Internet Appendix for "Matching Capital and Labor"

Jonathan B. Berk, Jules H. van Binsbergen, and Binying Liu

June 2, 2017

In this appendix, we summarize the steps we take in our best effort to create a clean, survival-bias free data set of mappings between mutual fund managers (firms) and mutual funds.

This appendix is organized as follows. In the next section, we describe the data sources we use and their shortcomings (the descriptions here are an elaboration of any summaries provided in the paper). In the third section we describe in detail our procedure for cleaning and merging multiple data sources on firms. In the final section we describe our procedure for cleaning and merging data on managers.

I. Our Data Sources

Our data originate from three sources. We use CRSP and Morningstar Principia for fund level information such as returns, assets under management, investment objective and holdings composition. A description of combining these two data sources can be found in the online appendix of Berk and van Binsbergen (2015). The two data sources also contain information about which fund managers manage each mutual fund. However we encountered several difficulties that prevent us from directly using this information.

- A manager is identified by various abbreviations of his name. Some entries list the full name of a manager, while others only use a manager's last name. Some versions of the same name contain a suffix or a title while others do not. This problem exists in both CRSP and Morningstar Principia.

- For over 5% of all fund observations (we define a fund observation as a unqiue fund $\times$ year-month combination) managers are identified as "Team Managed", or various other forms of this phrase.

- Many mutual funds in CRSP do not report managers. This problem exists but is less severe in Principia.

- Among those funds that report managers in both CRSP and Morningstar Princpia, the two sets of managers reported for the same fund during the same month are inconsistent in approximately 10% to 20% of all observations across the two databases even after taking into account the afore mentioned problems.

Because of these data limitations, we bring in a third data source: Morningstar Direct. Morningstar Direct is a database at the fund level that, for each fund in its record, lists the complete names of all managers that ever managed that fund since the fund's inception. The names are comprehensive in that the first, last, and middle names are all included and spelled out fully. More importantly, the same fund manager is always listed under the same name representation both over time and across funds. This allows us to be certain of a 1-to-1 mapping between names and individual managers. When multiple managers are managing a single fund, Morningstar Direct lists the names of all of these managers. By including these names, we eliminate a bias against co-managed funds.

Using the Morningstar Direct database alone, however, creates a different problem: the data do not cover the entire mutual fund space. Although the data source claims to include all active and deceased mutual funds, we find the number of funds included to be smaller than the combined CRSP and Principia sample. More importantly, this difference is more serious in the earlier sample (prior 1990), suggesting a possible survivorship bias.

A similar set of issues arises in the identification of mutual fund firms (fund families). In particular, the names used for a given firm constantly change in CRSP and Principia, whereas Morningstar Direct, which contains a clean list of firm names, does not cover the entire mutual fund space. As a solution, we incorporate the Morningstar Direct manager and firm data into the Morningstar Principia and CRSP data, using information in Morningstar Direct as a benchmark to clean manager and firm data in the other two data sets. We describe the detailed procedure that we employ in the next two sections.

A. Preview

We begin by creating firm identifiers. We start with Morningstar Direct, assuming that it contains the most accurate information. We then add data from Morningstar Principia that are not in Morningstar Direct, as long as adding that data does not introduce a contradiction with the Direct data. We then add in data from CRSP that are not in the combined Morningstar data set, as long as no contradiction is introduced. After firms are assigned identifiers, we perform a similar procedure to assign identifiers to managers. That is, we start with Morningstar Direct, and then add information from Principia and CRSP as long as this information does not contradict the more accurate data sources.

II. Cleaning and Merging Firm Data

A. Assigning Identifiers to Firms

In Morningstar Principia, firm names are stored under the variable *familyname*. However, each family name does not map 1-to-1 onto a unique firm. The same fund family (i.e. firm) may be reported under different name representations. For example: "IXIS Adivsor Funds" or "IXIS Advisors." It is therefore necessary to assign identifiers to family names, where the same identifier is assigned to multiple name representations of the same firm. We take the following steps to assign identifiers to firms in Princpia.

- We identify firms using the first word in *familyname*. For example, both firm names "IXIS Adivsor Funds" and "IXIS Advisors" are assigned the identifier "IXIS." The following case describes an exception.
  - If the first word of a *familyname* is generic (i.e. "The" or "American"), then the second word in the firm name is used instead of the first word as the identifier. If the second word is also generic, the third word is used, and so on.

Some observations do not report a firm name and therefore have missing identifiers. We use the following algorithm to fill them in if we can: if a fund has the same firm identifier at two different dates and its identifier is missing on a third date that is in between, then we consider the fund to have the same firm identifier for all three dates.

- Let A be an observation from Morningstar Principia for which the firm identifier is missing, B be an observation from the same fund as A but at a date prior to A and C be an observation from the same fund as A but a date after A. If the firm identifiers of B and C are the same, then A is also assigned the same identifier.

In Morningstar Direct, the official name of the firm (fund family) of each fund is always reported. There is, therefore, 1-to-1 mapping between a firm and its name. We simply use the official name of the firm as the identifier.

In CRSP, firm names are stored under the variable *mgmt-name* and firm identifiers are stored under the variable *mgmt-cd*. Because CRSP already assigns an identifier to each firm, we simply adopt it. Some observations have missing *mgmt-cd*. In such cases, we use the following algorithm to fill in missing identifiers if we can.

- Let A be an observation from CRSP for which the firm identifier is missing, B be an observation from the same fund as A but at a date prior to A and C be an observation from the same fund as A but at a date after A. If the firm identifiers of B and C are the same, then A is assigned the same identifier.

B. Combining Firm Info in Direct and Principia

In this subsection we combine firm information in the two Morningstar databases (Direct and Principia). We first match mutual funds in Principia with funds in Direct as follows.

- Two funds are matched if they share the same ticker.
- If multiple funds in Direct use the same ticker, a match is performed manually and the remaining unmatched funds' tickers are erased. Similarly, if multiple funds in Principia use the same ticker, a match is performed manually.
- For each of the remaining unmatched funds in Direct (they are unmatched mainly because they do not report a ticker), compare its fund name with the fund name of each unmatched fund observation in Principia.
- If over 80% of the words in the Principia fund name can be matched to the Direct fund name, we call the two funds potential matches. A fund in Direct can have multiple Principia potential matches and vice versa, and in these cases we manually check through all potential matched pairs to decide on a match.

In this subsection we use *the merged subset* to refer to the subset of fund observations that exist in both Direct and Principia. Each fund observation in the merged subset has two firm identifiers: one from Direct and one from Principia. Next we define the condition under which the two identifiers are judged to be consistent with each other.

- Let I be a firm identifier in Direct and J be a firm identifier in Principia. I and J are consistent if and only if J is a substring of I, and every fund observation in the merged subset that has J as its Principia firm identifier also has I as its Direct firm identifier (i.e. J is nested in I).

The algorithm simply says that if in the merged subset, whenever we see a Principia identifier, we always see the same Direct identifier on the same observation, then the two are consistent. We then combine the two Morningstar databases, keeping in mind that the Principia firm identifier is less accurate than Direct's. So if the addition of a Principia firm identifier in any way creates an inconsistency with any identifiers from Direct, the Principia identifier is not included in the combined data set. More specifically, we take the following steps to create the combined Morningstar data set. These steps need to be performed in the order they are listed.

- A fund observation from the merged subset is automatically included in the combined Morningstar data set, with Direct firm identifiers adopted.

- Take an observation in Principia but not in the merged subset. If it does not share the same Principia firm identifier with any observation in the merged subset, then this observation is included in the combined Morningstar data set. The Principia firm identifier is adopted.

- Take an observation in Principia but not in the merged subset, and whose Principia firm identifier is shared by at least one observation from the merged subset. If for every observation in the merged subset that uses this Principia identifier, the identifier is always consistent with the corresponding Direct firm identifier, then this observation is included in the combined Morningstar data set, the Direct firm identifier is adopted.

C. Combining Firm Info in Morningstar and CRSP

In this subsection we combine firm information in the Morningstar database with information in CRSP. At this point, we have a list of firm identifiers from Morningstar and a list of firm identifiers from CRSP, obtained through the procedures outlined in earlier subsections. Matching between the CRSP and Morningstar data sets at the fund level is outlined in the online appendix of Berk and van Binsbergen (2015).

We assume that the combined Morningstar database has more accurate firm level information than CRSP. Therefore, whenever Morningstar and CRSP are inconsistent with each other, we use Morningstar data to overrule CRSP data. In this subsection, we use *the merged subset* to refer to the subset of fund observations that exist in both Morningstar and CRSP. Each fund observation in the merged subset has two firm identifiers: one from Morningstar and one from CRSP. The condition under which the two identifiers are judged to be consistent with each other is as follows:

- Let I be a firm identifier in Morningstar and J be a firm identifier in CRSP. I and J are consistent if and only if J's corresponding firm name and I's corresponding firm name share at least one non-generic word in common, and every fund observation in the merged subset that has J as its CRSP firm identifier also has I as its Morningstar firm identifier.

We next merge the Morningstar and CRSP data sets to create a master data set. Each observation in this master data set has a valid firm identifier that we use in our paper. The merge proceeds using the following steps are performed in the order they are listed.

- A fund observation from the merged subset is automatically included in the master data set, with Morningstar firm identifiers adopted.

- Take an observation in CRSP but not in the merged subset. If it does not share the same CRSP firm identifier with any observation in the merged subset, then this observation is included in the master data set. The CRSP firm identifier is adopted.

- Take an observation in CRSP but not in the merged subset, and whose CRSP firm identifier is shared by at least one observation from the merged subset. If for every observation in the merged subset that uses this CRSP identifier, the identifier is always consistent with the corresponding Direct firm identifier, then this observation is included in the master data set. The Morningstar firm identifier is adopted.

We now have a master data set in which each fund observation is assigned a firm identifier. Each firm identifier maps one-to-one to a mutual fund firm. In the next section, we use this information to assign manager identifiers. From this point on, we restrict the Morningstar Principia and CRSP data sets to include only observations with clean firm identifiers (i.e. only observations that belong to this master data set).

III. Cleaning and Merging Manager Info

A. Cleaning Manager Names

Manager information is stored under the variable *managername* in Morningstar Principia. Because a fund can be co-managed by multiple fund managers, *managername* may contain information regarding multiple managers. We take the following steps to clean the *managername* variable for each fund observation. It is important to note that the phrase *managername* refers to a unique manager $\times$ fund $\times$ yearmonth combination. That is, even if the same last name "Smith" is reported to manage two different fund observations, we consider the two "Smith"s as two different manager names (as a starting point).

- Search for the key word "/." Each "/" represents the boundary between names of different managers when a fund is co-managed. Example: "Beck/Schor/Fenley." Whenever multiple managers are involved in the management of a single fund, Principia reports only the last names of these managers. We label a manager name for which only last name is avaliable as a *incomplete name*.

- If "/" is not present, then the fund observation either has only one manager or reports no manager. An observation with no manager is recognized using ".", " ", "-." A *managername* labeled as "Team Managed" or a variation of this phrase is also considered to report no manager.

- The remaining fund observations are each managed by a single manager. Most often a single manager name is presented by Principia in the format of "Lastname,

Firstname." For example: "Butler, Donald." We label a manager name with both first and last names as a *complete name*.

- In some cases, a middle initial is also provided along with the first and last names. For example:"Byrne, Susan M.." We keep track of the middle initial (if a middle initial is reported), which we will use for screening later in our algorithm.

- In some cases, a suffix is also provided including "Sr.", "Jr.", "I", "II", "III" etc. For example: "O'Boyle/Aster Jr.." Given that the suffix may differentiate managers with identical first names and last names, we keep track of the suffix (if a suffix is reported) and use it later in our code for screening.

- In some cases, a title is provided, such as "Ph.D.", "CPA", "CFA." Because a person's title can change over time, we ignore these.

- We ignore "(et. al.)."

- In some cases, different forms of the same first name are used. For example "Rob", "Robert", "Bob." We select one of these and standardize different forms of the same first name.

Although some manager names would otherwise be labeled incomplete (i.e. no first name), we make an exception and treat these names as complete under the following circumstances. A name that qualifies as an exception is considered complete despite having only a last name.

- For each last name reported in either CRSP or Morningstar database, count the number of distinct first names associated with it. If this number is 0 or 1, then consider any manager name with that last name to be an exception.

For each fund observation in Morningstar Principia we now have a set of corresponding manager names responsible for the fund's management. We remove fund observations for which not a single corresponding manager can be identified.

Manager names in CRSP are stored under the variable *mgr-name*. The steps we take to clean CRSP manager names includes those steps taken to clean = Principia manager names. However, the reporting of manager names in CRSP is even less standardized than in Principia. The following complications are found in CRSP but not in Morningstar Principia.

- Sometimes the company name is listed as the manager name. For example:"Aetna Fixed Income Group." We ignore these cases and assume no manager is reported.

- When a fund is co-managed, both the last name and the first name initial may be listed. For example: "W Holzer/N Bratt/A Ho." Such a name is also considered incomplete.
- A first name may be abbreviated. For example: "ED C. Spelman" for "Edmond C. Spelman."
- Some names are most likely misspelled. For example: "Sal Dosanto" and "Sal Disanto."
- A character, which appears to be the middle initial, may be listed at the beginning of a name rather than in between the first and last name. For example: "A James Ellman."
- The full middle name may be reported in place of a middle initial. For example: "Abigail Jones Feder."

We use the following algorithm to determine if a firm name is reported as the manager name.

- A firm name is reported as the manager name if the *mgr-name* variable and the *mgmt-name* variable share at least one non-generic word in common.

We also employ an algorithm to screen for name abbreviations and misspellings. Let A and B be two manager names. We consider B to be likely an abbreviation or misspelling of A if all of the conditions listed below are met. We then look through the several hundred suspicious cases manually to make a final judgement. In the case where an abbreviation or misspelling is confirmed, we remove manager name B from our data set and replace it with manager name A in all observations that list B.

- A and B manages the same fund at different points in time.
- Both A and B are classified as complete names.
- First names and last names of both A and B each contains at least 2 characters.
- One of the following conditions is met.
  - A's last name is identical to B's last name; B's first name is a word with n characters and is identical to the first n characters of A's first name.
  - A's last name (or first name) is identical to B's last name (or first name); B's first name (or last name) is identical to A's first name (or last name) after a character (any character) is removed from A's first name (or last name).

- A's last name (or first name) is identical to B's last name (or first name); B's first name (or last name) is identical to A's first name (or last name) after a character (any character) is removed each from A's first name (or last name) and from B's first name (or last name).

B. Assigning Identifiers to Managers

As stated in earlier sections, each Principia manager name does not map 1-to-1 onto a unique fund manager. The same manager maybe reported under different name representations. Because names are often abbreviated to last names, the same last name can also be used by different managers. In this subsection, we assign identifiers to managers, so that each identifier maps onto a unique manager. But before we do so, we first define the degree of fit between two clean manager names reported in Principia as follows.

- In a Class 1 fit, both manager names are complete (contain a last name and a first name) and both portions of the names are identical.
  - Furthermore, if both report middle initials, the two initials are identical.
  - AND if both report suffixes, the two suffixes also are identical.
- In a Class 2 fit, one manager name is complete and the other contains only the first name or only the last name. The incomplete manager name is a substring of the complete manager name.
- In a Class 3 fit, both manager names are incomplete (contains only the first name or only the last name). The two names are identical.

As a starting point, each manager name begins with a unique identifier. We first iteratively perform the following steps to refine our assignment of identifiers.

- Let A, B, C and D be four manager names.
- Assign the same identifier to A and B if A is a Class 1 fit of B and A and B both belong to the same firm.
- Assign the same identifier to A and B if B is a complete name , A is a Class 2 fit of B and A and B manage the same fund at different points in time.
  - Unless there exists a manager name C of the same fund such that C is a complete name and C Class 2 fits A but does not Class 1 fit B.
- Assign the same identifier to A and B if A is a Class 3 fit of B and A and B belong to the same fund.

– Unless there exist two manager names C and D of the same firm such that C and D are both full names, C Class 2 fits both A and B, D Class 2 fits both A and B but C does not Class 1 fit D.

It is important to note that in order to minimize error introduced into the matching process, we do not match two incomplete names within the same firm unless one of the three conditions above are met. In other words, if we observe two "Smith"s in the same firm, we cannot be sure that they refer to the same manager, unless 1) the two "Smith"s manage the same fund or 2) the same full name, say "John Smith", is also reported in the two funds that the two "Smith"s manage if they do not manage the same fund. We run the above three steps multiple iterations until no changes can be made from further iterations. Next, we assign common identifiers to managers who may have transfered across firms. This involves the following steps which need to be performed in the order they are listed (and all steps here need to be performed after the initial three steps listed above).

- Assign the same identifier to A and B if A Class 1 fits B, A is from firm I and B is from firm $J \neq I$ and there exist C in firm I and D in firm J such that C is matched to A, D is matched to B, and observation C is dated before D but within 24 months of D.
  - Unless there exist F in firm I and G in firm J such that F is matched to A, G is matched to B, and observation F is dated at the same time as or after D.
  - OR if by this criteria A (or B) can be simultaneously matched to observations from two different firms.

Intuitively, this condition simply says that if a complete name leaves a firm and joins another firm within two years, we consider these manager names to describe the same manager who moved across the two firms. Unless, however, it appears that this manager has left a firm and joined two other firms (almost) simultaneously, in which case we cannot be sure that the names in these three firms describe the same person.

Morningstar Direct provides the complete names (including first, middle and last name and title) for every manager it lists. It is unlikely that two managers will have the same first, middle, and last names and hold the same title. Therefore, the full names of Direct managers can be used as identifiers.

For CRSP, we simply follow the same set of steps as in Principia to assign identifiers to manager names.

C. Combining Manager Info in Direct and Principia

In this subsection we combine manager information in the two Morningstar databases. We assume that Direct has more accurate manager information than Principia. Therefore,

whenever Direct and Principia are inconsistent with each other, we use Direct data to overrule Principia data. In this subsection, we use *the merged subset* to refer to the subset of fund observations that exist in both Direct and Principia. Each fund observation in the merged subset has two corresponding sets of manager identifiers: one from Direct and one from Principia. We deetermine when a Direct manager identifier is consistent with a Principia manager identifier using the following rule.

- Let A be a manager identifier in Direct and B be a firm identifier in Principia. A and B are consistent if A's manager name incorporates one of B's manager names, and if every fund observation in the merged subset that is managed by B in Principia is managed by A in Direct.
  - Unless there exists another manager identifier C in Principia such that C and A are also consistent, and C and B are two different identifiers in Principia that manage the same fund at the same point in time.

We next merge the two Morningstar databases, keeping in mind that Principia manager data is less accurate than Direct's. If the addition of a Principia manager identifier creates an inconsistency with any identifiers from Direct, the Principia identifier is not included in the combined data set. More specifically, we take the following steps to create a combined Morningstar data set by performing the following steps in the order they are listed.

- A fund observation from the merged subset is automatically included into the combined Morningstar data set, with Direct manager identifiers adopted.
- Take an observation in Principia but not in the merged subset. If for every manager identifier A who manages this fund observation, A does not appear in the merged subset, then this observation is included in the Morningstar combined data set, and the Principia manager identifiers are adopted.
- Take an observation in Principia but not in the merged subset. Take any Principia manager identifier A who manages this observation and also manages an observation in the merged subset. If for every observation in the merged subset managed by A, I is always consistent with the corresponding Direct manager identifier, then this observation is included in the combined Morningstar data set, and the Direct manager identifiers are adopted.

We now have a data set that contains both Morningstar Principia and Morningstar Direct observations. A list of manager identifiers is mapped to each fund observation in this data set. Each manager identifier corresponds to a unique fund manager.

D. Combining Manager Info in Morningstar and CRSP

In this subsection we combine manager information in CRSP with information in the combined Morningstar data set. We assume that the combined Morningstar database has more accurate manager information than CRSP. Therefore, whenever Morningstar and CRSP are inconsistent with each other, we use Morningstar data to overrule CRSP data. In this subsection, we refer to the set of fund observations included in both Morningstar and CRSP as *the merged subset*. Each observation in the merged subset has two separate list of manager identifiers: one from Morningstar and one from CRSP. We use the following procedure to determine if a Morningstar manager identifier is consistent with a CRSP manager identifier.

- Let A be a manager identifier in Morningstar and B be a manager identifier in CRSP. A and B are consistent if one of A's manager names fits (Class 1, 2 or 3) one of B's manager names, and every fund observation in the merged subset that is managed by B in CRSP is managed by A in Morningstar.
  - Unless there exists another manager identifier C in CRSP such that C and A are also consistent, and C and B are two different identifiers in CRSP that manage the same fund at some point in time.

We apply the following steps to merge Morningstar and CRSP data. These steps need to be performed in the order they are listed.

- A fund observation from the merged subset is automatically included into the master data set, with Morningstar manager identifiers adopted.
- Take an observation in CRSP but not in the merged subset. If for every manager identifier I who manages this fund observation, I does not appear in the merged subset, then this observation is included in the master data set and the CRSP manager identifiers are adopted.
- Take an observation in CRSP but not in the merged subset. Take any CRSP manager identifier I who manages this observation and also manages an observation in the merged subset. If for every observation in the merged subset managed by I, I is always consistent with the corresponding Morningstar manager identifier, then this observation is included in the master data set. The Morningstar manager identifiers are adopted if avaliable. Otherwise the CRSP manager identifiers are adopted.

This master data set now includes fund observations from Morningstar Direct, Morningstar Principia, and CRSP. Each fund observation in this master data set contains a firm identifier and a non-empty list of manager identifiers. This is the data we use in our paper.

References

Berk, Jonathan B., and Jules H. van Binsbergen, 2015, Measuring skill in the mutual fund industry, *Journal of Financial Economics* 118, 1 – 20.