

Nonparametric Imputation of Missing Values for Estimating Equation Based Inference¹

DONG WANG and SONG XI CHEN

Department of Statistics, Ohio State University, Ames, IA

dongwang@osu.edu and songchen@osu.edu

¹MMA Y

We propose a nonparametric imputation procedure for data with missing values and establish an empirical likelihood inference for parameters defined by general estimating equations. The imputation is carried out multiple times via a nonparametric estimator of the conditional distribution of the missing variable given the always observable variable. The empirical likelihood is used to construct a profile likelihood for the parameter of interest. We demonstrate that the proposed nonparametric estimator achieves the same efficiency as the likelihood-based estimator.

Examples include the effort by *genenetwork.org* and other data depositories to combine genetics,

advantage of being more robust against model misspecification, although a correct model for the conditional distribution of the missing variable given the observed variable is often needed to attain the semiparametric efficiency bound.

Nonparametric methods have also been proposed for missing data. Titterton (1977) and Titterton and Mill (1983) considered kernel estimation of a multivariate density for data with incomplete observations. When the parameter of interest is the mean of a response variable which is subject to missingness, Cheng (1994) proposed using the kernel conditional mean estimator to impute the missing values. Hahn (1998) and Hirano, Imbens and Ridder (2003) studied the estimation of average treatment effects using nonparametrically estimated propensity scores. Since in treatment effect problems, the response of any unit can only be observed under one treatment, it is seen as a missing data problem. In survey statistics, Kim and Fuller (2004) proposed the fractional hot deck imputation method, in which multiple values are drawn from the same imputation cell as the missing observation, and a weight is assigned to each imputed value.

In this paper we consider estimation of parameters defined by a set of estimating equations in the presence of missing values. Estimating equation (Godambe, 1991; Boos, 1992) is a very general framework for statistical inference. When the parameter of interest is not directly related to the mean, the commonly used conditional mean based imputation via either a parametric (Yates, 1933) or nonparametric (Cheng, 1994) regression estimator may result in either biased estimation or reduced efficiency. This is especially the case for data with missing covariates.

We propose in this paper a nonparametric imputation procedure that generates multiple copies of missing values from a kernel estimator of the conditional distribution of the missing variables given the fully observable variables. This model-free (nonparametric) imputation is particularly suited for the wide range of parameters defined by estimating

equations. We then employ Owen (1988, 1990)'s empirical likelihood to formulate a non-parametric profile likelihood based on an extended sample which consists of the original data and the nonparametrically imputed values. Our use of empirical likelihood is largely encouraged by its attractive inferential features for estimating equations when there is no missing values by Qin and Lawless (1994), as well as its inference for a

Let $\delta_i = 1$ if x_i is observed and $\delta_i = 0$ if x_i is missing. Like Cheng (1994), Wang and Rao (2002) and others, we assume that δ and x are conditionally independent given z , namely the strongly ignorable missing at random proposed by Rosenbaum and Rubin (1983). As a result,

$$(\delta = 1 | z) = (\delta = 1 | x, z) =: p(z)$$

where $p(z)$ is called the propensity score and prescribes selection bias in the missingness if $p(z)$ is not a constant function.

Let $f(y | x_i)$ be the conditional distribution of y given $x = x_i$. Let

$$\hat{f}(y | x_i) = \frac{\sum_{l=1}^n \delta_l \left(\frac{x_l - x_i}{h} \right) \mathbb{I}(y \leq y_i)}{\sum_{j=1}^n \delta_j \left(\frac{x_j - x_i}{h} \right)} \quad (1)$$

be a kernel estimator of $f(y | x_i)$ based on the completely observed portion (no missing values) of the sample. Here $(\cdot)$ is a d_x -dimensional kernel function; h is a smoothing bandwidth satisfying $h^2 \rightarrow 0$ and $d_x h \rightarrow \infty$ as $n \rightarrow \infty$; and $\mathbb{I}(\cdot)$ is the d_y -dimensional indicator function. Here we concentrate on the case that both x and y are continuous random variables. Extension to discrete random variables can be readily made; see Section 5 for a case of binary random variable.

We propose to impute a missing x_i with a $\tilde{x}_i$ randomly generated from the estimated conditional distribution $\hat{f}(y | x_i)$. Effectively $\tilde{x}_i$ has a discrete distribution where the probability of selecting a x_i with $\delta_i = 1$ is

$$\frac{\{(\delta_i - \delta_i)\}}{\sum_{j=1}^n \delta_j \{(\delta_j - \delta_i)\}} \quad (2)$$

To control the variability of the estimating functions with imputed values, we make independent draws $\{\tilde{x}_i\}_{i=1}^n$ from $\hat{f}(y | x_i)$ and use

$$\tilde{f}(x_i) = \delta_i f(x_i) + (1 - \delta_i)^{-1} \sum_{i=1}^n (\delta_i \tilde{f}_i(x_i)) \quad (3)$$

as the estimation function for the i -th observation.

A popular method of imputation is to impute a missing y_i by the conditional mean of y given $x = x_i$ as proposed in Yates (1933) under a parametric regression model and in Cheng (1994) and Wang and Rao (2002) via the Nadaraya-Watson kernel estimator for the conditional mean. However, it may not work for other parameters, for instance, quantiles or correlation coefficients. Nor is it generally applicable to the case of missing covariates in a regression context. The proposed nonparametric imputation is generally applicable for any parameter defined by estimating equations. We are to show that when θ is a mean related parameter, the proposed imputation method leads to a parameter estimator that has the same efficiency as that obtained by conditional mean imputation.

"Curse of dimensionality" is an issue with kernel estimators. Indeed, the estimation accuracy of $\hat{f}(x_i)$ deteriorates as d_x increases. However, as demonstrated in Section 4, the curse of dimensionality does not pose any leading order effect on the estimation of θ as long as the bias of the kernel estimator is controlled by letting $\sqrt{n}h^2 \rightarrow 0$. There is one effect of the dimensionality though: when $d_x \geq 4$, controlling the bias requires a higher order kernel. Using a higher order kernel causes $\hat{f}(x_i)$ not being a bona fide conditional distribution and can cause a minor problem for the imputation. Wang and Chen (2006) propose a refinement of the imputation procedure to allow high order kernels and hence $d_x \geq 4$. To simplify our exposition, we confine ourselves in this paper to $d_x \leq 3$.

$$\text{EMP} \quad \text{ALL} \rightarrow \text{EL} \rightarrow \text{OOD}$$

The nonparametric imputation produces an extended sample $\{\tilde{y}_i\}_{i=1}^n$ where

$$\tilde{y}_i = \begin{cases} y_i & \text{if } \delta_i = 1; \\ (y_i, \{\tilde{y}_i\}_{i=1}^n) & \text{if } \delta_i = 0 \end{cases} \quad (4)$$

Let π_i represents the probability weight allocated to $\tilde{y}_i$. The empirical likelihood

is

$$L(\boldsymbol{\gamma}) = \sup \left\{ \prod_{i=1}^n \gamma_i : \gamma_i \geq 0, \sum_{i=1}^n \gamma_i = 1, \sum_{i=1}^n \gamma_i \tilde{\gamma}_i = 0 \right\}$$

where $\tilde{\gamma}$ is the adjustment to the original estimating function given in (3). This is the formulation of Qin and Lawless (1994) on adjusted estimating functions. By following the standard derivation of empirical likelihood (Qin and Lawless, 1994), the optimal γ_i is

$$\gamma_i = \frac{1}{1 + \lambda \tilde{\gamma}_i}$$

where λ is the Lagrange multiplier that satisfies

$$\sum_{i=1}^n \frac{\tilde{\gamma}_i}{1 + \lambda \tilde{\gamma}_i} = 0 \quad (5)$$

Let $\ell(\boldsymbol{\gamma}) = -\log\{L(\boldsymbol{\gamma})\}^{-n}$ be the log empirical likelihood ratio and $\hat{\boldsymbol{\gamma}}$ be the maximum empirical likelihood estimator that maximizes $L(\boldsymbol{\gamma})$.

The efficiency of $\hat{\boldsymbol{\gamma}}$ is studied in the next section which also includes a proposal for constructing confidence regions for $\boldsymbol{\gamma}$ based on the empirical likelihood ratio. The latter is largely motivated by the attractive features (natural shape and orientation as well as range respecting) of empirical likelihood confidence regions in the absence of missing values; see Hall and La Scala (1990) and Chen and Cui (2006).

$$\mathbf{M}_A = \mathbf{E}[\mathbf{L}']$$

Let $\boldsymbol{\gamma}_0$ denote the true parameter value. Write $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0)$. We define

$$\begin{aligned} \tilde{\boldsymbol{\gamma}} &= [\boldsymbol{\gamma} - \boldsymbol{\gamma}_0] + \{\boldsymbol{\gamma}_0\} \\ &= [\boldsymbol{\gamma} - \boldsymbol{\gamma}_0] + \{\boldsymbol{\gamma}_0\} \end{aligned}$$

and $\mathbf{L} = \left\{ \left(\frac{\partial \ell}{\partial \boldsymbol{\gamma}} \right)^{-1} \left(\frac{\partial \ell}{\partial \boldsymbol{\gamma}} \right) \right\}^{-1}$ at $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$.

more. Under the conditions given in the Appendixes $s \rightarrow \infty$ and $n \rightarrow \infty$,

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N(0, V)$$

in distribution with $V = \left(\frac{\partial^2 g}{\partial \theta^2} \right)^{-1} \sim^{-1} \sim^{-1} \left(\frac{\partial^2 g}{\partial \theta^2} \right)$

The estimator $\hat{\theta}$ is consistent for θ_0 and the potential selection bias in the missingness as measured by the propensity score $p(\cdot)$ has been filtered out. If there is no missing values, $\tilde{p} = p = 1$, which means that

$$V = \left\{ \left(\frac{\partial^2 g}{\partial \theta^2} \right) \left(\frac{\partial^2 g}{\partial \theta^2} \right)^{-1} \left(\frac{\partial^2 g}{\partial \theta^2} \right) \right\}^{-1}$$

This is the asymptotic variance of the maximum empirical likelihood estimator based on full observations given in Qin and Lawless (1994). Comparing the forms of V with and without missing values shows that the efficiency of the maximum empirical likelihood estimator based on the proposed imputation will be close to that based on full observations if either the proportion of missing data is low, that is when $p(\cdot)$ is close to 1, or if $\left\{ \frac{\partial^2 g}{\partial \theta^2} \right\}^{-1}$

To appreciate the proposal of letting the number of imputation $m \rightarrow \infty$, we note that when n is fixed, the $\mathbf{W}$ and $\tilde{\mathbf{W}}$ matrices used to define $\mathbf{V}$ have forms:

$$\mathbf{W} = \left\{ \mathbf{I}^{-1}(\mathbf{I} - \mathbf{I}^{-1}(\mathbf{I} - \mathbf{I})) \right\} \quad (\mathbf{I} - \mathbf{I}) + (\mathbf{I} - \mathbf{I}) (\mathbf{I} - \mathbf{I}) \quad \text{and}$$

$$\tilde{\mathbf{W}} = \left\{ (\mathbf{I} - \mathbf{I}) + \mathbf{I}^{-1}(\mathbf{I} - \mathbf{I}) \right\} \quad (\mathbf{I} - \mathbf{I}) + (\mathbf{I} - \mathbf{I}) (\mathbf{I} - \mathbf{I})$$

Hence, a larger m will reduce the terms in $\mathbf{W}$ and $\tilde{\mathbf{W}}$ which are due to a single nonparametric imputation. Our numerical experience suggests that $m = 20$ is sufficient for most situations.

Let us now turn our attention to the log empirical likelihood ratio

$$\mathcal{R}(\mathbf{0}) = 2'(\mathbf{0}) - 2'(\hat{\boldsymbol{\theta}})$$

Let $\mathbf{I}_r$ be the r -dimensional identity matrix. The next theorem shows that the log empirical likelihood ratio converges to a linear combination of independent

parameter is replaced by a plugged-in estimator as revealed by Hjort, McKeague and Van Keilegom (2004).

When $n \rightarrow \infty$, $\mathcal{R}(n) \rightarrow \{ \tau_1(n) \leq \tau_2(n) \} \stackrel{d}{\rightarrow} \mathcal{R}_1$ in distribution where

$$\tau_1(n) = \inf \{ t \in \mathbb{R} : \hat{F}_n(t) \leq \alpha \} + \inf \{ t \in \mathbb{R} : \hat{F}_n(t) \leq \alpha \}$$

and $\tau_2(n) = \inf \{ t \in \mathbb{R} : \hat{F}_n(t) \leq \alpha \} + \inf \{ t \in \mathbb{R} : \hat{F}_n(t) \leq \alpha \}$. This is the limiting distribution given in Wang and Rao (2002).

As confidence regions can be readily transformed to test statistics for testing a hypothesis regarding θ , we shall focus on confidence regions. There are potentially several methods for the construction of a confidence region for θ . One is based on an estimation of the covariance matrix and the asymptotic normality

method where the estimation of the conditional distribution is based on χ_{nc}^* .

3. Let $\ell^*(\hat{\gamma})$ be the empirical likelihood ratio based on the re-imputed data set χ_n^* , $\hat{\gamma}^*$ be the corresponding maximum empirical likelihood estimator, and $\mathcal{R}^*(\hat{\gamma}) = \frac{\ell^*(\hat{\gamma})}{\ell^*(\hat{\gamma}^*)}$.

it is based on the fact that

$$E\left\{\left(\frac{\delta_i - 1}{\delta_i}\right) \delta_i\right\} = 0$$

Based on the usual formulation of the generalized method of moments (GMM, Hansen, 1982), the weighted-GMM estimator for θ_0 considered in our simulation is

$$\hat{\theta}_T = \arg \min_{\theta} \left\{ \frac{1}{n_c} \sum_{i=1}^n \left(\frac{\delta_i - 1}{\delta_i} \right) \frac{1}{\sqrt{h(\delta_i)}} \right\}' A_T \left\{ \frac{1}{n_c} \sum_{i=1}^n \left(\frac{\delta_i - 1}{\delta_i} \right) \frac{1}{\sqrt{h(\delta_i)}} \right\}$$

where n_c is the number of complete observations, A_T is a nonnegative definite weighting matrix, and $\sqrt{h(\delta_i)}$ is a consistent estimator for $\sqrt{h(\delta_i)}$. The resulting GMM estimator $\hat{\theta}_T$ is efficient if A_T is chosen as $A_T = [E\{\left(\frac{\delta_i - 1}{\delta_i}\right)^2\}]^{-1}$.

$$d \quad () \equiv 0.65 \text{ for all } ;$$

$$c \quad () = 0.5 (\quad 0) + (\leq 0).$$

The missing mechanism (b) is missing completely at random; whereas the other two are missing at random and prescribe selection bias in the missingness.

Let μ_x and μ_y be the means, and σ_x^2 and σ_y^2 be the variances of x and y , respectively. In the construction of the empirical likelihood for (Owen, 1990), $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$ are treated as nuisance parameters.

Table 1 contains the bias and standard error of the four estimators considered based on 1000 simulations with the sample size $n = 100$ and 200 respectively. It also contains the empirical likelihood confidence intervals using the full observations, complete observations only, and the proposed nonparametric imputation method at a nominal level of 95%. They are all based on the proposed bootstrap calibration method with $B = 2000$. When using the nonparametric imputation method, $M = 20$ imputations were made for each missing x_i . The confidence intervals based on the weighted-GMM are calibrated using the asymptotic normal approximation with the covariance matrix estimated by the kernel method.

The results in Table 1 can be summarized as follows. The nonparametric imputation method significantly reduces the bias compared to inference based only on complete observations when the data are missing at random but not missing completely at random. The estimator based on the completely observed data suffers quite severe bias under missing mechanisms (a) and (c). The proposed imputation reduces the variance relative to the estimator

former is due to the selection bias and the latter is due to the normal approximation. The proposed confidence intervals have satisfactory coverages which are quite close to the nominal level 0.95.

5.2 *Generalized linear models with missing covariates*

In the second simulation study we consider missing covariates in a generalized linear model (GLM). We also take the opportunity to discuss an extension of the proposed imputation procedure to binary random variables. Commonly used methods in dealing with missing data

smaller than that based on the complete observations only; the proposed method also leads to a reduction in the mean square error by as much as 20% relative to the weighted-GMM. All three methods give similar mean squared errors for the parameter σ^2 . The confidence intervals based on only complete observations or the weighted-GMM tend to show notable undercoverage, while the proposed confidence intervals have satisfactory coverage levels for all parameters.

EMP AL DY

Microarray technology provides an powerful tool in molecular biology by measuring the expression level of thousands of genes simultaneously. One problem of interest is to test whether the expression level of genes is related to a traditional trait like body weight, food consumption, or bone density. This is usually the first step in uncovering roles that a gene plays in important pathways. The BXD recombinant inbred strains of mouse were derived from crosses between C57BL/6J (B6 or B) and DBA/2J (D2 or D) strains (Williams, Gu, Qi and Lu, 2001). Around one hundred BXD strains have been established by researchers at University of Tennessee and the Jackson Laboratory. A variety of phenotype data are accumulated for BXD mouse over the years (Pierce et al., 2004).

The trait that we consider is the fresh eye weight measured on 83 BXD strains by Zhai, Lu, and Williams (ID 10799, BXD phenotype data base). The Hamilton Eye Institute Mouse Eye M430v2 RMA Data Set contains measures of expression in the eye on 39,000 transcripts. It is of interest to test whether the fresh eye weight is related to the expression level of certain genes. However, the microarray data are only available for 45 out of the 83 BXD mouse strains for which fresh eye weights are all available. The most common practice is to use only complete observations and ignore missing values in the statistical inference. As demonstrated in our simulation, this approach can lead to inconsistent parameter estimators if there is a selection bias in the missingness. Even in the absence of

selection bias, the estimators are not efficient as only those complete observations are used.

We conduct four separate simple linear regression analysis of the eye weight on the expression level of four genes respectively. The genes are *H3071E5*, *Slc26a8*, *Tex9*, and *Rps16*, which are identified by the corresponding probe names in the microarray dataset. Here we have missing covariates in our analysis. The missing gene expression levels are imputed from a kernel estimator of the conditional distribution of the gene expression level given the fresh eye weight. The smoothing bandwidths were selected based on the cross-validation method, which is 1.5 for the first three genes in Table 3 and 1.8 for gene *Rps16*.

Table 3 reports empirical likelihood estimates of the intercept and slope parameters and their 95% confidence intervals based on the proposed nonparametric imputation and empirical likelihood. It also contains results from results

The following conditions are needed in the proofs of the theorems.

C1: The functions $\phi(\cdot)$, $\psi(\cdot)$ and $g(\cdot)$ all have bounded second partial derivatives, and $\inf_x \phi(\cdot) \geq \phi_0$ for some $\phi_0 > 0$.

C2: The estimating function $\psi(\cdot)$ has bounded second partial derivative with regard to θ , and $\|\psi(\theta_0)\|^4 < \infty$. In addition, $\phi^2(\cdot)$ is continuous in θ in a neighborhood of the true value θ_0 ; $\phi(\cdot)$, $\phi(\cdot)^3$, and $\phi^2(\cdot)$ are all bounded by some integrable functions in the neighborhood.

C3: The matrices Σ and $\tilde{\Sigma}$ are, respectively, positive definite with the smallest eigenvalue bounded away from zero, and $[\phi(\cdot)]$ has full column rank p .

C4: The kernel function K is a non-negative, symmetric and bounded probability density function with finite second moments.

C5: The smoothing bandwidth satisfies $n h_x \rightarrow \infty$, $\sqrt{n} h_x^2 \rightarrow 0$ as $n \rightarrow \infty$, and $p_x \leq 3$.

Assuming $\phi(\cdot)$ bounded away from zero in C1 implies that data cannot be missing with probability 1 anywhere in the domain of the X variable. Conditions C2 and C3 are required for empirical likelihood based inference with estimating equations. Conditions C4 and C5 are standard in kernel estimation, and that $\sqrt{n} h_x^2 \rightarrow 0$ is to

distribution, and then use the Cramer-Wold device to prove Lemma 1. Define

$$g_u(\cdot) = (u(\cdot, \cdot, 0) | \cdot = \cdot) \text{ and } \hat{g}_u(\cdot) = \frac{\sum_{i=1}^n \delta_i \left(\frac{x - X_i}{h} \right) u(\cdot, X_i, 0)}{\sum_{i=1}^n \delta_i \left(\frac{x - X_i}{h} \right)}$$

Now we have

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \left\{ \delta_i u(\cdot, X_i, 0) + (1 - \delta_i) \frac{1}{n} \sum_{j=1}^n u(\cdot, X_j, 0) \right\} \\ &= \frac{1}{n} \sum_{i=1}^n \delta_i \{ u(\cdot, X_i, 0) - g_u(\cdot) \} \\ & \quad + \frac{1}{n} \sum_{i=1}^n (1 - \delta_i) \left\{ \frac{1}{n} \sum_{j=1}^n u(\cdot, X_j, 0) - \hat{g}_u(\cdot) \right\} \\ & \quad + \frac{1}{n} \sum_{i=1}^n (1 - \delta_i) \{ \hat{g}_u(\cdot) - g_u(\cdot) \} + \frac{1}{n} \sum_{i=1}^n g_u(\cdot) \\ &:= T_n + A_n + T_n + g_n \end{aligned}$$

Note that T_n and g_n are sums of independent and identically distributed random variables.

Define $r^{\gamma}(\cdot) = (\cdot, \cdot)$ and $r^{\gamma}(\cdot) = \frac{1}{n} \sum_{j=1}^n \delta_j h(\cdot - X_j)$ as its kernel estimator, where

$h(\cdot) = -d_x(\cdot)$. Then,

$$\begin{aligned} T_n &= \frac{1}{n} \sum_{i=1}^n (1 - \delta_i) \frac{\frac{1}{n} \sum_{j=1}^n \delta_j h(\cdot - X_j) \{ u(\cdot, X_j, 0) - g_u(\cdot) \}}{r^{\gamma}(\cdot)} \\ & \quad + \frac{1}{n} \sum_{i=1}^n (1 - \delta_i) \{ \hat{g}_u(\cdot) - g_u(\cdot) \} \frac{r^{\gamma}(\cdot) - r^{\gamma}(\cdot)}{r^{\gamma}(\cdot)} \\ & \quad + \frac{1}{n} \sum_{i=1}^n (1 - \delta_i) \left\{ \frac{\frac{1}{n} \sum_{j=1}^n \delta_j h(\cdot - X_j) (g_u(\cdot) - g_u(\cdot))}{r^{\gamma}(\cdot)} \right\} \\ &:= T_{n1} + T_{n2} + T_{n3} \end{aligned}$$

Define

$$T_{n1} = \sum_{j=1}^n \{ T_{n1} | (\cdot, X_j, \delta_j) \} = \sum_{j=1}^n \delta_j \{ T_{n1} | (\cdot, X_j, \delta_j = 1) \} \quad (\text{A2})$$

to be a projection of T_{n1} . Then write $T_{n1} = T_{n1} + (T_{n1} - T_{n1})$. As

$$T_{n1} = \frac{1}{n} \sum_{i=1}^n (1 - \delta_i) \frac{\frac{1}{n} \sum_{j=1}^n \delta_j h(\cdot - X_j) \{ u(\cdot, X_j, 0) - g_u(\cdot) \}}{r^{\gamma}(\cdot)}$$

$$= \frac{1}{n} \sum_{j=1}^n \delta_j \{ u(\mathbf{x}_i, \mathbf{x}_j) - g_u(\mathbf{x}_j) \} \left\{ \frac{1}{n} \sum_{i=1}^n (1 - \delta_i) \frac{h(\mathbf{x}_i - \mathbf{x}_j)}{r^d(\mathbf{x}_i)} \right\}$$

$$T_{n1} = \frac{1}{n} \sum_{j=1}^n \delta_j \{ u(\mathbf{x}_i, \mathbf{x}_j) - T_{\mathbf{x}_j} \}$$

suggests that $T_{n1} = T_{n1} + p(-1=2)$. By standard argument, we can show that $T_{n2} = p(-\frac{1}{2})$. Derivations similar to those for T_{n1} can be used to establish $T_{n3} = p(-1=2)$. Thus, we have

$$\sqrt{n}T_n \rightarrow_d \left[0 \quad \left\{ (1 - p)^2 \frac{2}{g_u} \left(\frac{1}{2} \right) \right\} \right] \quad (A4)$$

in distribution, where $\frac{2}{g_u} \left(\frac{1}{2} \right) = \frac{2}{\int \{ u \left(\frac{1}{2} \right) \} dF}$.

Also note $\sqrt{n}T_n \rightarrow_d \left[0 \quad \left\{ \left(\frac{1}{2} \right) \frac{2}{g_u} \left(\frac{1}{2} \right) \right\} \right]$ and $\sqrt{n}T_n \rightarrow_d \left[0 \quad \left\{ g_u \left(\frac{1}{2} \right) \right\} \right]$ both in distribution. Further, it is straight forward to show that $(T_n T_n) = \left\{ (1 - p) \frac{2}{g_u} \left(\frac{1}{2} \right) \right\} + (1)$, $(T_n T_n) = 0$ and $(T_n T_n) = (1)$. It readily follows that

$$\sqrt{n}(T_n + T_n + T_n) \rightarrow_d \left[0 \quad \left\{ \frac{2}{g_u} \left(\frac{1}{2} \right) \right\} + \left\{ g_u \left(\frac{1}{2} \right) \right\} \right] \quad (A5)$$

in distribution.

Now we consider the asymptotic distribution of

$$A_n = \frac{1}{n} \sum_{i=1}^n (1 - \delta_i) \left\{ -1 \sum_{j=1}^n u \left(\frac{1}{2} \right) - \frac{1}{n} \sum_{j=1}^n u \left(\frac{1}{2} \right) \right\}$$

Given all the original observations, $(1 - \delta_i) \left\{ -1 \sum_{j=1}^n u \left(\frac{1}{2} \right) - \frac{1}{n} \sum_{j=1}^n u \left(\frac{1}{2} \right) \right\}$, $i = 1, 2, \dots, n$ are independent with conditional mean zero and conditional variance $(1 - \delta_i) \left\{ \frac{2}{g_u} \left(\frac{1}{2} \right) - \frac{1}{n} \sum_{j=1}^n \delta_j \frac{2}{g_u} \left(\frac{1}{2} \right) \right\}$. Here $\frac{2}{g_u} \left(\frac{1}{2} \right) = \frac{2}{\int \{ u \left(\frac{1}{2} \right) \} dF}$, $\frac{1}{n} \sum_{j=1}^n \delta_j \frac{2}{g_u} \left(\frac{1}{2} \right)$ is a kernel estimator of $\frac{2}{g_u} \left(\frac{1}{2} \right) = \frac{2}{\int \{ u \left(\frac{1}{2} \right) \} dF}$. By verifying Lyapounov's condition, we can show that conditioning on the original observations,

$$\sqrt{n}A_n \rightarrow_d \left[0 \quad \left\{ \frac{2}{g_u} \left(\frac{1}{2} \right) - \frac{1}{n} \sum_{j=1}^n \delta_j \frac{2}{g_u} \left(\frac{1}{2} \right) \right\} \right] \quad (A6)$$

in distribution. The conditional variance =

Also note $\sqrt{n}A_n \rightarrow_d \left[0 \quad \left\{ \frac{2}{g_u} \left(\frac{1}{2} \right) - \frac{1}{n} \sum_{j=1}^n \delta_j \frac{2}{g_u} \left(\frac{1}{2} \right) \right\} \right]$

By Lemma 1 of Schenker and Welsh (1988), as $\rightarrow \infty$ and $\rightarrow \infty$, $\sqrt{}$

$$:= T_{n1a} + T_{n1b} + T_{n1c} + T_{n1d}$$

III

Proof of Theorem 1: Using argument similar to that of Qin and Lawless (1994), it can be shown that as $n \rightarrow \infty$ and $m \rightarrow \infty$, with probability tending to 1, $L(\cdot)$ attains its maximum value at some point $\hat{\theta}$ within the open ball $\|\theta - \theta_0\| \leq \epsilon$, and $\hat{\theta}$ and $\hat{\theta} = \hat{\theta}(\hat{\theta})$ satisfy

$$l_n(\hat{\theta}) = 0 \quad l_m(\hat{\theta}) = 0$$

Taking the derivatives with regard to θ and η ,

$$\begin{aligned} \frac{\partial l_n(\theta_0)}{\partial \theta} &= \frac{1}{n} \sum_i \frac{\partial \log f(\tilde{y}_i | \theta)}{\partial \theta} & \frac{\partial l_n(\theta_0)}{\partial \eta} &= -\frac{1}{n} \sum_i \frac{\partial \log f(\tilde{y}_i | \theta)}{\partial \eta} \\ \frac{\partial l_m(\theta_0)}{\partial \theta} &= 0 & \frac{\partial l_m(\theta_0)}{\partial \eta} &= \frac{1}{m} \sum_i \left\{ \frac{\partial \log f(\tilde{y}_i | \theta)}{\partial \eta} \right\} \end{aligned}$$

Expanding $l_n(\hat{\theta})$, $l_m(\hat{\theta})$ at (θ_0, η_0) , we have

$$\begin{aligned} 0 &= l_n(\hat{\theta}) \\ &= l_n(\theta_0, \eta_0) + \frac{\partial l_n(\theta_0, \eta_0)}{\partial \theta} (\hat{\theta} - \theta_0) + \frac{\partial l_n(\theta_0, \eta_0)}{\partial \eta} (\hat{\eta} - \eta_0) + o_p(n^{-1/2}) \\ 0 &= l_m(\hat{\theta}) \\ &= l_m(\theta_0, \eta_0) + \frac{\partial l_m(\theta_0, \eta_0)}{\partial \theta} (\hat{\theta} - \theta_0) + \frac{\partial l_m(\theta_0, \eta_0)}{\partial \eta} (\hat{\eta} - \eta_0) + o_p(m^{-1/2}) \end{aligned}$$

where $n = n_1 + n_2$. Then we can write

$$\begin{pmatrix} \hat{\theta} \\ \hat{\eta} - \eta_0 \end{pmatrix} = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}^{-1} \begin{pmatrix} -\frac{\partial l_n(\theta_0, \eta_0)}{\partial \theta} + o_p(n^{-1/2}) \\ -\frac{\partial l_m(\theta_0, \eta_0)}{\partial \eta} + o_p(m^{-1/2}) \end{pmatrix}$$

where

$$n = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}$$

Proof of Theorem 2: **Notice that**

$$\mathcal{R}(\mathbf{0}) = 2 \sum_i$$

$$\rightarrow \mathbb{P}^* \left[0 \leq \left\{ \frac{2}{g_u(\cdot)} \left(\frac{x - X_i}{h} \right) \right\} + \left\{ g_u(\cdot) \right\} \right]$$

in distribution, where $\mathbb{P}^*(\cdot)$ and $\mathbb{P}(\cdot)$ represent the conditional expectation and variance given the original data respectively. Define

$$\hat{g}_u(\cdot) = \frac{\sum_{i=1}^n \delta_i \left(\frac{x - X_i}{h} \right) u(\cdot, i)}{\sum_{i=1}^n \delta_i \left(\frac{x - X_i}{h} \right)} \text{ and } \hat{g}_u^*(\cdot) = \frac{\sum_{i=1}^n \delta_i^* \left(\frac{x - X_i^*}{h} \right) u(\cdot, i^*)}{\sum_{i=1}^n \delta_i^* \left(\frac{x - X_i^*}{h} \right)}$$

Then

$$\begin{aligned} & \hat{g}_n^*(\cdot) + T_n^*(\cdot) + \hat{g}_n(\cdot) - \hat{g}_n(\cdot) - T_n(\cdot) - \hat{g}_n(\cdot) \\ &= \frac{1}{n} \sum_{i=1}^n \delta_i^* \{ u(\cdot, i^*) - \hat{g}_u(\cdot, i^*) \} - \frac{1}{n} \sum_{j=1}^n \delta_j \{ u(\cdot, j) - \hat{g}_u(\cdot, j) \} \\ &+ \frac{1}{n} \sum_{i=1}^n [(1 - \delta_i^*) \{ \hat{g}_u^*(\cdot, i^*) - \hat{g}_u(\cdot, i^*) \}] \\ &+ \frac{1}{n} \sum_{i=1}^n (1 - \delta_i^*) \{ \hat{g}_u(\cdot, i^*) - \hat{g}_u(\cdot, i^*) \} - \frac{1}{n} \sum_{j=1}^n (1 - \delta_j) \{ \hat{g}_u(\cdot, j) - \hat{g}_u(\cdot, j) \} \\ &+ \frac{1}{n} \sum_{i=1}^n \hat{g}_u(\cdot, i^*) - \frac{1}{n} \sum_{j=1}^n \hat{g}_u(\cdot, j) \\ &:= \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3 + \mathcal{I}_4 \end{aligned}$$

For both $\mathcal{I}_1$ and $\mathcal{I}_4$, we can apply the central limit theorem for bootstrap samples (e.g. Shao and Tu, 1985) to derive

$$\sqrt{n} \mathcal{I}_1 \rightarrow \mathbb{P}^* \left[0 \leq \left\{ \left(\frac{x - X_i}{h} \right) \frac{2}{g_u(\cdot)} \right\} \right] \text{ and } \sqrt{n} \mathcal{I}_4 \rightarrow \mathbb{P}^* \left[0 \leq \left\{ g_u(\cdot) \right\} \right] \quad (\text{A9})$$

in distribution. Also it can be shown $\mathcal{I}_2 = o_p(n^{-1/2})$. Use similar argument to (A3) to show

$$\begin{aligned} \mathcal{I}_3 &= \frac{1}{n} \sum_{i=1}^n \delta_i^* \{ u(\cdot, i^*) - \hat{g}_u(\cdot, i^*) \} \frac{1 - \left(\frac{x - X_i^*}{h} \right)}{\left(\frac{x - X_i^*}{h} \right)} \\ &- \frac{1}{n} \sum_{j=1}^n \delta_j \{ u(\cdot, j) - \hat{g}_u(\cdot, j) \} \frac{1 - \left(\frac{x - X_j}{h} \right)}{\left(\frac{x - X_j}{h} \right)} + o_p(n^{-1/2}) \end{aligned}$$

Then follow the proof for Lemma 1 and apply the bootstrap central limit theorem to conclude (A8).

For $\hat{\Delta}_n^*(\cdot)$, given the observations in the bootstrap sample that are not imputed, we have

✓

- Cheng, P. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association* , 81{87.
- Dempster, A. P., A. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Ser B* , 1{38.
- Devroye, L. P. and T. J. Wagner (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *the Annals of Statistics* , 231{239.
- Godambe, V. P. (1991). *Estimating Functions*. Oxford, U.K.: Oxford University Press.
- Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* , 315{331.
- Hall, P. and B. La Scala (1990). Methodology and algorithms of empirical likelihood. *International Statistical Review* , 109{127.
- Hansen, L. (1982). Large sample properties of generalized method of moment estimators. *Econometrica* , 1029{1084.
- Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* , 1161{1189.
- Hjort, N., I. McKeague, and I. Van Keilegom (2004). Extending the scope of empirical likelihood. *the Annals of Statistics*. Under revision.
- Ibrahim, J. G., M. Chen, S. R. Lipsitz, and A. H. Herring (2005). Missing-data methods for generalized linear models: a comparative review. *Journal of the American Statistical Association* , 332{346.

Kim, J. K. and W. Fuller (2004). Fractional hot deck imputation. *Bio metrika* , 559{578.

Kolaczyk, E. D. (1994). Empirical likelihood for generalized linear models. *Statistical Science* , 199{218.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* , 2nd edition. Hoboken, NJ, USA: Wiley.

McCullagh, P. and J. A. Nelder (1983). *Generalized Linear Models*. New York: Chapman and Hall.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Bio metrika* , 237{249.

Owen, A. (1990). *Empirical Likelihood*. New York: Chapman and Hall.

- Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* , 106{121.
- Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* , 41{55.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- Schenker, N. and A. H. Welsh (1988). Asymptotic results for multiple imputation. *The Annals of Statistics* , 1550{1566.
- Shao, J. and R. R. Sitter (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association* , 1278{1288.
- Shao, J. and D. Tu (1985). *The Confidence and Bootstrap*. New York: Springer Verlag.
- Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics* , 595{620.
- Titterton, D. M. (1977). Analysis of incomplete multivariate binary data by kernel method. *Biometrika* , 455{460.
- Titterton, D. M. and G. M. Mill (1983). Kernel-based density estimates from incomplete data. *Journal of the Royal Statistical Society. Series B* , 258{266.
- Wang, D. and S. X. Chen (2006). Nonparametric imputation of missing values for estimating equation based inference. Technical report, Iowa State University.
- Wang, J., R. W. Williams, and K. F. Manly (2003). WebQTL: Web-based complex trait analysis. *Neuroinformatics* , 299{308.

- Wang, Q. and J. N. K. Rao (2002). Empirical likelihood-based inference under imputation for missing response data. *The Annals of Statistics* , 896{924.**
- Williams, R. W., J. Gu, S. Qi, and L. Lu (2001). The genetic structure of recombinant inbred mice: high-resolution consensus maps for complex trait analysis. *Genome Biology* , research0046.1{0046.18.**
- Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Experimental Agriculture* , 129{142.**

= 100					
Methods	Bias	Std. Err.	MSE	Coverage	Length of CI
Full Observations	0.0018	0.0899	0.0081	0.937	0.3543
Missing Mechanism					
Complete Obs.	0.0597	0.1262	0.0195	0.863	0.4896
Weighted-GMM	0.0057	0.1105	0.0123	0.761	0.3568
N. Imputation	-0.0050	0.1013	0.0103	0.948	0.4888
Missing Mechanism $\checkmark$					
Complete Obs.	-0.0046	0.1160	0.0160	0.921	0.4562
Weighted-GMM	-0.0097	0.1068	0.0115	0.740	0.3057
N. Imputation	-0.0077	0.0991	0.0099	0.936	0.4239
Missing Mechanism c					
Complete Obs.	-0.1123	0.1480	0.0345	0.831	0.3594
Weighted-GMM	-0.0226	0.1141	0.0135	0.780	0.3569
N. Imputation	-0.0217	0.1050	0.0115	0.944	0.4264
= 200					
Methods	Bias	Std. Err.	MSE	Coverage	Length of CI
Full Observations	0.0047	0.0605	0.0037	0.949	0.2514
Missing Mechanism					
Complete Obs.	0.0727	0.0776	0.0113	0.849	0.3269
Weighted-GMM	0.0072	0.0755	0.0058	0.800	0.2479
N. Imputation	0.0076	0.0695	0.0049	0.953	0.3239
Missing Mechanism $\checkmark$					
Complete Obs.	-0.0007	0.0753	0.0057	0.957	0.3146
Weighted-GMM	-0.0067	0.0688	0.0048	0.793	0.2338
N. Imputation	-0.0004	0.0648	0.0041	0.957	0.2841
Missing Mechanism c					
Complete Obs.	-0.0905	0.1000	0.0182	0.782	0.3955
Weighted-GMM	-0.0035	0.0747	0.0060	0.773	0.2751
N. Imputation	-0.0055	0.0677	0.0046	0.946	0.2862

Table 1: Inference for the correlation coefficient with missing values. The four methods considered are empirical likelihood using full observations, empirical likelihood using

= 150					
Methods	Bias	Std. Err.	MSE	Coverage	Length of CI
$\beta_0 = -1$					
Full Observations	-0.0035	1.244	1.549	0.967	5.380
Complete Obs.	-1.622	1.489	4.847	0.901	6.429
Weighted-GMM	-0.3113	1.402	2.061	0.932	5.107
N. Imputation	0.0645	1.279	1.640	0.953	5.368
$\beta_1 = 1$					
Full Observations	0.0270	0.4270	0.1831	0.965	1.835
Complete Obs.	0.4070	0.4995	0.4152	0.908	2.308
Weighted-GMM	0.1200	0.4795	0.2443	0.935	1.722
N. Imputation	-0.0030	0.4346	0.1889	0.951	1.828
$\beta_2 = -1.5$					
Full Observations	-0.0766	0.5009	0.2568	0.976	2.172
Complete Obs.	-0.0664	0.5539	0.3112	0.975	2.506
Weighted-GMM	-0.0663	0.5589	0.3168	0.837	

Gene	Probe	Complete Observations Only (parametric)		Nonparametric Imputation (with empirical likelihood)	
Intercept					
<i>H3071E5</i>	1444597_at	-21.99	(-40.97, -2.998)	-15.69	(-37.02, 5.209)
<i>Slc26a8</i>	1441747_at	73.59	(49.45, 97.73)	67.28	(38.34, 95.87)
<i>Tex9</i>	1453360_a_at	-23.81	(-46.12, -1.507)	-14.66	(-38.57, 8.776)
<i>Rps16</i>	1455835_at	-13.52	(-31.08, 4.041)	-8.090	(-26.76, 10.18)
Slope					
<i>H3071E5</i>	1444597_at	10.16	(5.720, 14.59)	8.736	(2.688, 14.21)
<i>Slc26a8</i>	1441747_at	-6.352	(-9.294, -3.411)	-5.561	(-9.431, -1.471)
<i>Tex9</i>	1453360_a_at	5.101	(2.588, 7.613)	4.094	(0.8753, 6.979)
<i>Rps16</i>	1455835_at	6.766	(3.371, 10.16)	5.754	(1.948, 9.236)
Correlation Coe cient					
<i>H3071E5</i>	1444597_at	0.5757	(0.3395, 0.7436)	0.4426	(0.1321, 0.6977)
<i>Slc26a8</i>	1441747_at	-0.5533	(-0.7285, -0.3102)	-0.4319	(-0.6809, -0.0761)
<i>Tex9</i>	1453360_a_at	0.5296	(0.2996, 0.7124)	0.4024	(0.1013, 0.6846)
<i>Rps16</i>	1455835_at	0.5256	(0.2744, 0.7097)	0.4151	(0.0755, 0.6613)

Table 3: Parameter estimates and confidence intervals (shown in parentheses) based on a simple linear regression model using the parametric method with complete observations only and the empirical likelihood method using the proposed nonparametric imputation. For the parametric inference, the confidence intervals for the intercept and slope are obtained using quantiles of the t-distribution, and the confidence intervals for the correlation coefficient are obtained by Fisher's transformation. The four different genes are identified by the probe names.