The New Washdown Algorithm
(First Draft)

Ricardo Silva
Center for Automated Learning and Discovery
Carnegie Mellon University
`rbas@cs.cmu.edu`

Abstract

There are known methods for discovering pure linear measurement models given as input just the joint distribution of the observed variables and assumptions about the number of pure indicators per latent and normality. This paper extends this approach by describing a variation that is more computationally efficient, as well as providing a proof of its correctness.

1 The Washdown algorithm, revisited

In a previous report (Silva, 2002), we introduced the `Washdown` algorithm as an approach to find a correct pure measurement model for some unknown linear measurement/structural model. This report discusses a variation with possibly significant computational gains. In Section 2, a proof of correctness for `Washdown` is presented and more detailed implementation issues are discussed in Section 3.

The intuition behind the new `Washdown` algorithm is that we can achieve the same results using only pairwise comparisons of clusters. More than that, most comparisons can be made by evaluating one-factor models augmented by some indicator of other cluster. This can lead to very efficient performance when compared to the original formulation of this algorithm because of two main reasons:

- everytime some model is invalid and a variablel should be chosen for removal, we will be searching among a reduced set of candidates. In the original formulation of the algorithm, we always searched among all variables at any time;
- rarely we will need to fit two-factor models. One-factor models are simpler to fit, and there are known techniques to take advantage of its simplicity;

Table 1 introduces an apparently much more complicated version of `Washdown`, but in its essence it is nothing more than finding a new cluster at a time and then evaluating it against all other clusters that were already discovered. A new merging procedure was also considered. In principle, if two clusters can be merged, one of them could just be completed discarded since every node in it has some sort of impurity with respect to other node in the true graph. However, it can still be useful to keep these nodes till the end in order to have some extra indicators per latent (i.e., a node may start impure, but can later remain in the final graph if the source of its impurity was removed).

This divide-and-conquer approach actually simplifies the proof of correctness and provides more elegant implementations. The purification algorithm was also modified in order to simplify the description of the `Washdown` procedure. Table 3 describes a procedure to purify a two-factor model that seems to be too complex, but it is quite computationally efficient compared with the original algorithm found in (Silva, 2002).

The function $MMScore$ used in these algorithms is some goodness-of-approximation score for candidate graphs that are evaluated during the search process. It gets as inputs a pure latent variable graph and a joint density function over the observed variables of the graph. We define an idealized $MMScore$ function in Section 2.

2 Correctness of Washdown

Let a *set of constraints* $\mathbf{C}$ be a set of equations of the type $f(\mathbf{X}) = 0$, where $f(\mathbf{X})$ is polynomial function of a given set of variables $\mathbf{X}$, and there are no two functions $g(\mathbf{X})$ and $h(\mathbf{X})$ such that $g(\mathbf{X}) = 0$ and $h(\mathbf{X}) = 0$ belong to $\mathbf{C}$ and $f(\mathbf{X})$ is a function of g and h. *(Note: I will need a better definition here).* The definition of $MMScore$ relies in this definition of constraint. In the Gaussian case, a constraint is a function of the covariance matrix.

Let Θ_O be the vector of parameters that defines a joint probability distribution over a set $\mathbf{O}$ of observed variables. Elements of Θ_O are tied by a set of constraints $\mathbf{C}$. We make the following assumptions:

- the true model is a linear measurement/structural model with at least three pure indicators per latent, and the joint probability distribution over the observed variables is defined by Θ_O. Θ_O will be the covariance matrix over $\mathbf{O}$. *Our proof will rely on the fact that we know the population covariance matrix.*
- all constraints in $\mathbf{C}$ are entailed by a linear measurement/structural model G_T and a constraint is entailed in G_T if and only if it holds in Θ_O (i.e., the faithfulness assumption);

Let G be a linear structural equation model graph, and let G_T be the true measurement/structural graph corresponding to Θ_O. Define $MMScore(G, \mathbf{C})$ as the number of constraints among elements of Θ_O that are entailed by G but do not hold in $\mathbf{C}$. The intuition behind the correctness of Washdown/Purify lies in the identification of submodels G' in which $MMScore(G', \mathbf{C}) = 0$. In the demonstration, we constantly use the term "relatively pure node" for graphs $G \neq G_T$. That means that while a node may be a impurity in G_T, it is not in G. Notice also that if a graph has less than four indicators, it represents a satured model and no constraints will be entailed, i.e., $MMScore(G) = 0$ if G has less than four indicators.

The following lemma will be especially useful in proving the correctness of Washdown:

Lemma 1 *Let $G_{candidate}$ be a pure measurement/structural graph with freely correlated latents and a set of observed nodes $\mathbf{O_{candidate}} \subset \mathbf{O}$. Let P and I be two nodes that belong to the same cluster in the true graph G_T and also in $G_{candidate}$. Assume that the P and I are measures of the latent L in G_T and also (pure) indicators of the latent H in $G_{candidate}$. Assume that P is a pure measure of its latent with respect to other measures in G_T that are also included in $G_{candidate}$, and that I is a impure measure with respect to some other measures in G_T that are also included in $G_{candidate}$. Then, removing P will not lower the $MMScore$ of $G_{candidate}$ more than removing I. If we also assume also that there are at least three pure measures of the same latent L that are also children of H, then removing I will lower the $MMScore$ of $G_{candidate}$ more than removing P.*

Proof of Lemma 1:

In this proof, we will sometimes refer to constraints over a set of parameters $\Theta_\mathbf{V}$ as a constraint over the random variables that define those parameters, i.e., the symbol $C_i(\mathbf{V})$ is used to represent some constraint $C_i(\Theta_\mathbf{V})$.

First, let's proof that removing P will never eliminate more wrong constraints than removing I. For this, we have only to prove that for every constraint $C(P, \mathbf{X})$ that holds in $G_{candidate}$ but does not hold in G_T, there is also a constraint $H(I, \mathbf{X})$ under the same situation, where $\mathbf{X} \subset \mathbf{O} - \{P, I\}$. Initially, define the following mapping: for every $C(P, \mathbf{X})$, define $H(I, \mathbf{X})$ as a constraint where every appearance of a term σ_{px}, $x \in \mathbf{X}$, is substituted by a term σ_{Ix}. This is

```
function Washdown
inputs   a set O of random variables
         a joint probability distribution D over O
output   a partition of a subset of O
```

$N \leftarrow 0$
Candidates $\leftarrow$ **O**
repeat
  Discards $\leftarrow PurifyOneFactor(\mathbf{Candidates}, D, \mathbf{Candidates})$
  $T_{N+1} \leftarrow \mathbf{Candidates} - \mathbf{Discards}$
  if $MMScore(T_{N+1}, D) \neq 0$ **or** $|T_{N+1}| \leq 3$
    $T_{N+1} \leftarrow \mathbf{Candidates}$
    Discards $\leftarrow \emptyset$
  end if
  for `I = N, N - 1, ..., 1`
    $(\mathbf{D_I}, \mathbf{D_{N+1}}) \leftarrow PurifyTwoFactors(T_I, T_{N+1}, D)$
    if $\mathbf{Discards} \neq \emptyset$, $\mathbf{Discards} \leftarrow \mathbf{Discards} \cup \mathbf{D_{N+1}}$
    $T_I \leftarrow T_I - \mathbf{D_I}$
    if $I = N$, $T_{N+1} \leftarrow T_{N+1} \cup \mathbf{D_I}$
    $T_{N+1} \leftarrow T_{N+1} - \mathbf{D_{N+1}}$
  end for
  $M \leftarrow PossibleMerge(\{T_1, ..., T_N\}, T_{N+1}, D)$
  if $M \neq 0$
    $T_M \leftarrow T_M \cup T_{N+1}$
  else if $|T_{N+1}| \geq 3$
    $N \leftarrow N + 1$
  Candidates $\leftarrow$ **Discards**
until $\mathbf{Candidates} = \emptyset$
`return` $T_1, T_2, ..., T_N$

Table 1: The `Washdown` algorithm regroups sets of observed variables into different clusters by using `Purify` to identify which variables should not be in their current groups. The misplaced variables are then slid through all latents that may be required.

```
function PurifyOneCluster
inputs   a set of observed variables O
         a probability distribution D that has a marginal over O
         a subset V of O
output   a subset of O

Let G be a one-factor model over O
Discards ← ∅
while MMScore(G, D) > 0 and V ≠ ∅
   best_v ← argmin_{v∈V} MMScore(G − v, D)
   score ← MMScore(G − best_v, D)
   V ← V − best_v
   G ← G − best_v
   Discards ← Discards ∪ best_v
end while
return Discards
```

Table 2: The purification algorithm for choosing a subset of a graph G that is faithful to the marginal of some distribution D over a subset of the observed variables of G when there is just one cluster.

a one-to-one mapping. Our first goal is to show that if C that is entailed by $G_{candidate}$ but not by G_T, then H is not entailed by G_T either. It is trivial to see that if C holds in $G_{candidate}$, so does H, due to the equivalent graphical properties of P and I with respect to $\mathbf{X}$.

A constraint such as $C(P, \mathbf{X})$ is a polynom over $\Theta_{P,\mathbf{X}}$, as entailed by the treksum rule of linear models. Let λ_P be the label of the edge connecting P to its latent parent in $G_{candidate}$ and λ_I the corresponding edge label for I. One can rewrite $C(P, \mathbf{X})$ as a polynomial function over the edge labels and error variances, substitute every occurrence of λ_P for λ_I in that function, and write it back as a function of $\Theta_{\{I,\mathbf{X}\}}$. This process uniquely define a new constraint $H(I, \mathbf{X})$ out of $C(P, \mathbf{X})$.

Now, assume that $C(P, \mathbf{X})$ does not hold in G_T. If all treks connecting I in G_T to all elements in $\mathbf{X}$ go through the edge λ'_I connecting I to the same latent parent of P, then I will have the same graphical properties as P when defining parameters $\Theta_{\{I,\mathbf{X}\}}$ and relations among these parameters. Therefore, I will also not hold in G_T. Now, assume there are some other treks from I to elements of $\mathbf{X}$ that go through edges other than λ'_I. Rewriting I as a function of the edge labels and variances of G_T and by the additivity property of the trek rule for defining covariances in linear graphical models, one has:

$$H(I, \mathbf{X}) = M(I, \mathbf{X}) + N(I, \mathbf{X})$$

where M accounts for all the covariance that is explained through λ'_I and none produced by the other neighboring edges of I. It is easy to see that $M(I, \mathbf{X})$ is identical, minus some constant, to the left-hand side of $C(P, \mathbf{X})$ when written as a function of edges and variances of G_T and substuting every ocurrence of the label of the edge connecting P to its latent parent in G_T by λ'_I. Since M does not boil down to some constant, for H to be identically zero we have to have $N(I, \mathbf{X}) = -M(I, \mathbf{X}$. But since λ'_I appears only in the right side of that equation, this equality cannot hold unless for some specific values of the parameters. Therefore, for any invalid constraint that is entailed by P there is some unique invalid constraint entailed by I, and the number of invalid constraints entailed by P is at least as large as for I.

Now we have to prove that at least one invalid constraint is entailed by I, but not by P.

```
function PurifyTwoClusters
inputs   a set of observed variables O_1
         a set of observed variables O_2
         a probability distribution D that has a marginal over O_1 ∪ O_2
output   a list of two subsets of O_1 ∪ O_2

Let G_2 be a pure two-factor model with correlated latents, where the first
factor has O_1 as loadings and the second factor has O_2 as loadings
if MMScore(G_2, D) = 0
   return (∅, ∅)
D_1 ← ∅
C_1 ← O_1
for every element C_2 ∈ O_2
  D_0 ← PurifyOneCluster(C_1 ∪ C_2, D, C_1)
  C_1 ← C_1 − D_0
  D_1 ← D_1 ∪ D_0
end for
C_2 ← O_2 ∪ D_1
D_2 ← PurifyOneCluster(C_2, D, C_2)
C_2 ← C_2 − D_2
Let C_1 be some arbitrary element in C_1
D_2 ← D_2 ∪ PurifyOneCluster(C_2 ∪ C_1, D, C_2)
C_2 ← C_2 − D_2
Let G_2 be a pure two-factor model with correlated latents, where the first
factor has C_1 as loadings and the second factor has C_2 as loadings
while MMScore(G_2, D) > 0
   best_v ← argmin_{v ∈ C_2} MMScore(G_2 − v, D)
   C_2 ← C_2 − best_v
   G_2 ← G_2 − best_v
   D_2 ← D_2 ∪ best_v
end while
return (D_1, D_2)
```

Table 3: The purification algorithm for choosing a subset of a graph G that is faithful to the marginal of some distribution D over a subset of the observed variables of G when there are two clusters.

function	`PossibleMerge`
inputs	`an indexed set of sets of observed variables` $\mathbf{U}$ `a set of observed variables` $\mathbf{O}$ `a joint probability distribution` D `that has a marginal over all observed variables given as input`
output	`an index number`

if $\lvert\mathbf{O}\rvert < 2$, `return 0`
for $I = 1, 2, ..., \lvert\mathbf{U}\rvert$
  `Let` G `be a one-factor model over` $\mathbf{U}(I) \cup \mathbf{O}$
  if $MMScore(G, D) = 0$, **return** I
end for
`return 0`	

Table 4: A procedure to identify if two sets of indicators could belong to the same one-factor model.

First, let P_1, P_2 and P_3 be pure measures of the same latent (remember that we assume that there are at least three of them). Assume I is a intra-construct impurity (some I_2 is its child, which is also an indicator of the same cluster). We have that $\sigma_{P_1I_1}\sigma_{P_2I_2} - \sigma_{P_1P_2}\sigma_{I_1I_2} = 0$ holds in $G_{candidate}$ but not G_T. On the other hand, $\sigma_{P_1P_3}\sigma_{P_2I_2} - \sigma_{P_1P_2}\sigma_{P_3I_2} = 0$ holds in both situations, and therefore, the presence of I entails at least one more invalid constraint than P_3 (or, analogously, any other pure indicator of the same cluster). The same reasoning applies if I_1 and I_2 have correlated errors.

Now, assume that I is a cross-construct impure with some indicator I' of another cluster. Under $G_{candidate}$, $\sigma_{P_1P_2}\sigma_{I'I} - \sigma_{P_1I}\sigma_{P_2I'} = 0$. However, this is not entailed by G_T. On the other hand, $\sigma_{P_1P_2}\sigma_{P_3I'} - \sigma_{P_1P_3}\sigma_{P_2I'} = 0$ is entailed in both. Again, I implies at least one more constraint than P_3 (or any other pure indicator of the same cluster).

Now, assume that I is a latent-measured impurity. This case is analogous to the previous one (in other words, an extra trek from I to I' that was eliminated in $G_{candidate}$). $\square$

The following two lemmas will be necessary later to prove that every cluster of the original graph will appear at the end of the Washdown procedure:

Lemma 2 *Let $G_{candidate}$ be a one-factor graph and $\mathbf{P}$ a group of at least three variables in $G_{candidate}$ such that they are indicators of a same latent L in G_T, and they are relatively pure with respect to every variable in G_T that is in $G_{candidate}$. Then, if any variable is removed from $\mathbf{P}$ during purification, all elements of $\mathbf{P}$ will end up being removed (with the possible exception of one).*

Proof of Lemma 2:

Let $\#c(V)$ the number of constraints correctly entailed by $G_{candidate}$ that include a reference to variable V. Let P be some variable in $\mathbf{P}$ and W some variable in $\mathbf{W}$ such that every element of this set is the indicator of some latent $L_W \neq L$ in G_T, no element has correlated errors with any other element of $\mathbf{W}$, and there are no intra-construct impure measures in this set.

Due to symmetry, all variables in $G_{candidate}$ appear in the same number of entailed constraints. The variable that will be removed at any step (if the $MMScore$ of the graph is not zero, or it has less than five variables) is the one that appears in the least number of correctly entailed constraints, possibly tied with another one. Suppose that $\#c(P) \leq \#c(W)$, a necessary condition if some element of $\mathbf{P}$ is going to be removed.

Since in $G_{candidate}$ there are at least three relatively pure indicators of L, by Lemma 1 we can safely assume that there are no impure indicators of L in $G_{candidate}$, or otherwise an element in $\mathbf{P}$ would not be chosen to be removed. The number of correct entailed constraints with k variables that include P can be counted in the following way: the constraints that include P, $k-2$ elements of $\mathbf{P}$, and one element from $\mathbf{P} \cup \mathbf{W}$; the constraints that include P and $k-1$ elements of $\mathbf{W}$; the constraints that include P and at least one element not in $\mathbf{P} \cup \mathbf{W}$, which we will represent by $\#o_k(P)$. The total number of sets of k variables corresponding to these constraints is given by:

$$\#s_k(P) = \binom{|\mathbf{P}|-1}{k-2} \times (|\mathbf{P}| + |\mathbf{W}| - 1) + \binom{|\mathbf{W}|}{k-1} + \#o_k(P) \tag{1}$$

Analogously, we have

$$\#s_k(W) = \binom{|\mathbf{W}|-1}{k-2} \times (|\mathbf{W}| + |\mathbf{P}| - 1) + \binom{|\mathbf{P}|}{k-1} + \#o_k(W) \tag{2}$$

Since P is a pure measure and W may be impure with respect to some other indicators, if one assumes that $|\mathbf{P}| > |\mathbf{W}|$, it is straightforward to verify that $\#o_k(P) > \#o_k(W)$. Thus, assuming that $|\mathbf{P}| = |\mathbf{W}| + t$, $t > 0$, we have:

$$\begin{aligned} \#s_k(P) - \#s_k(W) \quad > \quad & \binom{|\mathbf{P}|-1}{k-2} \times (|\mathbf{P}| + |\mathbf{W}| - 1) + \binom{|\mathbf{W}|}{k-1} - \\ & \binom{|\mathbf{W}|-1}{k-2} \times (|\mathbf{W}| + |\mathbf{P}| - 1) - \binom{|\mathbf{P}|}{k-1} \end{aligned}$$

Now, summing over all the sets of size $k = 2$ to $k = |\mathbf{P}| + 1$ (which is the size of the largest subgraph that can entail only valid constraints), we have:

$$\begin{aligned} & \textstyle\sum_{k=2}^{k=|\mathbf{P}|+1} \#s_k(P) - \#s_k(W) > \\ & \textstyle\sum_{k=2}^{k=|\mathbf{P}|+1} \left[\binom{|\mathbf{P}|-1}{k-2} - \binom{|\mathbf{W}|-1}{k-2}\right] \times 2|\mathbf{W}| + \left[\binom{|\mathbf{W}|}{k-1} - \binom{|\mathbf{P}|}{k-1}\right] = \\ & \textstyle\sum_{k=0}^{k=|\mathbf{P}|-1} \left[\binom{|\mathbf{P}|-1}{k} - \binom{|\mathbf{W}|-1}{k}\right] \times 2|\mathbf{W}| + \left[\binom{|\mathbf{W}|}{k+1} - \binom{|\mathbf{P}|}{k+1}\right] = \\ & 2|\mathbf{W}| \left(2^{|\mathbf{P}|-1} - 2^{|\mathbf{W}|-1}\right) + \left(2^{|\mathbf{W}|} - 1 - 2^{|\mathbf{P}|} + 1\right) = \\ & (|\mathbf{W}| - 1)\left(2^{|\mathbf{P}|} - 2^{|\mathbf{W}|}\right) \geq 0 \end{aligned}$$

The number of correct entailed constraints with k variables that include some variable V will be a function of $\#s(V)$, which will be the same for any variable due to the symmetry of the one-factor model. Therefore, if $|\mathbf{P}| > |\mathbf{W}|$, we have that $\#c(P) > \#c(W)$, which is a contradiction. It follows that $|\mathbf{P}| \leq |\mathbf{W}|$.

Now, if we remove some element from $\mathbf{P}$ (because there is some invalid constraint in the graph), we will have that $|\mathbf{P}| < |\mathbf{W}|$. The sets $\#o_k(P)$ will get smaller or remain the same, while the sets $\#o_k(W)$ will not be affected. Performing a similar counting procedure, it follows that, after removing this element, $\#c(P) < \#c(W)$.

The fact that we removed one element from $\mathbf{P}$ will lead to a continuous removal of elements of $\mathbf{P}$, since $\#c(W)$ will never get to zero given that $|\mathbf{W}| > 3$ and the fact that, while there are at least two elements in $\mathbf{P}$, $MMScore(G_{candidate}) \neq 0$ and $\#c(P) < \#c(W)$, $P \in \mathbf{P}$, $W \in \mathbf{W}$. $\square$

Lemma 3 *Let $G_{candidate}$ be a one-factor pure latent graph. If at the end of purification, we have a cluster of n indicators where $n-1$ are indicators of some latent L in G_T, $n \geq 4$, and the MMScore of the purified candidate graph is zero, all the relatively pure indicators that were in $G_{candidate}$ before purification will remain after purification.*

Proof of Lemma 3:

If there are 3 or more relatively pure indicators of L in $G_{candidate}$, the result will follow directly from Lemmas 1 and 2. If there is just one pure indicator of L, the result will trivially hold.

Now, suppose there are only two relative pure indicators $\{P_1, P_2\}$ of L in $G_{candidate}$. Let $\{V_1, ..., V_{(n-1)}, K\}$ be the remaining variables after the purification of $G_{candidate}$, where $V_1, ..., V_{(n-1)}$ are indicators of L and K is not an indicator of L.

We will use here the same bijective function applied in the demonstration of Lemma 1 that maps a constraint referring to X_1 but not X_2 to a constraint referring to X_2 but not X_1. The goal is the same: proving that P_i cannot not be chosen to be discarded in place of some $V_j \neq P_1, P_2$ at any point of the purification process by demonstrating that the number of correct implied constraints containing P_i but not V_j should be always greater than those containing V_j but not P_i, or some of the assumptions of the lemma will not hold.

Suppose that we are at some stage of the purification process where P_i is chosen to be discarded. First, let's prove that there cannot be any cross-construct impurities between some indicator of L with respect to K in $G_{candidate}$. We know that all tetrad constraints entailed by the nodes $\{P_i, V_a, V_b, K\}$ are valid. Suppose that some other indicator I of L is cross-construct impure with respect to K. Then there will be some tetrad constraint entailed by the set $\{I, V_a, V_b, K\}$ that will be invalid. By a counting procedure analogous to the one used in the demonstration of Lemma 1, removing I will lower the *MMScore* of $G_{candidate}$ more than removing P_i, contrary to our assumptions.

Now, let's prove that there cannot be any intra-construct impure nodes among those indicators of L in $G_{candidate}$. We know that all tetrad constraints entailed by the nodes $\{P_i, V_a, V_b, K\}$ are valid. Suppose we have a pair of intra-construct impurities $\{I_1, I_2\}$. If I_1 is also intra-construct impure with respect to V_a, then there are invalid tetrad constraints entailed by the set $\{I_1, V_a, V_b, K\}$, which by our mapping procedure described in the demonstration of Lemma 1 will imply that removing I_1 will lower the *MMScore* by more than removing P_i, which is a contradiction. Now, suppose that I_1 is not a intra-construct impurity with respect to V_a, and as proved in the previous paragraph, I_1 is also not a cross-construct impurity with respect to K. Then, the set $\{P_i, V_a, I_1, K\}$ will entail only valid tetrad constraints. However, the set $\{I_2, V_a, I_1, K\}$ will not, which again implies that we should remove I_2 instead of P_i, contrary to our assumptions.

Suppose $X \neq K$ is not an indicator of L. If there are no two other nodes from the same cluster of X in $G_{candidate}$, then the only valid constraints with X are those referring to three or more set of indicators from another cluster, which can be at most as many as those referring to P_i. Since P_i participates in a constraint that cannot be mapped to X, this latter node would be removed in place of P_i.

Now suppose there are no triplets $\mathbf{X} = \{X_1, X_2, X_3\}$ of indicators that are not indicators of L such that there is no intra-construct impurities in the set. Then the only valid constraints where X_j can possibly participate are those containing three or more indicators of L. But our mapping procedure can count as many equivalent constraints containing P_i. On the other hand, we cannot substitute P_i by X_j in all tetrad constraints (correctly) entailed by the set $\{P_i, V_a, V_b, K\}$, unless $X_j = K$. But if there is no $X_j \neq K$ in $G_{candidate}$, then the *MMScore* of this graph will be zero and no node will be removed. Therefore, X_j should be removed instead of P_i, contrary to our assumptions.

So, we must have some triplet $\mathbf{X} = \{X_1, X_2, X_3\}$ that are not indicators of L and where there are no intra-construct impurities. If there is some triplet such that V_a is cross-construct impure with some element, then some tetrad constraint entailed by the set containing V_a and this triple will be invalid, while mapping this constraint by substituting V_a for P_i will not entail any invalid tetrad constraint. Therefore, V_a should be cross-construct pure with respect to every other node in $G_{candidate}$.

Now suppose V_a is latent-measure impure with respect to some triplet $\mathbf{X}$. There cannot be any other $V_b \neq V_a$ such that V_b is also a latent-measure impurity with respect to the same triplet, or otherwise they will have correlated errors, which is equivalent of being intra-construct impure. The set $\{P_i, P_j, V_b, X_t\}$ entail only valid tetrad constraints, but the set $\{V_a, P_j, V_b, X_t\}$ does not. Therefore, V_a must be latent-measure pure with respect to every triplet. But that means that V_a is relatively pure in the graph $G_{candidate}$, which contradicts our initial assumption of having only two such nodes. □

The following lemma complements the previous one. While at Lemma 3 we start assuming that we have already the purified graph, in the next one we start assuming we have a specific group of variables that will lead to such final purified graph.

Lemma 4 *Let $G_{candidate}$ be a one-factor model with more than three indicators and at least one group of relatively pure nodes of size three or more. The purification of $G_{candidate}$ will lead to a graph with $n \geq 4$ indicators and a $MMScore$ of zero, where $n-1$ of the observed variables are indicators of the same latent L in G_T.*

Proof of Lemma 4:

Suppose there is at least one group of three relatively pure measures and a fourth variable. Then there is a set of four variables (which here we will call *foursome*) that correctly entails at least one tetrad constraint. If a variable is not part of a foursome, it will not appear in any valid constraint. Therefore, at this point all variables that do not participate in any foursome will be eliminated. After that, all variables that remain will have to be part of some foursome, but impurities may still exist. Suppose one of the remaining variables, V, was chosen to be eliminated at some point of the purification. At least five variables should exist in the graph, or its $MMScore$ would be zero. Three of the remaining variables, $\{V_1, V_2, V_3\}$, belonged to a same foursome as V. Let X a fifth variable in $G_{candidate}$ that is not in $\{V_1, V_2, V_3, V\}$. Suppose V belonged to all foursomes that had X. It is easy to see that V participates in at least as many valid constraints as X. Since V forms a foursome with $\{V_1, V_2, V_3\}$ and X does not, then X will entail less valid constraints than V, which is a contradiction. Therefore, X belonged to some foursome that did not include V. In this case, there will still be a foursome in $G_{candidate}$ after we remove V, and all variables that are not part of foursomes will have to be eliminated again. This iteration will continue till we achieve a score of zero. A score of zero cannot be achieved if there is some impurity or at least two variables that belong to one cluster in G_T and two variables that belong to another cluster in G_T (there will be invalid tetrad constraints entailed $G_{candidate}$). Therefore, we should not have any impurity in the remaining n variables, and at most one variable that is not an indicator of the same latent as the other $n-1$ variables. □

The next lemma covers the other major case of graph used as input to Purify:

Lemma 5 *Let $G_{candidate}$ be a pure two-factor latent graph with latents T_1 and T_2, each having a set of indicators $\mathbf{I_1}$ and $\mathbf{I_2}$, respectively. Let $\mathbf{I_1} = \mathbf{P} \cup \mathbf{Q}$, where $\mathbf{P}$ is a set of pure indicators of some latent L_1 in G_T, $|\mathbf{P}| \geq 3$, and $\mathbf{Q}$ is a possibly empty set of indicators of L_1 that may be impure with respect to some element of $\mathbf{I_2}$. $\mathbf{Q}$ may also include at most one indicator of a latent in G_T that is not L_1. Finally, let $\mathbf{I_2}$ be a set of indicators of some latent L_2 in G_T possibly equal to L_1. The purification of $G_{candidate}$ will not remove any element of $\mathbf{P}$.*

Proof of Lemma 5:

In the purification of a two-factor graph, we build a one-factor auxiliar graph G' with a set of indicators $\mathbf{I'} = \mathbf{I_1} \cup \{I_0\}$, where $I_0 \in \mathbf{I_2}$. If some element of in $\mathbf{I_1}$ is impure with respect to I_0, according to Lemma 1 it will be removed instead of some pure indicator. If there are no impurities in G' and there is an element V in $\mathbf{I_1}$ that is not an indicator of L_1, it will imply more invalid constraints that any other indicator in the graph, since all invalid constraints entailed by G' include a reference to this element and all elements in $\mathbf{I}'$ participate in the same number of constraints; there are valid constraints (e.g., tetrad constraints with set of four indicators that does not include V) entailed by this graph. Other occasion where a node in $\mathbf{I_1}$ can be removed is in the last stage of this purification, where the two-factor graph in evaluated as a whole. But at this stage the only possible invalid constraints that can hold are those that include misplaced indicators both in $\mathbf{I_1}$ and $\mathbf{I_1}$, i.e., all invalid constraints will containg both misplaced indicators, while it is trivial to see there are valid constraints with a pure indicator P of $\mathbf{I_1}$ that cannot be mapped to a valid constraint containing references to this misplaced measure instead of P. In any situation, no pure indicator of L will be removed. □

Finally, the following theorem states the correctness of Washdown.

Theorem 1 *Given a set of random variables* $\mathbf{V}$ *for which there exists a subset whose marginal is generated by some almost pure linear latent model* G_T *with at least three pure indicators per latent and is parametrically described by the covariance matrix, and given a score function that tells which constraints relating elements of this covariance matrix are not satisfied, Washdown will return a correct almost pure linear model over a subset of* $\mathbf{V}$ *include all the latents.*

Proof of Theorem 1:

We will prove correctness by induction, first estabilishing the base case. `Washdown` starts by purifying a model composed of one latent and all observed variables attached as indicators. By the assumptions and Lemmas 3 and 4, some cluster T_1 with at least three pure indicators will remain. We do not know if there is an extra misplaced indicator in T_1, since a linear one-factor model of n variables with $n-1$ variables being pure indicators of a single latent, and one variable that is a pure indicator of another latent, is equivalent to a model with n variables coming from a single indicator. Let $G^{(2)}$ be the one-factor graph with indicators given by the initial observed variables that were discarded in the first purification.

Our induction step is as follows. Let $\mathbf{T} = \{T_1, ..., T_n\}$ be a given ordered list where: every cluster T_i contains at least three pure measurements of a given latent in G_T; there are no two clusters in $\mathbf{T}$ whose combined set of indicators are all indicators of a same latent in G_T; there are no intra-construct impurities within each cluster, and no inter-construct or latent-measured impurities between any pair of clusters in $\mathbf{T}$; all indicators in a specific cluster T_i in $\{T_1, ..., T_{n-1}\}$ are indicators of the same latent, and all the indicators, with the possible exception of one, in T_n are indicators of a same latent in G_T. Let $G^{(n+1)}$ be the non-empty current graph of variables that were not allocated in any T_i cluster in $\mathbf{T}$. Let $T_{(n+1)}$ the the one-factor model resulting from the purification of $G^{(n+1)}$. Let's use the notation $L(T_i)$ to indicate the latent in G_T which is the latent parent of the majority of indicators of T_i in the true graph.

Suppose first that $T_{(n+1)}$ is not a saturated model (i.e., it has four or more variables) and its $MMScore$ is zero. It means that is has at least three variables that fits a one-latent pure measurement model. A call to `PurifyTwoClusters` with T_n and $T_{(n+1)}$ as first and second arguments is then performed. Since T_n contains at least three pure measures, by Lemma 5 none of its pure measures will be thrown away. If there are any cross-construct impurities between elements of T_n and $T_{(n+1)}$, it will be detected when the corresponding variable in $T_{(n+1)}$ is tested against T_n. The same will happen for any intra-construct impurity that may exist if there is an indicator in $T_{(n+1)}$ that happens to be an impure indicator of $L(T_n)$.

If $L(T_{(n+1)}) = L(T_n)$, and there is one misplaced indicator in T_n (i.e., it is not an indicator of the same latent as the majority of the elements in T_n, which we know there will be at most one), the test of creating one-factor models with elements of T_n and one element of $T_{(n+1)}$ may not detect it, and the same will happen if there is a misplaced indicator in $T_{(n+1)}$. If there are misplaced indicators in both clusters, it is trivial to see that the final test of the two-factor model will detect and remove one (and only one) of them. In any case, after merging these two clusters in the consecutive step of `Washdown`, we keep the property that the last member of $\mathbf{T}$ contains at most one misplaced indicator. Also, any latent-measured impurity that may exist in $T_{(n+1)}$ with respect to T_n will be detected when we test the one-factor model with $T_{(n+1)}$ and one indicator of T_n.

If $L(T_{(n+1)})$ is the same latent parent of the (possible) misplaced indicator in T_n, the test of building a factor model with one element from $T_{(n+1)}$ will detect it, since we still can count more valid constraints among the other elements in T_n than with the misplaced indicator (in a reasoning analogous to other lemmas in this proof), it will be washed down. Trivially, this indicator will also be removed if $L(T_{(n+1)})$ is different from $L(T_n)$ or from the latent parent of the misplaced indicator.

If there is a misplaced indicator in $T_{(n+1)}$, the same reasoning will follow, guaranteeing that at most one misplaced indicator will remain in $T_{(n+1)}$, or $T_{(n+1)}$ will be merged with T_n. By iterating throught all other clusters in $\mathbf{T}$, we guarantee that no impurities between any pair of these clusters will hold after puryfing. By the merging procedure, we keep our property that no two clusters T_i and T_j will have the property $L(T_i) = L(T_j)$.

Now, we have also to guarantee that $\mathbf{T}$ keeps its property that every cluster should have at least three pure measures. Suppose that no indicators in $T_{(n+1)}$ manages to have some impure relation with any of the indicators in $\mathbf{T}$, but there is no set of three pure measures in $T_{(n+1)}$, i.e., all indicators in $T_{(n+1)}$ (with the possible exception of one or two) are impure with respect to some other variable in G_T which is possibly not in $\mathbf{T}$ yet. If we obtain a graph $T_{(n+1)}$ without three pure measures of $L(T_{(n+1)})$ but still not saturated and with a zero $MMScore$, one of the following events must have taken place:

- if $G^{(n+1)}$ had three pure measures of $L(T_{(n+1)})$, they should have appeared in $T_{(n+1)}$, according to Lemma 3. Since $G^{(n+1)}$ did not have three pure measures of $L(T_{(n+1)})$, it may be the case that there is some $T_i \in \mathbf{T}$ such that $L(T_i) = L(T_{(n+1)})$. In this situation, the merging procedure will eliminate $T_{(n+1)}$ by merging its indicators with T_i;
- now suppose that $G^{(n+1)}$ did not have three pure measures of $L(T_{(n+1)})$, and there is no $T_i \in \mathbf{T}$ such that $L(T_i) = L(T_{(n+1)})$. Since by assumption there should be at least three pure indicators of $L(T_{(n+1)})$ and no more than two are in G_T, the only place where one of the indicators should be is as the misplaced indicator of T_n. In this case, $G^{(n+1)}$ had two pure indicators of $L(T_{(n+1)})$, and by Lemma 3, $T_{(n+1)}$ will have these two indicators. After his first stage of `PurifyTwoClusters`, the misplaced indicator in T_n will be washed down back to $T_{(n+1)}$, and the result follows.

Now, suppose that $T_{(n+1)}$ has four indicators, but its $MMScore$ is different from zero. By Lemma 4, $G_{(n+1)}$ cannot have any subgraph that has three relatively pure indicators. It also means that there is at most one more cluster to be identified, because if there were two, by the arguments above, one had to have at least three measures in $G_{(n+1)}$ (since at least one of the two will have all of its indicators). At this point, at the first stage of `PurifyTwoClusters`, the algorithm will washdown any impurity or misplaced indicator of T_n back to $T_{(n+1)}$, which now was forced by the algorithm to contain all nodes from $G_{(n+1)}$. If there was a cluster left to be identified, now $T_{(n+1)}$ will have three pure indicators. Since there is now only one group of three relatively pure indicators, they will be kept, while all the others will be discarded and the algorithm terminated after we purify $T_{(n+1)}$ with respect to every other cluster in $\mathbf{T}$. If there was no cluster left to be identified, $T_{(n+1)}$ will get to the point of having only two indicators, which will be correctly discarded.

Now, suppose that $T_{(n+1)}$ has less than four indicators. If there is a cluster yet to be identified, certainly its two remaining indicators will be in $T_{(n+1)}$ and the same argument of the previous paragraph will follow. If there is no cluster to be identified, again these measures will be discarded or merged.

Finally, whenever we compare $T_{(n+1)}$ with T_i, $i < n$, if the only measures that can discarded from T_i are impure measures, which do not need to be washed down. The algorithm converges because, at every step, the number of elements in $G_{(n+1)}$ is less than in $G_{(n-1)}$ (since just a subset of $T_n \cup T_{(n+1)}$ will be washed down), which eventually will reach zero. $\Box$

3 Implementation issues

With one-factor models, it is possible to apply the procedure designed by Kano and Harada (2000) to fastly perform stepwise variable selection with a goodness-of-fitness criterion. It is not straightforward to change this algorithm to the case where latents are arbitrarily correlated, at least not in a sense where it will give clear computational advantages. However, as it can be seen in the description of the algorithm, we hardly need to perform a search of a node to be removed in a two-factor model. It is only necessary when the majority of the indicators (i.e., all but one) in each cluster are actually indicators of a single cluster in the true graph. And ideally we just need to remove one indicator to get a valid model.

Another good feature that could be explored is the fact that we have not only independent latents most of the time: we have just one latent. Also, during the `PurifyTwoClusters` procedure, we could also use techniques to fastly evaluate how to score a model when we add just one more latent.

4 Discussion

I did not make a formal connection of the idealized *MMScore* function and score functions such as χ^2: I am just assuming the intuitive concept of the χ^2 score as a test penalizing overconstrained models. There are still some other pending issues with χ^2 such as calculating the correct number of degrees of freedom for latent models, which as far as I know is still an open research question.

Also, it seems there is nothing in the definition of *MMScore* that precludes it from being a mere count of wrongly entailed tetrad constraints, without worrying about any other kind of constraints. This can lead to new variations and heuristics of this procedure.

References

Kano, Y and Harada, A. (2000) "Stepwise variable selection in factor analysis". *Psychometrika* 65(1), 7-22.

Silva, R. (2002) "The structure of the unobserved". Technical report CMU-CALD-02-102, Carnegie Mellon University, Pittsburgh, PA.