Notes on Imperfect Matching Paper
Rachel Anderson
August 13, 2019

1 Setup

The goal is to estimate θ_0 that satisfies the model

$$E_0[m(y_i, x_i, z_i; \theta_0)] \tag{1}$$

where the expectation is taken with respect to the joint distribution $f_0(y, x, z)$. Instead of observing a random sample (y_i, x_i, z_i) drawn from $f_0(y, x, z)$, the econometrician observes two random samples drawn from $f_0(y)$ and $f_0(x, z)$ independently.

The first sample contains $(x_i, z_i, w_i)_{i=1}^n$, where (x_i, z_i) are drawn from the marignal distribution $f_0(x, z)$. The second contains outcomes $(\{y_{i\ell}\}_{\ell=1}^{L_i}, w_i)_{i=1}^n$, that are linked to (x_i, z_i) by the identifier w_i, which also appears in the first dataset. The econometrician believes that only one of the $\{y_{i\ell}\}_{\ell=1}^{L_i}$ is drawn from the distribution $f(y_i^*|x_i^*, z_i^*)$; the other outcomes are spurious matches that arise because the identifying information w_i is insufficent to form unique data tuples.

To fix ideas, suppose the econometrician would like to estimate the impact of a single mother assistance program on participants' children's longevity (call this unknown parameter θ). The researcher collects two datasets; the first consists of aid program applications, with variables about mother and child characteristics (x_i^*, z_i^*) and the child's first and last names, and place and year of birth (w_i). The second data set contains death records (used to construct y_i^*), and is indexed by the individual's first and last names, and place and year of birth (w_i).

Although individuals with distinct names are unlikely to appear on multiple death records, individuals with common names like "John Smith" may be linked to multiple death records. Similarly, if the econometrician matches individuals using a subset of the variables in w_i (as in the case that females are matched by first name only, and place and year of birth), she is likely to find many possible links that are equally credible.

The following note formalizes the assumptions that are necessary to estimate the model (1) without dropping observations with non-unique matches.

Assumption 1. The observed (x_i, z_i, w_i) is a random sample drawn from the marginal distribution $f(x^*, z^*, w^*)$. The $\{y_{i\ell}\}_{\ell=1}^{L_i}$ is a random sample drawn from $f(y^*|w_i, L_i)$, so that $(x_i, z_i) = (x_i^*, z_i^*)$ and $y_{i\ell}$ are independent conditional on w_i for $y_{i\ell} \neq y_i^*$.

Assumption 2. There is exactly one $y_{i\ell} = y_i^*$, $\ell \in \{1, \ldots, L_i\}$ for all i. That is, we assume that one of the $\{y_{i\ell}\}_{\ell=1}^{L_i}$ is drawn from the marginal distribution $f(y_i^*|x_i^*, z_i^*) = f(y_i^*|x_i, z_i)$.

This assumption implies we can write $y_i^* = \sum_{\ell=1}^{L_i} s_{i\ell} y_{i\ell}$, where $s_{i\ell}$ is an unobserved latent variable that equals 1 if $(y_{i\ell}, x_i, z_i) = (y_i^*, x_i^*, z_i^*)$, and that equals 0 otherwise. Also, $\sum_{\ell=1}^{L_i} s_{i\ell} = 1$ for all i, and we can rewrite 1 as,

$$E\left[m(y_i^*, x_i^*, z_i^*)\right] = 0 \iff E\left[m(y_{i\ell}, x_i, z_i; \theta_0)|s_{i\ell} = 1\right] = 0 \qquad (1^*)$$

Assumption 3. The identifying variables w_i in (x_i, z_i, w_i) exactly match, or are sufficiently close to w_i in $(\{y_{i\ell}\}_{\ell=1}^{L_i}, w_i)$, such that the researcher cannot distinguish which $y_{i\ell} = y_i^*$ for $\ell \in \{1, \ldots, L_i\}$. In other words, the researcher behaves as if,

$$P(s_{i\ell} = 1|w_i, L_i) = \frac{1}{L_i}$$

Assumptions 1 and 3 rule out unobserved sample selection, in the sense that all individuals

with the same identifying information have equal probability of appearing in the sample. This assumption would be violated if, for example, higher income individuals have a greater probability of appearing in the sample, unless w_i includes income.

2 Estimating θ

The goal is to estimate (1^*) without observing the $s_{i\ell}$. One thing we can do is evaluate the moment conditions $m(\cdot)$ for each $y_{i\ell}$ in the sample. Hence, we consider (ignoring z_i for simplicity):

$$E\left[\sum_{\ell=1}^{L_i} m\left(y_{i\ell}, x_i; \theta\right) \middle| w_i, L_i, \right] = \sum_{\ell=1}^{L_i} \Big\{ E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 1] P(s_{i\ell} = 1|w_i, L_i) + E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 0] P(s_{i\ell} = 0|w_i, L_i) \Big\}$$

By Assumption 1, $P(s_{i\ell} = 1|w_i, L_i) = \frac{1}{L_i}$, so that

$$E\left[\sum_{\ell=1}^{L_i} m\left(y_{i\ell}, x_i; \theta\right) \middle| w_i, L_i, \right] = \frac{1}{L_i} \sum_{\ell=1}^{L_i} E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 1] + \frac{L_i - 1}{L_i} E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 0]$$

and random sampling implies the expectations are equal for all i,

$$E\left[\sum_{\ell=1}^{L_i} m\left(y_{i\ell}, x_i; \theta\right) \middle| w_i, L_i, \right] = E\left[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 1\right] + (L_i - 1) E\left[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 0\right]$$

Rearranging terms,

$$E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 1] = E\left[\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i; \theta)\middle| w_i, L_i\right] - (L_i - 1)E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 0]$$

Note that $E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 0]$ is an expectation with respect to the distribution of false matches. In this case, $y_{i\ell} \neq y_i^*$ is independent of $x_i^* = x_i$ conditional on w_i, so that

$$E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 0] = E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, x_i, s_{i\ell} = 0] \equiv g(w_i, L_i, x_i; \theta)$$

Finally, by the Law of Iterated Expectations,

$$E[m(y_{i\ell}, x_i; \theta)|s_{i\ell} = 1] = E\left[\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i; \theta)\right] - E[(L_i - 1)g(w_i, L_i, x_i, \theta)] \tag{2}$$

which we can approximate with the sample analog:

$$\hat{E}\left[m(y_{i\ell}, x_i; \theta)|s_{i\ell} = 1\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i; \theta) - \frac{1}{n}\sum_{i=1}^{n}(L_i - 1)\hat{g}(w_i, L_i, x_i; \theta) \tag{3}$$

where n is the number of observations in the (x_i, z_i, w_i) file, and $\hat{g}$ is a parametric or non-parametric estimate of $g(\cdot)$. Estimation proceeds as usual by applying GMM to 3 as the sample moments.

3 Large Sample Results

Define the estimator

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \; m_n(\theta, \hat{g})' \hat{W} m_n(\theta, \hat{g}) \tag{4}$$

$$m_n(\theta, \hat{g}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i, z_i; \theta) - (L_i - 1)\hat{g}(x_i, z_i, w_i, L_i; \theta) \tag{5}$$

where $\hat{g}(\cdot)$ is a nonparametric estimator of $g(\cdot)$. Let also $m(\theta) \equiv m(y_i, x_i, z_i; \theta)$.

Theorem 1 (Consistency). If the following assumptions are true,

1. $\left(x_i, z_i, w_i, \{y_{i\ell}\}_{\ell=1}^{L_i}\right)_{i=1}^{n}$ is a random sample, and $\{y_{i\ell}\}_{\ell=1}^{L_i}$ and (x_i, z_i) are i.i.d. samples conditional on w_i

2. $\hat{W} - W = o_p(1)$, W is positive semi-definite, $WE[m(y_i, x_i, z_i; \theta)] = 0$ has a unique solution on Θ at θ_0, with Θ a compact subset of $\mathcal{R}^d$.

3. (i) $m(y_{i\ell}, x_i, z_i, \theta)$ is continuous at each $\theta \in \Theta$ with probability 1;
   (ii) $E\left[\sup_{\theta\in\Theta} || \sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i, z_i, \theta)||\right] < \infty$;
   (iii) $\sup_\theta \|\hat{g}(x_i, z_i, w_i, L_i; \theta) - g(x_i, z_i, w_i, L_i; \theta)\| \xrightarrow{p} 0$.

then $\hat{\theta} \xrightarrow{p} \theta_0$.

Proof. Assumption 2 is the ID condition for GMM. To apply Theorem 2.1 in Newey & McFadden (1994), we verify uniform convergence of the sample objective function.

Define $\hat{Q}_n(\theta) = m_n(\theta, \hat{g})'\hat{W}m_n(\theta, \hat{g})$, and $Q_0(\theta) = E[m(\theta)]'WE[m(\theta)]$. Assumptions 1-3 and (7) imply

$$
\begin{aligned}
\sup_\theta \|m_n(\theta, \hat{g}) - E[m(\theta)]\| &\leq \sup_\theta \left\| \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i, z_i, \theta) - E\left[\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i, z_i; \theta)\right] \right\| \\
&\quad + \sup_\theta \left\| \frac{1}{n}\sum_{i=1}^{n}(L_i - 1)\hat{g}(x_i, z_i, w_i, L_i; \theta) - E\left[(L_i - 1)g(x_i, z_i, w_i, L_i; \theta)\right] \right\| \\
&\leq o_p(1) + \sup_\theta \left\| \frac{1}{n}\sum_{i=1}^{n}(L_i - 1)\left(\hat{g}(x_i, z_i, w_i, L_i; \theta) - g(x_i, z_i, w_i, L_i; \theta)\right) \right\| \\
&\quad + \sup_\theta \left\| \frac{1}{n}\sum_{i=1}^{n}(L_i - 1)g(x_i, z_i, w_i, L_i; \theta) - E[(L_i - 1)g(x_i, z_i, w_i, L_i; \theta) \right\| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}(L_i - 1)\sup_\theta \|(\hat{g}(x_i, z_i, w_i, L_i; \theta) - g(x_i, z_i, w_i, L_i; \theta))\| + o_p(1) = o_p(1)
\end{aligned}
$$

The remainder of the proof showing $\sup_\theta \|\hat{Q}_n(\theta) - Q(\theta)\|$ follows from the generic proof for GMM consistency. □

Note that the second part of Assumption 1 is necessary to derive the estimator. Assumption 2 is the standard identification assumption for GMM, and Assumption 3 assures uniform convergence of the objective function. If $\hat{g}$ is a sieve estimator, the conditions in Chen, Hong, and Tamer (2005) can be used to replace Assumption 3(iii).

Theorem 2 (Asymptotic Normality) If, in addition to Assumptions 1-3, the following are true,

4. $\theta_0 \in \text{int}(\Theta)$

5. $G(\theta) \equiv E\left[\nabla_\theta m(\theta)\right] = E\left[\frac{\partial}{\partial \theta'}\left\{\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i, z_i; \theta) - (L_i - 1)g(\cdot\,; \theta)\right\}\right]$ is continuous at θ_0 and $\sup_{\theta \in \mathcal{N}} \|\nabla_\theta m_n(\theta, \hat{g}) - G(\theta)\| \xrightarrow{p} 0$ where $\mathcal{N}$ is a neighborhood of θ_0

6. For $G = G(\theta_0)$, $G'WG$ is nonsingular

7. $E\left[m(y_i, x_i, z_i; \theta) m(y_i, x_i, z_i; \theta)\right]$ is finite, positive definite, which is also written as

$$E\left[\left\{\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i, z_i; \theta) - (L_i - 1)g(x_i, z_i, w_i, L_i; \theta)\right\}\left\{\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i, z_i; \theta) - (L_i - 1)g(x_i, z_i, w_i, L_i; \theta)\right\}'\right]$$

8. $\hat{g}(\cdot)$ satisfies regularity conditions (specifically, stochastic equicontinuity and mean-square continuity) so that $n^{1/2} m_n(\theta_0, \hat{g}) \Rightarrow \mathcal{N}(0, \Omega)$, with

$$\Omega = \text{Avar}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n}\left(\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i, z_i; \theta_0) - (L_i - 1)\hat{g}(x_i, z_i, w_i; \theta_0)\right)\right)$$

then $\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow \mathcal{N}(0, V)$, where

$$V = (G'WG)^{-1} G'W\Omega WG(G'WG)^{-1} \tag{6}$$

Proof. Mean-value expansion and first-order conditions imply:

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\theta_0) &= (\nabla_\theta m_n(\bar{\theta},\hat{g})'\hat{W}\nabla_\theta m_n(\hat{\theta},\hat{g}))^{-1}\nabla_\theta m_n(\hat{\theta},\hat{g})'\hat{W}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_n(\theta_0,\hat{g}) \\ &= (G'WG)^{-1}G'W\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_n(\theta_0,\hat{g}) + o_p(1)\end{aligned}$$

where

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_n(\theta_0,\hat{g}) = &\underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\sum_{\ell=1}^{L_i} m(y_{i\ell},x_i,z_i;\theta_0) - (L_i-1)g(x_i,z_i,w_i,L_i;\theta_0)\right)}_{\Rightarrow\mathcal{N}(0,E[m(\theta_0)m(\theta_0)'])} \\ &+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(L_i-1)\left(g(x_i,z_i,w_i,L_i;\theta_0) - \hat{g}(x_i,z_i,w_i,L_i;\theta_0)\right)\end{aligned}$$

To show that $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_n(\theta_0,\hat{g}) \Rightarrow \mathcal{N}(0,\Omega)$, we verify the conditions in Theorem 8.1 in Newey & McFadden (1994):

1. Linearization is immediate[1] because m_n is linear in $g(\cdot)$

2. Stochastic equicontinuity requires, for the true distribution of the data F_0 ,

$$\frac{1}{\sqrt{n}}\left[\sum_{i=1}^{n}(L_i-1)(g-\hat{g}) - \int(L_i-1)(g-\hat{g})dF_0\right] \xrightarrow{p} 0$$

3. (Mean-square differentiability) (i) there is $\delta(z)$ and a measure $\hat{F}$ s.t. $E[\delta(z)] = 0$, $E[\|\delta(z)\|^2] < \infty$, and for all $\|g-\hat{g}\|$ small enough, $\int(L_i-1)(g-\hat{g})dF_0 = \int\delta(z)d\hat{F}$

4. For the empirical distribution $\tilde{F}$, $\sqrt{n}\left[\int\delta(z)d\hat{F} - \int\delta(z)d\tilde{F}\right] \xrightarrow{p} 0$

Then $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_n(\theta_0,\hat{g}) \Rightarrow \mathcal{N}(0,\Omega)$, where $\Omega = \text{Var}[m(\theta_0)+\delta(z)]$. □

Assumptions 4-7 are the usual regularity and dominance conditions to ensure $\sqrt{n}$-

[1] with $G(z,g-g_0) = z_i(L_i-1)(g_0-\hat{g})$

normality of the GMM estimator of θ_0 when $g(\cdot)$ is known. Using instead a non-parametric estimator $\hat{g}$, such as a sieve estimator, requires additional regularity conditions, such as those outlined in the proof and in Chen, Hong and Tamer (2005).

Example: Linear Instrumental Variables

The model is

$$m(y_i, x_i, z_i; \theta) = z_i(y_i - x_i'\beta) \tag{7}$$

The $g(\cdot)$ nonparametric function becomes,

$$g(x_i, z_i, w_i, L_i; \theta) = z_i \left(E[y_{i\ell}|w_i, L_i] - x_i'\beta \right)$$

An estimate of $\hat{g}(\cdot)$ replaces $E[y_{i\ell}|w_i, L_i]$ with $\hat{E}[y_{i\ell}|w_i, L_i]$.

When the model is over-identified, i.e. $\dim(z_i) > \dim(x_i)$, the estimator is

$$\hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^{n} P z_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} P z_i \left(\sum_{\ell=1}^{L_i} y_{i\ell} - (L_i - 1)\hat{g}(w_i, L_i, \beta) \right)$$

where P is a weighting matrix, and choosing P is equivalent to choosing W in GMM. For identification, we replace Assumption 2 with

Assumption 2'. $PE[z_i x_i']$ exists and is nonsingular,

Additionally, we do not need Assumption 3(i) since $m(\cdot)$ is linear in β. Assumption 3(ii) will be satisfied if $E[z_i \sum_{\ell=1}^{L_i} y_{i\ell}] < \infty$.

Corollary (Consistency). If (i) $(x_i, z_i, w_i, \{y_{i\ell}\}_{\ell=1}^{L_i})_{i=1}^{n}$ is a random sample, and $\{y_{i\ell}\}_{\ell=1}^{L_i}$ and (x_i, z_i) are random samples conditional on w_i, (ii) $PE[z_i x_i']$ exists and is non-singular, (iii) $E[z_i \sum_{\ell=1}^{L_i} y_{i\ell}] < \infty$, and (iv) $\sup_\theta \|\hat{g}(w_i, L_i) - g(w_i, L_i)\| \xrightarrow{p} 0$, then $\hat{\beta} \xrightarrow{p} \beta_0$.

The statement is proved by verifying the assumptions of Theorem 1.

Corollary (Asymptotic Normality) If in addition to (i)-(iv) above, $\hat{g}(\cdot)$ satisfies regularity stochastic equicontinuity and mean-square differentiability, then $\sqrt{n}(\hat{\beta} - \beta_0) \Rightarrow \mathcal{N}(0, V)$, where V is defined as in (6) with $G = PE[z_i x_i']$ and

$$\Omega = \text{Avar}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i \left(\sum_{\ell=1}^{L_i} y_{i\ell} - (L_i - 1)\hat{g}(w_i, L_i, \beta)\right)\right)$$

As before, the asympotic variance formula depends on the choice of $\hat{g}(\cdot)$. In a later section, we derive exact formulas for a kernel estimator of $\hat{g}$.

4 Estimating Asymptotic Variance

As usual, a consistent estimator of (6) can be formed by plugging in estimators of the different pieces. An estimator of G can be formed in a straightforward way:

$$\hat{G} = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta m_n(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{\ell=1}^{L_i} \nabla_\theta m(y_{i\ell}, x_i, z_i; \hat{\theta}) - (L_i - 1)\nabla_\theta \hat{g}(x_i, z_i, w_i, L_i; \hat{\theta})\right)$$

which is consistent under the same conditions used for asymptotic normality of $\hat{\theta}$.

The more difficult term to estimate is the "score" variance Ω. One estimator, suggested in Newey & McFadden (1994), uses an estimator $\hat{\delta}(z)$ of the function $\delta(z)$,

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^{n} \left\{m(y_i, x_i, z_i; \hat{\theta}) + \hat{\delta}(z_i)\right\} \left\{m(y_i, x_i, z_i; \hat{\theta}) + \hat{\delta}(z_i)\right\}'$$

If, in addition to the assumptions of Theorem 1, $\frac{1}{n}\sum_{i=1}^{n} \left\|m_n(\hat{\theta}, \hat{g}) - m(\theta_0)\right\|^2 \xrightarrow{p} 0$, and $\frac{1}{n}\sum_{i=1}^{n} \left\|\hat{\delta}(z_i) - \delta(z_i)\right\|^2 \xrightarrow{p} 0$ then $\hat{\Omega} \xrightarrow{p} \Omega$ by Lemma 8.3 in Newey & McFadden (1994).

Hence, by Slutsky's and the continuous mapping theorem,

$$\hat{V} = (\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{\Omega}\hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1} \xrightarrow{p} (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1} = V$$

5 Bootstrap

The estimator proposed in this paper is a two-step semiparametric GMM estimator, where the first step involves estimating the control function g. As shown by Theorem 2, the resulting estimator is asymptotically linear and asymptotically normal, so it satisfies the regularity conditions in Mammen (1992) that guarantee validity of the nonparametric bootstrap. More specifically, the nonparametric bootstrap is consistent under the regularity conditions for a (potentially over-identified) standard GMM model specified in Hahn (1996, Theorem 1), when g is known.

The bootstrap is especially useful When $\hat{g}$ is estimated, as deriving Ω in the asymptotic variance formula may be difficult. In this case, the researcher should shrink the kernel bandwidth to reduce asymptotic bias (Horowitz, 2001).

Depending on the model, other bootstrap methods may be valid. For example, residual bootstrap (I think) would work for linear IV.

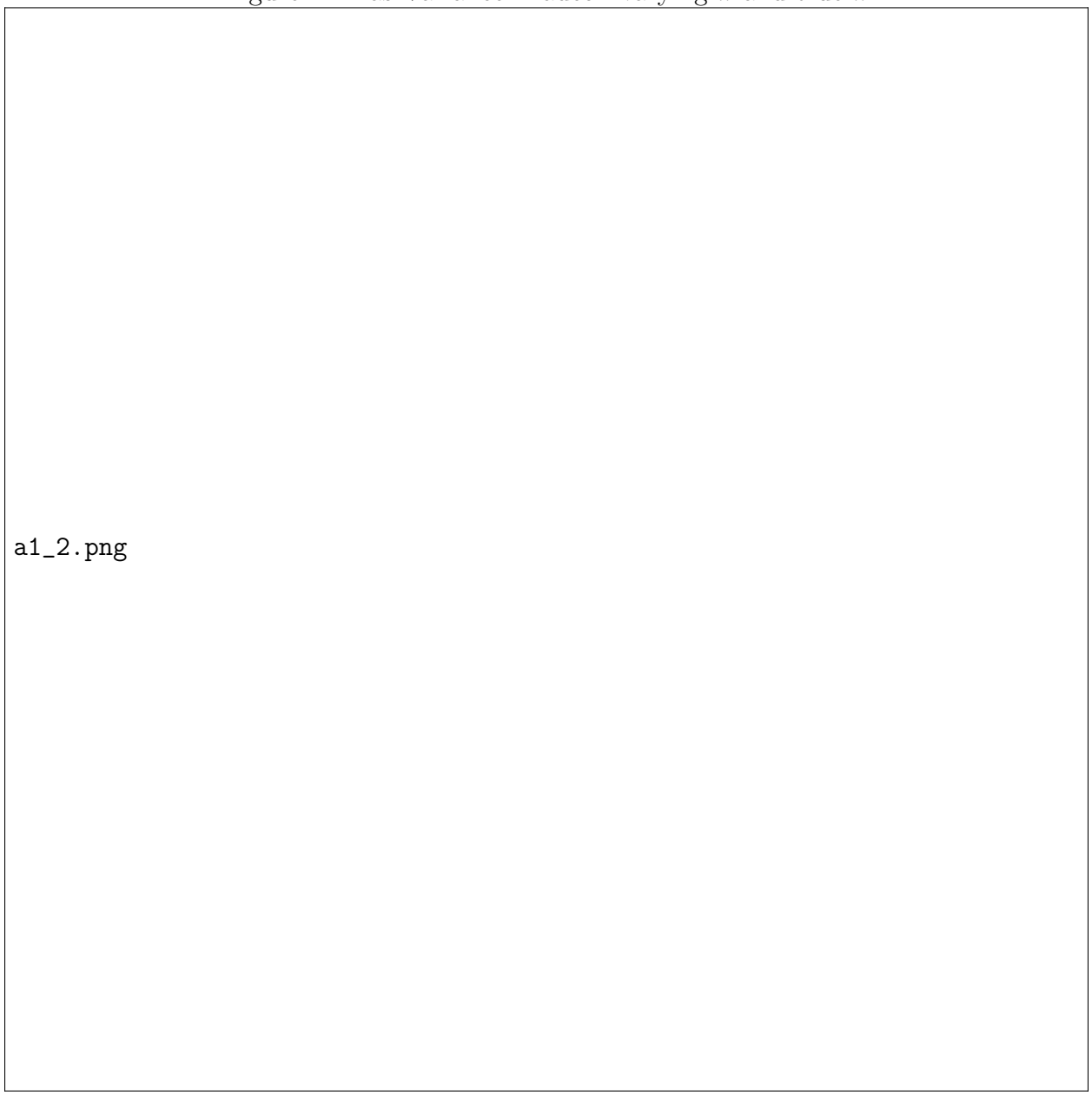
6 Bias-variance tradeoff

Suppose we are trying to estimate the mean of a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, however for each member of the sample i we collect two observations, X_{i1}, X_{i2}. Suppose we know that the first observation is drawn from the true distribution of X with probability π; with probability $1-\pi$, $X_{i1} \sim \mathcal{N}(\kappa, \omega^2)$. The two observations are dependent, so that when X_{i1} is drawn from the incorrect distribution, X_{i2} is drawn from the correct distribution, and vice versa. e For

$(\mu, \sigma^2) = (0, 1)$ and $(\kappa, \omega^2) = (0, 4)$ I solve for (a_2, a_3) that give an unbiased estimator of $\hat{\mu}$ using the formulas in the paper. For $n = 100$, $\pi \in (0.1, 0.9)$, and $a_1 \in \{1, \ldots, 10\}$, I plot the bias and variance of $\hat{\mu}$ that result from having beliefs $\hat{\pi} \neq \pi$.

When $a_1 = a_2 = a_3 = 1$, the bias is 0. Figure 1 shows my results for $a_1 = 2$.

Figure 1: Bias-Variance Tradeoff varying $\hat{\pi}$ and true π



7 Multiple x and multiple y

7.1 Single match within i

We observe $\{y_{i\ell}\}_{\ell=1}^{L_i}$ and $\{x_{ij}\}_{j=1}^{J_i}$ that share the same identifiers w_i. As before, we assume $P(y_{i\ell}$ and x_{ij} are a match $|w_i, L_i, J_i) = \frac{1}{L_i J_i}$. Interestingly, we need only condition on the joint value $L_i J_i$.

$$
\begin{aligned}
E\left[\sum_{\ell=1}^{L_i}\sum_{j=1}^{J_i} m(y_{i\ell}, x_{ij}; \theta)\middle| w_i, L_i, J_i\right] &= \sum_{\ell=1}^{L_i}\sum_{j=1}^{J_i} E\left[m(y_{i\ell}, x_{ij}; \theta)\right| \text{ true match}, w_i, L_i, J_i] \, P\left(\text{true match}|w_i, L_i, J_i\right. \\
&+ \sum_{\ell=1}^{L_i}\sum_{j=1}^{J_i} E\left[m(y_{i\ell}, x_{ij}; \theta)\right| \text{ false match}, w_i, L_i, J_i] \, P\left(\text{false match}|w_i, L_i, J_i\right. \\
&= \frac{1}{L_i J_i}\sum_{\ell=1}^{L_i}\sum_{j=1}^{J_i} E\left[m(y_{i\ell}, x_{ij}; \theta)\right| \text{ true match}, w_i, L_i, J_i] \\
&+ \frac{L_i J_i - 1}{L_i J_i}\sum_{\ell=1}^{L_i}\sum_{j=1}^{J_i} E\left[m(y_{i\ell}, x_{ij}; \theta)\right| \text{ false match}, w_i, L_i, J_i]
\end{aligned}
$$

Rearranging and using random sampling,

$$
\begin{aligned}
E\left[m(y_{i\ell}, x_{ij}; \theta)\right| \text{ true match}, w_i, L_i, J_i] &= E\left[\sum_{\ell=1}^{L_i}\sum_{j=1}^{J_i} m(y_{i\ell}, x_{ij}; \theta)\middle| w_i, L_i, J_i\right] \\
&- (L_i J_i - 1) E\left[m(y_{i\ell}, x_{ij}; \theta)\right| \text{ false match}, w_i, L_i, J_i]
\end{aligned}
$$

When $y_{i\ell}$ and x_{ij} are not associated with the same individual (but share a same w_i) they are independent conditional on w_i, L_i, J_i. Consider linear IV. Then:

$$
\begin{aligned}
E[z_{ij}(y_{i\ell} - x_{ij}'\beta)|w_i, L_i, J_i, \text{false match}] &= E[z_{ij}|w_i, L_i, J_i] E[y_{i\ell}|w_i, L_i, J_i] \\
&- E[z_{ij} x_{ij}'\beta|w_i, L_i, J_i]
\end{aligned}
$$

which can be estimated using $g_z(w)g_y(w) - g_{zx}(w)\beta$. Dimension of conditional expectations to estimate is larger.

Finally, using the law of iterated expectations

$$E\left[m(y_{i\ell}, x_{ij}; \theta)|\text{true match}\right] = E\left[\sum_{\ell=1}^{L_i}\sum_{j=1}^{J_i} m(y_{i\ell}, x_{ij}; \theta)\right] - E[(L_iJ_i - 1)g(w_i, L_i, J_i, \theta)] \quad (8)$$

If $L_i = 1$ or $J_i = 1$ this collapses to exactly what we had before.

Interesting to think about tradeoff in including i with large L_iJ_i vs. excluding. i.e. what is signal-noise (bias-variance) tradeoff involved in this decision. Should there be a threshold, similar to that for calling w_i the same? When is this method good?

7.2 Multiple matches within same i

If we believe that there are $\min\{J_i, L_i\}$ pairs among the observations with matching identifier w_i, then $P(y_{i\ell}, x_{ij} \text{ a match}|L_i, J_i, w_i) = \frac{\min\{J_i, L_i\}}{L_iJ_i}$, and we get

$$\begin{aligned} E\left[\sum_{\ell=1}^{L_i}\sum_{j=1}^{J_i} m(y_{i\ell}, x_{ij}; \theta)\middle| w_i, L_i, J_i\right] &= \frac{\min\{J_i, L_i\}}{L_iJ_i}\sum_{\ell=1}^{L_i}\sum_{j=1}^{J_i} E\left[m(y_{i\ell}, x_{ij}; \theta)|\text{ true match}, w_i, L_i, J_i\right] \\ &+ \frac{L_iJ_i - \min\{L_i, J_i\}}{L_iJ_i}\sum_{\ell=1}^{L_i}\sum_{j=1}^{J_i} E\left[m(y_{i\ell}, x_{ij}; \theta)|\text{ false match}, w_i, L_i, J_i\right] \end{aligned}$$

which implies the final moment condition,

$$E\left[m(y_{i\ell}, x_{ij}; \theta)|\text{true match}\right] = E\left[\min\{J_i, L_i\}\sum_{\ell=1}^{L_i}\sum_{j=1}^{J_i} m(y_{i\ell}, x_{ij}; \theta)\right] - E[(L_iJ_i - \min\{J_i, L_i\})g(w_i, L_i, J_i)] \quad (9)$$

How precise does $\hat{g}$ need to be?

I am concerned about how you might actually estimate g in applications.

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \left\{ \min_{\mathcal{Z}} \sum_{i=1}^{n} q(z_i; \theta) \right\}$$