U-statistic with side information

Ao Yuan\textsuperscript{a,}\textsuperscript{*}, Wenqing He\textsuperscript{b}, Binhuan Wang\textsuperscript{c}, and Gengsheng Qin\textsuperscript{c}
\textsuperscript{a}Howard University, Washington DC 20059, USA
\textsuperscript{b}University of Western Ontario, London, Ontario, Canada N6A 5B7
\textsuperscript{c}Georgia State University, Atlanta GA 30303, USA

Abstract

In this paper we study U-statistics with side information incorporated using the method of empirical likelihood. Some basic properties of the proposed statistics are investigated. We find that by implementing the side information properly, the proposed U-statistics can have smaller asymptotic variance than the existing U-statistics in the literature. The proposed U-statistics can achieve asymptotic efficiency in a formal sense and their weak limits admit a convolution result. We also find that the corresponding U-likelihood ratio procedure, as well as the U-empirical likelihood based confidence interval construction, do not benefit from incorporating side information, a result that is consistent with the result under the standard empirical likelihood ratio procedure. The impact of incorrect side information implementation in the proposed U-statistics is also explored. Simulation studies are conducted to assess the finite sample performance of the proposed method. The numerical results show that with side information implemented, the deduction of asymptotic variance can be substantial in some cases, and the coverage probability of the confidence interval using the U-empirical likelihood ratio based method outperforms that of the normal approximation based method, in particular in the cases when the underlying distribution is skewed.

Keywords
Efficiency; Information bound; Side information; U-statistic

1. Introduction

Since the pioneering work of Hoeffding [15], the U-statistics have been an active research field in statistics due to their wide range of applications. Hoeffding [16] established some fundamental properties of U-statistics which had close relationship with the V-statistic proposed by von Mises [37]. Berk [5] discovered the reverse martingale structure for U-statistic. Sen (e.g. [33]) made a number of contributions in this topic. Parallel to the result for V-statistics, Gregory [13] obtained the asymptotic distribution for degenerate U-statistics with rank two. The asymptotic distribution of U-statistics with arbitrary rank was developed by Janson [17] and Rubin and Vitale [32], etc. Borovskich [7] extended the results to Hilbert space. A detailed review and major historical developments in this field can be found in the book by Koroljuk and Borovskich [21], hereafter denoted as KB.

The empirical likelihood (EL) is one of the recent major developments in statistics. The original idea can be traced back to Thomas and Grunkemeier [35]. The work of Owen [25–
formally established the advantages and application scopes of this method, and paved the road of increasing popularity of EL due to the wide range of applications, the theoretical advantages, the simplicity of usage and the flexibility to incorporate auxiliary (or side) information in various forms. EL has been applied in various problems, for example, nonparametric confidence regions [9], the generalized linear model [20], survival analysis [1], density and quantile estimations [8,39], goodness-of-fit measure [3], nonparametric regression [10,29], marginal and conditional likelihood [30], ROC curve [31], econometrics [19], etc. It is well known that incorporating side information via empirical likelihood can reduce asymptotic variance of the estimators [28]. Motivated by this fact, we explore to incorporate side information into the U-statistic using the EL method, and expect that the new procedure can improve the performance of U-statistic under appropriate conditions.

It is also known that constructing confidence regions using EL ratio has various advantages than using normal approximation based method or bootstrap. For example, Wood et al. [38] and Jing et al. [18] have considered the EL method to U-statistics to construct confidence intervals without side information incorporated. We investigate the U-statistic to construct confidence intervals using the empirical likelihood by incorporating side information, and the resulting confidence intervals are compared with those based on normal approximation. Our method of formulating the weights of U-statistics is parallel to those in the EL, and is different from that in [38,18]. We find that by incorporating the side information properly, the proposed U-statistics will have smaller asymptotic variance than the existing U-statistics methods without side information. The proposed U-statistics can achieve asymptotic efficiency in a formal sense, and their weak limits admit a convolution result. We also find that the U-statistic EL based likelihood ratio procedure do not benefit from incorporating the side information asymptotically, a result that is consistent with the result under the standard empirical likelihood ratio procedure. The resulting coverage probability based on finite sample still outperforms that of the normal based approximation. The impact of incorrect side information incorporation is also explored.

In Section 2 we introduce the framework of the proposed U-statistics with side information incorporated, and investigate the basic asymptotic properties of the proposed U-statistics in Section 3. The U-empirical likelihood ratio with side information is formulated in Section 4. Examples and simulations results are given in Section 5 to illustrate the proposed method. All the relevant proofs are left in the Appendix.

2. Incorporating side information in U-Statistics

Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with unknown distribution function $F(x) = P(X \leq x)$. In this paper, we assume $X_i$ being random variable for simplicity, although there is no essential difference to extend it to the case of random vectors. Denote $X = (X_1, \ldots, X_m)'$, $(m \geq 2)$. Let $i = (i_1, \ldots, i_m)'$, $X_i = (X_{i1}, \ldots, X_{im})'$, $D_{n,m} = \{i : 1 \leq i_1 < \cdots < i_m \leq n\}$ denotes the collection of indices for the U-statistic of degree $m$. Let $C_m^n$ be the combination number of $m$ elements out of $n$, $x = (x_1, \ldots, x_m)'$, $F_m(x) = \prod_{j=1}^m F(x_j)$ and $F_{n, m}(x)$ be the empirical distribution function of $F_m$ based on the sample $x = (x_1, \ldots, x_m)'$, $\mathcal{F}_n = \{X_i : i \in D_{n,m}\}$, with mass $1/C_m^n$ at each point in $\mathcal{F}_n$. Given an $m$-variate symmetric kernel $h$, the U-statistic is defined as

$$U_n = \frac{1}{C_m^n} \sum_{i \in D_{n,m}} h(X_i) = E_{F_{n, m}} h(X).$$

$J Multivar Anal.$ Author manuscript; available in PMC 2013 May 21.
The goal is to estimate \( \theta = E_{F_m} h(X) \), where \( E_{F_m} \) denotes the expectation with respect to \( F_m \). It is known that the \( U \)-statistic \( U_n \) is the minimal variance unbiased estimator of \( \theta \) [34, p. 176].

Since the work of Owen [25], the empirical likelihood (EL) has gained increasing popularity due to its wide range of applications, simplicity to use and flexibility to incorporate auxiliary (or side) information. We examine here to combine both the EL method to flexibly incorporate side information and the \( U \)-statistics to achieve a better variance for the estimator.

We consider the set-up for EL as in [28]. Suppose the side information can be incorporated into the EL through a \( d \)-dimensional known function \( g(x) = (g_1(x), \ldots, g_d(x))' \) via the relationship

\[
E[g(X_i)] = 0,
\]

where \( E[\cdot] \) denotes the expectation with respect to \( F \). The EL is defined as

\[
L(F) = \prod_{i=1}^{n} w_i,
\]

where the \( w_i \) are the nonparametric maximum likelihood estimated empirical masses assigned to the observation \( X_i \). With the side information constraints, the EL is

\[
\max_{w} \prod_{i=1}^{n} w_i \text{subject to } \sum_{i=1}^{n} w_i = 1 \text{ and } \sum_{i=1}^{n} w_i g(X_i) = 0.
\]

Let \( t = (t_1, \ldots, t_d)' \) be the Lagrange multipliers corresponding to the constraint of \( g(\cdot) \), and as in [26], we get

\[
w_i = \frac{1}{n} \frac{1}{1 + t' g(X_i)},
\]

where \( t_j = t_j(X_1, \ldots, X_n)(j = 1, \ldots, d) \) are determined by

\[
\sum_{i=1}^{n} \frac{g(X_i)}{1 + t' g(X_i)} = 0.
\]

To combine the EL method and \( U \)-statistics, Wood et al. [38] considered a weighted \( U \)-statistic

\[
(C_n^{-1}) \sum_{1 \leq i_1 < \cdots < i_m \leq n} n^{m} w_{i_1} \cdots w_{i_m} h(X_{i_1}, \ldots, X_{i_m})
\]

with weight \( w(i_1, \ldots, i_m) = w_{i_1} \cdots w_{i_m} \) being estimated using EL procedure. Jing et al. [18] proposed a Jackknife EL for the \( U \)-statistic without side information considered. They first merge the \( C_n^m \) observed \( h(X_i) \)'s into a Jackknife sample, then treat this Jackknife pseudo
sample as a sample of $n$ i.i.d observations and apply the standard EL method for the mean to obtain the EL estimate for $U$-statistic.

In this paper, our goal is to estimate $\theta = E_{F_{m}}h(X)$ under the information constraints to incorporate side information in the form

$$E_{\hat{F}_{m}}g(X) = 0. \quad (1)$$

Without loss of generality $g(\cdot)$ is assumed symmetric with respect to its arguments (otherwise we can set $g(x_{1}, \ldots, x_{m}) = 1/m! \sum_{(p)} g(x_{i_{1}}, \ldots, x_{i_{m}})$ to make it symmetric, where the notation $\sum_{(p)}$ denote summation over the indices $(i_{1}, \ldots, i_{m})$ of all the permutations of $(1, \ldots, m)$). This function $g$ includes constraints $E_{F}(g(X_{1})) = 0$ as a special case by setting a componentwise product $g(X) = \prod_{j=1}^{m} g(X_{j})$. Some examples of $g(\cdot)$ will be given in Section 5 for illustration.

To formulate the proposed $U$-statistic we consider a different but a direct way to define the weights $w(i_{1}, \ldots, i_{m})'$. Let $w_{1} = F_{m}(\{X_{i}\})$ and $w = (w_{i} : i \in D_{n,m})$. Since the $w_{i}$'s are unknown (as is $F_{m}$), we maximizes the product of the $w_{i}$'s subject to appropriate constraints (they may not be independent of each other). Re-write the EL subject to the side information constraints as

$$\max_{w} \prod_{i \in D_{n,m}} w_{i}\text{subject to } \sum_{i \in D_{n,m}} w_{i} = 1 \text{ and } \sum_{i \in D_{n,m}} w_{i}g(X_{i}) = 0.$$  

We get, as in [26], that

$$w_{1} = (C_{m}^{-})^{-1} \frac{1}{1 + t'g(X_{1})}, \quad (2)$$

and $t = t_{n} = (t_{n1}, \ldots, t_{nd})'$ with $t_{nj} = t_{nj}(X_{1}, \ldots, X_{n})(j = 1, \ldots, d)$ being determined by

$$\sum_{i \in D_{n,m}} \frac{g(X_{i})}{1 + t'g(X_{i})} = 0. \quad (3)$$

For details regarding the existence of $t$ as the solution of (3) see, for example, the papers by Owen and others. The proposed weights for $U$-statistics are parallel to those in the EL, and simpler than some existing method in that there is no need to form a product of $m$ elements from $w_{1}, \ldots, w_{n}$ as in [38], nor to merge the data as in [18].

Similar to Hoeffding [15], for any kernel $h(\cdot)$ with $E_{F_{m}}(h(X)) < \infty$, let $h_{c}(x_{1}, \ldots, x_{d}) = E(h(X_{1}, \ldots, X_{m})|X_{1} = x_{1}, \ldots, X_{c} = x_{c})$, $h_{c}^{*} = h_{c} - \theta$ be its centered version $(c = 1, \ldots, m)$,

$$\hat{h}_{1}(x_{1}) = h_{1}^{*}(x_{1}), \quad \hat{h}_{2}(x_{1}, x_{2}) = h_{2}^{*}(x_{1}, x_{2}) - \hat{h}_{1}(x_{1}) - \hat{h}_{1}(x_{2}),$$

and in general,

$$J \text{ Multivar Anal. Author manuscript; available in PMC 2013 May 21.}$$
\[ \tilde{h}_c(x_1, \ldots, x_c) = h^c(x_1, \ldots, x_c) - \sum_{i=1}^{c} \tilde{h}_1(x_i) - \sum_{1 \leq i < j \leq c} \tilde{h}_{ij}(x_i, x_j) - \cdots - \sum_{1 \leq i < \cdots < j \leq c} \tilde{h}_{ij-1}(x_i, \ldots, x_{i-1}) = \int \cdots \int h_c(y_1, \ldots, y_c) \prod_{s=1}^{c} d(\delta_{y_s}(y_s)) - F(y_x), (c=1, \ldots, m), \]

where \( \delta_{x_s}(y_s) \) is the Dirac function, taking value 1 if \( y_s = x_s \) and 0 otherwise. The integration representation above can be found in KB. The \( \tilde{h}_c \)'s are called canonical forms of \( h \). If \( \tilde{h}_1 = \cdots = \tilde{h}_{k-1} = 0 \) and \( \tilde{h}_k \neq 0 \) (or equivalently \( \text{Var}(\tilde{h}_1) = \cdots = \text{Var}(\tilde{h}_{k-1}) = 0 \) and \( \text{Var}(\tilde{h}_k) \neq 0 \)), the \( U \)-statistic \( U_n \) with kernel \( h \) is said of rank \( k(1 \leq k \leq m) \). When \( k > 1 \) \( U_n \) is called degenerate, when \( k = m \) it is called complete degenerate. \( U_n \) has the following Hoeffding [16] representation

\[ U_n = \theta \sum_{c=k}^{m} C_n^c U_{nc} = (C_n^c)^{-1} \sum_{1 \leq i_1 < \cdots < i_c \leq n} \tilde{h}_c(X_{i_1}, \ldots, X_{i_c}). \]

Let \( \eta_c^2 = E[\tilde{h}_c^2](c=1, \ldots, m) \). \( U_n \) has the following variance formula [16]

\[ \text{Var}(U_n) = (C_n^c)^{-1} \sum_{c=1}^{m} C_n^c C_{n-m}^{c-\eta_c^2}. \]

Define \( g_c = (g_{c,1}, \ldots, g_{c,d})' \) with

\[ g_{c,j}(x_1, \ldots, x_c) = E_{\tilde{h}_c}(g_j(X_1, \ldots, X_c) | X_1 = x_1, \ldots, X_c = x_c), (j=1, \ldots, d; c=1, \ldots, m) \]

and the canonical forms \( \bar{g}_c = (\bar{g}_{c,1}, \ldots, \bar{g}_{c,d})' \) for \( g \) as,

\[ \bar{g}_c(x_1, \ldots, x_c) = \int \cdots \int g_{c,j}(y_1, \ldots, y_c) \prod_{s=1}^{c} d(\delta_{y_s}(y_s)) - F(y_x), (j=1, \ldots, d; c=1, \ldots, m). \]

Similarly, let \( q_c (c = 1, \ldots, m) \) be the canonical forms of \( g(h) = (g_1(h), \ldots, g_d(h))' \).

The canonical forms \( \bar{q}_c \) and \( \bar{q}_c (c = 1, \ldots, m) \) exist theoretically, but are unknown in practice since \( F \) is unknown. Let \( r_0 = \min \{ \text{rank}(g_{c1}), \ldots, \text{rank}(g_{cd}) \}, r = \text{rank}(h), r_1 = \min \{ \text{rank}(g_{c1}h), \ldots, \text{rank}(g_{cd}h) \}, \) and \( F_{nm} \) be the empirical distribution with mass \( w_i \) at the observation \( x_i \).

Using the weights \( w_i \)'s given in (2) and (3), we define the \( U \)-statistic with side information given by the constraints \( g \) as

\[ \bar{U}_n = \sum_{k \in D_{nm}} w_k h(X_k) = E_{\bar{F}_{nm}} h(X). \quad (4) \]
In comparison, the commonly used $U$-statistic $U_n$ has weight $(C_n^{-m})^{-1}$ at each observation $h(X_i)$, while with the EL formulation, the weights are replaced by $w_i$. In the following we investigate the basic asymptotic properties of $\tilde{U}_n$.

3. The asymptotic properties of $\tilde{U}_n$

In this section we study some basic asymptotic behavior of the proposed $U$-statistic, including its convergence, asymptotic distribution, uniform convergence, and asymptotic efficiency. The following conditions will be used in this section:

(C1). $\Omega = E[g(X)g'(X)]$ is positive definite.

(C2). $E\|g(X)\|_\alpha < \infty$ for some $\alpha > 0$ to be specified.

(C3). $EF_{\tilde{U}_n}|h(X)| < \infty$.

(C4). $EF_{\tilde{U}_n}h^2(X) < \infty$.

(C5). $EF_{\tilde{U}_n}\|g(X)\|^2|h(X)| < \infty$.

where $\|\cdot\|$ denotes the Euclidean norm. We note that (C2) with $\alpha \geq 4$ plus (C4) implies (C5).

3.1. Convergence rate of $\tilde{U}_n$

We first give a lemma to characterize the asymptotic form of the weight $w_i$’s, which will be used repeatedly in the asymptotic study.

Lemma. Assume (C1) and (C2) for $\alpha > 2m/r_0$, we have

i. $w_i \overset{a.s.}{=} \frac{1}{C_n} \left( 1 - g'(X_i)\Omega^{-1} \frac{1}{C_m} \sum_{j \in D_{n,m}} g(X_j) + \int g(X) + \int g(X) + \int g(X)^2 \right) O(p_n^2)$

where, $1_d = (1, \ldots, 1)$’ is the $d$-dimensional vector of 1’s, the $O(\cdot)$ terms are uniformly for all $x_i$’s and $i$’s, with

$$p_n = \begin{cases} O(n^{-1/2}(\log n)^{1/2}), & r_0 = 1; \\ o(n^{-r_0/2}\log n), & 1 < r_0 \leq m. \end{cases}$$

ii. $w_i = \frac{1}{C_n} \left( 1 - g'(X_i)\Omega^{-1} \frac{1}{C_m} \sum_{j \in D_{n,m}} g(X_j) + \int g(X) + \int g(X) + \int g(X)^2 \right) O(p_n^{-r_0})$

The $O_p(\cdot)$ terms above are uniformly for all the $x_i$’s and $i$’s.

Theorem 1. (i). Assume the conditions in the lemma plus (C3) and (C5), if $r = 1$, then

$$n^q(\tilde{U}_n - \theta) \rightarrow 0, a.s.\ for\ all q < 1/2.$$  

(ii). Assume conditions in the lemma plus (C4) and (C5), if $r > 1$, then
\( \alpha_n(\bar{U}_n - \theta) \to 0, (a.s.), \) where \( \alpha_n = \begin{cases} \frac{n^q}{\log n}, & \text{if } r_1 = r_o = 1; \\ \frac{n^{min(r_o, r_1)/2}}{\log n}, & \text{if } r_1 > r_o = 1; \\ \frac{n^{min(r_o, r_1)}}{\log n}, & \text{if } r_1 = r_o < r; \\ \min \left\{ \frac{r_o^{1/2}}{2r_o^{1/2}}, \frac{n^{min(r_o+1)/2}}{\log n} \right\}, & \text{if } r_o, r > 1. \end{cases} \)

(iii). Assume (C4) and conditions of Lemma (i), if \( r = 1, \) then with \( \sigma^2 \) given in Theorem 2 (i),

\[
\limsup_n \left( 2\sigma^2 \frac{\log \log n}{n} \right)^{-1/2} |\bar{U}_n - \theta| = 1, (a.s.).
\]

### 3.2. Asymptotic distribution of \( \bar{U}_n \)

Let \( J_1(h) \) be the Gaussian process indexed by \( h \in L^2(\mathbb{R}, \mathcal{B}, \mathcal{F}) \) with mean \( \mathbb{E}J_1(h) = 0 \) and covariance \( \text{Cov}(J_1(h), J_1(g)) = \int h(x)g(x)F(dx) \) for all \( h, g \in L^2(\mathbb{R}, \mathcal{B}, \mathcal{F}). \) Let \( W(\cdot) \) be the Gaussian random measure on \( L^2(\mathbb{R}, \mathcal{B}, \mathcal{P}) \) defined by \( W(A) = J_1(I_A), A \in \mathcal{B}, \) \( J_1(h) = \int h(x)W(dx) \) is called the Wiener–Itô integral of order 1. Generally, for \( h \in L^2(\mathbb{R}^r, \mathcal{B}^r, \mathcal{F}), \) the Wiener–Itô integral of order \( r \) is defined as

\[
J_r(h) = \int \cdots \int h(x_1, \ldots, x_r)W(dx_1) \cdots W(dx_r), \forall h \in L^2(\mathbb{R}^r, \mathcal{B}^r, \mathcal{F}),
\]

and its covariance is given by

\[
\text{Cov}(J_r(h), J_r(g)) = r! \int \cdots \int h(x_1, \ldots, x_r)g(x_1, \ldots, x_r)F(dx_1) \cdots F(dx_r), \forall h, g \in L^2(\mathbb{R}^r, \mathcal{B}^r, \mathcal{F}).
\]

For a vector function \( h = (h_1, \ldots, h_d)^\prime \) with \( h_j \in L^2(\mathbb{R}^r, \mathcal{B}^r, \mathcal{F})(j = 1, \ldots, d), \) define \( J_r(h) \) componentwisely as a \( d \)-dimensional random process. Denote \( \overset{D}{\rightarrow} \) for convergence in distribution.

**Theorem 2.** (i) Assume (C4) and conditions of the lemma, if \( r = 1, \) then

\[
\sqrt{n}(|\bar{U}_n - \theta|^2) \overset{D}{\rightarrow} N(0, \sigma^2), \sigma^2 = \begin{cases} \frac{n^q}{\log n}, & \text{if } r_1 = r_o = 1; \\ \frac{n^{min(r_o, r_1)}}{\log n}, & \text{if } r_1 > r_o = 1; \\ \frac{n^{min(r_o, r_1)}}{\log n}, & \text{if } r_1 = r_o < r; \\ \min \left\{ \frac{r_o^{1/2}}{2r_o^{1/2}}, \frac{n^{min(r_o+1)}}{\log n} \right\}, & \text{if } r_o, r > 1. \end{cases}
\]

where \( \eta_1^2 = \mathbb{E}_F h_1^2(X_1), \Omega_1 = \mathbb{E}_F (\mathbb{E}_F)^\prime g(X_1) h(x_i)h_i^\prime(X_1)), A = \mathbb{E}_F [g(X)|h(X)] \) and \( A_1 = \mathbb{E}_F [g(X_1)|h(X_1)] = \mathbb{E}_F [\tilde{g}(X_1)\tilde{h}(X_1)]. \)

(ii) Assume (C4), conditions of Lemma (ii) and \( r > 1, \) then

\[
n^{b/2}(|\bar{U}_n - \theta|^2) \overset{D}{\rightarrow} Z, \text{ where when } A \neq 0.
\]
when \( A = 0 \),

\[
\begin{align*}
& \begin{cases}
  b=r_o, & Z = -C_m A' \Omega^{-1} J_r_0 (\tilde{h}_r), \quad \text{if } r_o < r; \\
  b=r, & Z = C_m J(\tilde{h}_r) - A' \Omega^{-1} \tilde{g} (\tilde{r}), \quad \text{if } r_o = r; \\
  b=r, & Z = C_m J_1 (\tilde{h}_r), \quad \text{if } r_o > r;
\end{cases}
\end{align*}
\]

From Theorem 2 we see that the most interesting case is \( r = r_0 = r_1 = 1 \), in which \( \sigma^2 \) is asymptotic non-degenerate normal, with asymptotic variance being smaller than that of \( \sigma^2 \) is the same as that of \( U_{n} \) either when \( r_1 > 1, A = 0 \), or when \( r_0 > 1, A_1 = 0 \) and \( \Omega_1 = 0 \). Thus, for the side information to be of practical meaning, we need \( r = r_0 = r_1 = 1 \).

It is interesting to note that if we have full information about the parameter to be estimated, then we can “estimate” the parameter with perfection, i.e., its asymptotic variance being reduced to zero. As an artificial example, let \( a \) and \( b \) are nonzero known constants, \( \mu = E(X_1) \), and \( g \) is a known function, \( \mu \) must be known. We are to estimate \( \theta = E(X_1), \ldots, X_m) = am \mu \) using the \( U \)-statistic (4), with the \( w_i \)’s given by (2) and (3). In this case, \( \theta \) is already known as is \( \mu \), and we can “estimate” \( \theta \) with zero asymptotic variance, as in the following

**Corollary 1.** Assume \( \sigma^2 = \text{Var}(X_1) < \infty \), and \( h \) and \( g \) are as given in the above. Then Theorem 2 (i) holds with \( \sigma^2 = 0 \).

### 3.3. The optimality property of \( \hat{U}_n \)

To study the asymptotic efficiency of the estimators of \( \theta \), let \( \Theta|g \) be the information bound [6] for estimating \( \theta \) given the side information in \( g \), to be given in Theorem 3(i) below.

When the asymptotic variance of an estimator achieves this bound or equals to this bound up to a known multiplicative positive constant, the estimator is called asymptotically efficient. It is the limit version of the Cramer–Rao lower bound for variances of unbiased estimators. For Euclidean parameters without \( g \), the information (lower) bound is the inverse of Fisher information.

Suppose \( f(\cdot|\theta) \) is the density function of \( X \) given \( \theta \), \( \theta_n = \theta + n^{-1/2} b \) for some \( b \in C \), the complex plane. An estimator \( T_n = T_n(X_1, \ldots, X_n) \) is said to be regular, if under \( f(\cdot|\theta_n) \), \( W_n := \sqrt{n}(T_n - \theta_n) \to W \) for some random variable \( W \), and the result does not depend on the sequence \( \{\theta_n\} \). Let \( Z \oplus V \) denote the summation of two independent random variables \( Z \) and \( V \), \( I(\theta) \) be the Fisher information at \( \theta \), and \( Z \sim N(0, I^{-1}(\theta)) \). The convolution Theorem [14] states that for any regular estimator \( T_n \) with weak limit \( W \), there is a \( V \) such that

\[
W = Z \oplus V.
\]

This result further characterizes the weak limit of an asymptotic efficient estimator without side information: it is a normal random variable with mean zero and variance \( I^{-1}(\theta) \). Below
we obtain the information bound and convolution result for the proposed estimators with the presence of side information.

**Theorem 3.** Assume \( r = r_0 = 1 \), (C4) and conditions in the lemma, we have

i. \[
\mathbb{I}(\theta | g) = \eta_1^2 - A_1^\top \Omega_1^{-1} A_1.
\]

Thus, if we set \( g(x) = (g(x_1) + \cdots + g(x_m))/m \), then \( \text{rank}(g) = 1 \), \( A = mA_1 \), \( \Omega = m\Omega_1 \), \( \sigma^2 = m^2\mathbb{I}(\theta | g) \) and \( \bar{U}_n \) is efficient.

ii. Assume further that the density \( f(\cdot | \theta) \) of \( X \) has the second order continuous partial derivative with respect to \( \theta \), then for any regular estimator \( T_n \) with weak limit \( W \) of \( W_n := \sqrt{n}(T_n - \theta) \), \( W \) can be decomposed as, for some \( V \),

\[
W = Z \oplus V, \quad \text{with } Z \sim N(0, \mathbb{I}(\theta | g)).
\]

It is easy to see that any \( U \)-statistic with side information of the form \( \bar{U}_n \) is regular, thus is optimal in the sense of convolution under the conditions of Theorem 3. Without side information, the asymptotic variance of \( \sqrt{n}(U_n - \theta) \) is \( \eta_1^2 \); when side information presents, the asymptotic variance of \( \sqrt{n}(\bar{U}_n - \theta) \) is \( \eta_1^2 - A_1^\top \Omega_1^{-1} A_1 \) with a reduction of \( A_1^\top \Omega_1^{-1} A_1 \). From the proof of Theorem 3(i) we see that \( \mathbb{I}(\theta | g) \) is the length of the projection of \( \tilde{h}_1(X) \) onto \( \mathcal{g}_1(X)^\perp \), the linear span of the orthogonal complements of \( \mathcal{g}_1(X) \). Increasing the components in \( g \) (and thus in \( \tilde{g}_1 \)) shrinks the space \( \mathcal{g}_1(X)^\perp \), and shortens the length of the projection or increases the efficiency of \( \bar{U}_n \), or increasing the number of information constraints reduces the asymptotic variance of the \( U \)-statistic.

**Remark.** By Theorem 2(i) or Theorem 3(i), given a nominal level \( \alpha \), a level \((1 - \alpha)\) confidence interval of \( \theta \) can be obtained as \( [\bar{U}_n \pm n^{-1/2} \sigma \Phi^{-1}(1 - \alpha/2)] \), without using likelihood ratio, where \( \Phi^{-1}(\cdot) \) is the standard normal quantile function. Here \( \sigma \) is smaller with the presence of side information than no side information involved, hence the inference becomes more accurate.

### 3.4. The uniform SLLN and CLT of \( \bar{U}_n \)-processes

Let \( \bar{P}_{n,m}, P_{n,m}, P_m \) and \( P \) be the (random) probability measures induced by \( \bar{F}_{n,m}, F_{n,m}, F_m \) and \( F \) respectively. For a function \( h \), denote \( \bar{P}_{n,m,h} = \sum_{i \in D_{n,m}} w_i h(X_i) \), \( P_{m,h} = E_{P_m} h(X) \), \( \bar{G}_{n,m,h} = \sqrt{n}(\bar{P}_{n,m,h} - P_{m,h}) \) and \( G_{n,m,h} = \sqrt{n}(P_{n,m,h} - P_{m,h}) \). For fixed \( h \) and \( g \), we have shown that, under appropriate conditions,

\[
\bar{P}_{n,m,h} \rightarrow P_{m,h} = P h(1 \text{a.s.}) \text{ and } \bar{G}_{n,m,h} \stackrel{D}{\rightarrow} N(0, \sigma^2)
\]

with \( \sigma^2 = \sigma^2(h) = \mathbb{P}_1^{h_1} - P(\mathcal{g}_1 \mathcal{h}_1) \Omega_1^{-1} P(\mathcal{g}_1 \mathcal{h}_1) \). In contrast, \( \mathbb{G}_{n,m,h} \stackrel{D}{\rightarrow} N(0, \eta_1^2) \) with \( \eta_1^2 = P \mathbb{I}(\theta | g) \). Thus incorporating the side information \( g \) reduces the asymptotic variance by the amount of \( P(\mathcal{g}_1 \mathcal{h}_1) \Omega_1^{-1} P(\mathcal{g}_1 \mathcal{h}_1) \).

It is of interest to have a uniform version of the above SLLN and CLT over a class of functions \( \mathcal{H} \). The uniformity means supremum over \( \mathcal{H} \), which may or may not be measurable, thus the almost sure and weak convergence results here will be in the sense of outer measure \( P^* \) of \( P \)(cf. [36]; hereafter VW). When the corresponding quantity is
measurable, the convergence is automatically in the sense of the measure P itself. Nolan and Pollard [23,24] study the uniform SLLN and the CLT for U-process of order two. Giné and Zinn [12], Arcones and Giné [2] and Giné [11], among others, study other types of uniform problems in general situations. Here we explore the uniform laws for U-statistics under different conditions.

Let $\mathcal{H}$ be a class of functions satisfying (C4), and for any probability measure $Q$, denote $\|Qh\|_F = \sup_{h \in \mathcal{H}} |Qh|$. Let $L^\infty(\mathcal{H})$ be the space of functionals $z : \mathcal{H} \mapsto \mathbb{R}$ with norm $\|z\|_{\mathcal{H}} = \sup_{h \in \mathcal{H}} |z(h)|$. For an integer $m$-vector $k = (k_1, \ldots, k_m)$, a subset of $\mathbb{R}^m$ and a function $h : \mathcal{H} \mapsto \mathbb{R}$, denote $|k| = k_1 + \cdots + k_m$, and $\|x\|$ is the Euclidean norm for $x \in \mathbb{R}^m$. $C^M(\mathcal{H})$ is the set of functions $h : \mathcal{H} \mapsto \mathbb{R}$ with $\|h\|_s \leq M$.

Let $\mathcal{H}_1$ be the class of functions $h : \mathcal{H} \mapsto \mathbb{R}$ such that the restrictions $\mathcal{H}_1|_{I_j}$ belong to $C^M(I_j)$ for every $j$ and $M = \max_j M_j \leq \infty$. Let $\mathcal{H}_2$ be the class of convex functions $h : C \mapsto \mathbb{R}$ for some convex compact $C \subset \mathbb{R}^m$ such that $|h(x) - h(y)| \leq L\|x - y\|$ for all $0 < L < \infty$, all $x, y \in C$ and $h \in \mathcal{H}_2$, and $\sum_{i \in D_{n,m}} c_i h(x_i)$ is measurable for each $n$ and each $c_i \in \{-1, 1\}$. An envelope function $G$ of $\mathcal{H}$ is a function such that $|h(x)| \leq G(x)$ for all $x$, and $h \in \mathcal{H}$. Let $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$ with (C4) satisfied on $\mathcal{H}$ and $\lambda(\cdot)$ be the Lebesgue measure on $\mathbb{R}^m$. Let $\|\cdot\|$ denote weak convergence in $L^\infty(\mathcal{H})$.

**Theorem 4.** (i) Under the conditions of Theorem 1(i), for $\mathcal{H}$ defined above, assume that $\forall h \in \mathcal{H}$, $gh \in \mathcal{H}$ in the componentwise sense, $\mathcal{H}_1$ has a square integrable envelope function $H$, $\max_j \lambda(I_j) \leq \infty$, $\sum_{j=1}^\infty M_j^{1/2} P_m^{1/2}(I_j) < \infty$, and $\mathcal{H}_2$ is bounded. Then we have

$$\sup_{h \in \mathcal{H}} \|\hat{P}_{n,m}h - P_mh\| = 0, \text{ (a.s.)}. \quad \tag{1}$$

(ii) Under the conditions of Theorem 3(ii), assume $\mathcal{H}$ has a square integrable envelope function $H$, $\max_j \lambda(I_j) < \infty$, $\max_j \lambda(I_j) < \infty$, $m < 4$, $s > m/2$ for $\mathcal{H}_1$, and

$$\sum_{j=1}^\infty M_j^{2s/(s+2)} P_m^{s/(s+2)}(I_j) < \infty \quad \text{with } s = m/2. \quad \tag{2}$$

Then

$$\hat{G}_{n,m} D \Rightarrow G\in L^\infty(\mathcal{H}),$$

where $G$ is a Gaussian process indexed by $\mathcal{H}$, with $E_P(Gh) = 0$ and $\text{Cov}_p(Gh, Gq) = P(h_1 \tilde{q}_1) - P(\tilde{h}_1 \tilde{q}_1)\Omega_1^{-1} P(\tilde{h}_1 \tilde{q}_1)$ for all $h, q \in \mathcal{H}$.

Using results in [2,11], we can get many more results, below we only mention one.

**Corollary 2.** For a class of functions $\mathcal{H}$, assume that $\forall h \in \mathcal{H}, gh \in \mathcal{H}$; that $\mathcal{H}$ is a measurable VC-subgraph class of functions with envelope $H$ and $P_mH < \infty$, or $\forall \varepsilon > 0$, $N_{1/4}^\infty(\varepsilon, \mathcal{H}, P_m) < \infty$ (for definition, see p. 1512, [2]). Then

\[ J Multivar Anal. Author manuscript; available in PMC 2013 May 21. \]
4. The empirical likelihood ratio for U-statistics with side information

Next we define the empirical likelihood ratio for \( \theta \), and construct the confidence interval for \( \theta \) in the presence of side information. Let \( G(x|\theta) = (g'(x), h(x) - \theta)' \), we have \( E_{F_m} G(x|\theta) = 0 \). Without side information, the weights that maximize \( \prod_{i \in D_n} w_i \) subject to \( \sum_{i \in D_n} w_i = 1 \) are \( w_i = (C_n^m)^{-1} \) for all \( i \in D_n \); while the weights that maximize \( \prod_{i \in D_n} w_i \) subject to \( \sum_{i \in D_n} w_i = 1 \) and \( \sum_{i \in D_n} w_i G(X_i|\theta) = 0 \) are \( w_i = (C_n^m)^{-1} 1/(1 + t G(X_i|\theta)) \) and \( t \) is determined by (3) with \( g(\cdot) \) replaced by \( G(\cdot|\theta) \). Therefore we define the empirical log likelihood ratio of \( \theta \) with the presence of side information by

\[
R_c(\theta) = L_n(\theta)/(C_m^m) = C_n^m \prod_{i \in D_n} w_i,
\]

where

\[
L_n(\theta) = \max_{\sum_{i \in D_n} w_i = 1, \sum_{i \in D_n} w_i G(X_i|\theta) = 0} \prod_{i \in D_n} w_i
\]

and denote

\[
l(\theta) = -\log R_c(\theta) = \sum_{i \in D_n} \log [1 + t G(X_i|\theta)].
\]

Let \( \Lambda = E_{F_m} G(x|\theta)G'(X|\theta) = \begin{pmatrix} \Omega & A \\ A' \end{pmatrix} \) and \( \eta^2 = \text{Var}(h(X)) \); and \( \Lambda_1 = \text{Cov}(G_1) \), \( G_1 \) is the first canonical form (vector) of \( G \).

Note that when there is no side information, \( G(\cdot|\theta) \) reduces to \( h(\cdot) - \theta \), and \( t \) is a scalar determined by \( \sum_{i \in D_n} (h(X_i) - \theta)/[1 + t(h(X_i) - \theta)] = 0 \). The corresponding log-likelihood ratio is

\[
l_h(\theta) = \sum_{i \in D_n} \log [1 + t(h(X_i) - \theta)].
\]

**Theorem 5.** (i) Under conditions of Theorem 2 (i) or Theorem 3 (i), assume \( r_0 = 1 \) and \( \Lambda \) to be positive definite, then

\[
\frac{2n}{m^2 C_m^m} l(\theta) \xrightarrow{D} Z_{d_1+1}^{d_1} A_{1/2} A^{-1} A_{1/2}^{d_1+1}. Z_{d_1+1} \sim N(0, I_{d_1+1}).
\]

(ii) Assume (C4), then

\[
J \text{ Multivar Anal. Author manuscript; available in PMC 2013 May 21.}
\]
When \( m = 1 \), \( \Lambda_{1/2}^2 = \Lambda^\dagger \) and the above result for \( U \)-statistic automatically reduces to that for the common EL ratio, and the right hand side in Theorem 5(i) is \( \chi^2_{d+1} \) (see the corresponding result in Theorem 2 of Qin and Lawless [28]), therefore with side information incorporated in the likelihood ratio, the length of confidence region for \( \Theta \) cannot be reduced, this is an interesting contrast to the estimation with side information, in which the asymptotic variance is reduced. However, using the EL ratio, the shape of the confidence region is more natural than many other commonly used methods, such as the normal approximation, which are forced to be symmetric. The latter method may have poorer coverage probability because of the shorter interval length and its subjective shape.

Although side information is widely applied in practice to improve performance of estimators via the EL, the following Corollary 3 describes the effects when incorrect side information is used, thus side information should be used with care, and be justified properly before its use.

**Corollary 3.** If \( E_r f_m (x) = \delta \neq 0 \), then

i. Under conditions of Theorem 1 (i),

\[
\hat{U}_n - \theta \xrightarrow{a.s.} A^\dagger \Omega^{-1} \delta.
\]

ii. Under conditions of Theorem 2 (i),

\[
\sqrt{n}(\hat{U}_n - \theta - A^\dagger \Omega^{-1} \delta) \xrightarrow{D} N(0, \sigma^2).
\]

iii. If \( E_{F_m} G(x) = \delta \neq 0 \), then under conditions of Theorem 5 (i),

\[
-\frac{2n}{C_m} R_c(\theta) \xrightarrow{D} Z_{d+1}^2 \Lambda_1^{1/2} \Lambda^{-1}_2 \Lambda_1^{1/2} Z_{d+1}, Z_{d+1} \sim N(\sqrt{n} \Lambda_1^{-1/2} \delta, I_{d+1}),
\]

when \( \Lambda = \Lambda_1, Z_{d+1}^2 \Lambda_1^{1/2} \Lambda^{-1}_2 \Lambda_1^{1/2} Z_{d+1} = Z_{d+1}^2 (n \delta^2 \Lambda^{-1} \delta), \) the chi-squared distribution of degree \( d + 1 \) with noncentrality parameter \( n \delta^2 \Lambda^{-1} \delta \).

5. Examples and simulation studies

5.1. Examples

In this section we give some examples for illustration.

**Example 1.** For a given distribution \( F \), let \( \Theta(F) = \int (x - \mu)^2 dR(x) \) be the variance, where \( \mu \) is the mean. Let \( \mu_k \), \( k \geq 2 \) be the \( k \)-th moment of \( F \). For the kernel \( h(x_1, x_2) = (x_1 - x_2)^2/2 \),

we have \( \hat{h}_1(x_1) = [(x_1 - \mu)^2 - \theta]/2, \eta^2 = E(h^2) - \theta^2 = (\mu_4 + \theta^2)/2, \eta^2 = E(h^2) = (\mu_4 + \theta^2)/4 \).

Without side information, the asymptotic variance of \( U_n \) based on kernel \( h(x_1, x_2) \) is

\[
\sigma^2_{0} = 4 \eta^2 = \mu_4 - \theta^2, \]

which is the same as that for the sample variance estimator

\[
\hat{\sigma}^2_n := (n - 1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2.
\]

*J Multivar Anal.* Author manuscript; available in PMC 2013 May 21.
If we know that $F$ has median at 0: $F(0) = 1/2$, we take $g(x_1, x_2) = [k(x_1 \leq 0) + k(x_2 \leq 0)]/2 - 1/2$. Then $\tilde{g}_1(x_1) = [k(x_1 \leq 0) - 1/2]/2$, $\Lambda_1 = \text{E}(\tilde{g}_1) = [\int_{-\infty}^{0}(x - \mu)^2dF(x) - \theta^2]/4$, and $\Omega_1 = \text{E}(\tilde{g}_1^2) = 1/16$. So by Theorem 3(i), the asymptotic variance of $\tilde{U}_n$ is now

$$
\sigma^2 = \sigma_0^2 - \Lambda_1^{-1} \Omega_1^{-1} = 4\eta_1^2 - \left[ \int_{-\infty}^{0}(x - \mu)^2dF(x) - \sigma^2/2 \right]^2,
$$

a deduction of

$$
\left[ \int_{-\infty}^{0}(x - \mu)^2dF(x) - \sigma^2/2 \right]^2
$$

from $\sigma_0^2$.

**Example 2.** For the Wilcoxon one-sample statistic, $\Theta(F) = P_F(x_1 + x_2 \leq 0)$, the kernel for the corresponding $U$-statistic is $h(x_1, x_2) = k(x_1 + x_2 \leq 0)$, $h_1(x_1) = F(-x_1) - \theta$, $\eta_1^2 = \Theta(h_1(x_1)) = \int F^2(-x)dF(x) - \theta^2$. Without side information, the asymptotic variance of $U_n$ based on kernel $h(x_1, x_2)$ is

$$
\eta_1^2 = 4\eta_1^2.
$$

Suppose we know the distribution is symmetric about $a > 0$: $F(x - a) = 1 - F(a - x)$ for all $x$. Take $g(x_1, x_2) = [k(x_1 \leq 0) + k(x_1 \leq 2a) + k(x_2 \leq 0) + k(x_2 \leq 2a)]/2 - 1/2$, then $\tilde{g}_1(x_1) = [k(x_1 \leq 0) + k(x_1 \leq 2a)]/2 - 1/2$, $\Omega_1 = \int F(-a)dF(x)/2 - \int F(-x)dF(x)/2$. The deduction of asymptotic variance is $\Lambda_1^{-1}\Omega_1^{-1}$.

**Example 3.** For the Gini difference, $\Theta(F) = E_F|x_1 - x_2|$, the corresponding kernel for $U$-statistic is $h(x_1, x_2) = |x_1 - x_2|$. We have

$$
\tilde{h}_1(x_1) = \int_{x_1}^{\infty}xdF(x) - \int_{-\infty}^{x_1}xdF(x) - \theta, \quad \eta_1^2 = \left( \int_{x_1}^{\infty}xdF(x) - \int_{-\infty}^{x_1}xdF(x) \right)^2.
$$

Without side information, the asymptotic variance of $U_n$ based on kernel $h(x_1, x_2)$ is

$$
\eta_1^2 = 4\eta_1^2.
$$

If we know the distribution mean $\mu$, and take $g(x_1, x_2) = (x_1 + x_2)/2 - \mu$, then $\tilde{g}_1(x_1) = (x_1 - \mu)/2$, $\Omega_1 = \int(x - \mu)^2dF(x)/2 - \mu_1/2$, $\Lambda_1 = [\int_{x_1}^{\infty}xdF(x) - \int_{-\infty}^{x_1}xdF(x)]dF(x_1) - \theta]/2$. The deduction of asymptotic variance is $\Lambda_1^{-1}\Omega_1^{-1}$.

### 5.2. Simulation studies

Simulation studies are conducted to assess the finite-sample performance of the proposed methods in this section. These studies are based on Examples 1 and 2 in Section 5.1. We compare variance estimates of $U$-statistics with and without side information, and calculate the variance reduction under different sample sizes. We also compare various U-EL based and normal approximation-based confidence intervals for $\Theta$ in terms of coverage probability. Although side information does not have effect asymptotically, as indicated by Theorem 5, the finite sample property of constructing confidence intervals using the U-EL ratio is still of great interest, and will be compared with those obtained through normal approximation based method.

Based on the U-EL theory developed in Section 4, we can construct three U-EL based intervals for $\Theta$ as follows:

The first one, called EL1 interval, is defined as

$$
\left\{ \theta; \frac{2n}{m^2C_m^l}(\theta) \leq q_{1-\alpha} \right\}
$$

where $q_{1-\alpha}$ is the $(1 - \alpha)$-th quantile of the distribution of $\Lambda_1^{-1}Z_1\Lambda_1^{1/2}Z_1\Lambda_1^{1/2}$, $q_{1-\alpha}$ can be estimated by using the sample estimates of $\Lambda_1$ and $\Lambda$ and Monte Carlo method.
One can also approximate the quantile of the distribution of \( l(\theta) \) by using bootstrap method. Let \( \{ l_b(\tilde{\theta}); b = 1, \ldots, B \} \) \( (B \geq 200 \) is recommended) are \( B \) bootstrap replicates of \( l(\theta) \). Then, the second EL-based interval for \( \theta \), called EL2, is given by

\[
\left\{ \theta; l(\theta) \leq \frac{1}{B} \sum_{b=1}^{B} l_b(\tilde{\theta}) \right\},
\]

where \( l_b(\tilde{\theta}) \) is the \( b \)-th ordered value of \( l_b(\tilde{\theta}) \)'s, and \([x]\) represents the integer part of \( x \).

The third one, called EL3 interval, is constructed as follows:

\[
\left\{ \theta; c^* l(\theta) \leq \chi^2_{d+1, 1-\alpha} \right\},
\]

where \( c^* = \frac{d+1}{E(l(\theta))} \). This interval is motivated by the fact that the distribution of \( Z_{d+1}^{1/2} \Lambda^{-1/2} \Lambda^{1/2} Z_{d+1} \) can be approximated by a scaled chi-squares distribution, i.e.,

\[
c \cdot l(\theta) \xrightarrow{D} \chi^2_{d+1},
\]

where \( c \) is an unknown constant, and \( c \approx \frac{E(l(\theta))}{E(l(\theta))} = \frac{d+1}{E(l(\theta))} \).

The asymptotic normal distribution obtained in Theorem 2 can be used to construct two additional confidence intervals for \( \theta \), called AN1 and AN2 intervals, as follows:

\[
\left\{ \frac{\tilde{U}_n - z_{1-\alpha/2} \hat{\sigma}}{\sqrt{n}}, \frac{\tilde{U}_n + z_{1-\alpha/2} \hat{\sigma}}{\sqrt{n}} \right\}, \left\{ \tilde{U}_n - z_{1-\alpha/2} \hat{\sigma}^* / \sqrt{n}, \tilde{U}_n + z_{1-\alpha/2} \hat{\sigma}^* / \sqrt{n} \right\},
\]

where \( \hat{\sigma} \) is the estimate of \( \sigma \) by plugging the sample estimates of all population quantities in Theorem 2. \( \hat{\sigma}^* \) is the bootstrap estimate of \( \sigma \) based on \( B \) bootstrap samples. For computation consideration, we take \( B = 200 \) in the simulation studies.

Examples 1 and 2 in Section 5.1 are considered in the simulation study. In the first example, the underlying distribution is chosen to be a skewed distribution with median 0. Here we take \( X \sim \text{exp}(1) - \ln(2) \), the standard exponential distribution with a shifted center. Then \( EX = 1 - \ln 2 \), \( \text{Median}(X) = 0 \), and \( \theta = \text{Var}(X) = 1 \). In the second example, we consider a symmetric distribution with mean \( a \). We choose \( X \sim \mathcal{N}(1, 4) \), then \( \theta = \Phi(-\frac{1}{4}) \), where \( \Phi(x) \) is the cdf of the standard normal distribution. The simulation results are presented in Tables 1–4.

Tables 1 and 2 show the estimated asymptotic variances of \( U \)-statistics with or without side information respectively under sample size \( n = 50, 100, 150, 200 \). The reduction of variance is also calculated. The results are based on 1000 repetitions.

Tables 3 and 4 show the coverage probabilities of the EL-based intervals (EL1, EL2 and EL3) and the normal approximation-based intervals (AN1 and AN2) with side information.

The simulation results show that the proposed \( U \)-statistic \( \tilde{U}_n \) performs well in finite sample cases. From Tables 1 and 2 we can clearly see a reduction of the variance of estimating \( \theta \).
The variance reduction can be significant, as in Example 2, which shows that the proposed method could offer a more accurate estimation.

From the coverage probabilities in Tables 3 and 4, we see that the U-EL based confidence intervals work significantly better than the normal approximation-based confidence intervals when the underlying distribution is a skewed distribution (Example 1). When the underlying distribution is a symmetric distribution (Example 2), the performances of these methods are comparable. Furthermore, in most cases, bootstrap-based methods work better than plugin methods.

**Concluding remarks**—We studied a method to implement side information into the $U$-statistic, via the empirical likelihood approach, and investigated some asymptotic behavior of the proposed method. We show, for parameter estimation, the proposed $U$-statistic with side information has advantages, such as smaller asymptotic variance, over that without side information incorporated. We also explored the construction of confidence intervals using $U$-statistic based empirical likelihood ratio. Although such U-EL ratio does not benefit from side information asymptotically, our simulation studies show that the corresponding confidence intervals still out perform those based on normal approximation in finite sample cases. We also note that, if incorrect side information is incorporated, the resulting estimates can be seriously biased. Thus in practice the incorporation of side information should be justified properly.

**Acknowledgments**

This work is supported in part by the National Center for Research Resources at NIH grant 2G12RR003048. Dr. Gengsheng Qin’s work is supported in part by US NSA grant H98230-12-1-0228. Dr. Wenqing He’s work is partially supported by the Natural Sciences and Engineering Research Council of Canada.

**Appendix**

**Proof of the Lemma.** (i) As in [26], write $t = t_n = \rho_n e$ with $\rho_n \geq 0$ and $e = e(X_1, \ldots, X_n)$ a $d$-vector with $\|e\| = 1$. We first find the asymptotic order of $t_n$. Denote $b(t) = (C_n^n)^{-1} \sum_{i \in [n]} \| \frac{1}{t_i} g(X_i) \|$.

From the coverage probabilities in Tables 3 and 4, we see that the U-EL based confidence intervals work significantly better than the normal approximation-based confidence intervals when the underlying distribution is a skewed distribution (Example 1). When the underlying distribution is a symmetric distribution (Example 2), the performances of these methods are comparable. Furthermore, in most cases, bootstrap-based methods work better than plugin methods.

**Proof of the Lemma.** (i) As in [26], write $t = t_n = \rho_n e$ with $\rho_n \geq 0$ and $e = e(X_1, \ldots, X_n)$ a $d$-vector with $\|e\| = 1$. We first find the asymptotic order of $t_n$. Denote $b(t) = (C_n^n)^{-1} \sum_{i \in [n]} \| \frac{1}{t_i} g(X_i) \|$. Let $Z_n = \max \{ \| e' g(X_i) \| : i \in D_{n,m} \}$, and $\tilde{Z}_n = \max \{ \| e' g(X_i) \| : i \in \tilde{D}_{n,m} \}$ with $|D_{n,m}| = C_m^n$. Vi and $j$ have no common entry, and $X_i = (X_{i1}, \ldots, X_{im})$, $X_{ij}$ i.i.d. with $X_1$ for all $i_1, \ldots, i_m$ ($i = 1, \ldots, n$). Since $Z_n$ is a maximum over $C_m^n$ i.i.d. samples, while $Z_n$ is that over $C_n^n$ dependent samples come from the same distribution, for large $n$, we have $Z_n \preceq \tilde{Z}_n$ (a.s.). Since $E \| e' g(X) \|^{2a} < \infty$, $Z_n = o((C_m^n)^{1/2})$ (a.s.) as in [26], and so $Z_n = o((C_n^n)^{1/2})$ (a.s.). We have
Below we will show, for some \( 0 < c < C < \infty \), for all large \( n \), uniformly for all the \( X_i \)'s and \( i \)'s, 

\[
(A.1) \quad c < e' R_{1,n} e = \text{Ca.s. and } e' R_{2,n} = \begin{cases} O(n^{-1/2}(\log \log n)^{1/2}) \text{a.s. if } r_o = 1, \\ o(n^{-1/2}\log n) \text{a.s. if } r_o > 1. \end{cases}
\]

In fact, \( R_{1,n} \) is a (matrix valued) \( U \)-statistic with a.s. limit \( 0 < E[g(X)g'(X)] = \Omega < \infty \), where the “\( 0 < \)" is in the matrix positive definite sense and the “\( < \infty \)" is in the componentwise sense. Let \( 0 < \lambda_1 \leq \cdots \leq \lambda_d < \infty \) be all the eigenvalues of \( \Omega \), we have \( \Omega = Q \text{ diag}(\lambda_1, \ldots, \lambda_d) Q \) with \( Q \) being orthonormal. Denote \( \eta = Q e \), then \( \eta' \eta = 1 \). Then for large \( n \), \( R_{1,n} > \Omega/2 \) (a.s.) thus \( e' R_{1,n} e > e' \Omega e/2 = \eta' \text{ diag}(\lambda_1, \ldots, \lambda_d) \eta/2 \geq \lambda_1/2 =: c > 0 \) (a.s.). Similarly, for large \( n \), \( R_{1,n} < 2 \Omega \) (a.s.) and \( e' R_{1,n} e < 2 \lambda_d =: C \) (a.s.).

Since \( \|e\| = 1 \), we only need to prove the second assertion in (A.1) for \( R_{2,n} \). Note \( R_{2,n} \) is a (vector) \( U \)-statistic with kernel \( g(x) \) satisfying \( E[g(x)] = 0 \). Recall \( g_c \) is the canonical forms of \( g \), let \( R_{2,nc} \) be the corresponding Hoeffding forms of \( R_{2,n}(c = 1, \ldots, m) \). By the given condition we have \( E(\|g_c(X)\|^2) < \infty \), \( (c = r_o, \ldots, m) \), \( E(\|g(X)\|^{4/3}) < \infty \), and

\[
R_{2,n} = \sum_{c=r_o}^m C_{nc} R_{2,nc} \text{ in componentwise sense (component } j \text{ in } R_{2,nc} \text{ is zero for } c = r_o, \ldots, r_j -1 \text{ if } r_j := \text{rank}(g) > r_o). \text{ If } r_o = 1, \text{ let } \eta_{2,1}^2 := E(\|g_1(X)\|^2), \text{ by Theorem 9.1.1 in KB, we get}
\]

\[
\limsup \left( \frac{2m^2 \log \log n}{n} \right)^{-1/2} |R_{2,n}| = 1 \text{a.s.} \text{ or } R_{2,n} = O(n^{-1/2}(\log \log n)^{1/2}) \text{a.s.}
\]

If \( r_o > 1 \), by Lemma 9.2.1 in KB,

\[
\frac{n^{r_o/2}}{\log n} R_{2,nc} \to 0 \text{(a.s.)}, \text{ (c = } r_o, \ldots, m) \text{;and so } R_{2,n} = o(n^{-r_o/2}\log n) \text{(a.s.).}
\]
Now, since $R_{1,n} = O(1)$ (a.s.) with $0 < O(1) < \infty$ and $Z_n = O((C_{n}^{m})^{1/2}) = O(C^{m/2})$, we have, for $r_{0} = 1$, $\rho_{n}(1 + \rho_{n}Z_n) = O(e^{\sqrt{\rho_{n}}}) = O(n^{-1/2}(\log n)^{1/2})$ (a.s.), or $\rho_{n}(1 - O(n^{-1/2}(\log n)) = O(n^{-1/2}(\log n))$ (a.s.). For $r_{0} > 1$, $\rho_{n}(1 - O(n^{-1/2}(\log n))) = O(n^{-1/2}(\log n))$ (a.s.). Thus we have

$$||t|| = \rho_{n} = \begin{cases} O(n^{-1/2}(\log n)^{1/2}), & r_{0} = 1; \\ O(n^{-r_{0}/2}(\log n)), & 1 < r_{0} \leq m \end{cases} \text{ (a.s.)}.$$

Since for all $i$, $|t_{n}^{i}g(X_{i})| \leq ||t_{n}||Z_n = o(\rho_{n}n^{m/2}) \to 0$ (a.s.), thus for large $n$, $\max_{i \in D_{n,m}}|t_{n}^{i}g(X_{i})| < 1$ (a.s.), so we have

$$0 = \sum_{i \in D_{n,m}} \frac{g(X_{i})}{1 + t'g(X_{i})} = \sum_{i \in D_{n,m}} (g(X_{i}) (1 - t'g(X_{i}) ) + O(t'g(X_{i})^{2})) = \sum_{i \in D_{n,m}} (g(X_{i}) (1 - t'g(X_{i}) ) + ||g(X_{i})||^{2}O(\rho_{n}^{2})) \text{ (a.s.)},$$

or

$$\sum_{i \in D_{n,m}} g(X_{i})g'(X_{i}) = \sum_{i \in D_{n,m}} (||g(X_{i})||^{2}O(\rho_{n}^{2})).$$

Thus

$$t = \left( \sum_{i \in D_{n,m}} g(X_{i})g'(X_{i}) \right)^{-1} \sum_{i \in D_{n,m}} g(X_{i}) + O(\rho_{n}^{2}) \sum_{i \in D_{n,m}} g(X_{i})g'(X_{i}) = B_{n} + O(\rho_{n}^{2}) \sum_{i \in D_{n,m}} g(X_{i})g'(X_{i})^{-1} \sum_{i \in D_{n,m}} ||g(X_{i})||^{2}g(X_{i}) \text{ (a.s.)}.$$

We have already shown $R_{2,n} = (C_{n}^{m})^{-1} \sum_{i \in D_{n,m}} g(X_{i}) = O(\rho_{n})$ (a.s.). Also $g(\cdot)g'(\cdot)$ is non-degenerate by (C1), also since $m \geq 2$, $E||g(X)||^{4} < \infty$ by (C2), thus by the law of iterated logarithm (LIL) for $U$-statistics, $(C_{n}^{m})^{-1} \sum_{i \in D_{n,m}} g(X_{i})g'(X_{i}) = \Omega + O(n^{-1/2}(\log n)^{1/2})$ (a.s.), hence

$$B_{n} = [\Omega^{-1} + O(n^{-1/2}(\log n)^{1/2})] \sum_{i \in D_{n,m}} g(X_{i}) = \Omega^{-1} \sum_{i \in D_{n,m}} g(X_{i}) + O(\rho_{n}n^{-1/2}(\log n)^{1/2}) = O(\rho_{n}) \text{ (a.s.)}.$$

Similarly,

$$O(\rho_{n}^{2}) \sum_{i \in D_{n,m}} g(X_{i})g'(X_{i})^{-1} \sum_{i \in D_{n,m}} ||g(X_{i})||^{2}g(X_{i}) = O(\rho_{n}^{2}) \sum_{i \in D_{n,m}} g(X_{i})g'(X_{i})^{-1} ||g(X_{i})||^{2}g(X_{i}) = o(B_{n}) \text{ (a.s.)}.$$
so,

\[
 t = t_n = B_n + O(p_n^{-2}) = \Omega^{-1} \frac{1}{C_n^m} \sum_{j \in D_m} g(X_j) + O(p_n n^{-1/2} \log \log n) + O(p_n^2) \text{(a.s.)}
\]

From this we get, (a.s.),

\[
 w_1 = \frac{1}{C_n^m} \frac{1}{1 + t'g(X_i)} \\
 = \frac{1}{C_n^m} \left[ 1 - t'g(X_i) + \|g(X_i)\|^2 O(p_n^2) \right] \\
 = \frac{1}{C_n^m} \left( 1 - g'(X_i) \Omega^{-1} \frac{1}{C_n^m} \sum_{j \in D_m} g(X_j) \right) \\
 + \left[ t'g(X_i) + \|g(X_i)\|^2 O(p_n^2) \right].
\]

(ii) As in the proof of (i), we only need to show the results regardless of e. Treating \(gg'\) and \(R_{1,n}\) as vectors of length \(d\). Under the given conditions, note \(gg'\) is non-degenerate, by the central limit theorem (CLT) for \(U\)-statistics, \(\Xi\) is determined by \(gg'\). Similarly, for \(R_{2,n}\), since \(Eg(X) = 0\) if \(r_0 = 1\), by standard \(U\)-statistics theory, \(\sqrt{n}R_{2,n}\) is asymptotical normal, or \(R_{2,n} = Eg'(X) + O_{P}(n^{-1/2})\). If \(r_0 > 1\), note \(r_0 \leq m\), the given conditions implies \(E|g_c(x_1, \ldots, x_c)|^{2c/(2c-\sum_{j=1}^{c} \alpha_j)} < \infty\) for \(c = r_0, \ldots, m\). So, by Theorem 4.4.1 in KB, we have \(\|t_n\| = \rho_n = O_{P}(n^{-1/2})\).

Proof of Theorem 1. (i) By Lemma (i), we have

\[
 J Multivar Anal. Author manuscript; available in PMC 2013 May 21.
\]
By the given conditions and the SLLN of U-statistics, \( U_n \xrightarrow{a.s.} \theta \)

\[
\hat{U}_n = U_n = \left( \frac{1}{m} \sum_{i \in D_{n,m}} g'(X_i)h(X_i) \right) \Omega^{-1} \left( \frac{1}{m} \sum_{i \in D_{n,m}} g(X_i) \right) \\
+ O(\rho_n n^{-1/2}(\log n)^{1/2}) \\
\times \frac{1}{m} \sum_{i \in D_{n,m}} \hat{f}_i g(X_i) h(X_i) \\
+ O(\rho_n^2) \frac{1}{m} \sum_{i \in D_{n,m}} (1' g(X_i)) \\
+ \|g(X_i)\|^2) h(X_i) \text{ (a.s.)}
\]

where \( A = E F_m [g(X)h(X)] \), and

\[
U_{0,n} := \frac{1}{m} \sum_{i \in D_{n,m}} g(X_i) \xrightarrow{a.s.} E_{F_m} g(X) = 0, \quad U_{1,n} := \frac{1}{m} \sum_{i \in D_{n,m}} g(X_i) h(X_i) \xrightarrow{a.s.} A < \infty,
\]

where \( E = E_{F_m}[g(X)h(X)] \)

Since by the given conditions, \( E_{F_m}[\hat{f}_c(x)] \sim \infty \) and \( E_{F_m}[\|g_c(x)\|^2] \sim \infty \) for \( \gamma = cp(c-1) + 1 \), \( c = 1, \ldots, m \) and \( 1 < p < 2 \), so by Corollary 3.4.1 in KB, \( n^{-1/p+1} (U_n - \theta) \to 0 \) (a.s.), or \( n^q (U_n - \theta) \to 0 \) (a.s.) and \( n^q U_{0,n} \to 0 \) (a.s.) for all \( q < 1/2 \), and consequently, for all \( q < 1/2 \),

\[
n^q (U_n - \theta) = N^q (U_n - \theta) + O(n^q U_{0,n}) + O(n^q n^{-1/2-q} (\log n)^{1/2}) + O(n^q \rho_n^2) \to 0 \text{ (a.s.)}
\]

(ii) Use notations in (i), we have

\[
\hat{U}_n = U_n - \frac{1}{m} \sum_{i \in D_{n,m}} g'(X_i)h(X_i) \Omega^{-1} U_{0,n} + O(\rho_n n^{-1/2}(\log n)^{1/2}) U_{1,n} + O(\rho_n^2) U_{2,n} \text{ (a.s.)}
\]

Recall for any U-statistic \( U_n \) with rank \( r \) and canonical forms \( \hat{f}_c (c = r, \ldots, m) \) the following decomposition holds

\[
U_n - \theta = \sum_{c=r}^m \sum_{i_1 < \cdots < i_r \leq n} \hat{f}_c(x_{i_1}, \ldots, x_{i_r})
\]

Since \( \hat{E}_{c,c} < \infty \) for \( c = r, \ldots, m \), by Lemma 9.2.1 in KB, \( U_{n,c} = o(n^{-d_2} \log n) \) (a.s.), so \( U_n = o(n^{-d_2} \log n) \) (a.s.). Thus, when \( r_1 = r_0 = 1 \), \( U_{1,n} = O(1) \) (a.s.), \( U_{0,n} = o(n^{-q}) \) (a.s.) for any \( q < 1/2 \), and \( U_{2,n} = O(1) \) (a.s.) as its kernel is always non-degenerate, so by Lemma (i) we have

\[
\hat{U}_n - \theta = o(n^{-d_2/2} \log n) + o(n^{-q}) + O(n^{-1} \log n) + O(n^{-1} \log \log n) = o(n^{-q}), q < 1/2.
\]

J Multivar Anal. Author manuscript; available in PMC 2013 May 21.
When $r_1 > r_0 = 1$, by Lemma 9.2.1 in KB, $U_{1,n} = o(n^{-1/2} \log n)$ (a.s.), note $O(n^{-1} \log \log n) = o(n^{-1} \log n)$, and so

$$U_n - \theta = o(n^{-r/2} \log n)$$

$$+ o(n^{-r_1/2 + q} \log n)$$

$$+ o(n^{-(1+r/2)\log(n \log n)^{1/2}} + O(n^{-1} \log \log n))$$

$$= o(n^{-r_1/2, (r_1 + 1)/2} \log n)$$

$$= o(n^{-r_1/2, 2} \log n).$$

When $1 = r_1 < r_0$,

$$U_n - \theta = o(n^{-r/2} \log n) + o(n^{-(r_1/2) \log n}) + o(n^{-(1+r_1)/2} \log(n \log n)^{1/2}) + o(n^{-r_1}(\log n)^2) = o(n^{-r_1/2, r/2} \log n);$$

and when $r_1, r_0 > 1$,

$$U_n - \theta = o(n^{-r/2} \log n)$$

$$+ o(n^{-(r_1 + r)/2} (\log n)^2)$$

$$+ o(n^{-(1+r_1+r)/2} (\log n)^2 \log(n \log n)^{1/2})$$

$$+ o(n^{-r_1}(\log n)^2)$$

$$= o(\max (n^{-r/2} \log n, n^{-\min[(r_1 + r)/2, r_1]} (\log n)^2)).$$

(iii) Using Lemma (i) and notations in the proof of (i), we have, a.s.,

$$U_n - \theta = \frac{1}{c_{m,n}} \sum_{k \in D_{m,n}} (h(X_k) - \theta - A^T \Omega^{-1} g(X_k)) - \left(U_{1,n} - A \right) \Omega^{-1} U_{0,n} + O(\rho_n n^{-1/2} (\log \log n)^{1/2}) U_{1,n} + O(\rho_n^2 U_{2,n}).$$

Recall that $U_{0,n} = o(n^{-q})$, $U_{1,n} - A = o(n^{-q})$ (a.s.) for all $0 < q < 1/2$, and $U_{2,n} \to C_2$ (a.s.) for some $C_2 < \infty$. We have, a.s., for all $0 < q < 1/2$,

$$U_n - \theta = \frac{1}{c_{m,n}} \sum_{k \in D_{m,n}} (h(X_k) - \theta - A^T \Omega^{-1} g(X_k)) + o(n^{-2q}) + O(\rho_n n^{-1/2} (\log \log n)^{1/2}) + O(\rho_n^2).$$

By Theorem 9.1.1 of KB, the LIL holds for the first term above, and the above equation gives the desired result.

**Proof of Theorem 2.** (i) Using the fact that $U_{1,n} \to A$ (a.s.) and $U_{2,n} \to C_2$ (a.s.) for some $C_2 < \infty$ as proved in Theorem 1(i), thus by Lemma (ii),

$$\sqrt{n}(U_n - \theta) = \sqrt{n} \frac{1}{c_{m,n}} \sum_{k \in D_{m,n}} (h(X_k) - \theta - A^T \Omega^{-1} g(X_k)) - \sqrt{n}(U_{1,n} - A^T \Omega^{-1} U_{0,n} + O(\rho_n n^{-1/2}) U_{1,n} + O(\rho_n^2(n^{-r} - 1/2)) U_{2,n}).$$

The second term above is, for all $0 < q < 1/2$, $n^{1/2} O(n^{-2q}) = o_P(1)$; the third term above is $O_P(n^{-r/2})$ as $U_{1,n} \to A$ (a.s.) $< \infty$; and the last term above is $O_P(n^{-r/2(1/2)})$ as $U_{2,n} \to C$ (a.s.) for some $C < \infty$; Thus we only need to deal with the first term above.
Let
\[ H(x) = h(x) - \tilde{h}_1(x_1) - \cdots - \tilde{h}_1(x_m) - \theta, \quad G(x) = g(x) - \tilde{g}_1(x_1) - \cdots - \tilde{g}_1(x_m), \]

then \( H(\tilde{x}_1) = \mathbb{E}[H(X)|X_1 = \tilde{x}_1] = 0 \), i.e. \( H(x) \) is a degenerate kernel. Similarly, \( G(x) \) is degenerate, so is \( K(x) = H(x) - A'\Omega^{-1}G(x) \), with \( E_{\Omega,m}K(X) = 0 \) and \( \kappa = \text{rank}(K) \geq 2 \). Now we have
\[
\sqrt{n}(\hat{U}_n - \theta) = \sqrt{n} \sum_{i=1}^{n} \left( \tilde{h}_1(X_i) - \cdots - \tilde{h}_1(X_m) - A'\Omega^{-1}[\tilde{g}_1(X_i) + \cdots + \tilde{g}_1(X_m)] \right) + \sqrt{n} \sum_{i=1}^{n} \frac{K(X_i)}{\sqrt{C_m}} + O_p(n^{-1/2}).
\]

Let \( \tilde{K}_c \) be the canonical forms of \( K \), and \( \xi_c^2 = E_{\Omega,m}K_c(X) < \infty \) by the given conditions, \( c = r_k, \ldots, m \). So by Hoeffding's formula,
\[
\text{Var} \left( \sqrt{n} \sum_{i=1}^{n} K(X_i) \right) = \frac{m}{n} \sum_{c=r_k}^{C_m} (C_m)^{-1} \xi_c^2 = O(n^{-(r_1-1)}) \rightarrow 0,
\]
and so \( \sqrt{n} \sum_{i=1}^{n} K(X_i) \xrightarrow{P} 0 \). We get
\[
\sqrt{n}(\hat{U}_n - \theta) = \sqrt{n} \sum_{i=1}^{n} \left( \tilde{h}_1(X_i) - A'\Omega^{-1}\tilde{g}_1(X_i) \right) + o_p(1) + O_p(n^{-1/2}).
\]

Since \( \text{Var}(n\tilde{h}_1(X_1) - A'\Omega^{-1}\tilde{g}_1(X_1)) = \sigma^2 \) if \( r_0 = 1 \), and \( \text{Var}(\tilde{g}_1(X_1)) = m^2 \eta^2 \) when \( r_0 > 1 \) (in this case \( \tilde{g}_1(X_1) \equiv 0 \)). Now the result follows from the standard CLT and Slutsky's theorem.

(ii) We have, since \( U_{2,n} = O_p(1) \),
\[
\hat{U}_n - \theta = U_n - \theta - U_{1,n}J_r(\tilde{h}_1) + O_p(n^{-(r_0+1)/2})U_{1,n} + O_p(n^{-r_0}).
\]

By Theorem 4.4.2 in KB, \( n^{1/2}(U_n - \theta) \overset{D}{\to} C_mJ_r(\tilde{h}_1) \) or \( U_n - \theta = O_p(n^{-r_0/2}) \). Similarly in summary we have
\[
U_n - \theta = O_p(n^{-r_0/2}), \quad U_{1,n} = O_p(n^{-r_0/2}), \quad U_{1,n} - A = O_p(n^{-r_0/2}).
\]

Also, \( U_{1,n} \to A \) (a.s.), and when \( r_0 = 1 \), \( \sqrt{n}U_{0,n} \overset{D}{\to} N(0, m^2 \Omega_1) \).

First we consider the case \( A \neq 0 \). In this case, \( U_n - \theta = O_p(n^{-r_0/2}), U_{1,n} \Omega^{-1}U_{0,n} = O_p(n^{-r_0/2}) \) and \( O_p(n^{-(r_0+1)/2})U_{1,n} + O_p(n^{-r_0}) = o_p(n^{-r_0/2}) \).

Thus when \( r_0 < r \), we have
\[
r_0^{1/2}(\hat{U}_n - \theta) = - U_{1,n} \Omega^{-1}n^{-r_0/2}U_{0,n} + o_p(1) \overset{D}{\to} C_mA'\Omega^{-1}J_{r_0}(\tilde{g}_r).
\]
When $r_o = r$,

$$n^{r/2}(\hat{U}_n - \theta) = n^{r/2}(U_n - \theta) - U_1, \Omega^{-1} n^{r/2} U_{0,n} + o_p(1) \xrightarrow{D} C_m J_r(\hat{h}_r - A' \Omega^{-1} \bar{g}_r).$$

When $r_o > r$,

$$n^{r/2}(\hat{U}_n - \theta) = n^{r/2}(U_n - \theta) + o_p(1) \xrightarrow{D} C_m J_r(\hat{h}_r).$$

Now we consider the case $A = 0$, then $U_1, \Omega^{-1} U_{0,n} = O_p(n^{-(r_1 + r_o)/2})$, and

$$\hat{U}_n - \theta = U_n - \theta - U_1, \Omega^{-1} U_{0,n} + o_p(n^{-(r_1 + r_o)/2}) + o_p(n^{-r_1}).$$

When $r_o \leq \min\{r_1, r/2\}$, $\hat{U}_n - \theta = O_p(n^{-r_1/2})$, and its distribution needs more accurate expansion to evaluate. When $r < \min\{2r_o, r_1 + r_o\}$,

$$n^{-(r_1 + r_o)/2}(\hat{U}_n - \theta) = n^{r/2}(U_n - \theta) + o_p(1) \xrightarrow{D} C_m J_r(\hat{h}_r).$$

When $r_1 + r_o < r$ or $r_1 < r_o$,

$$n^{-(r_1 + r_o)/2}(\hat{U}_n - \theta) = n^{r/2}(U_n - \theta) + o_p(1) \xrightarrow{D} C_m J_r(\hat{h}_r).$$

When $r_1 + r_o = r$,

$$n^{-(r_1 + r_o)/2}(\hat{U}_n - \theta) = n^{r/2}(U_n - \theta) - n^{r/2} U_1, \Omega^{-1} n^{r/2} U_{0,n} + o_p(1) \xrightarrow{D} C_m C_m^r J_r(\hat{h}_r) \Omega^{-1} J_r(\bar{g}_r).$$

Proof of Corollary 1. In this case we have $h_1(X_1) = a[X_1 + (m - 1)\mu]$, $\bar{h}_1(X_1) = a(X_1 - \nu)$, $\eta_1^2 = a^2 \tau^2$. Also, $g_1(X_1) = b(X_1 - \mu) = \bar{g}_1(X_1)$, $A_1 = ab\tau^2$,

$$A = abE[\sum_{k=1}^{m} X_k - \mu][\sum_{k=1}^{m} X_k - \mu] = ab\tau^2, \quad \Omega = b^2 E[\sum_{k=1}^{m} X_k - \mu][\sum_{k=1}^{m} X_k - \mu] = mb^2 \tau^2$$

and $r_o = 1$. So by Theorem 2(i) we have $\sigma^2 = n^2(a^2 \tau^2 - 2d \tau^2 + \tau^2) = 0$.

Proof of Theorem 3. (i) Note $\Theta = E_{F_{\theta}} h(X)$. The information bound is for parameter of the form $E_{F_T}(s(X_1))$ for some $s(\cdot)$. Recall $\bar{h}_1(x_1) = E(h(x_1), ..., X_m|x_1) = E_F(h_1(X_1)) = \Theta$, thus we take $s(\cdot) = h_1(\cdot)$. Similarly, the constraint for computing the information bound should be a uni-variate function, we take it to be $g_1(x_1)$.

Let $f(x)$ be the density/mass function of $F(x)$ with respect to some dominating measure $\mu(x)$, denote $\gamma(\cdot) = \int h_1(x)f(x) d\mu(x) = \Theta$ as a functional of $f$, $\gamma(\cdot) = \gamma(\cdot)$ be the adjoint (evaluated at 1) of its pathwise derivative with respect to log $f$ (for definition, see, for example, [6]), $\gamma(\cdot) = E_F[g_1(X)]$ for the side information constraint, $\gamma(\cdot) = \gamma(\cdot)$ the adjoint (evaluated at 1) of its pathwise derivative, $L_{2,d,r}(f) = \{s(x) : s : R^d \to R^r, E_F[s(X) s'(X)] < \infty\}$, for $s_1 \in L_{2,d,r}(f)$ and $s_2 \in L_{2,d,r}(f)$, define the inner product (matrix) $\langle s_1, s_2 \rangle = E_F[s_1(X) s_2'(X)] = \int s_1(x)s_2'(x)f(x)d\mu(x)$, the norm (matrix) $\|s_1\|^2 = \langle s_1, s_1 \rangle$ and $\|s_1\|^{-2} = (\|s_1\|^2)^{-1}$ when $\|s_1\|^2$ is non-degenerate.
By Proposition A.5.2 in [6], we have \(\dot{\gamma} = h_1(X) - \theta = \tilde{h}_1(X)\) and \(\dot{\gamma}_1 = g_1(X) = \tilde{g}_1(X)\).
Let \(\Pi[u \cup v]\) be the projection of \(u\) onto \([v]\), the linear span of \(v\) with respect to \(f\) and \(\mu\), and \(\nu^f\) the orthogonal complement of \([v]\) with respect to \(f\) and \(\mu\). Without side information, the efficient influence function \(\mathcal{I}(X, \gamma(f))\) for estimating \(\gamma(f)\) is \(\mathcal{I}(X, \gamma(f)) = \dot{\gamma}(f)\) and the information bound is \(\|\mathcal{I}(X, \gamma(f))\|^2\). In the presence of side information \(\gamma_1(f)\), by Example 3.2.3 in [6], the efficient influence function \(\mathcal{I}(X, \gamma(f)|\gamma_1(f))\) for estimating \(\gamma(f)\) is \(\dot{\gamma}(f) = h_1(X) - A_1\Omega_1^{-1}\tilde{g}_1(X)\)

and the information (lower) bound for estimating \(\theta\), with side information \(g\), is

\[\mathbb{I}(\theta|g) = \mathbb{I}(X, \gamma(f)|\gamma_1(f)) = \mathbb{I}(X, \gamma(f)|\gamma_1(f)) = \mathbb{I}(X, \gamma(f)|\gamma_1(f)) = \mathbb{I}(\gamma_1(f), \gamma_1(f)).\]

Since \(m^2\) is a known positive constant, we can just divide \(\bar{U}_n\) by \(m\) so that its asymptotic variance is \(\mathbb{I}(\theta|g)\), and thus it is efficient.

Since \(\sigma^2 = \|\Pi[\dot{\gamma}(f)|\gamma_1(f)]\|^2 \geq 0\), with “=” if \(\dot{\gamma}(f) = \tilde{h}_1(X) \in [\gamma_1(f)]\), the linear span of \(\gamma_1(f) = \tilde{g}_1(X)\), or \(\theta\) is completely determined by \(\tilde{g}_1(X)\), which is impossible. Also

\[\|\mathcal{I}(X, \gamma(f)|\gamma_1(f))\|^2 = \mathbb{I}(X, \gamma(f)|\gamma_1(f)) = \mathbb{I}(X, \gamma(f)|\gamma_1(f)) = \mathbb{I}(\gamma_1(f), \gamma_1(f)) = A.\]

(ii) Let \(f(x|\theta, g)\) be the density function given the parameter \(\theta\) and the information constraint \(g\), \(S(x|\theta, g) = \partial \log f(X|\theta, g)/\partial \theta\) be the corresponding score function. The corresponding Fisher information is \(I(\theta|g) = \mathbb{I}(X, \gamma(f)|\gamma_1(f))\). Although \(S(x|\theta, g)\), hence \(I(\theta|g)\), is not directly available, the corresponding efficient influence function \(\mathcal{I}(X, \gamma(f)|\gamma_1(f))\) is given in (i), and we have the following relationship between the information bound \(\mathbb{I}(\theta|g)\) and the Fisher information \(I(\theta|g)\)

\[\mathbb{I}(\theta|g) = \mathbb{I}(X, \gamma(f)|\gamma_1(f)) = \mathbb{I}(X, \gamma(f)|\gamma_1(f)) = \mathbb{I}(X, \gamma(f)|\gamma_1(f)) = \mathbb{I}(\gamma_1(f), \gamma_1(f)) = A.\]

Let \(L(X^n|\theta, g) = \sum_{i=1}^n \log f(X_i|\theta, g)\) be the log-likelihood, we have the following locally asymptotic normality [22] of the likelihood ratio

\[\lambda_n := L(X^n|\theta_n) - L(X^n|\theta) = bV_n - b^2 I(\theta|g)/2 + o_p(1),\]

where \(V_n = n^{-1/2} \sum_{i=1}^n S(X_i|\theta, g) \overset{D}{\to} V \sim N(0, I(\theta|g)).\)

Let \(\Phi_Y(t) = E[\exp\{itY\}]\) be the characteristic function of a random variable \(Y\). We are to show \(L_n \overset{D}{\to} \chi^2(\theta)\). In fact, by assumption of regularity,

\[\phi_m(t) = E_{f(\cdot|\theta)}[\exp\{itW_n\}] = E_{f(\cdot|\theta)}[\exp\{it(W_n - \mu)\}] = E_{f(\cdot|\theta)}[\exp\{it(W_n - \mu) + \lambda_n\}] \to E[\exp\{it(W - \mu) + bV - b^2 I(\theta|g)/2\}],\]
where the last step above is by the same argument as in [4]. Since \( b \in C \) is arbitrary, take \( b = -it^{-1}(\theta | g) \), we get

\[
iti(W-b)+bV-b^2I(\theta | g)/2=iti(W-I^{-1}(\theta | g)V)-I^{-1}(\theta | g)^2/2=iti(W-I(\theta | g)V)-I(\theta | g)^2/2,
\]

thus

\[
\lim_n \phi_w(t)=E[ \exp(iti(W-I(\theta | g)V))\exp(-I(\theta | g)^2/2)]=\phi_{w-\log t}(t)\phi_t(t).
\]

Now take \( U = W - \Theta | g \) \( V \), the proof is complete.

**Proof of Theorem 4.** (i) Denote the related \( U \)-statistics as functions of \( h \), and note the \( O(\cdot) \) terms in the lemma are independent of \( h \). Note \( U_{1,n} \) is a functional of \( gh \), \( U_{2,n} \) is a functional of \( \Theta \) and \( U_n \) are functionals of \( h \), and \( U_{0,n} \) is a functional of \( g \). As in the proof of Theorem 1(ii), we have

\[
\text{sup}_{h \in \mathcal{H}} |U_{1,n}(gh)| \leq \text{sup}_{h \in \mathcal{H}} |U_{1,n}(h)| + \text{sup}_{h \in \mathcal{H}} \left| \frac{1}{d} g + \| g \|^2 h \right| < \infty, \text{ (a.s.)}
\]

Since \( U_{0,n}(g) \to 0 \) (a.s.) and is independent of \( h \), we only need to show, a.s.,

\[
\text{sup}_{h \in \mathcal{H}} |U_n(h) - \Theta(h)| \to 0; \text{ sup}_{h \in \mathcal{H}} |U_1,\Theta(g)| < \infty, \text{ and sup}_{h \in \mathcal{H}} |U_{2,\Theta}(1_d g + \| g \|^2 h)| < \infty.
\]

In fact, since \( gh \in \mathcal{H} \) for all \( h \in \mathcal{H} \), and \( U_{1,\Theta}(h) = U_{1}(h) \), we have \( \sup_{\mathcal{H}} |U_{1,\Theta}(gh)| \leq \sup_{\mathcal{H}} |U_{1,\Theta}(gh) - P_{\| g \|^2 h} + P_{\| g \|^2 h} - P_{\| g \|^2 h} + P_{\| g \|^2 h} - \Theta(h)| \to 0 \) (a.s.).

Similarly, since \( \| g \|^2 h \in \mathcal{H} \) for all \( h \in \mathcal{H} \), and \( U_{2,\Theta}(h) = U_{2}(h) \), we have \( \sup_{\mathcal{H}} |U_{2,\Theta}(1_d g + \| g \|^2 h)| < \infty \) (a.s.), if \( \sup_{\mathcal{H}} |U_{2}(h) - \Theta(h)| \to 0 \) (a.s.).

Now we only need to prove \( \sup_{\mathcal{H}} |U_{1}(h) - \Theta(h)| \to 0 \) (a.s.), (the class \( \mathcal{H} \) is then called P-Glivenko–Cantelli). Since the property of P-Glivenko–Cantelli is permanent for finite union of classes, we only need to prove this on \( \mathcal{H}_1 \) and \( \mathcal{H}_2 \) separately.

We first prove \( \mathcal{H}_1 \) is P-Glivenko–Cantelli. For \( e > 0 \), let \( \mathcal{N} \left( e, \mathcal{H}_1, L_1(P_m) \right) \) be the bracketing entropy of the class \( \mathcal{H}_1 \) with \( L_1(P) \) norm: \( \forall h \in \mathcal{H}_1, \| h \|_{P_m} = E_{P_m}[h] \). We first prove that if \( \mathcal{N} \left( e, \mathcal{H}_1, L_1(P_m) \right) < \infty \) for all \( e > 0 \), then the conclusion is true. In fact, given \( e > 0 \), since \( \mathcal{N} \left( e, \mathcal{H}_1, L_1(P_m) \right) < \infty \), there are finite many \( e \)-brackets \( [i, u_i] \) whose union covers \( \mathcal{H}_1 \) and such that \( P_m[u_i - l] < e \) for all \( i \). Then for any \( h \in \mathcal{H}_1 \), there is an upper bracket \( u_i \) such that
Consequently,

$$\sup_{h \in \mathcal{H}} (U_n(h) - P_m(h)) \leq \max_{i} (P_{n,m} - P_m) u_i + \epsilon.$$ 

Since by SLLN for U-statistics, \((P_{n,m} - P_m) u_i \to 0 \text{ (a.s.)})\), thus \(\lim_{n} \sup_{h \in \mathcal{H}} (U_n(h) - P_m(h)) \leq \epsilon \) (a.s.). Similarly, \(\lim_{n} \inf_{h \in \mathcal{H}} (U_n(h) - P_m(h)) \geq -\epsilon \) (a.s.). Since \(\epsilon > 0\) is arbitrary, we get \(\sup_{h \in \mathcal{H}} |U_n(h) - P_m(h)| \to 0 \) (a.s.).

Now we show \(N(\epsilon, \mathcal{H}, L_1(P_m)) < \infty\) for all \(\epsilon > 0\). Let \(I_j = \{x \in \mathbb{R}^m : \|x - I_j\| < 1\}\). By Corollary 2.7.4 in VW, for some constant \(K\) depending only on \(\nu\) and \(m\), for every \(\epsilon > 0\), \(\nu \geq 1\) and probability \(P\). The given condition implies that \(\max_{j} A(I_j) < \epsilon\). Take \(\nu = 1\), the right hand side above is finite by the given condition.

Now we show \(\sup_{h \in \mathcal{H}} |U_n(h) - \theta(h)| \to 0 \) (a.s.). Let, for \(\epsilon > 0\), \(N(\epsilon, \mathcal{H}, L_1(Q))\) be the entropy of \(\mathcal{H}\) without bracketing under norm \(L_1(Q)\) for some probability measure \(Q\) and \(N(\epsilon, \mathcal{H}, \|\cdot\|_{\infty})\) be that under norm \(\|\cdot\|_{\infty}\). Recall \(H\) is also an envelope function on \(\mathcal{H}\). Since \(L_1(Q) \leq \|\cdot\|_{\infty}\), \(N(\epsilon, \mathcal{H}, \|\cdot\|_{\infty}) \leq N(\epsilon \|H\|_{Q}, \mathcal{H}, \|\cdot\|_{\infty})\). Let \(M\) be the bound on \(\mathcal{H}\), and \(\mathcal{H} = \{(h \inf_{x \in C} h(x))/(M : h \in \mathcal{H}_2)\}\), then \(\mathcal{H}_2\) is the class of convex functions \(h : C \mapsto [0, 1]\) with Lipschitz constant \(L/M\), and \(N(\epsilon, \mathcal{H}_2, \|\cdot\|_{\infty}) = N(\epsilon M, \mathcal{H}_2, \|\cdot\|_{\infty})\). By Corollary 2.7.10 in VW, for any \(\epsilon > 0\),

$$\log N(\epsilon, \mathcal{H}_2, \|\cdot\|_{\infty}) \leq K(1 + L/M)^{\frac{m}{2}} M^{-\frac{m}{2}} e^{-m/2},$$

for any probability measure \(Q\) and \(K\) only depends on \(m\) and \(C\). Also, since \(H\) is an envelope function over \(\mathcal{H}_2 \neq \{0\}\), thus \(\inf_{Q} \|H_Q\| \geq \delta\) for some \(\delta > 0\) and the infimum is over \(\subseteq 0\) of all probability measures \(Q\) on \(C\), with \(\|H\|_{Q} < \infty\). Thus we have, for any \(\epsilon > 0\),

$$\log N(\epsilon \delta, \mathcal{H}_2, \|\cdot\|_{\infty}) \leq \log N(\epsilon \delta M, \mathcal{H}_2, \|\cdot\|_{\infty}) \leq K(1 + L/M)^{\frac{m}{2}} M^{-\frac{m}{2}} (\epsilon \delta)^{-m/2} < \infty,$$

also, \(\mathcal{H}_2\) is \(P\)-measurable by its definition, thus \(\mathcal{H}_2\) is \(P\)-Glivenko–Cantelli (cf. the statement in lines 5 to 3, p. 84 of VW).

(ii) The class \(\mathcal{H}\) with the stated property is called \(P\)-Donsker. First, it is apparent that for any \(k\) and \(h_1, \ldots, h_k \in \mathcal{H}\), \((\mathcal{G}_{n,m} h_1, \ldots, \mathcal{G}_{n,m} h_k) \rightarrow (\mathcal{H}_{h_1}, \ldots, \mathcal{H}_{h_k})\) for the Gaussian process \(\mathcal{G}_{n,m}\) as stated. So by Theorem 1.5.4 in VW, we only need to show that \((\mathcal{G}_{n,m})\) is asymptotically tight on \(\mathcal{H}\). Using Lemma (ii) and similar argument as in the proof of (i), since \(\sqrt{n} \rightarrow o(1)\) and \(\sqrt{n} - 1/2 \rightarrow o(1)\), we only need to show this for \(\{\mathcal{G}_{n,m}\}\), and
by Theorem 1.5.7 in VW, we only need to show that \( \{G_{n,m}\} \) is asymptotically equicontinuous and totally bounded on \( \mathcal{H} \). Below we will show that if
\[
\int_0^\infty \sup_{Q \in \mathcal{D}} \sqrt{\log N(\epsilon; \mathcal{H}, L_2(Q))} \, d\epsilon < \infty \quad (A.2)
\]
then \( \{G_{n,m}\} \) is asymptotically equicontinuous and totally bounded on \( \mathcal{H} \), where \( \mathcal{D} \) is the collection of all measures \( Q \) with \( \|H\|_Q < \infty \).

With (A.2), Theorem 2.5.2 in VW asserted the corresponding conclusion for empirical measures. Now we extend the result to \( U \)-statistics. For this, we point out that the symmetrization Lemma 2.3.1 in VW still holds for \( U \)-statistics, also Hoeffding’s inequality holds for \( U \)-statistics (Arcones and Giné [2, Proposition 2.3, p. 1501]), thus the proofs there are valid in our situation.

To check (A.2) on \( \mathcal{H} \), and we only need to check it for \( \mathcal{H}_1 \) and \( \mathcal{H}_2 \) separately. Using Corollary 2.7.4 in VW, we have
\[
\log N_{1}(\epsilon; \mathcal{H}_1, L_2(P)) \leq K e^{-\nu \left( \sum_{j=1}^{\infty} \lambda(I_j) \frac{1}{2\nu^2} M_j^{2\nu} P^{\frac{1}{2\nu}}(I_j) \right) \frac{1}{2\nu^2}},
\]
for all \( \epsilon > 0, \nu \geq m/\delta \). Since the given condition implies \( \max_j \lambda(I_j) < \infty \) in the above inequality choose \( \nu = m/\delta \), then \( \sum_{j=1}^{\infty} \lambda(I_j) \frac{1}{2\nu^2} M_j^{2\nu} P^{\frac{1}{2\nu}}(I_j) < \infty \) by the given condition, and since \( \nu < 2 \), we have
\[
\int_0^1 \sqrt{\log N_{1}(\epsilon; \mathcal{H}_1, L_2(P))} \, d\epsilon < \infty,
\]
hence by the statement in p. 85 in VW, \( \mathcal{H}_1 \) satisfies (A.2). The original statement in VW is for the integral \( \int_0^\infty \). Since \( \mathcal{H}_1 \) has a square integrable envelope function \( H \), so \( \forall h_1, h_2 \in \mathcal{H}_1, \|h_1 - h_2\|_{L_2(P)} \leq \max \|h\|_{L_2(P)} < 2\|H\|_{L_2(P)} < \infty \), i.e., \( \mathcal{H}_1 \) itself is a ball with radius no greater than \( 2\|H\|_{L_2(P)} \), or \( N_{1}(\epsilon; \mathcal{H}_1, L_2(P)) = 1 \) for \( \epsilon \geq 2\|H\|_{L_2(P)} \); thus its entropy is zero for \( \epsilon \geq 2\|H\|_{L_2(P)} \), so the integration \( \int_0^\infty \) is finite iff \( \int_0^1 \) is finite.

For \( \mathcal{H}_2 \), similarly as in the proof of (i), for some \( \eta > 0 \),
\[
\sup_{Q \in \mathcal{D}} \log N(\epsilon; \mathcal{H}_2, L_2(Q)) \leq K(1+L/M)^{m/2} M^{-m/2}(\epsilon \eta)^{-m/2}.
\]
Since \( m < 4 \), so
\[
\int_0^1 \sup_{Q \in \mathcal{D}} \sqrt{\log N(\epsilon; \mathcal{H}_2, L_2(Q))} \, d\epsilon \leq K^{1/2}(1+L/M)^{m/4} M^{-m/4} \eta^{-m/4} \int_0^1 \epsilon^{-m/4} \, d\epsilon < \infty,
\]
thus by (2.1.7) in VW, \( \mathcal{H}_2 \) is \( \mathcal{P} \)-Donsker.
Proof of Corollary 2. From proof of Theorem 4(i), we only need to show \( \sup_{h \in \mathcal{H}} |P_{n,m} - P_{n,m}^h| \to 0 \) (a.s.), which is true by Corollary 3.3, or 3.5 respectively in [2].

Proof of Theorem 5. (i) As in the proofs of the previous theorems, with \( g \) replaced by \( G \), since \( r_o = \min \{ \text{rank}(g_1), \ldots, \text{rank}(g_d), \text{rank}(h) \} = 1 \), by Lemma (ii) we have

\[
\begin{align*}
\omega_1 &= \frac{1}{C_n^{-1}} \left( 1 - G'(X_i|\theta) \Lambda^{-1} \frac{1}{C_n} \sum_{j \in D_{n,m}} G(X_j|\theta) + \left[ \frac{1}{m+1} G(X_i|\theta) + \|G(X_i|\theta)\|^2 \right] O_p(n^{-1}) \right),
\end{align*}
\]

and by standard \( U \)-statistics theory,

\[
\sqrt{n} \frac{1}{C_n} \sum_{j \in D_{n,m}} G(X_j|\theta) \xrightarrow{D} N(0, m^2 \Lambda_1), \frac{1}{C_n} \sum_{j \in D_{n,m}} G(X_j|\theta) G'(X_i|\theta) = \Lambda + O_p(n^{-1/2}).
\]

Also,

\[
t := \Lambda^{-1} \frac{1}{C_n} \sum_{j \in D_{n,m}} G(X_j|\theta) + O_p(n^{-1/2}), \frac{1}{C_n} \sum_{j \in D_{n,m}} \|G(X_i|\theta)\|^2 \xrightarrow{a.s.} E\|G(X|\theta)\|^2 < \infty,
\]

and \( \max \left| t' G(X_i|\theta) \right| = O_p(n^{-1/2} n^{m/2}) \to 0 \) since \( m^2 < r_o^2 / 2 = 1/2 \), so
This completes the proof since

\[ \sqrt{nm}^{-1} \Lambda_1^{-1/2} \frac{1}{C_n} \sum_{i \in D_{n,m}} G(X_i | \theta) \mathcal{D} \sim N(0, I_{r+1}). \]

(ii) This is a special case of (i).
References


Table 1

The asymptotic variance estimation of $U$-statistics. $X \sim \exp(1) - \ln(2)$.

<table>
<thead>
<tr>
<th>Method</th>
<th>$n = 50$</th>
<th>$n = 100$</th>
<th>$n = 150$</th>
<th>$n = 200$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without side information</td>
<td>8.5239</td>
<td>7.8569</td>
<td>7.3839</td>
<td>7.1557</td>
</tr>
<tr>
<td>With side information</td>
<td>8.4572</td>
<td>7.5524</td>
<td>7.2673</td>
<td>7.0791</td>
</tr>
<tr>
<td>Variance reduction</td>
<td>0.0667</td>
<td>0.3045</td>
<td>0.1165</td>
<td>0.0766</td>
</tr>
</tbody>
</table>

$J$ Multivar Anal. Author manuscript; available in PMC 2013 May 21.
### Table 2

The asymptotic variance estimation of $U$-statistics. $X \sim \mathcal{N}(1, 4)$.

<table>
<thead>
<tr>
<th>Method</th>
<th>$n = 50$</th>
<th>$n = 100$</th>
<th>$n = 150$</th>
<th>$n = 200$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without side information</td>
<td>0.2413</td>
<td>0.2208</td>
<td>0.2199</td>
<td>0.2203</td>
</tr>
<tr>
<td>With side information</td>
<td>0.0548</td>
<td>0.0526</td>
<td>0.0527</td>
<td>0.0572</td>
</tr>
<tr>
<td>Variance reduction</td>
<td>0.1865</td>
<td>0.1682</td>
<td>0.1673</td>
<td>0.1631</td>
</tr>
</tbody>
</table>
Table 3

Coverage probabilities of various 95% confidence intervals for $\theta$ with side information. $X \sim \exp(1) - \ln(2)$.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>EL1</th>
<th>EL2</th>
<th>EL3</th>
<th>AN1</th>
<th>AN2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 50$</td>
<td>0.942</td>
<td>0.949</td>
<td>0.994</td>
<td>0.783</td>
<td>0.784</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>0.929</td>
<td>0.950</td>
<td>0.991</td>
<td>0.858</td>
<td>0.878</td>
</tr>
<tr>
<td>$n = 150$</td>
<td>0.934</td>
<td>0.954</td>
<td>0.990</td>
<td>0.872</td>
<td>0.880</td>
</tr>
<tr>
<td>$n = 200$</td>
<td>0.950</td>
<td>0.949</td>
<td>0.989</td>
<td>0.898</td>
<td>0.904</td>
</tr>
</tbody>
</table>
Table 4

Coverage probabilities of various 95% confidence intervals for $\theta$ with side information. $X \sim \mathcal{N}(1, 4)$.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>EL1</th>
<th>EL2</th>
<th>EL3</th>
<th>AN1</th>
<th>AN2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 50$</td>
<td>0.876</td>
<td>0.983</td>
<td>0.918</td>
<td>0.956</td>
<td>0.945</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>0.882</td>
<td>0.981</td>
<td>0.931</td>
<td>0.961</td>
<td>0.947</td>
</tr>
<tr>
<td>$n = 150$</td>
<td>0.926</td>
<td>0.978</td>
<td>0.968</td>
<td>0.978</td>
<td>0.968</td>
</tr>
<tr>
<td>$n = 200$</td>
<td>0.942</td>
<td>0.986</td>
<td>0.984</td>
<td>0.970</td>
<td>0.954</td>
</tr>
</tbody>
</table>