $\sigma_n(\psi,\rho)/\hat{\sigma}_n$ converges to one uniformly in probability, that is, for all $\varepsilon > 0$

$$\sup_{(\psi,\rho)} \Pr_{(\psi\rho)}\left[\,\left|\, 1 - \sigma_n(\psi,\rho)/\hat{\sigma}_n\,\right| > \varepsilon\right] \to 0 \text{ as } n \to \infty \tag{23}$$

then, by the uniform version of Slutzky's Theorem, the t-statistic $n^{\frac{1}{2}}(\hat{\psi}_n - \psi)/\hat{\sigma}_n$ converges uniformly to a N(0,1) random variable, and thus the 'Wald' interval $C_n \equiv \hat{\psi}_n \pm z_{\alpha/2}\hat{\sigma}_n/\sqrt{n}$ is an

asymptotic $1-\alpha$ confidence interval for ψ where $z_{\alpha/2}$ is the $\alpha/2$ quantile of a standard normal distribution, and we have used the following definition.

Definition. C_n is an asymptotic $1-\alpha$ confidence interval for ψ if $\sup_{(\psi,\rho)} \mid \Pr_{(\psi,\rho)}[\psi \in C_n] - (1-\alpha) \mid \to 0$ as $n \to \infty$.

We required uniformity in our definition of an asymptotic confidence interval to be consistent with the usual definition of a non-asymptotic confidence interval. Specifically, by definition, at sample size n, a conservative $1-\alpha$ (non-asymptotic) confidence interval satisfies that for all $(\psi,\rho) \in \Psi \times \mathscr{R}$

$$\Pr_{(\psi,\rho)}[\psi \in C_n] \geq 1-\alpha.$$

Our definition of an asymptotic confidence interval satisfies the following consistency condition: if, for each sample size n, there is no conservative (non-asymptotic) $1-\alpha$ confidence interval for ψ, then no asymptotic $1-\alpha$ confidence interval for ψ exists. In contrast, if, in our definition of an asymptotic confidence interval, we had only required $\mid \Pr_{(\psi,\rho)}(\psi \in C_n) - (1-\alpha) \mid \to 0$ for all (ψ,ρ) without requiring uniformity, then the above consistency condition could be violated.

Note that $\hat{\psi}_n \pm z_{\alpha/2}\hat{\sigma}_n/\sqrt{n}$ will be an asymptotic $(1-\alpha)$ confidence interval for ψ if $\hat{\psi}$ is UANU and $\sigma_n(\psi,\rho)/\hat{\sigma}_n$ converges uniformly to one, even if $\hat{\psi}_n$ is not URG. If $\hat{\psi}_n$ is not URG, then even if $\sigma_n(\psi,\rho) \to \sigma(\psi,\rho)$ as $n \to \infty$ for all (ψ,ρ), this convergence cannot be uniform. Further, under mild regularity conditions, when $\hat{\psi}_n$ is UANU, the non-parametric bootstrap estimator $\hat{\sigma}_n$ of the standard error of $n^{\frac{1}{2}}\left(\hat{\psi}_n - \psi\right)$ will satisfy (23). Hence, if $\hat{\psi}_n$ is UANU, then $\hat{\psi}_n$ can be used to centre an asymptotic $(1-\alpha)$ Wald confidence interval for ψ using a bootstrap estimate of the standard error.

Remark. Let $\hat{\psi}_n$ be the mid-point of the interval $C_n(\alpha)$ where we write C_n as $C_n(\alpha)$ to emphasize dependence of the interval C_n on the nominal size $1-\alpha$. If the length of $C_n(\alpha)$ is converging to zero as $n \to \infty$ for each α then, if $C_n(\alpha)$ is either an asymptotic $1-\alpha$ or a conservative (non-asymptotic) $1-\alpha$ confidence interval for ψ, $\hat{\psi}_n$ is uniformly consistent for ψ. It follows from Theorem 3 that the continuously stratified sampling model $M(p,\pi,F_X)$ with unknown randomization probabilities does not admit asymptotic $1-\alpha$ or conservative non-asymptotic $1-\alpha$ confidence intervals for ψ. Similarly, by the Remark in Sec. 7.1.1, $1-\alpha$ intervals do not exist for the average treatment effect (ATE), $ATE \equiv E\{E(Y|R_1,X) - E(Y|R=0,X)\}$, in a randomized trial with dichotomous outcome Y and unknown randomization probabilities. It then follows by an argument similar to that given in the Proof of Corollary 3 that, as stated in the Introduction, a statistician who ignores the randomization probabilities cannot construct nominal 95 per cent confidence intervals for $\psi = E(Y)$ in the stratified random sampling model $M(P,F_X)$ or the ATE in a randomized trial that have both (i) an expected length which goes to zero in sample size, and (ii) a guaranteed expected actual coverage rate of at least 95 per cent over the ensemble of studies analysed by the statistician during his or her lifetime.

In contrast, in the random sampling model $M(\rho,F_x)$ with $\pi(x;\gamma_0)$ known, $C_n \equiv \hat{\psi}_{HT} \pm c(\alpha,n,\sigma)/n$ is, for each n, a non-asymptotic conservative $1-\alpha$ confidence interval for $\psi = E(Y)$ that satifies (i) and (ii), where $c(\alpha,n,\sigma) = (3\sigma)^{-1}\log(\alpha/2)\{-1+[1-18n\sigma/\log(\alpha/2)]^{\frac{1}{2}}\}$ with $\sigma = \inf\{\pi(x;\gamma_0); x \in [0,1]\}$. This interval C_n is derived using Bernstein's Inequality for i.i.d. bounded mean zero random variables χ_i with range in $[-M,M]$ and $v \geq n\,\mathrm{var}(\chi_i)$ specialized to $\chi_i = R_iY_i/\pi(X_i,\gamma_0) - \psi, M = \sigma^{-1}v = n\sigma^{-1}$.

Bernstein's Inequality: $P_r\left(\left|\sum_{i=1}^{n}\chi_i\right| > x\right) < 2\exp[-x^2/\{2(v+Mx/3)\}]$.

Similarly, by the Remark in Sec. 7.1.1, in a two-armed randomized trial with Bernoulli outcome Y with $\pi(x;\gamma_0)$ known,

$$C_n = n^{-1}\sum_{i=1}^{n}[R_iY_i/\pi(X_i;\gamma_0) - (1-R_i)Y_i/\{1-\pi(X_i,\gamma_0)\}] \pm c(\alpha,n,\sigma^*)/n$$

is, for each n, a non-asymptotic conservative $1-\alpha$ confidence interval for the *ATE* which satisfies (i) and (ii) above, where $\sigma^* \equiv \inf[\min\{\pi(x;\gamma_0), 1-\pi(x;\gamma_0)\}; x \in [0,1]]/2$.

We derived these new interval estimators because standard asymptotic intervals fail to satisfy (ii) for statisticians with primary involvement in studies with small spample size n.

8.2. Local and Global CODA Efficiency

We first provide some definitions and then describe the motivation that inspired them.

Definition. We say an estimator $\hat{\psi}$ of a finite dimensional parameter ψ in a semi-parametric model with likelihood $\mathscr{L}(\psi,\rho), (\psi,\rho) \in \Psi \times \mathscr{R}$, is a CODA estimator if $\hat{\psi}$ is UANU.

Definition. We say an estimator $\hat{\psi}$ in a model $\mathscr{L}(\psi,\rho)$ is locally CODA efficient at a submodel $\Psi \times \mathscr{R}^*, \mathscr{R}^* \subset \mathscr{R}$, if $\hat{\psi}$ is UANU on $\Psi \times \mathscr{R}$ (that is, $\hat{\psi}$ is CODA) and $\hat{\psi}$ is URG on $\Psi \times \mathscr{R}^*$with uniform asymptotic variance $I^{-1}_{\text{CODA}}(\psi,\rho)$.

Definition. We say $\hat{\psi}$ is globally CODA efficient if $\hat{\psi}$ is URG on $\Psi \times \mathscr{R}$ with uniform asymptotic variance $I^{-1}_{\text{CODA}}(\psi,\rho)$.

Motivation. Consider again the stratified random sampling model $M(p,F_X)$ with the sampling probabilities (that is, γ_0) known, X univariate and continuous, and $q(X;v) = E(Y \mid X) - \psi$ may be unsmooth. In Section 7.1.5 we noted that $\text{EIF}(\psi,v) = \text{EIF}_{\text{CODA}}(\psi,v) = A_1(q(v),\psi,\gamma_0) = \pi(X)^{-1}\{R(Y-\psi) - (R-\pi(X))q(X;v)\}$. Let $\hat{p}(X_i)$ be a histogram or kernel estimator of $E(Y \mid X_i)$. Let

$$\hat{\psi}_{\text{eff}} = n^{-1}\sum_i \pi(X_i)^{-1}[R_iY_i - (R_i - \pi(X_i))\,\hat{p}(X_i)] \tag{24}$$

be the solution to $0 = \sum_i \text{EIF}_{\text{CODA},i}(\psi,\hat{v}(\psi)) = \sum_i A_{1i}(q(\hat{v}(\psi)),\psi,\gamma_0)$ with $q(X;\hat{v}(\psi)) = \hat{p}(X) - \psi$. We will assume $q(X;v)$, although possibly unsmooth, is continuous. Then, under the regularity conditions described below, $\hat{\psi}_{\text{eff}}$ will be RAL with the influence function $\text{EIF}(\psi,v)$. That is, $\hat{\psi}_{\text{eff}}$ will be a globally efficient RAL estimator. (Indeed, results of Klaassen[29] suggest $\hat{\psi}_{\text{eff}}$ will under additional regularity conditions be a globally efficient RAL estimator if we only assume $q(X;v)$ is measurable). Further, we will show that $\hat{\psi}_{\text{eff}}$ is UANU with $\sigma_n^2(\psi,\rho)$ converging to $I^{-1}(\psi,\rho) = E_{\psi,\rho}\left[\text{EIF}(\psi,\rho)^2\right] = \psi(1-\psi) + \kappa(u,v)$ for each (ψ,ρ) where $\kappa(u,v) \equiv E_{\psi,\rho}\left[\pi(X)^{-1}(1-\pi(X))q(X;v)[1-q(X;v)]\right]$. However, $\hat{\psi}_{\text{eff}}$ is not URG and thus not uniformly asymptotically linear, since no uniformly consistent estimator of $\kappa(u,v)$ exists. This is a consequence of the fact that (i) at a given sample size n, we cannot estimate the 'wiggles' in $q(x;v)$ occurring within intervals of x of size less than $O(1/n)$, and thus, (ii) whatever be n, there will exist continuous functions $q(x;v)$ that are highly 'wiggly' on a scale less than $O(1/n)$ that cannot be well estimated.

Thus, when there exists a globally efficient RAL estimator $\hat{\psi}_{\text{eff}}$ but no globally URG estimator, we choose (as is clear from our definitions above) to say that global CODA efficiency is not attainable (since we cannot guarantee that the variance of $\hat{\psi}_{\text{eff}}$ is close to the CODA variance bound $I^{-1}_{\text{CODA}}(\psi,\rho)$ for all $\rho \in \mathscr{R}$ at any fixed sample size n).

Indeed, by $\hat{\psi}_{\text{eff}}$ UANU, both the theoretical interval $\hat{\psi}_{\text{eff}} \pm z_{\alpha/2}\sigma_n(\psi,\rho)/\sqrt{n}$ and the feasible interval $\hat{\psi}_{\text{eff}} \pm z_{\alpha/2}\hat{\sigma}_n/\sqrt{n}$ (with $\hat{\sigma}_n$ satisfying (23) and defined below) are asymptotic $(1-\alpha)$ confidence intervals for ψ. In contrast, the theoretical interval $\hat{\psi}_{\text{eff}} \pm z_{\alpha/2} I(\psi,\rho)^{-1/2}/\sqrt{n}$ is not an asymptotic $(1-\alpha)$ confidence interval for ψ since, for some $\varepsilon > 0$, at each sample size n, there will exist some law (ψ,ρ) (depending on n) such that the probability ψ lies in this theoretical interval under (ψ,ρ) will be less than $1-\alpha-\varepsilon$. This reflects the fact that $I(\psi,\rho)^{-1/2}$ underestimates $\sigma_n(\psi,\rho)$ at sample size n for laws (ψ,ρ) with $q(x;\nu)$ very wiggly.

However, let $\mathscr{R}^*$ be the subset of $\mathscr{R}$ such that $q(X;\nu)$ is Lipschitz α, $\alpha > 0$ with bound $c < \infty$, that is, $|q(X;\nu) - q(X^*;\nu)| \leq c|X - X^*|^{\alpha}$. We show below that $\hat{\psi}_{\text{eff}}$ is URG and thus locally CODA efficient on the submodel $\Psi \times \mathscr{R}^*$ because the Lipschitz assumption sets limits on how wiggly $q(X;\nu)$ can be. We now formalize this discussion.

Formal Definition of $\hat{p}(X_i)$. We choose a sample splitting technique that makes $\hat{p}_i(\cdot)$ independent of R_i. This insures that $\hat{\psi}_{\text{eff}}$ will be unbiased for ψ and avoids the need for $\hat{p}(\cdot)$ to lie in a Donsker class. Specifically, let $M_i = 1$ if $i \leq [n/2]$ and $M_i = 2$ if $i > [n/2]$, so $M_i = 1$ if subject i is in the first half sample and $M_i = 2$ otherwise. Let $h = h(n)$ be a histogram binwidth that depends on the sample size n *and* that divides $[0,1)$ into h^{-1} bins (intervals), $b_1,\ldots,b_{h^{-1}}$ where $b_\ell \equiv b_{n\ell} \equiv [(\ell-1)h, h\ell)$. Let $b_n(x) \equiv b(x)$ denote the interval b_ℓ in which x lies for $x \in [0,1)$. Then, for $m \in \{0,1\}$, $\hat{p}_n(x,m) = \sum_j I(M_j = m) Y_j R_j I[b(X_j) = b(x)] / \sum_j I(M_j = m) R_j I[b(X_j) = b(x)]$ is the half sample m histogram estimate of $E[Y \mid X = x]$. Then define $\hat{p}(X_i) \equiv \hat{p}_{ni}(X_i) \equiv \hat{p}_n(X_i, 1 - M_i)$ to be the histogram estimate of $E(Y \mid X_i)$ based on subjects in the half-sample not containing subject i. We then prove the following theorem in the Appendix.

Theorem 4. If, in the model $M(p,F_X)$ with X univariate, $\pi(X) > \sigma > 0$, $\infty > \sigma_x^* > f(x;\mu) > \sigma_x > 0$ for $x \in [0,1)$, and $q(x;\nu)$ only restricted by being continuous in x, if the binwidth $h(n) \to 0$ and $nh(n) \to \infty$ as $n \to \infty$, then:

(a) $\hat{\psi}_{\text{eff}}$ is a RAL estimator with influence function EIF (ψ,ρ);

(b) $\hat{\psi}_{\text{eff}}$ is UANU with $\sigma_n^2(\psi,\rho) = \psi(1-\psi) + E_{\psi,\rho}\left[\{1-\pi(X)\}\pi(X)^{-1}\{Y - \overline{p}_n(X)\}^2\right]$ with $\overline{p}_n(x) \equiv \int_{x\in b_n(x)} p(x;\theta)\pi(x) f(x;\mu)\,dx / \int_{x\in b_n(x)} \pi(x) f(x;\mu)\,dx$ and $\theta = (\psi,\nu)$;

(c) $\hat{\psi}_{\text{eff}}$ is URG on the submodel in which $q(x;\nu)$ is Lipschitz $\alpha, \alpha > 0$, with bound $c < \infty$;

(d) $\hat{\sigma}_n = \hat{\psi}_{\text{eff}}\left(1 - \hat{\psi}_{\text{eff}}\right) + n^{-1}\sum_i \{1-\pi(X_i)\}\pi(X_i)^{-1}\{Y_i - \hat{p}(X_i)\}^2$ satisfies equation (23).

Now consider the model $M(p,\pi_\sigma,F_X)$ in which the randomization probabilities $\pi(X;\gamma)$ are unknown but live in a smooth subset Γ_σ of Γ, and $q(x;\nu)$ still is only restricted to be continuous. Hence $s = \gamma$ and $h = \nu$. In this model, the CODA variance bound equals the usual variance bound and both are equal to that in model $M(\rho,F_X)$. Our informal conjecture (iv) of Section 7.3 states that a globally efficient RAL estimator should exist if $\mathscr{S}$ is sufficiently smooth. Specifically if $\mathscr{S} \equiv \Gamma_\sigma$ is such that $\pi(x;\gamma)$ is known to be Lipschitz $\alpha^* = 0.5 + \varepsilon^*$, $\varepsilon^* > 0$ with bound c^* then there exists an estimator $\hat{\psi}_{\text{eff}}$ that is (i) RAL with influence function EIF (ψ,ρ),(ii) UANU, and (iii) URG on the submodel where $q(x;\nu)$ is Lipschitz α with bound c, so local but not globally CODA efficient estimators exist. Specifically, $\hat{\psi}_{\text{eff}}$ is the estimator that replaces $\pi(X_i)$ in equation (24) by $\hat{\pi}(X_i)$ with $\hat{\pi}(X_i)$ defined and $\hat{p}(X_i)$ redefined as follows.

First redefine $M_i = 1$ if $i \leq [n/3]$, $M_i = 2$ if $[n/3] < i \leq [2n/3]$, $M_i = 3$ if $i > [2n/3]$ so the sample of n subjects is divided into three third-samples. Let $\hat{\pi}_n(x,m) = \sum_j I(M_j = m) R_j I[b(X_j) = b(x)] / \sum_j I(M_j = m) I[b(X_j) = b(x)]$ be the third sample m histogram estimate of $\pi(x) = \text{pr}[R = 1 \mid X = x]$. Let $\hat{p}_n(x,m)$ be as before except with $m \in \{1,2,3\}$ denoting third samples. Then let $\hat{p}(X_i) \equiv \hat{p}_{ni}(X_i) \equiv \hat{p}_n[X_i, M_i + 1 \,(\text{mod}3)]$ and $\hat{\pi}(X_i) \equiv \hat{\pi}_{ni}(X_i) \equiv \hat{\pi}_n[X_i, M_i + 2\,(\text{mod}3)]$

so $\hat{p}(X_i)$ and $\hat{\pi}(X_i)$ are computed from subsets of study subjects not containing subject i and distinct from one another.

Theorem 5. In model $M(p,\pi_\sigma,F_X)$ defined just above, if $h(n) = n^{-(1-\varepsilon)}$ with $0 < \varepsilon < \varepsilon^*/(0.5+\varepsilon^*)$ (i) $\hat{\psi}_{\text{eff}}$ is RAL with the EIF (ψ,ρ); (ii) $\hat{\psi}_{\text{eff}}$ is UANU with $\sigma_n^2(\psi,\rho)$ as in Theorem 4; (iii) $\hat{\psi}_{\text{eff}}$ is URG on the submodel in which $q(x;v)$ is Lipschitz α with bound $c < \infty$; and (iv) equation (23) is satisfied by $\hat{\sigma}_n$ of Theorem 4 with $\hat{\pi}(X_i)$ replacing $\pi(X_i)$

Suppose finally that X is multivariate and continuous in model $M(p,F_X)$ with the randomization probabilities known. Then we would have to construct $\hat{p}(X_i)$ using some dimension-reducing submodel. For example, we might fit the parametric regression model $E(Y \mid X,R=1) = \beta'X$ or the generalized additive model $E(Y \mid X,R=1) = \sum_{m=1}^{k} g_m(X_m)$ where $g_m(\cdot)$ is a smooth, Lipschitz α, univariate function with bound c. In either case, the resulting estimator $\hat{\psi}_{\text{eff}}$ would be locally CODA efficient. That is, $\hat{\psi}_{\text{eff}}$ would be URG with uniform asymptotic variance $I^{-1}(\psi,\rho)$ on the parametric or generalized additive submodel and would remain RAL and UANU off the submodel. From the above discussion, it is clear that with high-dimensional continuous X, it is possible to construct locally CODA efficient estimators at dimension-reducing submodels, but globally CODA efficient estimators cannot be constructed.

The above theorems motivate the following additional informal conjecture (v) concerning the CODA algorithm.

(v) Even if the CODA variance bound is finite, if the nuisance parameter $r_{\text{ce}} = (h_{\text{ce}},s_{\text{ce}})$ of $\text{ES}_{\text{CODA}}(\psi,r_{\text{ce}})$ depends on $h_{\text{ce}} \in \mathscr{H}_{\text{ce}}$ with $\mathscr{H}_{\text{ce}}$ an unsmoothed class, then no globally efficient CODA estimator of ψ will exist. Further, under regularity conditions, if $\mathscr{S}_{\text{ce}}$ is a sufficiently smoothed class, then (i) if $h_{\text{ce}} \neq \emptyset$, the estimator $\hat{\psi}_{\text{eff}}$ of informal conjecture (iv) will be locally CODA efficient and (ii) if $h_{\text{ce}} = \emptyset$, a global CODA efficient estimator solving $0 = \sum_i \text{ES}_{\text{CODA}}(\psi,\hat{s}_{\text{ce}}(\psi))$ will exist.

Example. As an example of this latter phenomenon in model $M(p_{\text{sur}},F_X)$ of Example 2 of Section 7.4, $\text{ES}_{\text{CODA}}(\psi,\rho) = \text{ES}_{\text{CODA}}(\psi) = A_{\text{cc}}(\psi)$ and the complete case logistic regression estimator $\hat{\psi}_{\text{cc}}$ solving $\sum_i A_{\text{cc},i}(\psi) = 0$ is globally CODA efficient.

Remark. The regularity conditions referred to in conjecture (v) include strong conditions such as bounds on higher-order moments of the observed random variable O. Such strong conditions are appropriate for our purposes, since our goal is to distinguish the difficulties in estimation caused by the curse of dimensionality and lack of smoothness from those caused by distributions with heavy tails and unbounded support.

9. CONDITIONAL INFERENCE

Heretofore, we have only been referring to unconditional inference. However, in the continuously stratified random sampling model $\mathscr{M}(p)$ with the marginal law of X and γ_0 known, $W \equiv \{(R_i,X_i);i=1,\ldots,n\}$ is ancillary for $\theta = (\psi,v)$ and thus for ψ since $L(\theta) = f(Z \mid W;\theta)f(W)$ and $f(W)$ is known. Here, $Z = \{Y_iR_i;i=1,\ldots,n\}$. If we adopt the conditionality principle, inference should be performed conditional on ancillary statistics.

Now $\hat{\psi}_{\text{HT}}$ is not asymptotically unbiased for ψ conditional on the ancillary statistic W.[16] This raises the question of whether any estimators of ψ in model $\mathscr{M}(p)$ remain asymptotically unbiased conditional on the ancillary statistic W. One can calculate that among all the estimating functions $A_2(g,g^\dagger,\psi,\gamma_0,\mu_0)$ in $\Lambda^\perp$, the only one that has mean zero given W is the EIF with $g \equiv g^* \equiv q(v)$. That is, unconditional efficiency is equivalent to asymptotic unbiasedness conditional on the ancillary statistic W. It follows that only those estimators which attain

the CODA variance bound unconditionally remain conditionally asymptotically unbiased with probability approaching one (w.p.a. 1). Specifically, suppose again X is univariate and continuous and $q(x;v)$ may be unsmooth, only restricted by being continuous. Consider the estimator $\tilde{\psi}_{\text{eff}} = n^{-1}\sum_i R_i \pi(X_i;\gamma_0)^{-1}\{Y_i - \hat{p}(X_i)\} + \int \hat{p}_{ni}(x) f(x;\mu_0)\,dx$ solving the EIF equation $0 = \sum_i A_{2i}(q(\hat{v}(\psi)), q(\hat{v}(\psi)), \psi, \gamma_0, \mu_0)$. Here $\hat{p}(X_i) = \hat{p}_{ni}(X_i)$ is the half-sample histogram estimator of $E(Y \mid X_i)$. Ritov and Robins[19] prove theorem 6.

Theorem 6.

(a) $\tilde{\psi}_{\text{eff}}$ is a globally efficient RAL estimator. Given W, with probability approaching 1 (w.p.a. 1), $\tilde{\psi}_{\text{eff}}$ is asymptotically normal and unbiased with asymptotic variance equal to $I^{-1}(\psi,\rho) \equiv E_{\psi,\rho}\left[A_2(q(v), q(v), \psi, \gamma_0, \mu_0)^{\otimes 2}\right]$.

(b) Unconditionally, $\tilde{\psi}_{\text{eff}}$ is UANU. However, $\tilde{\psi}_{\text{eff}}$ is not UANU conditional on W (w.p.a. 1).

(c) Both unconditionally and conditional on W (w.p.a. 1), $\tilde{\psi}_{\text{eff}}$ is URG with uniform asymptotic variance $I^{-1}(\psi,\rho)$ on the submodel $\Psi \times \mathscr{R}^*$ in which $q(x;v)$ is Lipschitz α with bound c.

If we adopt the conditionality principle, we would, at a minimum, like our estimators to be conditionally UANU (w.p.a. 1) given the ancillary statistic W. Unfortunately, when $q(X;v) \equiv E(Y \mid X) - \psi$ is not necessarily smooth, this desire cannot be satisfied globally (that is, for all $q(X;v)$). The best we can do is to construct estimators $\tilde{\psi}_{\text{eff}}$ that are (i) guaranteed to be unconditionally UANU and locally CODA efficient for ψ at smooth submodels and (ii) are also conditionally UANU on the submodel (w.p.a.1). Further, even if we forgo uniformity, it will not be possible to construct an estimator $\tilde{\psi}_{\text{eff}}$ that will be asymptotically normal and unbiased for ψ along all sequences $\{w_n\}$ where w_n is a realization of W at sample size n.

The type of incompatibility between conditional and unconditional inference that we describe is, in a sense, an asymptotic version of the ancillarity paradoxes described in references 30 and 31.

APPENDIX

Proof of Theorem 3

The proof is analogous to that of Theorem 1 in Ritov and Bickel.[20] We fix the law (θ_1, γ_1). Without loss of generality, we let $p(X;\theta_1) = 0.5 + 4ad$ with $a \equiv 0.1$ and $d \equiv 0.1$ and $\pi(X;\gamma_1) = 0.25$. Fix $\varepsilon > 0$. To prove part (i), we are going to display a set of laws $P^* = \{(\theta,\gamma)\}$ such that for all $(\theta,\gamma) \in P^*$, $\max \mid p(x;\theta_1) - p(x;\theta) \mid < \varepsilon$ and $\max \mid \pi(x,\gamma_1) - \pi(x,\gamma) \mid < \varepsilon$ with the following properties. For any estimator $\hat{\psi}_n$, and any sequence a_n as in Remark 1 of Sec. 5, there exists (θ_0,γ_0) in P^* depending on $\hat{\psi}_n$, such that $a_n \mid \hat{\psi} - \psi_0 \mid$ does not converge to 0 in (θ_0,γ_0)-probability. The construction requires a number of steps which include a proof of part (ii). For $k, k = 1,2,\ldots$, define $P_k^* = \{(\theta_{k\ell},\gamma_{k\ell}), \ell = 1,\ldots,2^k\}$ as the set of all laws that can be constructed by the following process:

1. Divide $(0,1)$ into 2^k intervals of length $1/2^k$.
2. Subdivide each interval into two subintervals – a left subinterval and a right subinterval.
3. Consider two choices for $p(x)$ and $\pi(x)$ on the subintervals.
   Choice 1: $\pi(x) = (0.25 + d)$ and $p(x) = (0.5 + a)$ on the left subinterval, while $\pi(x) = (0.25 - d)$ and $p(x) = (0.5 - a)$ on the right subinterval.
   Choice 2: Same as choice 1, except with left and right interchanged.
4. We obtain a single law $(\theta_{k\ell},\gamma_{k\ell}) \equiv (p_{k\ell}(x), \pi_{k\ell}(x))$ by making either choice 1 or choice 2 on each of the 2^k intervals.

Thus, there are 2^{2^k} such laws that could be so generated. Let P_k^* be the set of these laws. Note that for each law in P_k^*, $\psi_{k\ell} = \psi(\theta_{k\ell}) = 0.5$.

We now argue for, $k > 3\log n/\log 2$, the sample does not distinguish between the following two possibilities.

> Possibility (a): $\pi(x)$ is a constant 0.25; $p(x)$ is the constant $0.5 + 4ad$ (and thus $\psi = 0.5 + 4ad$).
> Possibility (b): $\pi(x)$ is $\gamma_{k\ell}$ and then $\psi = 0.5$ with the specific choice ℓ selected at random as follows. For each of the 2^k intervals, the selection of choice (1) or choice (2) has been made independently by 2^k flips of a fair coin.

For sample size n, for all k such that $2^k > n^3$, the probability that any 2 of the n observations lie in the same interval out of the 2^k possible is less than or equal to $O(1/n)$. (This is true, since if we divide $(0,1)$ into n^3 equal intervals, the probability of finding two observations in a single interval is $O_p(1/n)$.) However, (a) and (b) can only be distinguished by observing more than one observation in any interval in order to see whether one subinterval or the other is being oversampled. (Note that in possibility (a), for indistinguishability from possibility (b), $p(x)$ must be $0.5+4ad$ since, if the data were really generated by possibility (b), and we incorrectly assumed, as in possibility (a) that $\pi(x)$ is the constant 0.25, then we could use the data to conclude that the mean ψ of Y must be $0.5+4ad$ since $(0.5+a)(0.25+d)+(0.5-a)(0.25-d)/\{(0.25+d)+(0.25-d)\} = 0.5+4ad$.) By choosing $\delta < ad$, we conclude that equation (11) and thus equation (9) and Theorem 3(ii) are true. To make the last result local, one could rechoose $a = d = .1\varepsilon/2$].

We are now ready to describe the set of laws P^*. Let $c_m = \beta\left[m(\log m)^2\right]^{-1}$. Note that for small enough β, $\sum_m c_m < \varepsilon/2$. Given a sequence $\ell(1), \ell(2), \ldots$ with $1 \le \ell(m) \le 2^{2^m}$ and a sequence $b_1, b_2, \ldots$ with $b_m \in \{0,1\}$, we consider a law (θ, γ) given by

$$p(x;\theta) = 0.5 + 4ad + \sum_{m=1}^{\infty} c_m \{b_m p_{m,\ell(m)}(x) + (1-b_m)(0.5-4ad)\} \tag{25}$$

$$\pi(x;\gamma) = 0.25 + \sum_{m=1}^{\infty} c_m \{b_m \pi_{m,\ell(m)}(x) + (1-b_m)0.25\}. \tag{26}$$

$P^* = \{(\theta,\gamma)\}$ is the set of laws generated as the choices of b_m and $\left(p_{m,\ell(m)}, \pi_{m,\ell(m)}\right)$ vary.

Consider now the Bayesian estimate of ψ when the b_m's are chosen by independent flips of a fair coin and $\ell(1), \ell(2), \ldots$ are independent (and independent of $b_1, b_2, \ldots$), such that $\ell(m)$ is uniform distributed on $1, 2, \ldots, 2^{2^m}$. If the sample size is n take the loss function to be $1(|\hat{\psi} - \psi(\theta)| > adc_{m(n)})$, for $m(n) > \lceil 6\log n \rceil$ (where $\lceil x \rceil$ is the smallest integer not smaller than x). Let $h_j(\theta,\gamma)$ be equal to (θ,γ) except b_j with $b_j' = 1 - b_j$. We have already argued that with sample size n with probability of $1 - O(n^{-1})$ there are no two observations in any of the relevant intervals with length less then n^{-3}, and hence, given the data, the difference between the *a posteriori* probability of $h_{m(n)}(\theta)$ and θ converges to 0, but for any θ, $|\psi(h_m(\theta)) - \psi(\theta)| > 4adc_m$. We can conclude that Bayes estimator will 'fail' and that

$$\liminf_n \liminf_{\hat{\psi}} \int P_{(\theta,\gamma)}\left[|\hat{\psi} - \psi(\theta)| > adc_{m(n)}\right] dF(\theta,\gamma) \ge 1/2 . \tag{27}$$

Suppose there was an estimator $\hat{\psi}$ such that

$$\sup_{\theta_0,\gamma_0} \lim_n P_{(\theta_0,\gamma_0)}\left[|\hat{\psi} - \psi(\theta_0)| > adc_{m(n)}\right] \to 0.$$

Then we will get from the dominated convergence theorem that

$$\lim_n \int P_{(\theta,\gamma)} \left[| \hat{\psi} - \psi(\theta) | > adc_{m(n)} \right] dF(\theta,\gamma) \to 0$$

contradicting (27). Hence the best achievable rate is no better than $(\log\log n)^2 \log n$.

To prove part (iii), we construct now a consistent estimator. Let $W_i = (X_i, R_i, R_i Y_i)$, $i = 1,\ldots,n$ be the sample. Let $N = \sum_i R_i$, and let $W_{(i)} = (X_{(i)}, 1, Y_{(i)})$, $i = 1,\ldots,N$, be the sub-sample of observations with $R_i = 1$ ordered such that $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(N)}$. Define the estimator:

$$\hat{\psi} = \sum_{i=1}^{N} Y_{(i)}(X_{(i)} - X_{(i-1)})$$

where $X_{(0)} = 0$. Let $\Xi = \{X_1,\ldots,X_n, R_1,\ldots,R_n\}$. It is to see that

$$E(\hat{\psi} \mid \Xi) = \sum_{i=1}^{N} p(X_{(i)})(X_{(i)} - X_{(i-1)}) \xrightarrow{p} \psi$$

by $p(x)$ Riemann integrable since $\max_i\{X_{(i)} - X_{(i-1)}\} \xrightarrow{p} 0$. Since

$$\begin{aligned} \operatorname{var}(\hat{\psi} \mid \Xi) &= \sum_{i=1}^{N} p(X_{(i)})[1 - p(X_{(i)})](X_{(i)} - X_{(i-1)})^2 \\ &\leq \max\{X_{(i)} - X_{(i-1)}\} \sum_{i=1}^{N} p(X_{(i)})[1 - p(X_{(i)})](X_{(i)} - X_{(i-1)}) \\ &\xrightarrow{p} 0 \end{aligned}$$

we conclude that $\hat{\psi}$ is consistent.

Proof of Theorem 5.

Theorem 4 is proved in a similar manner but its proof is easier and thus is excluded. By some algebra $n^{\frac{1}{2}}\left(\hat{\psi}_{\text{eff}} - \psi\right) = n^{-\frac{1}{2}} \sum_i \chi_i + V_1 + V_2$, $\chi \equiv Y - \psi + \{R/\pi(X) - 1\}\{Y - \overline{p}_n(X)\}$, $V_2 = V_{21} + V_{22} + V_{23}$, $V_{2m} = -n^{-\frac{1}{2}} \sum_i \eta_{mi}$, $\eta_{mi} \equiv [R_i/\pi(X_i) - R_i/\hat{\pi}(X_i)]\left[Y_i - \overline{p}_n(X_i)\right] I(M_i = m)$, $V_1 = V_{11} + V_{12} + V_{13}$, $V_{1m} = n^{-\frac{1}{2}} \sum_i \zeta_{mi}$, $\zeta_m = \{R/\hat{\pi}(X) - 1\}\{\hat{p}_n(X) - \overline{p}_n(X)\} I(M = m)$.

Throughout the following argument, it is understood that we are conditioning on the event Q that $\hat{p}_n(x,m)$ and $\hat{\pi}_n(x,m)$ are defined (that is, have non-zero denominators) for all $x \in [0,1), m \in \{1,2,3\}$. This is justified since if $h(n) \to \infty$, $\sup_{(\psi,\rho)} \Pr_{(\psi,\rho)}[Q^c] > \varepsilon \to 0$ as $n \to \infty$ for $\varepsilon > 0$ where Q^c is the complement of the event Q. For notational convenience, we shall suppress the dependence of $p(x) = E(Y \mid X = x)$, $\pi(x)$ and $f(x)$ on (ψ, v), γ, and μ respectively.

We first show $\hat{\psi}_{\text{eff}}$ is UANU. By χ bounded and $\operatorname{var}(\chi) = \sigma_n^2(\psi,\rho)$ we have, by the Lindeberg–Feller Central Limit Theorem, $\sup_{(\psi,\rho)} \left| \Pr\left[n^{-\frac{1}{2}} \sum_i \chi_i / \sigma_n(\psi,\rho) > t \right] - \phi(t;1) \right| \to 0$ as $n \to \infty$. Hence $\hat{\psi}_{\text{eff}}$ is UANU if we can show that V_{2m} and $V_{1m}, m = 1,2,3$, converge uniformly to zero. To do so, note that since $\pi(x;\gamma)$ is Lipschitz $0.5 + \varepsilon^*$, a straightforward bias calculation gives $E\left[|\pi(x;\gamma) - \hat{\pi}_n(x,m)|\right] = O\left(h^{0.5+\varepsilon^*}\right)$ uniformly over $\Psi \times \mathscr{R}$. Similarly, $\operatorname{var}[\hat{\pi}_n(x,m)] = O(1/nh)$ uniformly.

Now V_{2m} converges to zero uniformly because, with $O_{-m} = \{O_j; M_j \neq m\}$ representing the other two-thirds of the data, $E\left[n^{\frac{1}{2}}\eta_{mi}\,|O_{-m}\right]$ and $\text{var}\left[\eta_{mi}\,|O_{-m}\right]$ converge uniformly to zero. Write $\hat{\pi}_n(x, m+2\,(\text{mod}3))$ as $\hat{\pi}(x)$. Then $E\left[n^{\frac{1}{2}}\eta_{mi}\,|O_{-m}\right] = n^{\frac{1}{2}}E[\int\{1-\pi(x)/\hat{\pi}(x)\}\{p(x)-\overline{p}_n(x)\}$ $f(x)\,dx] = n^{\frac{1}{2}}\int\{p(x)-\overline{p}_n(x)\}f(x)\,dx = o(1)$ since (i) by definition of $\hat{\pi}(x)$ and $\overline{p}_n(x)$, $\int\{\pi(x)/\hat{\pi}(x)\}\{p(x)-\overline{p}_n(x)\}f(x)\,dx = 0$ and
(ii) $n^{\frac{1}{2}}\int\{p(x)-\overline{p}_n(x)\}f(x)\,dx$ is uniformly $o(1)$ by $\pi(x;\gamma)$ Lipschitz $0.5+\varepsilon^*$ and $h = n^{-(1-\varepsilon)}$. Further, $\text{var}\left[\eta_{mi}\,|O_{-m}\right] = \int\pi(x)^{-1}\{1-\pi(x)\}\left[p(x)(1-p(x)) + \left(p(x)-\overline{p}_n(x)\right)^2\right]f(x)$ $\left[\{\pi(x)-\hat{\pi}(x)\}^2/\hat{\pi}^2(x)\right]dx$ which is uniformly $O_p\left(h^{2(0.5+\varepsilon^*)}\right) + O_p(1/nh) = o_p(1)$.

To show that the V_{1m} converges to zero uniformly, note $E\left[n^{\frac{1}{2}}\zeta_{mi}\mid O_i, \hat{\pi}(X_i)\right] = n^{\frac{1}{2}}\{E[\hat{p}_n(X_i)\mid O_i, \hat{\pi}(X_i)] - \overline{p}_n(X_i)\} = o_p(1)$ uniformly and $\text{var}\left[\zeta_{mi}\,|O_i, \hat{p}(X_i)\right] = O(1/nh) = o_p(1)$ uniformly.

Now to prove $\hat{\psi}_{\text{eff}}$ is RAL with influence function $\text{EIF}(\psi,\rho)$, it follows from Proposition 3.3.1 of reference 8 that we only need to show that $\sigma_n(\psi,\rho) \to \text{EIF}(\psi,\rho)$ for each (ψ,ρ) which follows from the fact that $\overline{p}_n(x) \to p(x)$ as $n \to \infty$ and dominated convergence.

To prove $\hat{\psi}_{\text{eff}}$ is URG on $\psi \times \mathscr{R}^*$, it suffices to show that $\sigma_n(\psi,\rho) \to \text{EIF}(\psi,\rho)$ uniformly on this submodel. This follows from dominated convergence and the fact that $\overline{p}_n(x) \to p(x)$ uniformly on $\Psi \times \mathscr{R}^*$. Finally, a similar argument can be used to show $\hat{\sigma}_n$ satisfies equation (23).

ACKNOWLEDGEMENTS

Jim Berger, Jon Wellner, Richard Gill, and Sander Greenland provided valuable help. Dr. Robins support for this research was provided in part by grant R01-A132475 from the National Institutes of Health.

REFERENCES

1. Robins, J.M. and Rotnitzky, A. 'Recovery of information and adjustment for dependent censoring using surrogate markers', in: Jewell, N., Dietz, K., Farewell, V. and Boston, M.A. (eds.) AIDS Epidemiology – Methodological Issues, Birkhäuser, 1992, pp. 297–331.
2. Robins, J.M., Mark, S.D. and Newey, W.K. 'Estimating exposure effects by modelling the expectation of exposure conditional on confounders', *Biometrics*, **48,** 479–495 (1992).
3. Robins, J.M., Rotnitzky, A. and Zhao L-P. 1994. 'Estimation of regression coefficients when a regressor is not always observed', *Journal of the American Statistical Association*, **89,** 846–866 (1994).
4. Robins, J.M. and Rotnitzky, A. 'Semiparametric efficiency in multivariate regression models with missing data', *Journal of the American Statistical Association*, **90,** 122–129 (1995).
5. Robins J.M. 'Locally efficient median regression with random censoring and surrogate markers', in: Jewell, N.P. *et al.* (eds), *Lifetime Data: Models in Reliability and Survival Analysis*, Proceedings of the 1994 Conference on Lifetime Data Models in Reliability and Survival Analysis, Boston, MA, 1995, pp. 263–274.
6. Rubin, D.B. 'Inference and missing data', *Biometrika*, **63,** 581–592 (1976).
7. Rubin, D.B. 'The use of propensity scores in applied Bayesian inference', Bernardo, J.M., *et al. Bayesian Statisitics 2*, (ed.), North-Holland/Elsevier, Amsterdam, 1985, pp. 463–472.
8. Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. *Efficient and Adaptive Inference in Semiparametric Models*, Johns Hopkins University Press, Baltimore, MD, 1993.
9. Pepe, M.S. 'Inference using surrogate outcome data and a validation sample', *Biometrika*, **79,** 355–366 (1992).
10. Rotnitzky, A. and Robins, J.M. 'Semiparametric regression estimation in the presence of dependent censoring'. *Biometrika*, **82,** 805–820 (1995).

11. Tsiatis, A.A. 'Estimating regression parameters using linear rank tests for censored data'. *Annals of Statistics*, **18,** 354–372 (1990).
12. Ritov, Y. 'Estimation in a linear regression model with censored data'. *Annals of Statistics*, **18,** 303–328 (1990).
13. Buckley, J. and James, I. 'Linear regression with censored data'. *Biometrika*, **66,** 429–436 (1979).
14. Ying, Z., Jeung, S.H. and Wei, L.J. 'Survival analysis with median regression models'. *Journal of the American Statistical Association*, **90,** 178–185 (1995).
15. Rosenbaum, P.R. 'Conditional permutation tests and the propensity score in observational studies' *Journal of the American Statistical Association*, **79,** 565–574 (1984).
16. Robins, J.M. 'Estimation of the time-dependent accelerated failure time model in the presence of confounding factors'. *Biometrika,* **79,** 321–334 (1992).
17. Horvitz, D.G. and Thompson, D.J. 'A generalization of sampling without replacement from a finite universe', 6, **47,** 663–685 (1952).
18. Berger and Wolpert *The Likelihood Principle*, Springer-Verlag, 1984.
19. Ritov, Y. and Robins, J.M. 'Semiparametric models without smoothness', (manuscript in preparation) (1995).
20. Ritov, Y. and Bickel, P. 'Achieving information bounds in non- and semi-parametric models', *Annals of Statistics*, **18,** 925–938 (1990).
21. Klaassen, C.A.J. 'Non-uniformity of the convergence of location estimators', *Second Prague Symposium on Asymptotic Statistics*, 251–258 (1983).
22. Bickel, P.J. and Klaassen, C.A.J. 'Empirical Bayes estimation in functional and structural models and uniformly adaptive estimation of location', *Advances in Applied Mathematics*, **7,** 55–69 (1986).
23. Donoho, D. 'One-sided inference about functionals of a density', *Annals of Statistics*, **16,** 1390–1420 (1988).
24. Cox and Hinkley *Theoretical Statistics*, Chapman and Hall, London (1974).
25. Robins, J.M., Hsieh, F-S. and Newey, W. 'Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates', *Journal of the Royal Statistical Society, Ser. B*, **57,** 409–424 (1995).
26. Edwards, A.W.F. Likelihood, Cambridge University Press, Cambridge, England (1972).
27. Newey, W., Hsieh, F. and Robins, J. 'Bias corrected semi-parametric estimation', Technical Report, Massachusetts Institute of Technology, Department of Economics, Cambridge, MA (1993).
28. Bickel, P.J. 'On adaptive estimation' *Annals of Statistics*, **10,** 647–671 (1982).
29. Klaassen, C.A.J. 'Consistent estimation of the influence function of locally asymptotically linear estimators', *Annals of Statistics*, **15,** 1548–1562 (1987).
30. Brown, L.D. 'An ancillarity paradox which appears in multiple linear regression (with discussion)' *Annals of Statistics*, **18,** 471–536 (1990).
31. Foster, D.P. and George, E.I. 'A simple ancillarity paradox', *Scandinavian Journal of Statistics* (1996).
32. Hansen, M.H., Madow, W.G. and Tepping, B.J. 'An evaluation of model-dependent and probability-sampling inferences in sample surveys', *Journal of the American Statistical Association*, **78,** 776–793 (1983).
33. Rosenbaum, P.R. and Rubin, D.B. 'The central role of the propensity score in observational studies for causal effects', *Biometrika*, **70,** 41–55 (1983).
34. Godambe V.P. and Thompson M.E., 'Philosophy of Survey Sampling' in Foundations of Probability Theory, Statisticsl Inference and Statistical Theories of Science eds. Harper W, Hooker C., and Reidel D., Univ of Western Ontario Series in Philosophy of Science Vol. V1, P. 102–122.