

Murphy and van der Vaart (MV) provide an elegant characterization of conditions (such as smoothness conditions) under which the profile likelihood for the finite dimensional parameter  $\theta$  in a semiparametric model is approximately quadratic in large samples and the asymptotic distribution of the MLE and profile likelihood ratio test are respectively normal and chi-squared. However, Robins and Ritov (RR;

1997) have argued that, if an analyst plans to use a statistic's asymptotic distribution to approximate its unknown exact distribution, then, in many important high-dimensional models, MV's conditions, even when true, should not be imposed, because the nice asymptotic behavior of the profile likelihood does not reflect its poor moderate sample behavior. In Sections 1–6 of this discussion, we show that when RR's advice is followed and MV's conditions are not imposed, all likelihood-based methods of inference will fail;

---

James M. Robins is Professor of Epidemiology and Biostatistics and Andrea Rotnitzky is Associate Professor of Biostatistics, Harvard School of Public Health, Boston, MA 02115. Mark van der Laan is Associate Professor, School of Public Health, University of California, Berkeley, CA. This research was supported by NIH grants R01AI32475 and R01GM48704-07.

---

© 2000 American Statistical Association  
Journal of the American Statistical Association  
June 2000, Vol. 95, No. 450, Theory and Methods

nevertheless, nonlikelihood inference based on locally efficient estimating equations will often succeed. In Section 8 we discuss an unpublished manuscript of Donglin Zeng and Susan Murphy, wherein they attempt to "save" likelihood inference.

## 1. THE MODELS

We consider models  $M(\mathcal{K} \times \Gamma)$  indexed by infinite-dimensional parameters  $(\kappa, \gamma) \in \mathcal{K} \times \Gamma$  for  $n$  iid copies of a random vector  $X$  with factorized likelihood function  $\mathcal{L}_n(\kappa, \gamma) = \mathcal{L}_{n1}(\kappa)\mathcal{L}_{n2}(\gamma)$  with respect to a dominating measure  $\nu$  that suffer from the "curse of dimensionality" in the sense that the parameter space  $\mathcal{K} \times \Gamma$  is very large. Here,  $\mathcal{L}_n(\kappa, \gamma)$  is the product of the  $n$  unit-specific likelihood contributions  $\mathcal{L}(\kappa, \gamma) = \mathcal{L}_1(\kappa)\mathcal{L}_2(\gamma)$ . We wish to make inference regarding a finite dimensional functional  $\theta \equiv \theta(\kappa)$ . If we write  $\kappa = (\theta, \omega)$  with  $\theta \in \Theta$  and  $\omega \in \Omega$ , then  $\eta \equiv (\omega, \gamma)$  with  $\eta \in \mathcal{N} = \Omega \times \Gamma$  denotes the infinite-dimensional nuisance parameter. The parameters  $\theta$  and  $\omega$  need not be variation independent; that is,  $\mathcal{K}$  need not equal  $\Theta \times \Omega$ .

*Example 1: Semiparametric Regression.* Assume that  $Y = \theta R + h(V; \omega_1) + \varepsilon$ , with  $\varepsilon$  independent of  $(R, V)$ ,  $\varepsilon \sim N(0, 1)$ ,  $R$  Bernoulli, and  $V$  highly multivariate and continuous. The likelihood has factors  $\mathcal{L}_{n1}(\theta, \omega) = \prod_{i=1}^n \phi(Y_i - \theta R_i - h(V_i; \omega_1))f(V_i; \omega_2)$  and  $\mathcal{L}_{n2}(\gamma) = \prod_{i=1}^n \pi(V_i; \gamma)^{R_i} \{1 - \pi(V_i; \gamma)\}^{1-R_i}$ , where  $\phi(\cdot)$  is the standard normal density. Here we assume that  $\gamma$  indexes the set  $\Gamma$  of functions  $\pi(V; \gamma)$  taking values in  $(c, 1-c)$  for a small positive  $c$ ,  $\omega_1$  indexes the set  $\Omega_1$  of uniformly bounded continuous functions  $h(V; \omega_1)$ ,  $\omega_2$  indexes the set  $\Omega_2$  of densities with compact support, and  $\mathcal{K} = \Omega_1 \times \Omega_2$ . Generalizations of this model can be used to analyze a randomized experiment with an additive effect  $\theta$  of the treatment  $R$  on the outcome  $Y$  and with randomization probabilities  $\pi(V; \gamma) = \text{pr}(R = 1|V; \gamma)$  that depend on a vector  $V$  of pretreatment variables. If the treatment effect is not additive, the model in Example 2a below can be used.

*Example 2: Coarsened at Random Missing-Data Models.* Let  $L$  denote a subject's full data and let  $X = (R, c_R(L))$  denote the observed data where  $c_R(\cdot)$  is a known coarsening (i.e., function) of  $L$  depending on  $R$  and  $R$  indicates what part of  $L$  is observed. Let  $M_{\text{ful}}(\mathcal{K})$  denote a semiparametric model for the law of  $L$  with likelihood  $\mathcal{L}_{n, \text{ful}}(\kappa) = \prod_{i=1}^n f(L_i; \kappa)$  dominated by a given measure  $\nu_L$  and, unless otherwise noted, let  $\Gamma$  denote all laws of  $R$  given  $L$  dominated by some measure and satisfying the coarsened at random (CAR) restriction that  $f(R|L; \gamma)$  is a function  $s(X; \gamma)$  of the observed data (Gill, van der Laan, and Robins 1997; Heitjan and Rubin 1991; Jacobsen and Keiding 1995). This induces a model  $M(\mathcal{K} \times \Gamma)$  for  $X$  with  $\mathcal{L}_{n1}(\kappa) = \prod_{i=1}^n \int_{\{l: c_R(l) = c_R(L_i)\}} f(l; \kappa) d\nu_L(l)$  and  $\mathcal{L}_{n2}(\gamma) = \prod_{i=1}^n s(X_i; \gamma)$ .

*Example 2a: Continuously Stratified Random Sampling Model.* Let  $L = (Y, V)$ ,  $X = (R, c_R(L))$ ,  $c_R(L) = (V, RY)$ ,  $Y$  and  $R$  Bernoulli,  $V$  highly multivariate and con-

tinuous,  $\theta$  the mean of  $Y$ , and  $f(L; \kappa) = \zeta(V; \kappa_1)^Y [1 - \zeta(V; \kappa_1)]^{1-Y} f(V; \kappa_2)$ . Here  $\kappa_1$  indexes the set  $\mathcal{K}_1$  of continuous functions  $\zeta(V; \kappa_1) = \text{pr}(Y = 1|V; \kappa_1)$  with range in  $(0, 1)$ ,  $\kappa_2$  indexes the set  $\mathcal{K}_2$  of densities with compact support, and  $\mathcal{K} = \mathcal{K}_1 \times \mathcal{K}_2$ . In this model, CAR implies that  $\text{pr}(R = 1|L; \gamma) \equiv \pi(V; \gamma)$  depends only on  $V$ . Here we let  $\gamma$  index the set  $\Gamma$  of functions  $\pi(V; \gamma)$  with range  $(c, 1)$  for some  $c > 0$ .

## 2. DIFFICULTIES IN ESTIMATION

Let  $\rho^* \equiv (\kappa^*, \gamma^*)$  denote the  $\rho \equiv (\kappa, \gamma)$  generating the data and let  $\mathcal{R} = \mathcal{K} \times \Gamma$ . Ritov and Bickel (1992) and RR studied many models  $M(\mathcal{K} \times \Gamma)$ , including all of those that we consider in this discussion, which have the following properties: (a) The semiparametric variance bound (SVB)  $\bar{I}_\rho^{-1}$  for  $n^{1/2}$ -consistent estimators of  $\theta$  is finite, and yet no estimator is consistent for  $\theta$  uniformly over  $\rho^* \in \mathcal{R}$ ; (b) no estimator of  $\theta$  attains a pointwise (i.e., nonuniform) rate of convergence of  $n^\alpha$  at all  $\rho^* \in \mathcal{R}$  for any  $\alpha > 0$ ; (c) no "valid"  $1 - \alpha$  interval estimator for  $\theta$  exists. By valid, we mean that under all  $\rho^* \in \mathcal{R}$ , (1) the coverage is at least  $(1 - \alpha)$  at each sample size  $n$  and (2) the length goes to 0 in probability with increasing sample size.

### 2.1 The Curse of Dimensionality

In Examples 1 and 2a, there do exist uniformly  $n^{1/2}$ -consistent estimators and valid confidence intervals with length  $O_p(n^{-1/2})$  in submodels  $M(\mathcal{K}_{\text{sub, smooth}} \times \Gamma)$ , which impose the additional assumption that  $h(V; \omega_1)$  and  $\zeta(V; \kappa_1)$  are locally smooth (i.e., differentiable to a suitably high order with bounded derivatives) in  $V$ . However, when  $V$  is high dimensional, RR argued that even when the submodel  $M(\mathcal{K}_{\text{sub, smooth}} \times \Gamma)$  is known to be correct, asymptotics based on the larger model  $M(\mathcal{K} \times \Gamma)$  that does not assume smoothness provides a more relevant and appropriate guide to moderate sample performance. For example, with moderate size samples, there do not exist interval estimators that perform well (in the sense of being narrow enough to be substantively useful while covering  $\theta$  at near nominal level) under all laws allowed by the smooth submodels. This is due to the curse of dimensionality; in high-dimensional models with moderate sample sizes, local smoothness assumptions, even when true, are not useful, because essentially no two units will have  $V$  vectors close enough to one another to allow the "borrowing of information" necessary for smoothing.

### 3. ESTIMATION IN MODELS WITH $\gamma^*$ KNOWN

We now consider inference about  $\theta$  in the (sub)model  $M(\mathcal{K})$  in which  $\gamma^*$  is known. This is of interest because in the designed experiments of Examples 1 and 2a,  $\gamma^*$  is usually known to the analyst.

#### 3.1 Profile Likelihood Inference

Suppose that profile likelihood inference for  $\theta = \theta(\kappa)$  is to be based on a working or sieve likelihood  $\mathcal{L}_n^*(\kappa, \gamma) = \mathcal{L}_{n1}^*(\kappa)\mathcal{L}_{n2}^*(\gamma)$  with  $(\kappa, \gamma) \in \mathcal{K}_n \times \Gamma_n$ . Then inference for  $\theta$  is the same regardless of the known value  $\gamma^*$  of  $\gamma$ . RR

referred to any inferential method with this property as strict factorization-based (SFB). A SFB estimator is then one that is not a function of the known value  $\gamma^*$  of  $\gamma$ . In model  $M(\mathcal{K})$  any method that obeys the likelihood principle is SFB (Robins and Wasserman 2000). Because any SFB estimator in model  $M(\mathcal{K})$  can also serve as an estimator in model  $M(\mathcal{K} \times \Gamma)$ , it follows that the aforementioned properties (a)–(c) also hold for any SFB estimator in model  $M(\mathcal{K})$ . In particular, (a)–(c) are true for profile-likelihood inference.

### 3.2 Non-SFB Inference

If model  $M(\mathcal{K} \times \Gamma)$  satisfies condition A given later, then, in model  $M(\mathcal{K})$  there exist non-SFB estimators that do not suffer from (a)–(c). In fact, for these models (1) there exist non-SFB estimators  $\hat{\theta}_n(\gamma^*)$  depending on the known  $\gamma^*$  that are not only uniformly  $n^{1/2}$  consistent but also uniformly asymptotically normal (UAN), and (2) the Wald interval  $C_n(\gamma^*) = \hat{\theta}_n(\gamma^*) \pm z_{\alpha/2} \hat{\sigma}_n / \sqrt{n}$  is a uniform  $1 - \alpha$  asymptotic confidence interval with length  $O_p(n^{-1/2})$ , which implies that a valid  $1 - \alpha$  interval estimator exists. Here  $\hat{\sigma}_n/n^{1/2}$  is, for example, the nonparametric bootstrap estimate of the standard error of  $\hat{\theta}_n(\gamma^*)$  and  $z_\alpha$  is the  $\alpha$  quantile of a  $N(0, 1)$ . By UAN, we mean that there exists a sequence  $\sigma_n(\rho)$ , such that, for all  $t$ ,  $\sup_{\rho^* \in \mathcal{R}} |Pr_{\rho^*}[\{ \sqrt{n}/[\sigma_n(\rho^*)] \}(\hat{\theta}_n(\gamma^*) - \theta^*) < t] - \Phi(t)| \rightarrow 0$  as  $n \rightarrow \infty$ , where  $\Phi$  is the standard normal cdf. Importantly, in moderate-size samples,  $C_n(\gamma^*)$  will generally cover  $\theta^*$  with probability close to nominal. We stress uniform procedures, because uniformity is necessary (although not sufficient) to link asymptotic behavior to moderate sample behavior. For example, if an estimator of  $\theta$  is consistent but not uniformly consistent, then there does not exist a sample size  $n_0$ , not depending on  $\rho^*$ , at which the difference between the estimator and  $\theta$  is guaranteed to be small with high probability. Uniformity is not sufficient because, as in model  $M(\mathcal{K}_{\text{sub,smooth}} \times \Gamma)$ , the required sample size  $n_0$  may be enormous.

Let  $L_2^0(\rho)$  be the Hilbert space with covariance inner product of random vectors of the dimension of  $\theta$  with mean 0 and finite variance matrix under  $\rho$ . Let  $T_\rho(\Omega) \subset L_2^0(\rho)$  be the tangent space (i.e., closed linear span of scores) for the nuisance parameter  $\omega$  when the data is generated under  $\rho$  and let  $T_\rho(\Omega)^\perp$  be its orthogonal complement in  $L_2^0(\rho)$ . We shall restrict attention to models satisfying the following condition. More general models are considered by Robins and Ritov (1997).

**Condition A.** There exists a (possibly improper) subset  $\mathcal{W}_\rho(\Omega) = \{U_g(\theta, \gamma); g \in \mathcal{G}\}$  of  $T_\rho(\Omega)^\perp$  indexed by functions  $g$  ranging over a set  $\mathcal{G}$  that is comprised solely of unbiased estimating functions  $U_g(\theta, \gamma)$ . That is,  $\mathcal{W}_\rho(\Omega) = \mathcal{W}_{\theta\gamma}(\Omega)$  does not depend on the value of  $\omega$  generating the data.

Condition A is satisfied in the models of Examples 1 and 2a. For example, in 2a, we can take  $\mathcal{W}_\rho(\Omega) = \{U_g(\theta, \gamma) = gR\pi(V; \gamma)^{-1}(Y - \theta); g \in R^1\}$ . In fact, condition A is always satisfied in CAR models in which  $\theta$  is the mean of a random variable  $b(L)$  and  $M_{\text{ful}}(\mathcal{K})$  is nonparametric (i.e., contains all laws dominated by  $\nu_L$ ). In example 1, we can

take  $\mathcal{W}_\rho(\Omega) = \{U_g(\theta, \gamma) = [Y - \theta R]g(V)[R - E_\gamma(R|V)]; g(\cdot)$  an arbitrary function}. More generally, suppose there exists  $\gamma^\dagger$  and  $H_g(\theta)$  such that  $E_{\kappa\gamma^\dagger}[H_g(\theta)] = 0$ . Then Condition A holds with  $U_g(\theta, \gamma) = H_g(\theta)\mathcal{L}_2(\gamma^\dagger)/\mathcal{L}_2(\gamma)$ , when  $\text{support}(\kappa, \gamma^\dagger) \subset \text{support}(\kappa, \gamma)$  for all  $(\kappa, \gamma) \in \mathcal{K} \times \Gamma$ . Here  $\text{support}(\kappa, \gamma)$  is the support of  $X$  under  $(\kappa, \gamma)$ . To see why the properties (1) and (2) hold for model  $M(\mathcal{K})$  when condition A is true, note that under regularity conditions such as uniform bounds on higher-order moments, the estimator  $\hat{\theta}(\gamma^*) \equiv \hat{\theta}_n(\gamma^*)$  solving  $\sum_{i=1}^n U_{g,i}(\theta, \gamma^*) = 0$  for  $U_g(\theta, \gamma) \in \mathcal{W}_\rho(\Omega)$  is UAN.

### 4. LOCAL EFFICIENCY

Because of the factorization of the likelihood into a  $\gamma$  part and a  $\kappa$  part, in model  $M(\mathcal{K} \times \Gamma)$  (a) the tangent space  $T_\rho(\Gamma) = T_\gamma(\Gamma)$  for  $\gamma$  at  $\rho$  does not depend on  $\kappa$ ; (b) if condition A is satisfied, the orthogonal complement  $T_\rho(\mathcal{N})^\perp$  to the tangent space for the nuisance parameter  $\eta = (\omega, \gamma)$  includes  $\mathcal{W}_\rho(\mathcal{N}) = \{\tilde{U}_g(\theta, \omega, \gamma) = U_g(\theta, \gamma) - \Pi_{\kappa\gamma}[U_g(\theta, \gamma)]; U_g(\theta, \gamma) \in \mathcal{W}_\rho(\Omega)\}$ , where  $\Pi_{\kappa\gamma}[D]$  is the projection of  $D$  on  $T_\gamma(\Gamma)$ ; and (c) the efficient score  $\tilde{l}_\rho \in T_\rho(\mathcal{N})^\perp$  for  $\theta$  at  $\rho$  is the same in models  $M(\mathcal{K})$  and  $M(\mathcal{K} \times \Gamma)$ .

When  $\mathcal{L}_2(\gamma)$  is a partial likelihood that is unrestricted (i.e., nonparametric) as  $\gamma$  varies over  $\Gamma$ , Robins (1999, Theorem 3.2) provides a closed form expression for  $\Pi_{\kappa\gamma}[U_g(\theta, \gamma)]$ . For example, suppose for a sequence  $\{(A_k, H_k)\}$ ,  $\mathcal{L}_2(\gamma) = \prod_k f[A_k | H_k; \gamma]$ . Then if  $H_{k+1}$  includes  $(A_k, H_k)$  as components,  $\mathcal{L}_2(\gamma)$  is a partial likelihood and  $\Pi_{\kappa\gamma}[D] = \sum_k E_{\kappa\gamma}[D | A_k, H_k] - E_{\kappa\gamma}[D | H_k]$  when  $\Gamma$  is nonparametric.

Suppose now that  $\mathcal{W}_\rho(\Omega) = \{U_g(\theta, \gamma); g \in \mathcal{G}\}$  is such that for some  $g_{\text{opt}, \rho} \in \mathcal{G}$  (possibly depending on  $\rho$ ),  $\tilde{U}_{g_{\text{opt}, \rho}}(\theta, \omega, \gamma)$  is the efficient score  $\tilde{l}_\rho$  for  $\theta$  at  $\rho$ . In such a case we can specify a submodel  $M(\mathcal{K}_{\text{sub}})$  with parameter space  $\mathcal{K}_{\text{sub}}$  for  $\kappa = (\theta, \omega)$  small enough that the regularity conditions of MV hold. Let  $\hat{\omega}(\theta)$  and  $\hat{\kappa}$  respectively be the profile MLE of  $\omega$  and MLE of  $\kappa$  in model  $M(\mathcal{K}_{\text{sub}})$ . Let  $\tilde{U}_g(\theta, \gamma; \hat{\kappa}) = U_g(\theta, \gamma) - \Pi_{\hat{\kappa}\gamma}[U_g(\theta, \gamma)]$ . Then, under further regularity conditions, the estimators  $\hat{\theta}_{\text{loceff}}(\gamma^*)$  and  $\hat{\theta}_{\text{loceff}}(\gamma^*)$  solving  $B_{\text{loceff}}(\theta, \gamma^*) \equiv \sum_{i=1}^n \tilde{U}_{i, g_{\text{opt}, \theta, \hat{\omega}(\theta), \gamma^*}}(\theta, \hat{\omega}(\theta), \gamma^*) = 0$  and  $\hat{B}_{\text{loceff}}(\theta, \gamma^*) = \sum_{i=1}^n \tilde{U}_{i, g_{\text{opt}, \hat{\kappa}, \gamma^*}}(\theta, \gamma^*; \hat{\kappa}) = 0$  are UAN under model  $M(\mathcal{K})$  and are locally semiparametric efficient in model  $M(\mathcal{K})$  at the submodel  $M(\mathcal{K}_{\text{sub}})$ . This means that they have asymptotic variance equal to the SVB for model  $M(\mathcal{K})$  when the model  $M(\mathcal{K}_{\text{sub}})$  is correct. Indeed, they are locally curse of dimensionality appropriate (CODA) efficient as defined later. For expositional ease, we focus on  $\hat{\theta}_{\text{loceff}}(\gamma^*)$  but all results apply equally to  $\hat{\theta}_{\text{loceff}}(\gamma^*)$ ; when  $\theta$  and  $\omega$  are not variation-independent,  $\hat{\theta}_{\text{loceff}}(\gamma^*)$  will often be easier to compute than  $\hat{\theta}_{\text{loceff}}(\gamma^*)$ .

**Example 1 (Continued).**  $\tilde{U}_g(\theta, \omega, \gamma) = \{Y - \theta R - h(V; \omega_1)\}\{R - E_\gamma[R|V]\}g(V)$  and  $g_{\text{opt}, \rho}(V) = 1$ . Hence  $\tilde{l}_\rho = E_{\omega_2}[\pi(V; \gamma) - \pi^2(V; \gamma)]$ . A typical choice for  $\mathcal{K}_{\text{sub}} \equiv \Theta \times \Omega_{\text{sub}, 1} \times \Omega_2$  would be to impose a parametric or a smooth generalized additive model for  $h(V; \omega_1)$ .

*Example 2 (Continued).* Suppose in model  $M_{\text{ful}}(\mathcal{K})$  the orthogonal complement  $T_{\kappa, \text{ful}}(\Omega)^\perp = T_{\theta, \text{ful}}(\Omega)^\perp = \{U_{\text{ful},g}(\theta), g \in \mathcal{G}\}$  to the nuisance tangent space  $T_{\kappa, \text{ful}}(\Omega)$  for  $\omega$  at  $\kappa$  does not depend on  $\omega$ . Here the set  $\mathcal{G}$  indexes the elements of  $T_{\theta, \text{ful}}(\Omega)^\perp$ . A sufficient condition for this is that  $\theta$  is the mean of some  $b(L)$  and  $M_{\text{ful}}(\mathcal{K})$  is nonparametric in which case  $U_{\text{ful},g}(\theta) = (b(L) - \theta)g$  with  $\mathcal{G} = R^1$ . Let  $\Delta_g = 1$  if  $U_{\text{ful},g}(\theta)$  is a function of  $X$  and  $\Delta_g = 0$  otherwise. To define  $\mathcal{W}_\rho(\Omega) = \{U_g(\theta, \gamma); g \in \mathcal{G}\}$  and  $\mathcal{W}_\rho(\mathcal{N}) = \{\tilde{U}_g(\theta, \omega, \gamma)\}$ , we need the following definitions given by Bickel et al. (1993). Let  $\mathbf{g}_\kappa(\cdot)$  denote the missing-data score operator  $E_\kappa(\cdot|L)$  and let  $\mathbf{m}_\rho(\cdot)$  denote the nonparametric information operator  $E_\gamma\{E_\kappa(\cdot|X)|L\}$ . Let  $\mathbf{m}_\rho^{-1}$  be the (possibly generalized) inverse of  $\mathbf{m}_\rho$ . If, for all  $\gamma \in \Gamma$ ,  $\text{pr}(\Delta_g = 1|L; \gamma) = \pi_g(L; \gamma)$  is bounded away from 0 with probability 1, then we define  $U_g(\theta, \gamma) = \Delta_g U_{\text{ful},g}(\theta)/\pi_g(L; \gamma)$  (Robins and Rotnitzky 1992). If  $\pi_g(L; \gamma)$  is not positive with probability 1, then, provided  $U_{\text{ful},g}(\theta)$  is in the range of  $\mathbf{m}_{\theta, \omega^\dagger, \gamma}$ , take  $U_g(\theta, \gamma) = \mathbf{g}_{\theta, \omega^\dagger, \gamma}\{\mathbf{m}_{\theta, \omega^\dagger, \gamma}^{-1}(U_{\text{ful},g}(\theta))\}$ , where  $\omega^\dagger \in \Omega$  is arbitrary (van der Laan, Robins, and Gill 2000). In either case it can be shown that  $U_g(\theta, \gamma) \in T_\rho(\Omega)^\perp$  and  $\tilde{U}_g(\theta, \omega, \gamma) = \mathbf{g}_\kappa\{\mathbf{m}_\rho^{-1}(U_{\text{ful},g}(\theta))\} \in T_\rho(\mathcal{N})^\perp$ . Further if  $\tilde{l}_\rho$  is a score then  $\tilde{l}_\rho = \tilde{U}_{g_{\text{opt}, \rho}}(\theta, \omega, \gamma)$  for some  $g_{\text{opt}, \rho} \in \mathcal{G}$ . When  $M_{\text{ful}}(\mathcal{K})$  is nonparametric, all  $g \in \mathcal{G}$  lead to the same estimator, and the dependence on  $g$  can be suppressed in the notation. In that case the only remaining issue is computation of  $\mathbf{m}_\rho^{-1}(\cdot)$ . When  $\mathcal{L}_2(\gamma)$  is a partial likelihood (which includes monotone missing-data patterns), Robins and Rotnitzky (1992) and Robins (1999, Eq. 15) provided a closed-form expression. When  $\mathcal{L}_2(\gamma)$  is not a partial likelihood, van der Laan (1996a) showed how to calculate  $\mathbf{m}_\rho^{-1}(\cdot)$  iteratively by successive approximation (Kress 1989); that is,  $\mathbf{m}_\rho^{-1}(U)$  is the limit of the sequence  $D_j = d_j(L)$  where  $D_j = U + D_{j-1} - \mathbf{m}_\rho(D_{j-1})$  with  $D_0$  an arbitrary starting random variable. Robins and Wang (1998) provided a worked example that also involved computing  $g_{\text{opt}, \rho}$  for a semiparametric model  $M_{\text{ful}}(\mathcal{K})$ .

*Example 2a (Continued).* Here  $M_{\text{ful}}(\mathcal{K})$  is nonparametric,  $U_{\text{ful}}(\theta) = Y - \theta$ ,  $\Delta = R$ ,  $\pi(L; \gamma) = \pi(V; \gamma)$ ,  $\mathbf{m}_\rho^{-1} = \pi(V; \gamma)^{-1}U_{\text{ful}}(\theta) - \{\pi(V; \gamma)^{-1} - 1\}E_{\kappa_1}[U_{\text{ful}}(\theta)|V]$ , and  $\tilde{U}(\theta, \omega, \gamma) = \Delta\pi(V; \gamma)^{-1}U_{\text{ful}}(\theta) - \{\Delta\pi(V; \gamma)^{-1} - 1\}E_{\kappa_1}[U_{\text{ful}}(\theta)|V]$ .

## 5. LOCAL VERSUS GLOBAL EFFICIENCY

We first present some asymptotic results, then describe their relevance for moderate sample performance. Consider model  $M(\mathcal{K})$  with  $\gamma$  known and equal to  $\gamma^*$ . In Example 1, for each  $\theta$  we can estimate  $h(V; \omega_1)$  by a multivariate nonparametric kernel regression of  $Y - \theta R$  on  $V$ , and in Example 2a we can estimate  $\zeta(V; \kappa_1)$  by a multivariate kernel regression of  $Y$  on  $V$  among subjects with  $R = 1$ , where in each case the bandwidth  $b(n)$  satisfies  $b(n) \rightarrow 0$  and  $nb(n) \rightarrow \infty$  as  $n \rightarrow \infty$ . RR showed that the estimator  $\hat{\theta}_{\text{eff}}(\gamma^*)$  solving  $B_{\text{eff}}(\theta, \gamma^*) = 0$  is UAN with  $\sigma_n^2(\rho^*)$  converging to the SVB  $\tilde{I}_{\rho^*}^{-1}$  for  $\theta$  for all  $\rho^* \in \mathcal{R}$ , where  $B_{\text{eff}}(\theta, \gamma^*)$  is a slightly modified version of  $B_{\text{loceff}}(\theta, \gamma^*)$  that uses sample splitting techniques. This implies that  $\hat{\theta}_{\text{eff}}(\gamma^*)$  is globally semiparametric efficient in the sense of

Bickel et al. (1993). However, RR also showed that with  $\gamma^*$  fixed and known, the convergence of  $\sigma_n^2(\rho^*)/\tilde{I}_{\rho^*}^{-1}$  to 1 is not uniform over  $\kappa^* \in \mathcal{K}$  and, moreover, there can be no UAN estimator for which the convergence is uniform. RR defined a UAN estimator to be globally CODA efficient if and only if the convergence of  $\sigma_n^2(\rho^*)/\tilde{I}_{\rho^*}^{-1}$  to 1 is uniform.

RR argued that global CODA efficiency is a more appropriate concept of global asymptotic efficiency than the Bickel et al. definition in the sense that it is natural to demand of a globally efficient estimator  $\hat{\theta}_{\text{globeff}}(\gamma^*)$  that the associated theoretical interval  $\hat{\theta}_{\text{globeff}}(\gamma^*) \pm z_{\alpha/2}\tilde{I}_{\rho^*}^{-1/2}/\sqrt{n}$  that uses the efficient variance but an asymptotic  $(1 - \alpha)$  confidence interval for  $\theta$  uniformly for  $\kappa^* \in \mathcal{K}$ . This requires the estimator to be globally CODA efficient. For example, because in the models of Examples 1 and 2a there exists no globally CODA efficient estimator, it follows that for each  $\gamma^*$  and some  $\varepsilon > 0$ , at each sample size  $n$  there exists some law  $\kappa^* \in \mathcal{K}$  (that changes with  $n$ ) such that the probability that  $\theta^*$  lies in  $\hat{\theta}_{\text{eff}}(\gamma^*) \pm z_{\alpha/2}\tilde{I}_{\rho^*}^{-1/2}/\sqrt{n}$  under  $\rho^*$  is less than  $1 - \alpha - \varepsilon$ . This reflects the fact that, due to the nonuniformity of the convergence,  $\tilde{I}_{\rho^*}^{-1}$  underestimates  $\sigma_n(\rho^*)$  at sample size  $n$  under distributions  $\rho^*$  for which the functions  $h(V; \omega_1)$  of Example 1 and  $\zeta(V; \kappa_1)$  of Example 2a are sufficiently wiggly.

Now  $\hat{\theta}_{\text{eff}}(\gamma^*)$  is globally CODA efficient and  $\hat{\theta}_{\text{eff}}(\gamma^*) \pm z_{\alpha/2}\tilde{I}_{\rho^*}^{-1/2}/\sqrt{n}$  is a uniform asymptotic  $(1 - \alpha)$  interval on the submodels  $M(\mathcal{K}_{\text{sub}, \text{smooth}})$ , which impose only the additional assumption that  $h(V; \omega_1)$  and  $\zeta(V; \kappa_1)$  are locally smooth in  $V$ . However, when  $V$  is high-dimensional, even when the submodel  $M(\mathcal{K}_{\text{sub}, \text{smooth}})$  is known to be correct, the asymptotics based on the larger model  $M(\mathcal{K})$  provides a more relevant and appropriate guide to moderate sample performance. For example, with moderate size samples, for any estimator  $\hat{\theta}(\gamma^*)$ , there will exist laws  $\rho^*$  in  $M(\mathcal{K}_{\text{sub}, \text{smooth}})$  such that  $\hat{\theta}(\gamma^*) \pm z_{\alpha/2}\tilde{I}_{\rho^*}^{-1/2}/\sqrt{n}$  covers  $\theta^*$  with probability much less than the nominal  $(1 - \alpha)$ .

Under regularity conditions, the estimator  $\hat{\theta}_{\text{loceff}}(\gamma^*)$  solving  $B_{\text{loceff}}(\theta, \gamma^*) = 0$  is a locally CODA efficient estimator in model  $M(\mathcal{K})$  at the submodel  $M(\mathcal{K}_{\text{sub}})$ . By this we mean  $\hat{\theta}_{\text{loceff}}(\gamma^*)$  is UAN in model  $M(\mathcal{K})$  and  $\sigma_n^2(\rho^*)/\tilde{I}_{\rho^*}^{-1}$  converges to 1 uniformly for  $\kappa^* \in \mathcal{K}_{\text{sub}}$ . When, as in our examples, there are no globally efficient CODA estimators, local CODA efficiency is the best that can be hoped for. The bootstrap interval  $\hat{\theta}_{\text{loceff}}(\gamma^*) \pm z_{\alpha/2}\hat{\sigma}_n/\sqrt{n}$  will be an uniform asymptotic  $(1 - \alpha)$  interval on model  $M(\mathcal{K})$ , with length uniformly close to that of  $\hat{\theta}_{\text{loceff}}(\gamma^*) \pm z_{\alpha/2}\tilde{I}_{\rho^*}^{-1/2}/\sqrt{n}$  on the submodel  $M(\mathcal{K}_{\text{sub}})$ ; in addition, the latter interval is an uniform asymptotic  $(1 - \alpha)$  interval for  $\rho^*$  restricted to  $M(\mathcal{K}_{\text{sub}})$ . Most important, if  $M(\mathcal{K}_{\text{sub}})$  is sufficiently small, then in moderate-size samples, the intervals  $\hat{\theta}_{\text{loceff}}(\gamma^*) \pm z_{\alpha/2}\hat{\sigma}_n/\sqrt{n}$  will be narrow enough to be substantively useful with actual coverage close to the nominal  $(1 - \alpha)$  under all laws  $\rho^*$  contained in model  $M(\mathcal{K})$  and will have length nearly equal to that of the interval  $\hat{\theta}_{\text{loceff}}(\gamma^*) \pm z_{\alpha/2}\tilde{I}_{\rho^*}^{-1/2}/\sqrt{n}$  for  $\rho^* \in M(\mathcal{K}_{\text{sub}})$ . For example, this will generally be the case for submodels  $M(\mathcal{K}_{\text{sub}})$  that assume that  $h(V; \omega_1)$  and  $\zeta(V; \kappa_1)$  fol-

low particular parametric or smooth generalized additive models.

## 6. LOWER-DIMENSIONAL SUBMODELS OF $\Gamma$ WHEN $\gamma$ IS UNKNOWN

In many settings,  $\gamma^*$  is not known but the analyst assumes that  $\gamma \in \Gamma_{\text{sub}} \subset \Gamma$ . Let  $\hat{\gamma} \in \Gamma_{\text{sub}}$  be an estimate of  $\gamma^*$ . Then the estimator  $\hat{\theta}_{\text{loceff}}(\hat{\gamma})$  solving  $B_{\text{loceff}}(\theta, \hat{\gamma}) = 0$  will be a locally efficient estimator in the model  $M(\mathcal{K} \times \Gamma_{\text{sub}})$  at the submodel  $M(\mathcal{K}_{\text{sub}} \times \Gamma_{\text{sub}})$ , provided that the size of  $\Gamma_{\text{sub}}$  is small enough so that  $\hat{\gamma}$  converges to  $\gamma^*$  sufficiently fast.

*Examples 1 and 2a (Continued).* Let  $\Gamma_{\text{sub}} \subset R^q$  be the parameter space for  $\gamma$  in a regular parametric submodel for the distribution of  $R$  given  $V$ . The parametric MLE  $\hat{\gamma}$  maximizes  $L_{n2}(\gamma)$  over  $\Gamma_{\text{sub}}$ .

*Example 2b: Independently Censored Multivariate Survival Data.* Suppose that  $\mathbf{L} \equiv \mathbf{T} \equiv (T_1, \dots, T_m)$  and  $\mathbf{C} \equiv (C_1, \dots, C_m)$  are continuous multivariate failure and censoring times and we observe  $\mathbf{X} = (\mathbf{Y}, \tau) = (\mathbf{R}, c_{\mathbf{R}}(\mathbf{L}))$ , where  $\mathbf{Y}, \tau, \mathbf{R}$ , and  $c_{\mathbf{R}}(\mathbf{L})$  have components  $Y_j = \min\{T_j, C_j\}$ ,  $\tau_j = I(Y_j = T_j)$ ,  $R_j = C_j$  if  $\tau_j = 0$  and  $R_j = \infty$  otherwise, and  $c_{\mathbf{R}}(\mathbf{L})_j = L_j$  if  $R_j = \infty$  and 0 otherwise. We take as our parameter of interest  $\theta = \text{pr}(\mathbf{T} > \mathbf{t})$  for a given vector  $\mathbf{t}$ , and we allow  $M_{\text{ful}}(\mathcal{K})$  to be nonparametric. Thus  $U_{\text{ful}}(\theta) = I(\mathbf{T} > \mathbf{t}) - \theta$ ,  $\Delta = I\{\mathbf{C} > \mathbf{T}^*\}$  and  $\pi(\mathbf{L}; \gamma) = \text{pr}(\mathbf{C} > \mathbf{T}^* | \mathbf{L}; \gamma)$ , where  $\mathbf{T}^*$  has components  $T_j^* = \min(T_j, t_j)$ . We complete the model by taking  $\Gamma_{\text{sub}}$  to be all distributions with  $\mathbf{C}$  independent of  $\mathbf{T}$ . Then  $\pi(\mathbf{L}; \gamma) = S(\mathbf{T}^*; \gamma)$ , where  $S(\mathbf{t}; \gamma) \equiv \text{pr}(\mathbf{C} > \mathbf{t}; \gamma)$ .

To construct a locally semiparametric estimator, we take  $M_{\text{ful}}(\mathcal{K}_{\text{sub}})$  to be the semiparametric gamma frailty (or, alternatively, a fully parametric) model for  $\mathbf{L}$  and  $S(\mathbf{t}; \hat{\gamma})$  to be the Dabrowska (1988) estimator of the censoring distribution. Then  $\hat{\theta}_{\text{loceff}}(\hat{\gamma})$  improves on previous estimators; unlike the (repaired) MLE of van der Laan (1996a,b) but like the Dabrowska, Prentice-Cai (1992), and Bickel (Dabrowska 1988) estimators, it does not require smoothing and so will perform well in moderate-size samples even if  $m$  is large, say 7 or 8. Like van der Laan's MLE but unlike all other previous estimators,  $\hat{\theta}_{\text{loceff}}(\hat{\gamma})$  can be efficient (nearly efficient) in the independent censoring model  $M(\mathcal{K} \times \Gamma_{\text{sub}})$  even when the components of  $\mathbf{T}$  are highly dependent, whenever the model  $M_{\text{ful}}(\mathcal{K}_{\text{sub}})$  for  $\mathbf{L}$  is correct (nearly correct). In general,  $m_p^{-1}$  can be computed by successive approximation. Suppose, however, that there is a common censoring time; that is,  $C_1 = \dots = C_m = C$ . Then, missingness is monotone and  $m_p^{-1}$  can be explicitly computed. In this setting with  $T_j \leq T_{j+1}$  with probability one and  $t_j \leq t_{j+1}$ , Gill et al. (1997), Lin, Sun, and Ying (1999), and Satten and Datta (2000) considered the inefficient estimator  $\hat{\theta}_{\text{ineff}}(\hat{\gamma})$  solving  $0 = \sum_{i=1}^n U_i(\theta, \hat{\gamma})$  with  $U(\theta, \hat{\gamma}) = \Delta\{I(\mathbf{T} > \mathbf{t}) - \theta\}/S(\mathbf{T}_m^*; \hat{\gamma})$  where  $S(u, \hat{\gamma})$  is the Kaplan Meier estimator for censoring;  $\hat{\theta}_{\text{loceff}}(\hat{\gamma})$  solves  $0 = \sum_{i=1}^n \tilde{U}_i(\theta, \hat{\omega}(\theta), \hat{\gamma})$  where  $\tilde{U}(\theta, \omega, \hat{\gamma}) = U(\theta, \hat{\gamma}) + \int_0^\infty dM_C(u; \hat{\gamma})\{S(u; \hat{\gamma})\}^{-1} E_{\theta\omega}\{I(\mathbf{T} > \mathbf{t}) - \theta\}\{\min(T_j^*, u), j = 1, \dots, m\}$ ,  $dM_C(u) = I(C = u, T_m^* \geq u) - d\Lambda_C(u)$ ,

$\hat{\gamma})I(C \geq u, T_m^* \geq u)$ , and  $\Lambda_C(u; \hat{\gamma}) = -\ln S(u; \hat{\gamma})$ . Van der Laan, Hubbard, and Robins (1999) studied  $\hat{\theta}_{\text{loceff}}(\hat{\gamma})$  which, for this model, is easier to compute than  $\hat{\theta}_{\text{loceff}}(\hat{\gamma})$ .

## 7. DOUBLE ROBUSTNESS

Because when  $\gamma^*$  is unknown, the analyst cannot be certain that  $\gamma^* \in \Gamma_{\text{sub}}$  or that  $\kappa^* \in \mathcal{K}_{\text{sub}}$ , the best that can be hoped for is an estimator that is UAN in the union model that assumes that  $\rho^*$  lies in either  $M(\mathcal{K} \times \Gamma_{\text{sub}})$  or  $M(\mathcal{K}_{\text{sub}} \times \Gamma)$ . We refer to such an estimator as doubly robust or doubly protected. It follows from Lemma 1 that, under suitable regularity conditions,  $\hat{\theta}_{\text{loceff}}(\hat{\gamma})$  is doubly robust, provided that the model  $M(\mathcal{K} \times \Gamma)$  is convex in  $\gamma$ . Convexity in  $\gamma$  means that, as in Examples 1 and 2, if the laws indexed by  $(\kappa, \gamma_1)$  and  $(\kappa, \gamma_2)$  are in  $M(\mathcal{K} \times \Gamma)$ , then so is any mixture of the two. Furthermore,  $\hat{\theta}_{\text{loceff}}(\hat{\gamma})$  is locally CODA efficient in the union model at the intersection submodel  $M(\mathcal{K}_{\text{sub}} \times \Gamma_{\text{sub}})$ . Thus in Example 2b,  $\hat{\theta}_{\text{loceff}}(\hat{\gamma})$  is UAN in the union model that assumes that either (a) the data are CAR and  $\mathbf{T}$  follows a gamma frailty model or (b)  $\mathbf{T}$  and  $\mathbf{C}$  are independent. In contrast,  $\hat{\theta}_{\text{ineff}}(\hat{\gamma})$  will be inconsistent unless (b) holds.

*Lemma 1.* Suppose that  $M(\mathcal{K} \times \Gamma)$  is convex in  $\gamma$ . Then for any  $\omega^\dagger$  and  $\gamma^\dagger$  such that  $\text{support}(\kappa, \gamma) \subset \text{support}(\kappa, \gamma^\dagger)$  for all  $\gamma \in \Gamma$ ,  $E_{\theta, \omega, \gamma}\{\tilde{U}(\theta, \omega^\dagger, \gamma)\} = E_{\theta, \omega, \gamma^\dagger}\{\tilde{U}(\theta, \omega^\dagger, \gamma^\dagger)\} = 0$  where the dependence on  $g$  has been suppressed in the notation.

*Proof.* By  $E_{\theta, \omega, \gamma}\{\tilde{U}(\theta, \omega, \gamma^\dagger)\} = E_{\theta, \omega, \gamma^\dagger}\{\tilde{U}(\theta, \omega, \gamma^\dagger)\} f(X; \theta, \omega, \gamma)/f(X; \theta, \omega, \gamma^\dagger)$  and  $E_{\theta, \omega, \gamma^\dagger}\{\tilde{U}(\theta, \omega, \gamma^\dagger)\} = 0$ , we have  $E_{\theta, \omega, \gamma}\{\tilde{U}(\theta, \omega, \gamma^\dagger)\} = E_{\theta, \omega, \gamma^\dagger}\{\tilde{U}(\theta, \omega, \gamma^\dagger) S(\theta, \omega, \gamma^\dagger)\}$ , where  $S(\theta, \omega, \gamma^\dagger) = f(X; \theta, \omega, \gamma)/f(X; \theta, \omega, \gamma^\dagger) - 1$ . But  $S(\theta, \omega, \gamma^\dagger) \equiv S(\gamma^\dagger)$  is an element of  $T_{\gamma^\dagger}(\Gamma)$ , because convexity in  $\gamma$  implies that under  $\gamma^\dagger$ ,  $\{f(X; \theta, \omega, \gamma^\dagger)[1 + \phi S(\theta, \omega, \gamma^\dagger)]; \phi \in (-\varepsilon, \varepsilon)\}$  is, for suitably small  $\varepsilon$ , a one-dimensional parametric submodel of  $\Gamma$  indexed by  $\phi$  with true value  $\phi = 0$  and score equal to  $S(\theta, \omega, \gamma^\dagger)$ . But this implies that  $E_{\theta, \omega, \gamma^\dagger}\{\tilde{U}(\theta, \omega, \gamma^\dagger) S(\theta, \omega, \gamma^\dagger)\} = 0$ , because  $\tilde{U}(\theta, \omega, \gamma^\dagger) \in T_{\gamma^\dagger}(\Gamma)^\perp$ . Further,  $E_{\theta, \omega, \gamma}\{\tilde{U}(\theta, \omega^\dagger, \gamma)\} = E_{\theta, \omega, \gamma}\{U(\theta, \gamma) - \Pi_{\kappa, \gamma}[U(\theta, \gamma)]\} = 0$  by  $U(\theta, \gamma)$ , an unbiased estimating function, and  $\Pi_{\kappa, \gamma}[U(\theta, \gamma)] \in T_\gamma(\Gamma)$ .

## 8. CAN LIKELIHOOD-BASED INFERENCE BE SAVED?

In an unpublished manuscript, Zeng and Murphy studied the model of Example 2a and proposed a sieve likelihood  $\mathcal{L}_n^*(\kappa, \gamma) = \mathcal{L}_{n1}^*(\kappa, \gamma)\mathcal{L}_{n2}^*(\gamma)$  with  $\kappa \in \mathcal{K}_n$  and  $\gamma \in \Gamma_n$ , where they artificially add  $\gamma$  to  $\mathcal{L}_{n1}^*(\kappa)$ . Zeng and Murphy provided a particular sieve  $\mathcal{K}_n$  that depends on  $n$  for which the sieve MLE  $\hat{\theta}(\gamma^*) \equiv \theta\{\hat{\kappa}(\gamma^*)\}$  with  $\hat{\kappa}(\gamma^*)$  maximizing  $\mathcal{L}_{n1}^*(\kappa, \gamma^*)$  over  $\kappa \in \mathcal{K}_n$  is locally efficient for  $\theta$  in model  $M(\mathcal{K})$ . Because the sieve MLE depends on the known value  $\gamma^*$  of  $\gamma$ , it is not an SFB estimator and thus violates the likelihood principle. Thus these authors have succeeded in saving "likelihood-based" inference, but only in the narrow computational sense that they have constructed a locally efficient estimator by maximization of a sieve likelihood function. Scharfstein, Rotnitzky, and Robins (1999, page 1141) proposed a fixed sieve  $\mathcal{K}_n = \mathcal{K}_{\text{sub}}$  for which the

MLE of  $\theta$  is also locally efficient in model  $M(\mathcal{K})$ . They used as their sieve the fixed model  $M(\mathcal{K}_{\text{sub}})$  for the observed data  $X$  induced by the model  $M_{\text{ful}}(\mathcal{K}_{\text{sub}})$  characterized by the linear logistic model  $\text{pr}(Y = 1|V; \kappa_1, \gamma^*) = \{1 + \exp[-d_1(V; \kappa_{1,1}) - \kappa_{1,2}/\pi(V; \gamma^*)]\}^{-1}$ , where  $d_1(V; \kappa_{1,1})$  is a known function of an unknown  $p_1 - 1$ -dimensional parameter  $\kappa_{1,1}$  and  $\kappa_{1,2}$  is an unknown scalar. The sieve MLE  $\hat{\theta}_{\text{sieve}}(\gamma^*) \equiv n^{-1} \sum_{i=1}^n \text{pr}(Y = 1|V_i; \hat{\kappa}_1(\gamma^*), \gamma^*)$  is locally efficient in model  $M(\mathcal{K})$  at the submodel  $M(\mathcal{K}_{\text{sub}})$ , because  $\hat{\theta}_{\text{sieve}}(\gamma^*)$  is algebraically identical to  $\hat{\theta}_{\text{loceff}}(\gamma^*)$  based on model  $M(\mathcal{K}_{\text{sub}})$ . Here  $\hat{\kappa}_1(\gamma^*)$  is the  $\kappa_1$  maximizing the linear logistic likelihood among units with  $R = 1$ .

However, we conjecture without proof that for many models, likelihood inference cannot be "saved" even in this narrow sense. This conjecture comes from our pondering the following examples. Consider the model  $M(\mathcal{K} \times \Gamma_{\text{sub}})$  in Example 2a of Section 6. Scharfstein et al. (1999) showed that  $\hat{\theta}_{\text{sieve}}(\hat{\gamma})$ , with  $\hat{\gamma}$  maximizing  $\mathcal{L}_{n2}^+(\gamma) \equiv \mathcal{L}_{n2}(\gamma)$  over  $\Gamma_{\text{sub}}$ , is identically equal to  $\hat{\theta}_{\text{loceff}}(\hat{\gamma})$  and thus is locally efficient in model  $M(\mathcal{K} \times \Gamma_{\text{sub}})$  at the submodel  $M(\mathcal{K}_{\text{sub}} \times \Gamma_{\text{sub}})$ . However, we do not regard  $\hat{\theta}_{\text{sieve}}(\hat{\gamma})$  as likelihood-based, because it is not obtained by maximizing a likelihood function. Rather, in our fixed sieve model, the sieve MLE is  $\theta\{\hat{\kappa}_{\text{sub}}\}$ , where  $\hat{\kappa}_{\text{sub}}$  is the value of  $\kappa$  obtained by maximizing  $\mathcal{L}_{n1}^+(\kappa, \gamma)\mathcal{L}_{n2}^+(\gamma)$  over  $(\kappa, \gamma)$  in  $\mathcal{K}_{\text{sub}} \times \Gamma_{\text{sub}}$ . Unfortunately,  $\theta\{\hat{\gamma}_{\text{sub}}\}$  fails in the sense that it is inconsistent if either  $\kappa^* \notin \mathcal{K}_{\text{sub}}$  or  $\gamma^* \notin \Gamma_{\text{sub}}$ . Next, consider the extension of model  $M(\mathcal{K})$  of Example 2a to a two-stage stratified random sampling design. Specifically, redefine  $\mathbf{V} = (V_0, V_1)$ ,  $\mathbf{R} = (R_1, R_2)$ , with  $V_j, j = 1, 2$ , highly multivariate and continuous;  $R_j$  Bernoulli,  $R_2 = R_1 R_2$ ; and  $c_{\mathbf{R}}(\mathbf{L}) = (V_0, R_1 V_1, R_2 Y)$ . Here CAR implies that  $\text{pr}(R_2 = 1|\mathbf{L}, R_1 = 1; \gamma^*) = \pi_2(\mathbf{V}; \gamma^*)$  and  $\text{pr}(R_1 = 1|\mathbf{L}; \gamma^*) = \pi_1(V_0; \gamma^*)$ . In contrast with the one-stage design discussed earlier, we have failed to find a sieve likelihood for which the MLE  $\hat{\theta}_{\text{sieve}}(\gamma^*)$  is a locally efficient estimator of  $\theta^*$  (see Robins 2000 for further discussion).

#### ADDITIONAL REFERENCES

- Dabrowska, D. M. (1988), "Kaplan-Meier Estimator on the Plane," *The Annals of Statistics*, 16, 1475-1489.
- Gill, R. D., van der Laan, M. J., and Robins, J. M. (1997), "Coarsening at Random: Characterizations, Conjectures and Counterexamples," in *Proceedings of the First Seattle Symposium on Survival Analysis*, eds. D. Y. Lin and T. R. Fleming, New York: Springer, pp. 255-294.
- Hastie, T. J., and Tibshirani, R. J. (1995), *Generalized Additive Models*, New York: Chapman and Hall.
- Heitjan, D. F., and Rubin, D. B. (1991), "Ignorability and Coarse Data," *The Annals of Statistics*, 19, 2244-2253.
- Jacobsen, N., and Keiding, N. (1995), "Coarsening at Random in General Sample Spaces and Random Censoring in Continuous Time," *The Annals of Statistics*, 23, 774-786.
- Kress, R. (1989), *Linear Integral Equations*, Berlin: Springer-Verlag.
- Lin, D. Y., Sun, W., and Ying, Z. (1999), "Estimation of the Gap Distribution," *Biometrika*, 86, 350-359.
- Prentice, R. L., and Cai, J. (1992), "Covariance and Survivor Function Estimation Using Censored Multivariate Failure Time Data," *Biometrika*, 79, 495-512.
- Robins, J. M. (1999), "Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference," in *Statistical Models in Epidemiology: The Environment and Clinical Trials*, eds. M. E. Halloran and D. Berry, NY: Springer-Verlag, pp. 95-134.
- , (2000), "Robust Estimation in Sequentially Ignorable Missing Data and Causal Inference Models," *Proceedings of the 1999 Joint Statistical Meetings* (to appear).
- Robins, J. M., and Rotnitzky, A. (1992), "Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers," in *AIDS Epidemiology—Methodological Issues*, eds. N. Jewell, K. Dietz, and V. Farewell, Boston: Birkhäuser, pp. 297-331.
- Robins, J. M., and Wang, N. (1998), Discussion of the articles by Forster and Smith and Clayton et al., *Journal of the Royal Statistical Society*, Ser. B, 60, 91-93.
- Robins, J. M., and Wasserman, L. (2000), "Conditioning, Likelihood, and Coherence: A Review of Some Foundational Concepts," *Journal of the American Statistical Association*, (to appear).
- Satten, G., and Datta, S. (2000), "Marginal Estimation for Multistage Models: Waiting Time Distributions and Competing Risks Analysis," University of Georgia, Department of Statistics Tech Report #STA 00-04.
- Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999), Rejoinder to Comments on "Adjusting for Nonignorable Dropout Using Semiparametric Nonresponse Models," *Journal of the American Statistical Association* 94, 1135-1146.
- van der Laan, M. (1996a), "Efficient Estimator of the Bivariate Survival Function and Repairing NPMLE," *The Annals of Statistics*, 24, 596-627.
- (1996b), "Efficient and Inefficient Estimation in Semiparametric Models," CWL tract 114, Centre of Computer Science and Mathematics, Amsterdam.
- van der Laan, M. J., Hubbard, A. E., and Robins, J. M. (1999), "Locally Efficient Estimation of a Multivariate Survival Function in Longitudinal Studies," *Journal of the American Statistician*, (submitted).
- van der Laan, M., Robins, J. M., and Gill, R. D. (2000), "Locally Efficient Estimation in Censored Data Models: Theory and Examples," technical report, University of California, Berkeley, Dept. of Biostatistics.