DISCUSSION PAPER SERIES

IZA DP No. 1042

Sequential Matching Estimation of Dynamic Causal Models

Michael Lechner

March 2004

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Sequential Matching Estimation of Dynamic Causal Models

Michael Lechner
SIAW, University of St. Gallen,
CEPR, ZEW and IZA Bonn

Discussion Paper No. 1042
March 2004

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

IZA Discussion Paper No. 1042
March 2004

ABSTRACT

Sequential Matching Estimation of Dynamic Causal Models[*]

This paper proposes sequential matching and inverse selection probability weighting to estimate dynamic causal effects. The sequential matching estimators extend simple, matching estimators based on propensity scores for static causal analysis that have been frequently applied in the evaluation literature. A Monte Carlo study shows that the suggested estimators perform well in small and medium size samples. Based on the application of the sequential matching estimators to an empirical problem - an evaluation study of the Swiss active labour market policies - some implementational issues are discussed and results are provided.

JEL Classification: C40

Keywords: dynamic treatment effects, nonparametric identification, causal effects, sequential randomisation, programme evaluation, panel data

Corresponding author:

Michael Lechner
Swiss Institute for International Economics and Applied Economic Research (SIAW)
University of St. Gallen
Dufourstr. 48
9000 St. Gallen
Switzerland
Email: Michael.Lechner@unisg.ch

[*] Financial support from the Swiss National Science Foundation (projects 4043-058311 and 4045-050673) is gratefully acknowledged. Part of the data originated from a database generated for the evaluation of the Swiss active labour market policy together with Michael Gerfin. I am grateful to the Swiss State Secretariat for Economic Affairs (*seco; Arbeitsmarktstatistik*) and the Swiss Federal Office for Social Security (*BSV*) for providing the data and to Dragana Djurdjevic, Michael Gerfin and Heidi Steiger for their substantial input in preparing them. I thank Ruth Miquel for helpful comments on several issues raised in the paper. The paper has been presented in seminars at Cambridge University, Mannheim University and the Tinbergen Institute, Amsterdam. I thank participants for helpful comments. The usual disclaimer applies.

1 Introduction

The sample selection problem in its static version has received considerable attention in the microeconometrics and statistics literature concerned with uncovering causal effects of interventions (e.g. Heckman, 1979, Heckman and Robb, 1985, Holland, 1986, Roy, 1951, Rubin, 1973, 1974, 1977). The fundamental problem is that we desire a comparison of two (or more) different states of the world (with one outcome for each state) and have to perform this comparison using the units (individual, firms, etc.) that actually are observed in that state. If there are factors that jointly influence selection into the different states and the variables used to measure the (causal) effect of being in one state or the other, then a (unadjusted) comparison of means of these outcome variables in the different states do not estimate causal effects. In this case different methods of adjustment may be used to recover the causal effects. Which of those methods is appropriate depends on the specific 'nature' of the connection of the outcome process to the selection process that may differ from one application to the other. The surveys by Angrist and Krueger (1999) and Heckman, LaLonde, and Smith (1999) give comprehensive overviews.

The 'good data case' (all variables that jointly influence selection and outcomes are observable) received considerable attention in applied as well as methodological studies. This case is at issue in this paper as well. In particular matching methods, which implicitly or explicitly form comparison groups to adjust for the difference in observable characteristics related to the selection process, are popular (e.g. Rubin, 1973, Rosenbaum and Rubin, 1985, Deheija and Wahba, 1999, 2002, Heckman, Lalonde, and Smith, 1999, Smith and Todd, 2001, Lechner, 1999, 2002a). There are recent advances to improve the understanding of the asymptotic properties of various estimators used in applications (Abadie and Imbens, 2002, Hahn, 1998, Heckman, Ichimura, and Todd, 1997, Heckman, Ichimura, Smith, and Todd, 1998, Hirano, Imbens, and Ridder, 2003, Ichimura and Linton, 2001, among others).

However, the static model may not be able to address all selection issues that occur in applications. Suppose, for example, one is interested in the effect on female labour supply of giving birth to two children (sequence 2) compared to giving no birth (sequence 1) in a given period of time. Apparently, selection occurs with respect to the first and second fertility decision (ignore twins). However, the second decision may well depend on the effect of the first birth on subsequent labour supply (an intermediate outcome). Since even in the ‘good data case’ selection cannot be ‘controlled’ for in the beginning of the sequence (because plans may be changed depending on the intermediate outcomes), the static model is not flexible enough to handle selection biases in such situations. Similar problems occur in other fields, too, for example, when evaluating the effects of sequences of training programmes.

There are several ways of handling such dynamic selection problems in observational studies. When the outcome variable is a duration variable, like months of unemployment, Abbring and van den Berg (2003) suggest modelling the hazard rate into unemployment with a selectivity correction. However, duration modelling has the disadvantage that usually a key identifying assumption is the multiplicative proportional hazard rate condition that is in many cases hard to motivate from substantive considerations about the selection process. Furthermore, taking account of intermediate outcomes (others than those coming directly from duration dependence) is not trivial.[1]

Other candidates of parametric models that allow intermediate outcomes to some extent are classical dynamic panel data models with sample selection (e.g. the survey by Arellano and Honoré, 2001). Besides the sensitivity with respect to misspecification, another issue is that such models provide only indirect estimates of the causal effects, and in many cases the connection between the estimated coefficient and the causal parameters of interest is not straightforward.[2]

Another approach to dynamic selection problems is to ignore intermediate outcomes and treat the sequence participation as being determined from the start. The problem is considerably simplified. Thus, this approach allows using the estimation methods available for the static causal model. For example, Arulampalam and Booth (2001) analyse the effects of multiple training events within 10 years by modelling them

[1] For the use of duration models analysing experiments see Ham and LaLonde (1996).

[2] The work by Heckman and Robb (1985) and Heckman and Hotz (1989) address the issue of how to use panel data to correct for selection effects and thus identify causal effects. However, with respect to the definition of the treatments their approach is static.

as outcomes of a count data process. In such a process there is no role for any intermediate outcomes to determine selection.

Much of the literature on (static) causal effects in the ‘good data case’ is devoted to reduce the role of functional form assumptions and to leave the heterogeneity of the effects across units unrestricted. To exploit these advantages of the static evaluation literature for the dynamic selection problem some papers ‘twist’ the static causal model by defining effects in a way that avoids most of the dynamic selection problems. Then standard matching techniques are used for estimation. For example, evaluating programmes of the active labour market policy Gerfin and Lechner (2002) circumvent the problem of a single individual participating in several programmes sequentially by estimating the effect of the first programme only. Similarly, Li, Propert, and Rosenbaum (2001) and Sianesi (2001) are estimating the effect of a delay of a treatment. Both papers are based on matching approaches for binary treatments. Although the actual implementation of the matching algorithms used are fairly different, both papers estimate the effect of ‘waiting’ by matching those people who at a given time ‘face the risk’ of treatment but do not participate to those joining the treatment, and then averaging over the distribution of the start dates of the treatment. However, ‘waiting’ essentially defines a dynamic treatment sequence if ‘leaving the queue’ is allowed during waiting. Bergemann, Fitzenberger, and Speckesser (2001) analyse the effect of programme sequences using a combination of a matching and difference-in-difference estimator. Again, their approach consists in sequentially using the static evaluation model. Since the causal effects of interest cannot be formally defined using the static framework, the conditions required for identification remain unclear.[3]

Robins (1986) suggests an explicitly dynamic causal framework that is subsequently used in applications in epidemiology and biostatistics. Although he focuses on specific sequences and identifying conditions that would be unconventional in econometrics (as well as using a parametric estimation framework), he seems to be the first who explicitly formalises causal effects of dynamic sequences using potential out-

[3] Recently, Miquel (2003) discussed identification by IV and differences-in-differences in an explicitly dynamic model of treatment sequences.

comes and allowing for intermediate outcomes to determine the next state of the sequence. His approach is extended by Gill and Robins (2001) to the case of continuous covariates and treatments.

Recently, Lechner and Miquel (2001, LM01 further on) extend Robins' (1986) framework to allow comparisons of more general sequences and selection processes. They establish notation and assumptions that are more common in econometrics. Focussing on the 'good data case' (selection on observables, including intermediate outcomes), LM01 discuss identification conditions, denoting them as the weak and strong dynamic conditional independence assumptions (W-DCIA, S-DCIA). They show that under S-DCIA no specific problems arise. For W-DCIA the endogeneity problem stemming from the endogeneity of intermediate outcomes leads to a loss of identification even if all selection variables are observable, but interesting causal parameters are still identified.

This paper proposes estimators for the model of LM01 that retain the flexible properties of the estimators commonly used in the static evaluation literature, namely that they are robust to functional form assumptions, do not restrict effect heterogeneity, and are fairly easy to compute. They are a sequential version of matching-on-the-propensity score estimators. Furthermore, a particular variant of inverse probability weighted estimators is discussed as well. In a Monte Carlo study some finite sample properties of these estimators are exemplified and compared to other variants of matching estimators. Furthermore, they are applied to the evaluation of the Swiss active labour market programmes. This empirical part illustrates several issues concerning the implementation of sequential matching estimators.

Section 2 outlines the dynamic causal framework suggested by LM01: The notation is introduced and the basic identification conditions are restated. To focus ideas the basic model only is presented (an initial period, two subsequent periods, and two treatments each period). In Section 3 the estimation problem is explained and sequential matching as well as sequential inverse probability weighted estimators are proposed. The Monte Carlo study in Section 4 contains simulations for two data generating processes, one fulfilling S-DCIA and one fulfilling W-DCIA only. Section 5 presents the application and Section 6 concludes. Appendix A addresses issues that arise with sequential propensity score methods with multiple treatments as used in the empirical part. Appendix B details the sequential matching protocol used in the application and Appendix C discusses some properties of the inverse-probability weighted estimators.

2 The dynamic causal model

This section introduces the dynamic causal model proposed by LM01 and rephrases their identification conditions based on sequential selection on observables. Since it is sufficient to use a three-periods-two-treatments model to discuss all relevant issues that distinguish the dynamic from the static model, this section shows the results only for this basic version of the model.[4]

2.1 The variables and the definition of the effects to be estimated

Time periods indexed by *t* and τ $(t, \tau \in \{0,1,2\})$. The vector of random variables $S = (S_0, S_1, S_2)$ describes the treatment received by a member of the population.[5] In period 0 everybody receives the same treatment, i.e. is in the same state $S_0 = 0$. From period 1 S_t can take two values. A particular realisation of S_t is denoted by $s_t \in \{0,1\}$. Furthermore, denote the history of variables up to period *t* by a bar below a variable, e.g. $\underline{s}_2 = (s_1, s_2)$.[6] In period 1 a member of the population can be observed in exactly one of two treatments (0, 1). In period 2 she participates in one of four treatment sequences $((0,0),(1,0),(0,1),(1,1))$. Therefore, every individual 'belongs' to exactly one 'short' sequence defined by s_1 and another 'long' sequence defined by $\underline{s}_2$. To sum up, in the three-periods-two-treatments example LM01 consider 6 different overlapping potential outcomes corresponding to 2 mutually exclusive states

[4] For the finite-number-of-periods-finite-number-of-treatments model the reader is referred to LM01 instead.

[5] The notation follows the spirit of Rubin (1974) and Robins (1986). The terms *treatment* and *state* are used interchangeably. In the following members of the population are sometimes called *individuals* for simplicity. Capital letters usually denote random variables and small letters denote specific values of the random variable.

[6] To differentiate between different sequences a letter (e.g. *j*) may be used to index a sequence like $\underline{s}_t^j$. Furthermore, the initial period is ignored when denoting different sequences.

defined by treatment status in period 1, plus 4 mutually exclusive states defined by treatment status in period 1 and 2 together.

Variables used to measure the causal effects of the treatment, i.e. the potential outcomes, are indexed by treatments and denoted by $Y^{\underline{s}_t} = (Y_0^{\underline{s}_t}, Y_1^{\underline{s}_t}, Y_2^{\underline{s}_t})$. Potential outcomes are measured at the end of each period, treatment status is measured in the beginning of each period. For each length of a sequence, one of the potential outcomes is observable and denoted by Y_t. The observation rules are defined in equation (1):

$$Y_t = S_1 Y_t^1 + (1 - S_1) Y_t^0 = S_1 S_2 Y_t^{1,1} + S_1 (1 - S_2) Y_t^{1,0} + (1 - S_1) S_2 Y_t^{0,1} + (1 - S_1)(1 - S_2) Y_t^{0,0} . \qquad (1)$$

Next, variables that may influence treatment selection and potential outcomes, often called *attributes* or *confounders*, are defined and denoted by *X*. Because treatment status may influence the realisations of these variables (introducing some *endogeneity*), there are potential values of these variables as well ($X^{s_1} = (X_0^{s_1}, X_1^{s_1})$). $X_1^{s_1}$ may contain $Y_1^{s_1}$ or functions of it. The *K* dimensional vector X_t is observable at the same time as Y_t (thus it is observed only after the selection S_t is realised[7]). The observation rule for X_t is analogous to the one for the potential outcomes given in equation (1).

Interest is in the estimation of the mean causal effect of a sequence of treatments defined up to period 2 ($\underline{s}_2^k$) compared to another sequence of the same length ($\underline{s}_2^l$) for a particular population and for outcomes of period 2 (or later). This effect is denoted by $\theta_2^{\underline{s}_2^k, \underline{s}_2^l}$.[8] $\theta_2^{\underline{s}_2^k, \underline{s}_2^l}$ may be of interest for several subpopulations: Mean causal effects can be constructed for subpopulations defined by variables not influenced by the treatment, because causal statements that are conditional on the effects of the treatments usually do not have useful interpretations. Thus, conditioning the effects on *X* requires 'exogeneity' assumptions that are discussed below. However, conditioning on treatment status is interesting, because it allows similar com-

[7] Therefore, there is no role for X_2 in a two period model.

[8] Causal effects of sequences of length 1 can be analysed within the static model of potential outcomes.

parisons like comparing the well-known effects of treatment on the treated (ATET) and treatment on the nontreated of the static framework. The definition of such average causal effects is given in equation (2):

$$\theta_2^{\underline{s}_2^k,\underline{s}_2^l}(\underline{s}_\tau^j) := E(Y_2^{\underline{s}_2^k} \mid \underline{S}_\tau = \underline{s}_\tau^j) - E(Y_2^{\underline{s}_2^l} \mid \underline{S}_\tau = \underline{s}_\tau^j),\ 0 \le \tau \le 2,\ k \ne l,\ k,l \in \{1,2,3,4\},\ j \in \{1,...,2^\tau\}. \quad (2)$$

To consider $\theta_2^{\underline{s}_2^k,\underline{s}_2^l}(\underline{s}_\tau^j)$ a causal effect the standard assumptions of the Rubin (1974) model are necessary, like the Stable Unit Treatment Value Assumption (SUTVA), implying that the effects of treatment on one person do not depend on the treatment choices of others. $\theta_2^{\underline{s}_2^k,\underline{s}_2^l}$ $[= \theta_2^{\underline{s}_2^k,\underline{s}_2^l}(s_0 = 0)]$ is called the dynamic average treatment effect (DATE). Accordingly, $\theta_2^{\underline{s}_2^k,\underline{s}_2^l}(\underline{s}_2^k)$ is termed DATE on the treated (DATET). Note that there are cases in-between, like $\theta_2^{\underline{s}_2^k,\underline{s}_2^l}(\underline{s}_1^l)$, when the population is defined by participating in a sequence shorter than the one evaluated. Furthermore, the effects are symmetric when defined for the same population in the sense of $\theta_2^{\underline{s}_2^k,\underline{s}_2^l}(\underline{s}_\tau^j) = -\theta_2^{\underline{s}_2^l,\underline{s}_2^k}(\underline{s}_\tau^j)$, but that $\theta_2^{\underline{s}_2^k,\underline{s}_2^l}(\underline{s}_\tau^k) \ne -\theta_2^{\underline{s}_2^l,\underline{s}_2^k}(\underline{s}_\tau^l)$, otherwise. Table 1 summarises the notation introduced so far.

Table 1: Summary of notation and definitions

Symbol	Meaning	Timing within period
$t = 0,1,2$	time periods	--
$S = (0, S_1, S_2)$	RV: treatment	beginning
s_1, $\underline{s}_2 = (s_1, s_2)$	specific sequence of treatments until period 1 or 2	beginning
$s_t \in \{0,1\}$	*2* exclusive treatments in each period	beginning
$Y^{\underline{s}_t} = (Y_1^{\underline{s}_t}, Y_2^{\underline{s}_t})$	RV: potential outcomes	end
$Y = (Y_1, Y_2)$	RV: observable outcomes	end
$X^{\underline{s}_t} = (X_0^{\underline{s}_t}, X_1^{\underline{s}_t})$	RV: potential confounders	end
$X = (X_0, X_1)$	RV: observable confounders	end
$\theta_2^{\underline{s}_2^k,\underline{s}_2^l}(\underline{s}_\tau^j)$	mean causal effect of $\underline{s}_2^k$ compared to $\underline{s}_2^l$ for those participating in $\underline{s}_\tau^j$	end

RV: Random variable.

2.2 Identification of the effects of dynamic treatments

Suppose there is an infinitely large random sample ($\{s_{1i}, s_{2i}, x_{0i}, x_{1i}, y_{1i}, y_{2i}\}_{i=1,\ldots,N}$) from the population $S_0 = 0$ that is defined by the corresponding random variables (S_1, S_2, X_0, X_1, Y_1, Y_2). LM01 explore the identifying power of assumptions that may be termed 'good data', 'selection on observables' or 'conditional independence' assumptions. More precisely, they assume that out of the variables that determine treatment status in each period the sample contains those that are related to the relevant potential outcomes. If they are not influenced by the treatment in a static model conditioning the observed outcomes on these variables removes all selection bias (see Rubin, 1977). The difference between the static model and the dynamic causal model is that the latter allows for a second type of selection bias, because intermediate outcomes might influence the decision to continue a sequence, in other words, the intermediate outcomes may not be considered exogenous anymore.

Assume that in the beginning of each period the researcher has sufficient knowledge to assume that assignment to treatment is independent of potential outcomes conditional on that information.[9] This information may be influenced by intermediate outcomes. Assumption W-DCIA of LM01 formalises what LM01 call the WEAK DYNAMIC CONDITIONAL INDEPENDENCE ASSUMPTION.[10]

Assumption W-DCIA (weak dynamic conditional independence assumption)

a) $Y_2^{0,0}, Y_2^{1,0}, Y_2^{0,1}, Y_2^{1,1} \amalg S_2 \mid S_1 = s_1, \underline{X}_1 = \underline{x}_1$;[11]

[9] The following assumptions relate to identification of all treatment effects defined in Section 2. Hence, all potential outcomes are involved. If the comparison desired involves fewer potential outcomes, or is based on a subpopulation defined by a specific treatment sequence of interest, then fewer potential outcomes need to appear in the following assumptions. For the sake of brevity, we do not mention this issue below anymore, but the required changes will be obvious (as they are in the static framework).

[10] Note that the assumptions WDCIA and SDCIA (below) are somewhat stronger than necessary, but have the virtue of being easily explained in terms relating to the selection process (see LM 04 for details).

[11] $A \amalg B \mid C = c$ means that *each element* of the vector of random variables *B* is independent of the random variable *A* conditional on the random variable C taking a value of c. $(A) \amalg B \mid C = c$ means that the *joint distribution* of the elements of *A* is independent of *B* conditional on *C* = *c*.

b) $Y_2^{0,0}, Y_2^{1,0}, Y_2^{0,1}, Y_2^{1,1} \coprod S_1 \mid X_0 = x_0$;

c) $1 > P(S_1 = 1 \mid X_0 = x_0) > 0$, $1 > P(S_2 = 1 \mid \underline{X}_1 = \underline{x}_1, S_1 = s_1) > 0$; $\forall \underline{x}_1 \in \underline{\chi}_1$, $\forall s_1 : s_1 \in \{0,1\}$.

Part a) states that conditional on the previous treatment, observable outcomes and confounding variables, the potential outcomes are independent of selection in period 2 (S_2). Part b) states that conditional on some exogenous variables X_0 potential outcomes are independent of assignments in period 1 (S_1). These assumptions are valid for all values of x_0 and x_1 in a given set of interest $\underline{\chi}_1$. Part c) is the usual ***common support requirement*** (CSR), basically stating that all sequences to be evaluated must have a positive probability of occurring in all strata defined by the values of x_0 and x_1 that are in the set of interest $\underline{\chi}_1$. Obviously, only sequences that are feasible can be evaluated. If W-DCIA is combined with an initial condition for the confounding variables ($X_0 = X_0^0 = X_0^1 = X_0^{1,1} = X_0^{1,0} = X_0^{0,1} = X_0^{0,0}$) and some regularity, then THEOREM 1 of LM01 shows that $\theta_2^{\underline{s}_2^k, \underline{s}_2^l}$, $\theta_2^{\underline{s}_2^k, \underline{s}_2^l}(s_1^j)$ and $\theta_2^{(s_1^k, s_2^k),(s_1^k, s_2^l)}(s_1^k, s_2^j)$ ($\forall s_1^k, s_2^k$, s_1^l, s_2^l , $s_1^j, s_2^j \in \{0,1\}$) are identified. In plain words, pair-wise comparisons of all sequences are identified, but only for groups of individuals defined by their treatment status in period 1, or on average in the population. The relevant distinction between the populations defined by treatment state in the first and subsequent periods is that in the first period treatment choice is random conditional on exogenous variables (the result of the initial condition for S_0 , X_0), whereas in the second period, randomisation into these treatments is conditional on variables already influenced by the first part of the treatment. Specific comparisons are identified for populations defined by treatment status in both periods, if populations and treatment sequences are the same in the first period.

To identify the effects of two different sequences defined for a subpopulation given by treatment status in both periods, additional 'exogeneity' assumptions are required. The strong conditional independence assumption (S-DCIA) states that conditional on X_0 , knowing S_1 does not help to predict the potential outcomes given a value of the observed X_1. The direct and testable implication of this assumption is $X_1 \amalg S_1 \mid X_0 = x_0$. Hence, implies that the confounders are not Granger-caused by previous treatments.

Assumption S-DCIA (strong dynamic conditional independence assumption)

a) $Y_2^{0,0}, Y_2^{1,0}, Y_2^{0,1}, Y_2^{1,1} \amalg S_2 \mid \underline{S}_1 = s_1, \underline{X}_1 = \underline{x}_1$;

b) $Y_2^{0,0}, Y_2^{1,0}, Y_2^{0,1}, Y_2^{1,1}, (Y_2^{0,0}, X_1), (Y_2^{1,0}, X_1), (Y_2^{0,1}, X_1), (Y_2^{1,1}, X_1) \amalg S_1 \mid X_0 = x_0$;

c) $1 > P(S_2 = 1 \mid \underline{X}_1 = \underline{x}_1, S_1 = s_1) > 0$, $1 > P(S_1 = 1 \mid X_0 = x_0) > 0$; $\quad \forall \underline{x}_1 \in \underline{\chi}_1$, $\forall s_1 : s_1 \in \{0,1\}$.

ASSUMPTION S-DCIA allows for outcomes of previous treatments (*predetermined endogenous* variables) to appear in the conditioning set, and is still strong enough to identify all effects. However, these variables have to fulfil a strong exogeneity requirement in the sense that if they are influenced by the treatment in period 1, this influence must not have an impact on the potential outcomes of interest in period 2.

3 Estimation

Presuming identification by ASSUMPTIONS W-DCIA or S-DCIA this section discusses the general structure of the estimation problem and proposes matching estimators that are fairly close to the estimators frequently used in static evaluation studies based on models with selection on observables (e.g. Dehejia and Wahba, 1999, 2002, Heckman, LaLonde, and Smith, 1999, Smith and Todd, 2001, Lechner, 1999, 2002a). Estimators based on (sequential) reweighting using the inverse selection probabilities are discussed as well. It is shown how the general structure of adjustments to the estimators of the static model looks like. The focus on those classes of estimators is for simplicity only. All the usual estimators available could be adjusted to the dynamic context in the same way the sequential matching and inverse probability weighted estimators are adjusted.[12]

[12] The so-called regression imputation estimators as discussed for example by Frölich (2001, 2004), Hahn (1998), and Heckman, Ichimura, and Todd (1998) are efficient but suffer from the unsolved problem of optimally choosing the ‘tuning’ parameters, like bandwidths in the nonparametric regression steps. The estimators proposed here avoid these problems.

This section concentrates on the crucial ingredients of the effects to be estimated, namely the respective counterfactual expectations and their relation to the observable outcomes. Given this connection, the construction of estimators for the effects is straightforward. The first subsection reviews estimation of counterfactuals that can be analysed within the usual static framework, whereas the second subsection addresses issues concerning counterfactuals that require a dynamic framework.

3.1 Same treatment and conditioning sequences in period 1 – a static problem

The complexity of the estimation problem depends on the similarity between the sequences used to index the potential outcomes and those defining the population of interest. If the sequences coincide ($E(Y_2^{s_1} \mid S_1 = s_1) = E(Y_2 \mid S_1 = s_1)$, $E(Y_2^{s_1,s_2} \mid S_1 = s_1, S_2 = s_2) = E(Y_2 \mid S_1 = s_1, S_2 = s_2)$), the sample mean in the corresponding subsample is a natural nonparametric estimator.[13]

The estimation of counterfactuals defined for sequences identical to the conditioning set in period 1 is a well-known estimation problem and is extensively discussed in the literature about static causal models. As shown by LM01 the typical estimands have the following form:

$$E[Y_2^{s_1,s_2} \mid S_1 = s_1, S_2 = 1 - s_2] = \underset{\underline{X}_1}{E}[E(Y_t \mid S_1 = s_1, S_2 = s_2, \underline{X}_1) \mid S_1 = s_1, S_2 = 1 - s_2] =$$

$$= \underset{p_2^{s2|s1}(\underline{X}_1)}{E}[E(Y_t \mid S_1 = s_1, S_2 = s_2, p_2^{s_2|s_1}(\underline{X}_1)) \mid S_1 = s_1, S_2 = 1 - s_2],$$

$$p_2^{s_2|s_1}(\underline{x}_1) := P(S_2 = s_2 \mid \underline{X}_1 = \underline{x}_1, S_1 = s_1),\ s_1, s_2 \in \{0,1\}. \qquad (3)$$

For this type of estimand with a binary treatment many different estimators are suggested in the literature and applied in empirical studies. Such estimators exploiting the conditional independence assumption (CIA) are termed *matching methods* in the survey by Heckman, LaLonde, and Smith (1999), because they are based on comparing different observations with different treatment status but similar values of

[13] Without loss of generality we consider the potential outcomes of period 2 only. However the outcome $Y_t^{s_1,s_2}$ may be evaluated in any period after the treatment (*t* > 2) as long as those potential outcomes do not influence S_2, S_1.

(X_0, X_1) or the respective conditional selection probabilities $p_2^{s_2|s_1}(\underline{x}_1)$.[14] Rosenbaum and Rubin (1983) introduced the frequently applied principle to condition on the conditional participation probabilities (so-called propensity scores) instead of the control variables directly as a way to reduce the dimension of the estimation problem. Abadie and Imbens (2002), Hahn (1998), Heckman, Ichimura, and Todd (1998), and Ichimura and Linton (2001) investigate the asymptotic distributions of different types of matching estimators. Hirano, Imbens, and Ridder (2003) and Hernan, Brumback, and Robins (2001) focus on the asymptotic distribution of particular variants of inverse probability-weighting estimators. Among others, Frölich (2001, 2004) and Smith and Todd (2001) discuss practical and small sample issues for different types of such estimators.[15]

3.2 Different treatment and conditioning sequences in period 1 – the dynamic problem

3.2.1 Relation between potential and observable outcomes

If the sequences of interest differ in the first period, then LM01 show that the estimand identified by S-W-DCIA is given by equation (4):

$$E(Y_2^{\underline{s}_2^k} \mid \underline{S}_1 = \underline{s}_1^j) = \underset{X_0}{E}\{\underset{X_1}{E}[E(Y_2 \mid \underline{S}_2 = \underline{s}_2^k, \underline{X}_1 = \underline{x}_1) \mid S_1 = s_1^k, X_0 = x_0] \mid S_1 = s_1^j\} =$$

$$= \underset{p_1^{s_1^k}(X_0)}{E}\{\underset{p_2^{s_2^k|s_1^k}(\underline{X}_1)}{E}[E(Y_2 \mid \underline{S}_2 = \underline{s}_2^k, \underline{p}_2^{s_2^k|s_1^k, s_1^k}(\underline{x}_1) \mid S_1 = s_1^k, p_1^{s_1^k}(X_0)] \mid S_1 = s_1^j\},$$

$$p_1^{s_1}(x_0) := P(S_1 = s_1 \mid X_0 = x_0),\ \underline{p}_2^{s_2^k|s_1^k, s_1}(\underline{x}_1) := [p_2^{s_2^k|s_1^k}(\underline{x}_1), p_1^{s_1}(x_0)],\ s_1^k, s_2^k, s_1^j, s_1 \in \{0,1\}. \quad (4)$$

[14] Note that the statistics literature uses the term *matching estimation* in a more restrictive way. It includes only estimators that actually form comparison groups based on similarity of treated and controls with respect to (X_0, X_1), or the corresponding probabilities (e.g. Rosenbaum and Rubin, 1983).

[15] For more details see the excellent survey by Imbens (2003).

The previous estimation principles can be applied here as well. However, the reweighting has to be performed sequentially. In a first stage a regression of Y_2 on $\underline{X}_1$ (or $\underline{p}_2^{s_2^k|s_1^k,s_1^k}(\underline{x}_1)$) in the subsample of $\underline{S}_2 = \underline{s}_2^k$ is performed, leading to $E(Y_2 | \underline{S}_2 = \underline{s}_2^k, \underline{X}_1 = \underline{x}_1)$ (or $E(Y_2 | \underline{S}_2 = \underline{s}_2^k, \underline{p}_2^{s_2^k|s_1^k,s_1^k}(\underline{X}_1))$. Within each strata of X_0 (or $p_1^{s_1^k}(x_0)$) in the subpopulation $S_1 = s_1^k$, this regression function is averaged according to the distribution of X_1 (or $p_2^{s_2^k|s_1^k}(\underline{x}_1)$) in each such stratum. These averages are functions of X_0 only. Finally, this function is averaged over the distribution of X_0 (or $p_1^{s_1^k}(x_0)$) in $S_1 = s_1^j$, leading to a sequential matching estimator to be discussed in greater detail later on.

Of course, if S-DCIA holds, then (4) holds as well, but S-DCIA also implies the following simpler expression:

$$E(Y_2^{\underline{s}_2^k} | \underline{S}_1 = \underline{s}_1^j) = \underset{X_0,X_1}{E}[E(Y_2 | \underline{S}_2 = \underline{s}_2^k, \underline{X}_1 = \underline{x}_1) | S_1 = s_1^j] =$$

$$= \underset{\underline{p}_2^{s_2^k|s_1^k,s_1^k}(\underline{X}_1)}{E}[E(Y_2 | \underline{S}_2 = \underline{s}_2^k, \underline{p}_2^{s_2^k|s_1^k,s_1^k}(\underline{x}_1) | S_1 = s_1^j\},$$

$$p_1^{s_1}(x_0) := P(S_1 = s_1 | X_0 = x_0),\ \underline{p}_2^{s_2^k|s_1^k,s_1}(\underline{x}_1) := [p_2^{s_2^k|s_1^k}(\underline{x}_1), p_1^{s_1}(x_0)],\ s_1^k, s_2^k, s_1^j, s_1 \in \{0,1\}. \quad (5)$$

This expression gives raise to a one matching procedure as in static models. In principle, comparing estimators based either on (4) or (5) can form the basis of a test for additional exogeneity assumptions implied by S-DCIA.

For the average effect in the population the same principle applies (equation (6)):

$$E(Y_2^{\underline{s}_2^k}) = \underset{X_0}{E}\{\underset{X_1}{E}[E(Y_2 | \underline{S}_2 = \underline{s}_2^k, \underline{X}_1 = \underline{x}_1) | S_1 = s_1^k, X_0 = x_0]\} =$$

$$= \underset{p_1^{s_1^k}(X_0)}{E}\{\underset{p_2^{s_2^k|s_1^k}(\underline{X}_1)}{E}[E(Y_2 | \underline{S}_2 = \underline{s}_2^k, \underline{p}_2^{s_2^k|s_1^k,s_1^k}(\underline{X}_1)) | S_1 = s_1^k, p_1^{s_1^k}(X_0)]\};\ s_1^k, s_2^k \in \{0,1\}. \quad (6)$$

Finally, when S-DCIA holds, $E(Y_2^{\underline{s}_2^k} \mid \underline{S}_2 = \underline{s}_2^j)$, $s_1^k \neq s_1^j$, is identified as well:

$$E(Y_2^{\underline{s}_2^k} \mid \underline{S}_2 = \underline{s}_2^j) = \underset{X_1, X_0}{E}\{E(Y_2 \mid \underline{S}_2 = \underline{s}_2^k, \underline{X}_1) \mid \underline{S}_2 = \underline{s}_2^j\} =$$

$$= \underset{P(\cdot)}{E}\{E[Y_2 \mid \underline{S}_2 = \underline{s}_2^k, P(\underline{S}_2 = \underline{s}_2^k \mid \underline{X}_1, \underline{S}_2 \in \{\underline{s}_2^k, \underline{s}_2^j\})] \mid \underline{S}_2 = \underline{s}_2^j\}\, ; s_1^k, s_2^k, s_1^j, s_2^j \in \{0,1\}\,. \qquad (7)$$

Again, this estimation problem has the same structure as the typical static estimation problem based on CIA. The only difference is that it is of the multiple treatment type, because four different sequences are involved (cf. Imbens, 2000, Lechner, 2001a). Note that if S-DCIA is valid, then the results for the coarser conditioning sets $S_1 = s_1^j$ and $S_0 = 0$ can either be obtained by a weighted mean of the effects for populations defined by period 2 treatment (e.g. $E(Y_2^{\underline{s}_2^k} \mid S_1 = s_1^j)$ = $E(Y_2^{\underline{s}_2^k} \mid S_1 = s_1^j, S_2 = 1)$ $P(S_2 = 1 \mid S_1 = s_1^j)$ + $E(Y_2^{\underline{s}_2^k} \mid S_1 = s_1^j, S_2 = 0)$ $[1 - P(S_2 = 1 \mid S_1 = s_1^j)]$, or by 'matching' directly (or weighting the regression function) according to the redefined target distribution ($S_1 = s_1^j$ or $S_0 = 0$).

Appendix A discusses some issues that come up for non-binary S_1 and S_2. However, since the general structure of the estimation problem does not change, the main part of the text focuses on the binary case.

3.2.2 Sequential estimators of $E(Y^{\underline{s}_2^k} \mid S_1 = s_1^j)$ and $E(Y^{\underline{s}_2^k})$ under W-DCIA

In this section estimation of the counterfactuals that cannot be estimated by the usual 'static' matching methods, namely $E(Y^{\underline{s}_2^k} \mid S_1 = s_1^j)$ and $E(Y^{\underline{s}_2^k})$ under W-DCIA (see equations (4) and (6)) is discussed.[16] The focus is on estimators conditioning on propensity scores, but the same principles apply to estimators that condition directly on the respective control variables. The Monte Carlo study below runs both types of estimators. All proposed estimators have a similar structure, because they are computed as weighted

[16] Estimators consistent under W-DCIA are consistent under the stronger assumption S-DCIA, but not conversely.

means (with weights $\tilde{w}_i$, w_i) of a function of the outcome variables ($g_2^{\underline{s}_2^k}(\underline{p}_2^{\underline{s}_2^k|s_1^k,s_1^k}(\underline{x}_{1,i}))$) observed in subsample $\underline{S}_2 = \underline{s}_2^k$ (the treated population):

$$\widehat{E(Y_2^{\underline{s}_2^k} \mid S_1 = s_1^j)} = \sum_{i \in \underline{s}_2^k} w_i^{\underline{s}_2^k, s_1^j} \; g_2^{\underline{s}_2^k}(\underline{p}_2^{\underline{s}_2^k|s_1^k,s_1^k}(\underline{x}_{1,i})); \quad \tilde{w}_i^{\underline{s}_2^k, s_1^j} \geq 0; \quad \sum_{i \in \underline{s}_2^k} \tilde{w}_i^{\underline{s}_2^k, s_1^j} = 1; \tag{8}$$

$$\widehat{E(Y_2^{\underline{s}_2^k})} = \sum_{i \in \underline{s}_2^k} w_i^{\underline{s}_2^k} \; g(\underline{p}_2^{\underline{s}_2^k|s_1^k,s_1^k}(\underline{x}_{1,i})); \qquad w_i^{\underline{s}_2^k} \geq 0; \quad \sum_{i \in \underline{s}_2^k} w_i^{\underline{s}_2^k} = 1. \tag{9}$$

$g_2^{\underline{s}_2^k}(\underline{p}_2^{\underline{s}_2^k|s_1^k,s_1^k}(\underline{x}_{1,i}))$ is constructed to have mean $E[Y_2 \mid \underline{S} = \underline{s}_2^k, \underline{p}_2^{\underline{s}_2^k|s_1^k,s_1^k}(\underline{X}_{1,i}) = \underline{p}_2^{\underline{s}_2^k|s_1^k,s_1^k}(\underline{x}_{1,i}))]$, at least asymptotically. The weights reweight this function towards the target distribution $f(\underline{x}_{1,i} \mid S_1 = s_1^j)$ for equation (8) and $f(\underline{x}_{1,i} \mid S_0 = 0)[= f(\underline{x}_{1,i})]$ for equation (9). Two types of weights are considered according to whether they are estimated by sequential matching or by sequential inverse choice probabilities. Common choices for $g_2^{\underline{s}_2^k}(\underline{p}_2^{\underline{s}_2^k|s_1^k,s_1^k}(\underline{x}_{1,i}))$ in the matching literature would be either y_i ('direct' or 'pair' matching) or a (kernel) regression estimate of y_i on $p_2^{\underline{s}_2^k|s_1^k}(\underline{x}_{1,i})$ and $p_1^{s_1^k}(x_{0,i})$. The latter has the advantage of leading to an asymptotically more efficient estimator if the bandwidth is chosen appropriately. This gain should be particularly large when many of the weights are zero using the direct approach. It has however the disadvantage that it is subject to the course of dimensionality (the number of conditioning variables increases linearly with the number of periods) and that there appears to be no generally applicable theory of optimally choosing the bandwidth.

Note that there are related estimators known from the matching literature for the weights as well as for $g_2^{\underline{s}_2^k}(\underline{p}_2^{\underline{s}_2^k|s_1^k,s_1^k}(\underline{x}_{1,i}))$ that can be adjusted to the dynamic framework, like 'blocking' or stratification estimators (Rosenbaum and Rubin, 1984, Dehejia and Wahba, 1999, see again Imbens, 2003, for a comprehensive discussion). For the sake of brevity they are ignored.

3.2.2.1 Sequential matching estimators (SM)

The idea of the matching estimators is to sequentially adjust the covariate distribution of the treatment population towards the target population so as to mimic the sequential conditional expectations appearing in expressions (8) and (9). The first step is the same for both effects and consist in finding for every member of $S_1 = s_1^k$ a member of $\underline{S}_2 = \underline{s}_2^k$ with very similar (the same) values of $p_2^{s_2^k|s_1^k}(\underline{x}_{1,i})$ and $p_1^{s_1^k}(x_{0,i})$. Note that one observation in the treatment population may be matched to many or to none of the intermediate target population $S_1 = s_1^k$ (matching with replacement). In the second step every member of $S_1 = s_1^j$ ((8)) or $S_0 = 0$ ((9)) is to be paired with a member of $S_1 = s_1^k$ with very similar (same) values of $p_1^{s_1^k}(x_{0,i})$. The positive weights that are attached to some or all members of $\underline{S}_2 = \underline{s}_2^k$ coming from step 1 are then updated depending on how often an observation in $\underline{S}_2 = \underline{s}_2^k$ is matched to an observation of the target population via the intermediate matching step. This procedures leads to the following estimators:

$$w_i^{\underline{s}_2^k, s_1^j, SM} = \frac{1}{N^{s_1^j}} \sum_{n \in s_1^j} \sum_{m \in s_1^k} v_1[p_1^{s_1^k}(x_{0,n}), p_1^{s_1^k}(x_{0,m}); \cdot]\, v_2[\underline{p}_2^{\underline{s}_2^k|s_1^k, s_1^k}(x_{1,m}), \underline{p}_2^{\underline{s}_2^k|s_1^k, s_1^k}(x_{1,i}); \cdot]; \ \forall i \in \underline{S}_2 = \underline{s}_2^k; \ (10)$$

$$w_i^{\underline{s}_2^k, SM} = \frac{1}{N} \sum_{n=1}^{N} \sum_{m \in s_1^k} v_1[p_1^{s_1^k}(x_{0,n}), p_1^{s_1^k}(x_{0,m}); \cdot]\, v_2[\underline{p}_2^{\underline{s}_2^k|s_1^k, s_1^k}(x_{1,m}), \underline{p}_2^{\underline{s}_2^k|s_1^k, s_1^k}(x_{1,i}); \cdot]; \quad \forall i \in \underline{S}_2 = \underline{s}_2^k. \ (11)$$

$N^{s_1^j}$ denotes the number of observations for which $S_1 = s_1^j$. The function $v_1[p_1^{s_1^k}(x_{0,n}), p_1^{s_1^k}(x_{0,m}); \cdot]$ is defined to be one if $p_1^{s_1}(x_{0,m})$ is closest to $p_1^{s_1}(x_{0,n})$ of all observations belonging to the subsample defined by $S_1 = s_1^k$, and zero otherwise. Similarly $v_2[\underline{p}_2^{\underline{s}_2^k|s_1^k, s_1^k}(x_{1,m}), \underline{p}_2^{\underline{s}_2^k|s_1^k, s_1^k}(x_{1,i}); \cdot]$ is one if observation *i* is closest to observation *m* in terms of $p_2^{\underline{s}_2^k|s_1^k}(\underline{x}_{1,i})$ and $p_1^{s_1^k}(x_{0,i})$, and zero otherwise. Similarity could for example be defined by the Mahalanobis metric. Note that the weight of observation *i* is 0 if it is not matched to any member of the target population (even if it is matched in the first step). On the other extreme, if observation *i* would be matched to every member of the target population then its weight would be 1. Particular issues on how to implement such an estimator (for example on how to insure that the common support conditions are satisfied) are discussed in Sections 4 and 5. A specific variant of this estimator that is used in the empirical part (Section 5) is detailed in Appendix B.

One of the special features of this estimator is that the number of propensity scores to match on increases linearly with the number of periods. Abadie and Imbens (2002) show that in static models a matching estimator with up to two continuous variable is $\sqrt{N}-$consistent and asymptotically normally distributed. However, if it is based on a fixed number of matches (not increasing with sample size), then it does not reach the semiparametric efficiency bound derived by Hahn (1998). If the number of continuous covariates is larger than 2 (as for $T > 2$), then matching estimators remain consistent, but because the bias vanishes at a lower rate than $\sqrt{N}$, they are not asymptotically normally distributed (around the true value). Abadie and Imbens (2002) suggest a correction based on nonparametric regressions of the conditional mean function of the outcome to eliminate the bias and show that the bias corrected estimate is asymptotically normally distributed independent of the number of continuous covariates.[17] In our framework it is clear how the bias correction procedures could be implemented in the last matching step. However, it is not straightforward how to address biases resulting from mismatches in intermediate steps.[18]

[17] These results are obtained for ‘direct’ matching estimators without using the propensity score. However, one could conjecture that matching on the propensity score is approximately like matching on a continuous variable if the dimension of the independent variables in the propensity score is large enough; even if all independent variables are discrete (the problems of matching occur because with continuous variables exact matches have probability zero).

[18] The bias adjusted estimators suffer from the additional problem that they must be based on the same consistent estimates of the nonparametric regression functions as the regression imputation estimators. However, Abadie and Imbens (2002) argue that due to the local nature of the bias correction when combined with matching, the results are not sensitive to bandwidth choices, whereas the regression imputation estimators are.

3.2.2.2 Sequential inverse probability weighting estimators (SIPW)

Reweighting by the inverse selection probabilities is another way to obtain the appropriate weights.[19] The weights have the general structure such that every observation of the treatment population s_2^k is divided by its conditional probability of being selected into treatment and multiplied by its conditional probability of being in the counterfactual state. The weights for quantities identified under the different DCIA assumptions are given in equations (12) to (15). Note that common scale factors - covering relative sample sizes such the weights sum up to one - are omitted.

$$w_i^{s_2^k,IP} = \frac{1}{\hat{p}_2^{s_2^k|s_1^k}(\underline{x}_{1,i})\hat{p}_1^{s_1^k}(x_{0,i})}; \qquad \forall i \in \underline{S}_2 = \underline{s}_2^k; \qquad (12)$$

$$w_i^{s_2^k,s_1^k,IP} = \frac{1}{\hat{p}_2^{s_2^k|s_1^k}(\underline{x}_{1,i})}; \qquad \forall i \in \underline{S}_2 = \underline{s}_2^k. \qquad (13)$$

$$w_i^{s_2^k,s_1^j,IP} = \frac{\hat{p}_1^{s_1^j}(x_{0,i})}{\hat{p}_2^{s_2^k|s_1^k}(\underline{x}_{1,i})\hat{p}_1^{s_1^k}(x_{0,i})}; \qquad \forall i \in \underline{S}_2 = \underline{s}_2^k; \qquad (14)$$

$$w_i^{s_2^k,\underline{s}_2^j,IP} = \frac{\hat{p}_2^{s_2^j|s_1^j}(\underline{x}_{1,i})\hat{p}_1^{s_1^j}(x_{0,i})}{\hat{p}_2^{s_2^k|s_1^k}(\underline{x}_{1,i})\hat{p}_1^{s_1^k}(x_{0,i})};; \qquad \forall i \in \underline{S}_2 = \underline{s}_2^k; \qquad (15)$$

The derivations of the weights as well as the scale factors are given in Appendix C.[20]

4 A Monte Carlo study

As a small sample check for the above suggested estimators a Monte Carlo study is conducted. It is based on two different data generating processes (DGP) fulfilling the weak and the strong DCIA. The design

[19] See for example Hirano, Imbens, and Ridder (2003) and Rosenbaum (1987).

[20] Similar estimators can be found in the literature on attrition and sample selection (e.g. Horvitz and Thomson, 1952, Nevo, 2003, Robins and Rotnitzky, 1995, Robins, Rotnitzky, Zhao, 1995, and Wooldridge, 2003).

aims at a very stylised image of typical DGP’s to be found in potential applications while keeping it as simple as possible to keep computation time manageable.

4.1 Data generating processes

The DGP’s are constructed so that they exhibit nonlinearities, heterogeneous effects, dynamics and selectivity problems to highlight the general approach of the dynamic treatment effect literature. The main ingredients into the data generating process (detailed in Table 2) are specifications of the potential outcome and selection equations. The potential outcomes in the first period are generated by functions of an exogenous regressor and an error term. The potential outcomes in the second period are functions of the potential outcomes of the first period plus error terms. The functions chosen differ for the different potential outcomes. The error terms of the potential outcomes are all mutually correlated, so that the error terms of the second period potential outcome equations are correlated with the *regressors* of these equations. The selection equation in the first period is given by an indicator function of the exogenous confounder that appears in the potential outcome equation for period 1 plus a white noise normal error.

Table 2: Specification of the data generating processes

Selection equation t = 1	$S_1 = \underline{1}(2 + X_0 + U_7 > 0)$, $X_0 \sim N(2,1)$
Outcomes equations t = 1	$Y_1^1 = [(X_0)^2]^{0.75} + U_1$, $Y_1^0 = X_0 + U_2$
Outcomes equations t = 2	$Y_2^{11} = Y_1^1 + U_3$, $Y_2^{10} = [(Y_1^1/4)^2]^{0.75} + U_4$, $Y_2^{01} = \ln(\vert Y_1^0 \vert) + U_5$, $Y_2^{00} = \sqrt{\vert Y_1^0 \vert} + U_6$
Selection equation t = 2: DPG 1 (S-DCIA)	$S_2 = \underline{1}(-0.5 - S_1 + 0.5\tilde{X}_1 + U_8 > 0)$; $\tilde{X}_1 \sim N(2,1)$
Selection equation t = 2: DGP 2 (W-DCIA)	$S_2 = \underline{1}(-1 - S_1 + 0.5 Y_1 + U_8 > 0)$
Distribution of error terms	$(U_1,...,U_8) \sim N(0,\cdot)$, $Var(U_1,...,U_8) = (1,...,1)$,
	$Corr(U_1,...,U_8) = \begin{pmatrix} .3 & .2 & .2 & .2 & .2 & 0 & 0 \\ & .2 & .2 & .2 & .2 & 0 & 0 \\ & & .3 & .3 & .3 & 0 & 0 \\ & & & .3 & .3 & 0 & 0 \\ & & & & .3 & 0 & 0 \\ & & & & & 0 & 0 \\ & & & & & & 0 \end{pmatrix}$

Note: If not explicitly stated otherwise, the specification relates to both DGP’s. All draws of random numbers (using Gauss 3.2.32) are independent across observations and replications.

While all other equations are the same for both DGP’s the selection equation in the second period differs. For DGP 1 it is modelled as a probit indicator function with the realised selection in period 1 and an exogenous variable as explanatory variables. DGP 2 substitutes the exogenous variable by the observed outcome after period 1. The error terms of the selection equations are independent of the error terms of the potential outcome equations, so that both DGP's are based on ***selection on observables***. Because all variables determining selection are not related to outcomes of the treatment DGP 1 fulfils W-DCIA and S-DCIA. The crucial difference between the weak version and the strong versions of DCIA is that the strong version requires $F(Y_2^{s_1^k,s_2} \mid X_1, S_1 = s_1^l, X_0)$ = $F(Y_2^{s_1^k,s_2} \mid X_1, X_0)$ to hold. It is fulfilled by DGP 1, because X_1 is a regressor that is independent of all potential outcomes. In DGP 2 X_1 becomes the outcome of period 1, Y_1, which is related to potential outcomes of period 2 by the chosen autoregressive specification of the outcome equations and the period 1 selection rule. Therefore, DGP 2 violates the strong version of DCIA, because of the endogeneity of one of the variables determining selection in period 2. Nevertheless, DGP 2 satisfies the conditions for W-DCIA (all variables influencing selection and potential outcomes are observable).

Functional forms and coefficients in the outcome equations of both DGP's have been chosen such that treatment effects exhibit heterogeneity and state dependence. The coefficients of the selection equations are tuned to produce marginal participation probabilities $P(S_1 = 1)$ and $P(S_2 = 1)$ of about 50%.

Tables 3 and 4 give basic unconditional descriptive statistics for both DGP's. The mean of the potential outcomes differ substantially from each other and thus from the observed outcomes. They exhibit considerable individual heterogeneity leading to heterogeneity of the effects. Unconditionally, potential outcomes, effects and observed outcomes are highly correlated with both selection variables, the correlation for S_2 being higher for DGP 2 due to the different specification of the selection equation in period 2. Note that although X_1 is uncorrelated with potential outcomes, it is correlated with observed outcomes and thus exhibits the classical features of an instrumental variable.

Table 3: Some descriptive statistics for the data generating process for DGP 1

	Mean	Std.	Correlations (x 100)											
			Y_2^{10}	Y_2^{01}	Y_2^{00}	$\theta_2^{11,10}$	$\theta_2^{11,01}$	$\theta_2^{11,00}$	Y_2	Y_1	S_2	S_1	X_0	$\tilde{X}_1$
Y_2^{11}	3.1	2.6	68	43	49	86	89	83	62	84	-15	45	78	0
Y_2^{10}	.84	1.3	-	43	45	23	54	49	60	62	-10	29	51	0
Y_2^{01}	1.4	1.2		-	53	28	-2	15	48	38	-6	16	27	0
Y_2^{00}	.45	1.5			-	35	28	-8	48	47	-8	22	37	0
$\theta_2^{11,10}$	2.3	1.9				-	82	77	42	70	-14	40	69	0
$\theta_2^{11,01}$	1.7	2.3					-	84	45	74	-14	42	73	0
$\theta_2^{11,00}$	2.7	2.3						-	40	66	-13	37	66	0
Y_2	1.5	2.0							-	63	33	34	49	16
Y_1	2.9	2.3								-	-22	62	85	0
S_2	.50	.50									-	-35	-20	32
S_1	.50	.50										-	57	0
X_0	2.0	1.0											-	0
$\tilde{X}_1$	2.0	1.0												-

Note: Sample statistics based on one sample of *N* = 200.000. $P(S_2 = 1, S_1 = 1)$ = .17; $P(S_2 = 0, S_1 = 1)$ = .33; $P(S_2 = 1, S_1 = 0)$ = .34; $P(S_2 = 0, S_1 = 0)$ = .16.

Table 4: Some descriptive statistics for the data generating process for DGP 2

	Mean	Std.	Correlations (x 100)										
			Y_2^{10}	Y_2^{01}	Y_2^{00}	$\theta_2^{11,10}$	$\theta_2^{11,01}$	$\theta_2^{11,00}$	Y_2	Y_1	S_2	S_1	X_0
Y_2^{11}	6.1	6.8	83	4	14	98	83	96	29	32	17	17	30
Y_2^{10}	1.4	1.8	-	8	24	70	66	75	43	45	23	21	37
Y_2^{01}	4.4	4.5		-	75	3	-53	-17	16	10	10	4	7
Y_2^{00}	1.4	1.9			-	10	-30	-15	40	36	22	16	28
$\theta_2^{11,10}$	4.7	5.4				-	82	95	21	25	13	15	25
$\theta_2^{11,01}$	1.7	8.1					-	91	16	21	9	12	21
$\theta_2^{11,00}$	4.7	6.9						-	17	22	10	13	22
Y_2	3.9	3.7							-	76	81	39	62
Y_1	2.9	2.3								-	47	62	85
S_2	.50	.50									-	14	39
S_1	.50	.50										-	57
X_0	2.0	1.0											-

Note: Sample statistics based on one sample of *N* = 200.000. $P(S_2 = 1, S_1 = 1)$ = .28; $P(S_2 = 0, S_1 = 1)$ = .22; $P(S_2 = 1, S_1 = 0)$ = .21; $P(S_2 = 0, S_1 = 0)$ = .28.

The sample sizes considered in the simulations are *N* = 400, 1600, and 6400, respectively. They cover a reasonable range given recent labour market applications of matching methods (e.g. Gerfin and Lechner, 2002, or Sianesi, 2001). Furthermore, given the few confounding variables the sample sizes should be large enough to observe convergence properties of the estimators.

4.2 Estimators

Five different estimators are considered in the simulations. Four of them are nearest-neighbour type matching estimators, two of them sequential and two of them one-step estimators; two of the matching estimators that match on an estimated propensity score, two estimators match directly on the confounding variables.[21] When there is more than one variable to match on, the distance metric is the Mahalanobis distance with weight matrix computed in the sample of the particular target population. The remaining estimator uses estimated sequential probabilities – as defined in the previous section - to directly reweight the treated observations towards the target population.

The sequential matching estimators (SM) using propensity scores as well as the inverse probability-weighted estimator (SIPW) are based on conditional probabilities estimated by a probit model ($P(S_2 = 1 | S_1 = 1, X_1)$, $P(S_2 = 1 | S_1 = 0, X_1)$, $P(S_1 = 1 | X_0)$) in the respective subsample defined by treatment status.[22] Both one-step-matching estimators are based on binary probits for the probability of being in the particular subsample conditional on being either in the subsample for which outcomes are observed or in the target population (for example $P(S_2 = 1, S_1 = 1 | X_1, X_0)$ for estimating EY_2^{11}, or $P(S_2 = 1, S_1 = 1 | \underline{S}_2 \in \{(0,0),(1,1)\}, X_1, X_0)$ for estimating $E(Y_2^{11} | S_2 = 0, S_1 = 0)$). All variables explaining both selection steps are included in these specifications. Thus, we generally expect the one-step estimators to be inconsistent for DGP 2 (with the exception of $E(Y_2^{11} | S_2 = 0, S_1 = 1)$, $E(Y_2^{11} | S_1 = 1)$, etc.), because they have the same structure as the propensity score estimators proposed for the static multiple treatment model that assumes exogenous confounders. Furthermore, to check the effects of estimating the (higher dimensional) balancing scores in situations in which the dimension of *X* is actually small

[21] $E(Y_2^{11} | S_2 = 1, S_1 = 1)$ is estimated only by one estimator, the (unweighted) sample mean, and thus has some 'ideal' properties that we would like the matching estimators to have as well (unbiasedness, efficiency, consistency and $\sqrt{N}$ – convergence).

[22] *X₁* denotes the 'regressor' in the probit model that would either be Y_1 (DGP 1) or $\tilde{X}_1$ (DGP 2). $\tilde{X}_1$ needs not to be included in the probit estimation because it is not a confounding variable. Nevertheless, it is included to keep the dimension of the estimation problem the same for both DGP's.

enough to match directly on the *X* variables, the two direct matching estimators proceed exactly as the propensity score matching estimators, but instead of using the scores, they use the explanatory variables of the scores directly.

A detailed matching protocol for $E(Y^{s_2^k} \mid S_1 = s_1^l)$ - $E(Y^{s_2^l} \mid S_1 = s_1^l)$ can be found in Appendix B (with obvious extensions when other effects are of interest or the one-step or direct matching estimators are used). However, to speed up computation checks for common support - which should not be necessary because the variables explaining selection are continuous - have not been performed for the SM estimators, neither have very small or very large probabilities been trimmed for the SIPW estimator. Some remarks about this protocol are warranted:

First, note that matching is with replacement. Every step of the matching sequence is like matching in a static framework, so that all the different estimators discussed in the literature with their merits and drawback are potential candidates. Note that here matching involves several probabilities so that there is the question about how to define 'closeness'. It seems to be common practise in the propensity score matching literature to use the Mahalanobis distance.

Next, some issues arise from the sequential nature of matching for example to obtain $E(Y^{s_2^k} \mid S_1 = s_1^l)$: By choosing observations as matches with similar values of the probabilities instead of the same values, it may happen that the probabilities attached to observations in early matching steps ***change over different sequential matching steps*** due to imprecise matching. To prevent this happening every matched comparison observation in period 2 could be recorded with the values $\hat{p}_1^{s_1^l}$ of the observation it is matched to, instead of its own. Hence the 'history' of the match, or in other words the characteristics of the reference distribution, does not change when the next match occurs in the subsequent period.

Furthermore, to compute $E(Y^{\underline{s}_2^k} \mid S_1 = s_1^l)$ the only information that is needed for the $N^{s_1^l}$ participants in s_1^l is $\hat{p}_{1,i}^{s_1^k}$. Similarly, for participants in $\underline{s}_2^k$, all probabilities of the type $\hat{\underline{p}}_{2,i}^{s_2^k|s_1^k,s_1^k}$ are required. For participants in s_1^k but not in $\underline{s}_2^k$ only $\hat{p}_{1,i}^{s_1^k}$ is needed, and so on.

To estimate $E(Y^{\underline{s}_2^l} \mid S_1 = s_1^l)$ instead of $E(Y^{\underline{s}_2^k} \mid S_1 = s_1^l)$ (part B in Table B.1) the only change in the previous matching protocol is that the initial matching step on $\hat{p}_{1,i}^l$ is redundant in this case. When interest is in the average effect in the population ($E(Y^{\underline{s}_2^k})$), then the whole population plays the role of the first reference group (instead of s_1^l). In this case in the matching step based on $\hat{p}_{1,i}^k$ all participants in s_1^k will also be matched to themselves, as well as selected participants in s_1^k will be matched to the participants in the remaining treatments in the first period.

When matching is on the propensity score instead of matching directly on the confounding variables there is the issue of selecting a probability model. It seems that so far even in the static model the literature has not addressed this thoroughly. There are some results for specific nonparametric approaches as in Hirano, Imbens, and Ridder (2003), but the general consensus seems to be that a flexibly specified (and extensively tested) parametric model, like a logit or a probit model, is sufficiently rich and that the choice of the model does not really matter (see for example the Monte Carlo results by Zhao, 2000).

There is one issue, which has not been discussed so far because it is not the focus of this paper, but important in practise: estimation of the standard errors. In the applied evaluation literature there seems to be two common ways to estimate standard errors: First, compute the standard errors conditional on the weights. When the weights are based on estimated propensity scores, the uncertainty of the estimation step is typically ignored. Alternatively naïve bootstraps are used (including the estimation step of the propensity scores) and standard errors (or confidence intervals) are computed from the bootstrap distribution. However, there appears to be no proof available that indeed the conditions for the consistency of the bootstrap (it is also not clear that the standard error is the best quantity for which to perform the bootstrap, because it is not asymptotically pivotal). In the Monte Carlo study the standard errors are computed using these two methods. However, due to computing time restrictions the bootstrap (200 replications) is used for the smallest sample only.

4.3 Results

Tables 5 (DGP 1) and 6 (DGP 2) present the results of the simulations. They are based on 1000 replications. Since causal effects are usually estimates as the difference of the estimated (counterfactual) means of the potential outcomes in the respective population of interest, the tables focus on the estimated mean of the counterfactual outcomes only. Reporting the results for all subpopulations would require too much space. Therefore, both tables consider only three populations, namely the participants in the treatment in both periods ($\underline{S}_2 = (1,1)$), participants in the first period ($S_1 = 1$), and the population. The results for other populations defined by treatment status are qualitatively identical.

Detailed results are given for the sequential matching-on-the-estimated-propensity-score estimator that appears to be of major interest (true mean; bias of estimators for mean and standard errors; standard deviation, skewness, kurtosis, root mean squared error and median absolute error of mean estimator). In its bootstrap version shown for N=400 the respective line contains the same statistics for the estimator defined to be the mean of the bootstrap distribution, with one exception. The bias of the standard error relates to the bias of the standard error for the bootstrap compared to the Monte Carlo standard error of the non-bootstrapped sequential matching estimator given in the line above the bootstrap. For the sake of brevity mean squared errors only are given for the other estimators.[23]

For DGP 1 all estimators are almost unbiased. Even for N = 400 the bias is very small. The sequential (as well as all other) matching estimators appear to be fairly close to the normal distribution for all sample sizes considering skewness and kurtosis observed in the Monte Carlo study. When the sample size quadruples the standard errors are reduced by about half. In conclusion, the matching estimators appear to be $\sqrt{N} -$ convergent.

Next, we compare the various estimators according to their mean squared errors, beginning with the matching estimators. In many cases the matching estimators based on the estimated propensity scores have lower RMSE than the corresponding matching estimators based directly on the variables determining selection. This result may seem surprising, because there is only one regressor in the each probit, so that the propensity score does not reduce the dimension of the conditioning variables in this setting. This result is however in line with the findings of Hirano, Imbens, and Ridder (2003), although the latter (theoretical) results are obtained for a different class of estimators. Comparing the sequential and one-step matching estimators, no systematic differences appear for DGP 1 (they will differ drastically for DGP 2).

Table 5: Results for DGP 1

				Sequential propensity score matching							seq. on X	one PSM	one on X	prop. weight
Pot. outc.	Pop-ulation	N	true	bias	std.	skew.	kurt.	bias std.	MAE x 10			RMSE x 10		
Y_2^{11}	11	400	4.26	-.01	.32	.15	2.96	-.01	2.26	3.23	--	--	--	--
	BS	400		.01	.31	-.09	3.15	*.01*	2.02	3.13	--	--	--	--
		1600		.00	.16	-.01	3.08	.00	1.05	1.53	--	--	--	--
		6400		.00	.08	-.01	3.05	-.00	.54	.83	--	--	--	--
	1	400	4.26	-.03	.28	.06	2.86	.09	1.87	2.79	4.75	--	--	3.23
	BS	400		-.02	.26	-.11	2.89	*-.02*	1.73	2.61	3.96	--	--	2.46
		1600		-.02	.13	-.14	3.08	.06	.87	1.30	2.19	--	--	1.17
		6400		-.01	.07	.04	2.89	.02	.46	.69	1.15	--	--	.62
	all	400	3.12	.06	.34	.20	3.35	.15	2.05	3.41	3.54	3.43	3.38	4.28
	BS	400		.08	.28	.04	2.96	*-.05*	1.96	2.96	3.29	3.01	3.35	4.15
		1600		.01	.17	.01	3.12	.10	1.16	1.75	1.84	1.77	1.71	2.52
		6400		-.01	.09	-.00	3.13	.05	.65	.97	1.03	.98	.94	1.41
Y_2^{10}	11	400	1.22	.01	.21	.09	3.32	.02	1.43	2.10	2.57	--	--	1.43
	BS	400		-.00	.17	.12	3.02	*.01*	1.12	1.70	2.16	--	--	1.36
		1600		.01	.11	-.01	2.95	.01	.73	1.07	1.28	--	--	.70
		6400		.00	.06	-.00	3.01	.00	.36	.54	.66	--	--	.36
	1	400	1.22	-.00	.13	-.05	3.05	.01	.85	1.30	1.45	--	--	1.20
	BS	400		-.01	.12	-.04	2.98	*-.01*	.76	1.16	1.35	--	--	1.11
		1600		-.00	.06	.00	3.06	.00	.42	.63	.71	--	--	.57
		6400		-.00	.03	.00	3.10	.00	.22	.33	.36	--	--	.30
	all	400	.84	.00	.15	.02	3.28	.04	1.02	1.56	1.61	1.58	1.50	1.57
	BS	400		.00	.13	-.06	2.95	*-.01*	.88	1.39	1.40	1.34	1.33	1.45
		1600		-.00	.08	.07	3.20	.02	.54	.84	.86	.87	.77	.85
		6400		-.00	.05	.04	3.15	.01	.31	.46	.47	.47	.41	.43

Table 5 to be continued.

[23] The complete set of results (all estimators and other subpopulations) is available on request from the author.

Table 5: continued

				Sequential propensity score matching							seq.	one	one	prop.
											on X	PSM	on X	weight
Pot.	Pop-	N	true	bias	std.	skew.	kurt.	bias	MAE x			RMSE x 10		
outc.	ulation							std.	10					
Y_2^{01}	11	400	1.55	-.02	.27	-.25	3.75	.00	1.78	2.71	2.56	2.96	2.56	2.31
	BS	400		-.05	.20	.09	3.07	*-.01*	1.41	2.61	2.22	2.45	2.22	2.29
		1600		-.02	.14	-.10	2.97	.01	.99	1.43	1.40	1.59	1.40	1.32
		6400		-.00	.07	.08	3.04	.00	.52	.76	.75	.86	.75	.79
	1	400	1.55	-.02	.25	-.15	3.54	.01	1.55	2.47	2.65	2.67	2.25	2.29
	BS	400		-.04	.20	.14	3.28	*-.03*	1.30	2.04	2.23	2.20	2.06	2.27
		1600		-.01	.14	-.20	3.15	.01	.91	1.38	1.49	1.46	1.25	1.33
		6400		-.00	.08	-.02	3.42	.01	.52	.79	.82	.81	.71	.80
	all	400	1.37	-.01	.15	-.17	3.27	.01	.99	1.53	1.62	1.57	1.43	1.51
	BS	400		-.02	.14	.02	3.14	*-.01*	.90	1.39	1.47	1.41	1.40	1.55
		1600		-.01	.08	-.18	3.15	.00	.56	.84	.90	.86	.78	.84
		6400		-.00	.05	-.04	3.16	.00	.32	.46	.48	.46	.43	.51
Y_2^{00}	11	400	.76	-.05	.37	-.25	3.39	.08	2.35	3.70	4.43	4.10	4.43	3.44
	BS	400		-.09	.30	-.15	3.56	*-.03*	1.99	3.09	3.82	3.39	3.83	3.16
		1600		-.01	.22	.00	3.69	.04	1.37	2.18	2.51	2.35	2.51	2.06
		6400		-.01	.12	-.05	3.31	.03	.76	1.16	1.39	1.26	1.14	1.31
	1	400	.76	-.04	.35	-.27	3.51	.08	2.28	3.62	3.97	3.75	3.60	3.46
	BS	400		-.09	.30	-.09	3.69	*-.04*	2.04	3.08	3.26	3.15	3.19	3.18
		1600		-.01	.22	.04	3.84	.05	1.36	2.18	2.21	2.19	2.00	2.04
		6400		-.01	.12	-.04	3.50	.04	.78	1.21	1.21	1.20	1.10	1.31
	all	400	.45	-.02	.25	-.12	3.08	.04	1.71	2.57	2.73	2.51	2.56	2.49
	BS	400		-.04	.21	-.09	3.40	*-.03*	1.48	2.22	2.39	2.22	2.35	2.34
		1600		-.00	.14	.02	3.29	.03	.91	1.39	1.46	1.37	1.34	1.37
		6400		-.01	.07	-.03	3.36	.02	.50	.74	.77	.74	.72	.87

Note: *Bias*: Mean of estimated effect - true effect. *Std.*: Standard deviation observed in Monte Carlo. *Skew.*: Skewness. *Kurt.*: Kurtosis. *RMSE*: Root mean squared error. *MAE*: Median absolute error. *Bias std.*: Mean of estimated standard deviation – *Std.*. *Seq. on X.*: Sequential matching using the appropriate control variables instead of the propensity score. *One PSM*: One step propensity score matching. *One on X*: One-step matching using the appropriate control variables instead of the propensity score. In cases when the sequential and the one-step estimators coincide, the one-step estimators are not given. *BS*: Bootstrap estimates are based on 200 bootstrap samples of *N* draws (with replacement) in the respective sample. The *bias of the bootstrap standard errors* does not relate to the mean of the bootstrap distribution, but they relate to the sequential matching estimator given one line above. All other numbers given in the line *BS* relate to the estimator defined as the mean of the bootstrap distribution.

In many cases the (untrimmed) SIPW estimator has the lowest RMSE of all estimators considered for *N=400*. When the sample size increases this advantage disappears because the estimator converges slower than $\sqrt{N}$. In fact, the kurtosis of the estimator in the simulations increases fairly dramatically with sample size suggesting that the higher order moments (or even the RMSE) of this estimator may not exist.[24] Trimming very small and very large probabilities may be necessary to improve the convergence properties. This is also confirmed by considering the median absolute error that shows more pronounced reduc-

[24] These findings are confirmed by simulations for *N*=25600 which are computed for the IPW estimator only.

tions with increased sample size than the RMSE. A detailed investigation into this issue is however beyond the scope of this paper.

For the smallest sample the mean of the bootstrap distribution dominates (with very rare exceptions) the non-bootstrapped estimators. Since the gains are sometimes considerable and they seem to exist for all estimators, further research on bootstrapping in the dynamic matching framework seems to be warranted.

If S-DCIA is not valid as in DGP 2, $E(Y_2^{00} \mid S_2 = 1, S_1 = 1)$ and $E(Y_2^{01} \mid S_2 = 1, S_1 = 1)$ are not identified. Furthermore, the one-step estimators are inconsistent for $E(Y_2^{00})$, $E(Y_2^{01})$, $E(Y_2^{10})$, $E(Y^{11})$ as well as for $E(Y_2^{01} \mid S_1 = 1)$ and $E(Y_2^{00} \mid S_1 = 1)$. In Table 6 the entries of RMSE's for inconsistent estimators are shaded.

Table 6: Results for DGP 2

				Sequential propensity score matching							seq. on X	one PSM	one on X	prop. weight
Pot. outc.	Pop-ulation	N	true mean	bias	std.	skew.	kurt.	bias std.	MAE x 10			RMSE x 10		
Y_2^{11}	11	400	8.55	.00	.22	.09	2.81	.00	1.52	2.21	--	--	--	--
	BS	400		.01	.23	.08	2.99	*.00*	1.46	2.27	--	--	--	--
		1600		-.01	.11	.16	2.80	.00	.75	1.11	--	--	--	--
		6400		-.00	.06	.01	3.04	.00	.37	.55	--	--	--	--
	1	400	7.27	.04	.23	.05	2.86	.13	1.64	2.40	2.39	--	--	4.00
	BS	400		.09	.22	.02	3.14	*-.01*	1.49	2.33	2.34	--	--	3.77
		1600		.01	.12	.-.06	3.14	.07	.79	1.23	1.23	--	--	2.41
		6400		-.00	.06	-.07	3.13	.04	.41	.61	.61	--	--	1.49
	all	400	6.08	.48	.35	-.03	3.16	.22	4.83	5.92	5.68	4.32	3.89	7.92
	BS	400		.58	.29	-.04	2.85	*-.08*	5.80	6.44	6.38	4.44	4.33	7.79
		1600		.30	.20	-.06	3.16	.16	3.05	3.59	3.42	2.08	1.82	6.34
		6400		.18	.12	-.02	3.03	.11	1.82	2.16	2.05	1.61	1.57	4.65
Y_2^{10}	11	400	2.31	-.17	.35	-.03	3.40	.05	2.67	3.91	3.91	--	--	4.51
	BS	400		-.21	.29	-.12	3.30	*-.06*	2.50	3.60	3.63	--	--	4.08
		1600		-.10	.22	-.14	3.66	.01	1.53	2.42	2.42	--	--	3.17
		6400		-.07	.13	-.09	3.24	.02	.92	1.43	1.42	--	--	2.94
	1	400	1.82	-.09	.21	.00	3.68	.02	1.58	2.28	2.28	--	--	3.21
	BS	400		-.11	.18	-.08	3.04	-.03	1.42	2.13	2.13	--	--	2.74
		1600		-.05	.12	-.20	3.42	.01	.89	1.37	1.37	--	--	2.17
		6400		-.04	.07	.00	3.23	.01	.52	.80	.79	--	--	2.21
	all	400	1.43	-.05	.16	-.03	3.07	.02	1.22	1.72	1.68	2.33	2.27	2.46
	BS	400		-.06	.15	-.08	2.80	-.06	1.11	1.63	1.58	2.34	2.29	2.08
		1600		-.04	.10	-.01	3.23	.00	.67	1.02	1.00	1.86	1.84	1.50
		6400		-.03	.05	.02	2.76	.00	.39	.57	.54	1.62	1.60	1.54

Table 6 to be continued.

Table 6: continued

				Sequential propensity score matching							seq. on X	one PSM	one on X	prop. weight
Pot. outc.	Pop-ulation	N	true mean	bias	std.	skew.	kurt.	bias std.	MAE x 10		RMSE x 10			
Y_2^{01}	11	400	4.80	.52	.49	.38	3.08	.02	5.00	7.22	7.17	9.07	7.17	3.45
	BS	400		.46	.37	.07	3.29	-.02	4.59	5.92	6.11	7.60	6.11	3.39
		1600		.71	.43	.27	3.16	.03	7.03	8.32	7.74	9.37	7.74	2.54
		6400		.84	.34	.03	3.01	.05	8.29	9.09	8.35	9.42	8.35	2.24
	1	400	4.59	.00	.26	-.10	3.45	.02	1.65	2.56	2.50	5.78	5.72	2.34
	BS	400		.01	.22	-.04	3.08	-.03	1.50	2.21	2.18	5.07	5.04	2.34
		1600		-.01	.15	.01	3.02	.01	.95	1.45	1.41	5.88	6.03	1.35
		6400		-.00	.08	.00	3.09	.00	.55	.82	.81	6.23	6.47	.90
	all	400	4.38	-.00	.17	-.11	3.09	.02	1.13	1.71	1.70	3.08	3.27	1.66
	BS	400		.00	.16	-.09	2.93	-.01	1.06	1.61	1.61	2.80	3.30	1.75
		1600		-.01	.09	-.01	2.97	.01	.64	.95	.93	2.96	3.14	.90
		6400		-.00	.05	.13	3.04	.00	.34	.51	.50	3.11	3.30	.58
Y_2^{00}	11	400	1.95	.34	.58	.08	3.22	.28	4.57	6.67	6.36	9.31	6.36	4.54
	BS	400		.26	.46	.09	3.42	-.03	3.57	5.27	5.02	7.02	5.02	4.08
		1600		.53	.51	.06	3.32	.24	5.32	7.32	6.89	9.74	6.89	3.56
		6400		.67	.45	.10	3.01	.26	6.61	8.06	7.40	9.55	7.40	2.74
	1	400	1.68	-.10	.34	.01	3.26	.13	2.36	3.52	3.42	5.76	5.28	3.63
	BS	400		-.11	.28	.04	2.84	-.06	2.09	3.01	2.99	4.76	4.32	3.41
		1600		-.03	.20	.14	2.94	.08	1.32	1.99	1.93	5.98	5.73	2.34
		6400		-.02	.12	.11	3.45	.05	.79	1.17	1.14	6.05	6.20	1.57
	all	400	1.35	-.06	.21	.09	3.61	.07	1.65	2.23	2.16	3.06	3.00	2.51
	BS	400		-.07	.18	.06	2.88	-.03	1.30	1.94	1.92	2.56	2.50	2.39
		1600		-.02	.12	.12	3.02	.05	.79	1.17	1.14	3.04	3.06	1.56
		6400		-.01	.07	.11	3.17	.03	.45	.67	.66	3.03	3.22	1.07

Note: See note below Table 5. RMSE's relating to estimators that are inconsistent for the particular effect are shaded.

Table 6 shows that all estimators are severely biased for the unidentified effects. For the identified effects for which the sequential estimators differ to the one-step estimators, the former are either unbiased or exhibit a bias that is disappearing with increasing sample size. With one exception, the RMSE for the inconsistent one-step estimators is always larger than for the sequential ones. The single exception is the estimation of $E(Y^{11})$, for which the sequential estimators are biased, although the bias is getting smaller the larger the sample. The bias of the one-step estimator is smaller for these sample sizes and thus the RMSE is smaller. However, comparing the systematic development of the bias with increasing sample size (.48, .30, .18 for the sequential estimators and .22, .02, -.10 for the one-step estimators) may indicate that for larger sample sizes the one-step estimators are dominated by the sequential ones (the bias of the sequential matching estimator may eventually disappear, but its speed seems to be below $\sqrt{N}$ which could be an indication of the problems pointed out by Abadie and Imbens, 2002, discussed in the previous section).

Comparing the consistent estimators does not reveal substantial differences other than those already mentioned in the discussion of Table 5. In conclusion, the Monte Carlo study confirms the theoretical considerations and shows that in most cases the estimators have reasonable finite sample properties.

Comparing the bias of the 'conventional' and bootstrapped standard errors leads to the general conclusion that the 'conventional' standard errors are systematically estimated somewhat too large (but decreasing with sample size), thus giving conservative inference. The bootstrapped standard errors are almost always too small. This finding already appeared in Lechner (2002b). Future research will show whether for example the 'match-to-treated-to-the-treated' estimate for standard errors suggested by Abadie and Imbens (2002) performs better. However, since this is a problem that is not akin to the specifics of dynamic matching estimation, it is not pursued further.

5 An empirical application

In this section the sequential matching estimator is applied to a real world situation to show that it can provide useful results and to discuss issues of implementation that may arise in applications.[25]

5.1 The estimation problem and the data available

The study by Gerfin and Lechner (2002, GL02 henceforth) serves as an example for this exercise. They are interested in the effects of different components of the Swiss active labour market policies on subsequent labour market outcomes using a rich administrative individual database coming from the merged records of the Swiss unemployment insurance system and the public pension system. Because they lack a dynamic framework GL02 estimate the effects of beginning the first programme in an unemployment spell.[26] 8 different types of programmes are considered. They argue extensively that the data is informative enough to make the conditional independence assumption plausible.[27]

[25] Different estimators lead to different issues in implementation (e. g. trimming for the SIPW estimator). Therefore, for the sake of brevity only the sequential matching estimator conditional on the propensity scores is discussed.

[26] The endogeneity problem of programme duration and subsequent programme participation arises because both are most likely influenced by the effect of beginning the first programme.

Using the dynamic framework it is possible to go beyond this study and take account of the endogeneity of programme duration and subsequent participation. The interest is in the effects of being two periods in different states for individuals entering unemployment in the last quarter of 1997.[28] The different states used are unemployment (U), training courses (C), employment programmes (E) and receiving a temporary wage subsidy (T). The period is defined as an interval of two months.[29] Therefore, the treatment occurs between January and April 1998. For example considering sequences like EE compared to CC allows assessing the effects of four months of employment programmes compared to four months of courses. Obviously this approach would allow investigating the effects of programme combinations as well, for example by comparing CE to UU or EE. As another example it would be possible to check the effect of waiting for participation by comparing e.g. UC or UCC to CC. For the sake of brevity only effects of the types CC, EE, TT and UU are considered.

The outcome variables of interest are the probability of unsubsidised employment and monthly earnings between May 1998 and the end of the observation period in December 1999. Taking the identification arguments of GL02 for beginning the first programme for granted, S-DCIA in this context either require that sequence participation are determined before the start of their first component, or that the participation in the second period does essentially not depend on the outcome of the first part of the sequence. Since individuals may leave sequences (or are not even allocated to sequences beforehand) and because the labour market effects of the first part of the sequence may cause this behaviour (particularly attending a second short course, but also leaving programmes intended for a longer period), this assumption does hardly appear to be tenable. However, if the intermediate outcomes that determine the next step in the sequence are observable, then W-DCIA is plausible.

[27] For all details about the programmes evaluated and the data used the reader is referred to GL02. This application relies on the same population as GL02 but uses the extended version of the sample described in Gerfin, Lechner, and Steiger (2002).

[28] Of course much longer sequences could be considered but are not discussed for the sake of brevity.

Table 7: Selected descriptive statistics of control variables and intermediate outcomes

Variables	Means / shares in % in subsamples							
Subpopulations:	U	UU	C	CC	E	EE	T	TT
	Control variables (measured in Dec. 1997 or before)							
Insured earnings in CHF	3943	3904	4008	3932	3592	3517	4023	4067
Chances to find a job (employability): no information	9	8	5	5	8	7	11	14
very easy	8	8	6	5	3	4	6	6
easy	17	16	19	18	10	12	18	18
medium	55	56	59	60	58	58	56	54
difficult	9	10	11	10	17	15	7	7
special case	1	1	1	1	4	4	1	2
Sanction days without benefits (current spell; until Dec. '97)	3.1	3.2	3.8	3.2	3.5	3.3	2.6	2.3
	Intermediate outcome variables (measured in February 1998)							
Employed	13	4	2	1	1	0	6	1
Earnings in CHF	509	164	1283	51	33	0	229	30
Chances to find a job (employability): no information	2	2	2	2	0	0	3	3
very easy (or some employment)	20	12	8	5	4	4	12	8
easy	16	16	17	16	13	14	20	21
medium	52	57	60	62	59	59	56	59
difficult	9	11	13	14	20	19	7	7
special case	1	2	1	1	3	4	1	1
Sanction days without benefits Jan. + Feb. '98	1.2	1.3	.9	.8	.5	.4	.5	.6
Sample size	7982	5122	573	316	118	99	790	382

Note: Descriptive statistics of the outcome variables after April 1998 are contained in Tables 8 and 9. The sample is based on the same selection criteria of GL02, but constrained to those entering unemployment in the forth quarter of 1997.

Table 7 gives some descriptive statistics on some important control variables, treatments, and intermediate outcomes (see GL02, for a more complete account of the variables available) for the different subsamples of interest. The control variables (X_0) considered in this table are the monthly earnings in the previous job, the subjective valuation of the caseworkers as well as the number of sanction days without benefit (imposed by the caseworkers on unemployed violating the rules). The subjective valuation gives an assessment of the employability of the unemployed. This assessment may be changed by the caseworker at the monthly interview of the unemployed. Comparing the descriptive statistics across subsamples defined by treatment status reveals that participants in employment programmes are the group having the worst a priori chances on the labour market whereas participants in the temporary wage subsidies appear to be the group with the best a priori chances.

[29] Any bimonthly period in which an individual participates at least 2 weeks in a programme (C, E or T) is considered as programme participation. In the rare event of participation in 2 programmes in the same period, the individual is defined to participate in the longest of those programmes.

The lower panel in Table 7 shows results for some intermediate outcomes (X_1), like employment status, earnings, the subjective valuation of the caseworkers, and new sanction days. Comparing the variables across treatments confirms basically the previous conclusion. However, this part of Table 7 also shows that participants leaving the sequence may be markedly different than participants staying.

The intermediate outcome variable employment points to a potential problem that might occur in an empirical study using the dynamic framework: the occurrence of absorbing states violates the common support condition. In other words, if individuals experience intermediate outcomes that will prohibit them from future participation in the sequence, it is not possible to evaluate the effects of the longer sequences for such (groups of) individuals. Here it is not clearly obvious whether this is a problem or not, although the problem seems to be most pronounced for the group of nonparticipants: Even if somebody is employed there is in principle the possibility to participate in some programme if other conditions are satisfied. Furthermore, the employment variable (as well as earnings) is computed from the pension records and may pick up pension contributions for minor employment during the unemployment spell.[30]

5.2 The sequential matching estimator in practise

The estimator used in this empirical example follows closely the matching protocol outlined in Section 4. Due to the computing time necessary for bootstraps the standard errors presented are not obtained from the bootstrap distribution.

5.2.1 The estimation of the conditional participation probabilities

Conditional probabilities are estimated at each step in each subsample using binary probit models, which are (sequentially) subjected to specification tests. In total 6 binary probit models presented in Table 8 cover all selection equations that are of interest for the effects of the treatments defined above. The specifications follow broadly specifications by GL02 and Gerfin, Lechner, and Steiger (2002) adjusted for the specifics of the dynamic framework and the different aggregation levels and sample sizes. The exclusion restrictions used are mainly motivated by the dynamic structure (X_1 does not influence S_1 conditional on X_0). Furthermore, the sample sizes and the variation of the dependent variable in some cases require omitting variables from the specification. The latter problem is particularly pronounced for the transition from E to EE with only 19 observations leaving the sequence at this stage.

Table 8: Probit specifications and estimates of the probit coefficients

Variable		U vs. C	U vs. E	U vs. T	UU if U	CC if C	EE if E	TT if T
Age in years / 10		-*.04*	-.00	.02	*.06*	-.00	-.44	-.00
Female		-*.16*	.26	-*.22*	-.06	.05	.02	.17
Mother tongue not German / French / Italian		-.20	*.31*	*.14*	.04	.23	0	.11
Mother tongue G / F / I, not main language in own canton		-.09	-.09	-.16	-.04	.14	.16	.19
French mother tongue		-*.26*	.18	-*.12*	-.07	*.41*	-.65	.25
Italian mother tongue		-.16	-*.33*	-.12	.09	.59	0	-.03
Nationality*:*	foreign with yearly permit	-.03	*.26*	.08	.11	.02	.40	-.03
Foreign Languages*:*	other Swiss language	.05	.26	*.13*	-.03	.30	-1.14	.06
Job position	very low	.03	0	-.03	-*.11*	.16	.58	.02
Qualification level:	skilled (highest)	-.17	-.13	-.03	-.04	-.00	1.11	-.22
	unskilled (lowest)	.01	-.11	.07	.04	-.19	-.62	.00
Chances to find a job:	no information	*.25*	-.01	-.03	-.00	-.06	0	.37
(reference category: medium)	very easy	.11	*.36*	.06	-.23	.52	0	-.08
	easy	.02	.28	-.01	-.06	.22	0	.10
	difficult	.05	-.11	*.13*	-.00	-.05	0	.15
	special case	.07	-.14	.07	-.02	0	0	0
Chances to find a job :	no inform. Feb ‘98	X	X	X	.06	0	0	0
(reference category: medium)	very easy Feb ‘98	X	X	X	.19	-.61	.32	.05
	easy Feb ‘98	X	X	X	.00	-.31	.71	-.09
	difficult Feb ‘98	X	X	X	-.01	-.02	.32	-.19
	special c. Feb ‘98	X	X	X	.05	0	0	0
Looking for job with X % of full time job, Feb. ’98 (0-1)		X	X	X	*1.95*	-.52	-.52	*1.12*
Unemployment-status:	full time UE	-.20	-.14	-.15	-.52	-1.00	-.71	-.16
	part time UE	.15	-.11	-.04	.13	*-1.43*	0	.38
No information on desired full or part time job		.03	-.00	*.11*	*1.54*	-.67	0	*1.14*
Desired = previous occupation, 3-digit level, Feb. ‘98		X	X	X	-*.09*	.11	.44	.14
Previous occupation:	metals	*.26*	.15	0	.06	-.76	0	0
	painting, technical drawing	0	0	.12	0	0	0	-*.33*
	entrepreneurs, senior officials, justice	0	0	*.51*	0	0	0	-.14
	office and computer	-*.24*	.14	*.27*	-.04	.01	0	.06
	retail trade	-.10	.44	.24	.08	-.38	0	.17
	science	-*.45*	.16	0	-.02	-.07	0	0

Table 8 to be continued.

[30] Ideally, in such an application the intermediate outcome variable one would like to observe is ‘number and quality of job offers received’. However, if the caseworker is aware of the job offers received she should factor this information into her assessment of employability.

Table 8: continued

Variable		U vs. C	U vs. E	U vs. T	UU if U	CC if C	EE if E	TT if T
Monthly insured earnings (in last job) in CHF / 1000		-.02	.09	-.02	*-.04*	-.01	-.26	.05
Employed Feb. '02		X	X	X	*-1.16*	-.58	0	*-1.62*
Current unemployment spell is first spell		-.02	.01	.12	*.07*	.31	*.80*	-.07
Sanction days without benefits (current spell) / 10		.01	.02	.05	.03	-.03	-.09	-.05
Sanction days without benefits / 10 Jan. + Feb. '98		X	X	X	.02	-.20	0	.17
Duration of UE spell (days / 10)		*-.78*	-.43	0	.12	*-.13*	-.05	0
Unemployment benefits / 1000	in 1996	-.01	.00	0	*.01*	-.01	0	0
	in 1997	*.01*	*-.01*	*.01*	.00	-.00	.00	.01
Month of entry into social security system		*-.46*	-.02	.18	.09	*.72*	-1.07	-.36
Share of employment 1988-1997		-.35	.59	*-.49*	*-.34*	.57	0	*.75*
Average duration of UE spells 1988-1997 / 10		-.80	.73	1.21	-.01	*3.95*	0	-.31
Subsidized temporary job before Dec. 97		0	0	*-.38*	0	0	0	.19
Size of town (previous employment) < 30.000		-.03	-.05	-.03	-.14	0	0	0
Region of placement office in rural area		.00	.11	-.07	.07	.01	-.20	-.15
Inflow rate to long-term unemployment (by region of PO)		.53	.41	*.56*	.19	.37	-1.98	.25
Region (reference Central)	west	-.01	.08	*-.20*	*-.20*	*-.57*	0	.06
	east	0	*.67*	0	0	0	0	0
	Zurich	*.18*	*.69*	*.12*	*.24*	-.24	.29	*-.41*
	south-west	*.55*	*.68*	*-.24*	-.08	-.68	0	-.13
	north-west	*.42*	.43	*.21*	.14	*-.74*	-.38	-.03
	Ticino	*.56*	*.82*	*.37*	.20	*-1.01*	0	.19
Additional regional effects by canton: Bern		.24	.02	.14	*.37*	.02	0	-.21
French main language in canton		.11	-.15	*.39*	*.47*	-.18	0	-.13
Subsample		U or C	U or E	U or T	U	C	E	T
Number of observations in subsample		8512	8100	8772	7982	574	118	790
Dependent variable		U	U	U	UU	CC	EE	TT
Mean of dependent variable in subsample		.94	.98	.91	.64	.55	.83	.48

Note: Binary probit model estimated on the respective subsample. All specifications include an intercept and are subjected to specification tests (omitted variables, non-normality). If not stated otherwise, all information in the variables relates to the last day in December 1997. *Exclusion restrictions*: 0: Variables omitted from specification. X: Variable not temporarily prior to dependent variable. ***Bold*** letters in ***italics*** denote significance at the 1% level. **Bold** letters denote significance at the 5% level. *Italics* denote significance at the 10% level.

5.2.2 Common support

The next remark about the matching protocol concerns the common support. The definition of the region of common support - as defined on the reference distribution for which the effect is desired - has to be adjusted period by period with respect to the conditioning variables of this period. The matching estimator allows easily tracing back the impact of this procedure on the reference distribution. Suppose that for the choice of treatments in period 2 a lack of support is detected. Suppose furthermore that the reference distribution is defined with respect to the treatment status in period 1. By the virtue of sequential matching, the observations without support in period 2 are related to some specific observations in period 1, namely those they are matched to in the matching steps before period 2. Therefore, the adjustment of the distribution in period 1 due to the lack of support in any other period is immediate.

Table 9: Estimation results for gross monthly earnings (in CHF) in December 2002

Sequences $\underline{s}_2^1$ $\underline{s}_2^0$	Target pop. s_1	$E(Y_0 \mid \underline{S}_2 = \underline{s}_2^1)$, $E(Y_0 \mid \underline{S}_2 = \underline{s}_2^0)$, $E(Y_0 \mid S_1 = s_1)$	$N_{\underline{s}_2^1}$, $N_{\underline{s}_2^0}$, N_{s_1}	Observations deleted for common support			Mean of *Y0* in deleted sub-sample	Concentration of weights in %	$E(Y_t \mid \underline{S}_2 = \underline{s}_2^1)$, $E(Y_t \mid \underline{S}_2 = \underline{s}_2^0)$, $E(Y_t \mid S_1 = s_1)$		$E(Y_t^{\underline{s}_2^1} \mid S_1 = s_1)$, $E(Y_t^{\underline{s}_2^0} \mid S_1 = s_1)$, $\theta_t^{\underline{s}_2^1,\underline{s}_2^0}(s_1)$	
				t=1	t=2	in % of s_1						
										std.		std.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
UU		3904	5122		266			27	2171	(34)	2326	(67)
CC		3932	316	164	268			46	2699	(135)	3014	(241)
	U	3943	7982			12	4166		2387	(30)	*-688*	(250)
UU		3904	5122	0	27			24	2171	(34)	2484	(144)
CC		3932	316		9			24	2699	(135)	2986	(161)
	C	4012	573			6	4396		2705	(104)	**-502**	(216)
UU		3904	5122		238			27	2171	(34)	2307	(63)
EE		3517	99	774	55			49	2069	(208)	2053	(380)
	U	3943	7982			13	3009		2387	(30)	254	(386)
UU		3904	5122	0	3			16	2171	(34)	1662	(207)
EE		3517	99		3			21	2069	(208)	2003	(223)
	E	3592	118			5	4019		2120	(189)	-341	(304)
UU		3904	5122		223			27	2171	(34)	2299	(64)
TT		4067	382	55	468			39	2595	(119)	2360	(187)
	U	3943	7982			9	4069		2387	(30)	-61	(198)
UU		3904	5122	0	27			25	2171	(34)	2243	(127)
TT		4067	382		40			26	2595	(119)	2475	(146)
	T	4023	790			8	3867		2676	(85)	-232	(194)
EE		3517	99		3			20	2069	(208)	2179	(233)
TT		4067	382	1	8			30	2595	(119)	2167	(329)
	E	3943	118			8	3757		2481	(189)	11	(404)
EE		3517	99	133	6			45	2069	(208)	2120	(377)
TT		4067	382		43			25	2595	(119)	2495	(148)
	T	4023	790			23	4422		2676	(85)	-375	(405)

Note: The sequences are defined on a bimonthly basis. *U*: Unemployed; *C, E*: Participating in a programme of the active labour market policy (Course or Employment programme); *T*: Receiving a temporary wage subsidy. Earnings are coded as 0 if individuals receive a temporary wage subsidy, participate in a programme, or are unemployed.

Concentration of weights: Share of the largest weights (10%) to total weights. *Y0* denotes 'insured monthly earnings' in CHF that are used by the unemployment insurance to calculate unemployment benefits. Since they are computed from the data coming from the UE insurance, they may not be perfectly comparable to the outcome variable that is calculated from the pension records. *Std.*: Standard error of estimated mean (conditional on the weights and without adjustment for estimated scores).

Bold letters in ***italics*** denote significance at the 1% level. **Bold** letters denote significance at the 5% level. *Italics* denote significance at the 10% level. Estimates for the same parameter may differ across different comparisons because the common support is defined with respect to $\theta_t^{\underline{s}_2^1,\underline{s}_2^0}(s_1)$.

Table 9 contains the results of the estimation as well as descriptive statistics about the data used in the specific comparisons and operational characteristics of the sequential matching estimation.

Columns 5 and 6 of Table 9 show how many observations are deleted at each step when the common support is enforced. The total share deleted (col. 7) varies from 5% for the comparison of unemployment with employment programmes for the (small) population of 1st period participants in employment programmes and 23% for the comparison of the sequences that appear to have the most dissimilar participants, namely EE and TT. However, even in this case the share of observations deleted depends very much on whether the small (E, 8%) or the large target population is considered (T, 23%). In the latter case column 8 containing the previous earnings of those deleted, suggesting that for many of the participants in T with higher earnings no adequate matches could be found. Therefore, enforcing the common support requirement redefine the underlying population for which this effect is defined considerably and thus the results for the original effect are not very reliably estimated for this specific comparison.[31]

5.2.3 Weights, adjustments, and sample sizes

The next remark concerns an issue arising because matching is with replacement: if few observations have a very large weight when computing the weighted means in the matched comparison groups, a very noisy estimator results. This effect can be fairly easily spotted with the type of matching used by analysing the weights directly. Column 9 contains the 10 % concentration ratio of the weights in the two subsamples of treated observations (after adjusting for common support).[32] The results show systematically higher ratios when small treated groups are matched to large (and diverse) target populations, such as EE to U.

Comparing the mean earnings of both treatment groups with the earnings of the target group (col. 3) gives a rough idea whether we expect matching to adjust mean observed post-treatment earnings upwards or downwards. Indeed these adjustments appear to move the estimators in the right direction in almost all

[31] Lechner (2001b) discusses alternative methods of accounting for common support problems, in particular bounding the parameter of interest that could not be estimated due to the common support problem. In principle these suggestions can be adapted to the dynamic framework in a straightforward manner. Note also that tighter or less tight conditions than those appearing in the matching protocol may be imposed to define the common support.

[32] The 10% ratio shows the share of the target population covered by those 10% of the treatment population that have the largest weights. In case of random assignment this ratio would be about 10. If only a single observation is matched to all members of the target population, this ratio would be 100.

cases although the magnitudes appear sometimes to be surprising given the pre-treatment differences. However, there are differences in other confounders that are not displayed in the table. A further observation from this table concerns standard errors and sample sizes necessary for a reliable dynamic analysis. In fact in many cases the standard errors are so large that effects of more than 800 CHF per month would be needed for significance. However such large effects appear to be rather implausible for the programmes under investigation.

5.2.4 Results

To shed some light on the development of the effects over time, Table 10 presents results for outcome variables 1, 6, 17 months after the sequence of interest. Furthermore, since it is the ‘official’ objective for the active labour market policy to increase reemployment chances of the unemployed, an outcome variable measuring whether an individual is employed in that particular month is considered as well.

Table 10: Dynamics of $\theta_t^{\underline{s}_2^1,\underline{s}_2^0}(s_1)$

Seq. $\underline{s}_2^1$ - $\underline{s}_2^0$	Target pop.	$E(Y_t \mid \underline{S}_2 = \underline{s}_2^1)$			$E(Y_t \mid \underline{S}_2 = \underline{s}_2^0)$			$E(Y_t^{\underline{s}_2^1} \mid S_1 = s_1)$			$E(Y_t^{\underline{s}_2^0} \mid S_1 = s_1)$			$\theta_t^{\underline{s}_2^1,\underline{s}_2^0}(s_1)$		
		5/01	9/01	9/02	5/01	9/01	9/02	5/01	9/01	9/02	5/01	9/01	9/02	5/01	9/01	9/02
								Earnings (0 if not employed)								
UU-CC	U	1345	1913	2533	867	1728	2683	1504	2034	2706	964	1883	2905	539	150	-198
	C							1260	1851	2785	1002	1861	2963	258	-10	-177
UU-EE	U				371	1112	2163	1442	1953	2619	129	1184	2243	*1313*	769	376
	E							914	1447	1918	358	1041	2046	*556*	406	-127
UU-TT	U				1368	2124	2921	1531	2020	2653	1239	1908	2711	294	112	186
	T							1531	2070	2680	1389	2125	2791	141	-54	-110
EE-TT	E							324	1149	2175	1386	2331	2557	-1062	*-1182*	-382
	T							245	1244	2199	1490	2295	2978	*-1244*	*-1050*	-779
								Employment								
UU-CC	U	22	31	44	14	28	46	24	34	48	17	30	48	*7*	4	0
	C							20	29	51	16	30	50	4	-1	1
UU-EE	U				8	25	39	23	32	46	2	27	44	*21*	5	3
	E							15	21	41	7	23	36	8	-2	6
UU-TT	U				22	35	52	25	34	46	20	30	49	*6*	4	-2
	T							22	32	44	22	35	52	0	-3	-8
EE-TT	E							7	25	36	24	43	45	-17	*-17*	-9
	T							5	26	38	27	40	53	*-21*	*-15*	-9

Note: See note below Table 9. For the definition of the outcome variables individuals are considered as not employed if they receive a temporary wage subsidy or participate in a programme.

The results broadly confirm the previous findings in the sense (i) that employment programmes appear to be the least effective, (ii) that it is very hard to pin down effects of training courses, and (iii) that wage subsidies look like the only programme that seems to be somewhat successful. But again the variability of the estimates is a problem. It is however obvious that the negative effects of the employment programmes become smaller or even disappear after 17 months, which could suggest that they are basically due to some lock-in effect: Most employment programmes run for six months and during this time job search activities of participants are most likely much lower than outside a programme.

6 Conclusion

This paper proposed and discussed sequential matching and inverse selection probability weighted estimators that can be used to estimate the causal effects defined within the dynamic causal model introduced by Robins (1986) and extended by Lechner and Miquel (2001). The sequential matching estimators mimic the simple, well known and frequently applied matching estimators based on so-called propensity scores that are popular among empirical researchers for the static causal model. A small Monte Carlo study revealed that the suggested estimators perform well in small and medium size samples. Using an application of the sequential matching estimators to an empirical problem, namely an evaluation study of the Swiss active labour market policies, some implementational issues are discussed and results are provided.

Future work should be directed at extending the rigours proofs of the asymptotic distribution by Abadie and Imbens (2002) to the sequential versions of the matching estimators. Furthermore, other nonparametric sequentially weighted regression type estimators that have been proposed for the static model can be developed for the dynamic context as well and compared. Finally, efforts should be spent in obtaining more reliable inference perhaps by adapting the ‘match-the-treated-to-treated-and-the-controls-to-the-controls’ methods proposed by Abadie and Imbens (2002). Another option would be to develop more sophisticated bootstrap procedures.

References

Abadie, A., and G. Imbens (2002): “Simple and Bias-Corrected Matching Estimators for Average Treatment Effects”, *mimeo*.

Abbring, J. H., and G. van den Berg (2002): "Dynamically Assigned Treatments: Duration Models, Binary Treatment Models, and Panel Data Models", mimeo.

Abbring, J. H., and G. van den Berg (2003): “The Non-Parametric Identification of Treatment Effects in Duration Models”, forthcoming *Econometrica*.

Angrist, J. D., and A. B. Krueger (1999): “Empirical Strategies in Labor Economics”, in O. Ashenfelter and D. Card (Hrsg.), *Handbook of Labor Economics*, Vol. III A, Amsterdam: North-Holland, 1277-1366.

Arellano, M., and B. Honoré (2001): “Panel Data Models: Some Recent Developments”, in J.J. Heckman and E. Leamer, *Handbook of Econometrics*, Vol. V., Ch. 53, Amsterdam: North-Holland, 3229-3296.

Arulampalam, W., and A. L. Booth (2001): “Learning and Earning: Do Multiple Training Events Pay? A Decade of Evidence from a Cohort of Young British Men”, *Economica*, 68, 379-400.

Bergemann, A., B. Fitzenberger, and S. Speckesser (2001): “Evaluating the Employment Effects of Public Sector Sponsored Training in East Germany: Conditional Difference-in-Differences and Ashenfelter’s Dip”, *mimeo*.

Dehejia, R. H., and S. Wahba (1999): “Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programs,” *Journal of the American Statistical Association*, 94, 1053-1062.

Dehejia, R. H., and S. Wahba (2002): "Propensity Score-matching Methods for Nonexperimental Causal Studies", *Review of Economics and Statistics*, 84, 151-161.

Frölich, M. (2001): “Nonparametric Covariate Adjustment: Pair Matching versus Local Polynomial Matching”, *mimeo*.

Frölich, M. (2004): "Finite Sample Properties of Propensity-Score Matching and Weighting Estimators", forthcoming in *Review of Economics and Statistics*, 86(1).

Gerfin, M., and M. Lechner (2002): “Microeconometric Evaluation of the Active Labour Market Policy in Switzerland,” *The Economic Journal*, 112, 854-893.

Gerfin, M., M. Lechner, and H. Steiger (2002): "Does subsidised temporary employment get the unemployed back to work? An econometric analysis of two different schemes", *Discussion Paper, Department of Economics, University of St. Gallen*.

Gill, R. D., and J. M. Robins (2001): "Causal Inference for Complex Longitudinal Data: the continuous case", *The Annals of Statistics*, 1-27.

Hahn, J. (1998): “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects”, *Econometrica*, 66, 315-331.

Ham, J. C., and R. J. LaLonde (1996): "The Effect of Sample Selection and Initial Conditions in Duration Models: Evidence from Experimental Data on Training", *Econometrica*, 64, 175-205.

Heckman, J. J. (1979): “Sample Selection Bias as a Specification Error”, *Econometrica*, 47, 153-161.

Heckman, J. J., H. Ichimura, and P. E. Todd (1998): “Matching as an Econometric Evaluation Estimator”, *Review of Economic Studies*, 65, 261-294.

Heckman, J. J., H. Ichimura, J. A. Smith, and P. Todd (1998): "Characterisation Selection Bias Using Experimental Data", *Econometrica*, 66, 1017-1098.

Heckman, J. J., R. J. LaLonde, and J. A. Smith (1999): “The Economics and Econometrics of Active Labor Market Programs”, in O. Ashenfelter and D. Card (eds.): *Handbook of Labor Economics*, Vol. III A, Amsterdam: North-Holland, 1865-2097.

Heckman, J.J., and R. Robb (1985): ”Alternative Methods of Evaluating the Impact of Interventions”, in: J. J. Heckman and B. Singer (eds.), *Longitudinal Analysis of Labour Market Data*, New York: Cambridge University Press, 156-245.

Heckman, J.J., and V.J. Hotz (1989): "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training", *Journal of the American Statistical Association*, 84, 862-880.

Hernan, M. A., B. Brumback, and J. M. Robins (2001): "Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Trials," *Journal of the American Statistical Association*, 96, 440-448.

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", *Econometrica*.

Holland, P.W. (1986): “Statistics and Causal Inference”, *Journal of the American Statistical Association*, 81, 945-970, with discussion.

Horvitz, J. L., and D. Thomson (1952): "A Generalisation of Sampling without Replacement from a Finite Population", *Journal of the American Statistical Association*, 47, 663-685.

Ichimura, H., and O. Linton (2001): “Trick or Treat: Asymptotic Expansions for Some Semiparametric Program Evaluation Estimators”, *mimeo*.

Imbens, G. W. (2000): “The Role of the Propensity Score in Estimating Dose-Response Functions”, *Biometrika*, 87, 706-710.

Imbens, G. W. (2003): "Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review", *mimeo.*

Lechner, M. (1999): “Earnings and Employment Effects of Continuous Off-the-job Training in East Germany after Unification,” *Journal of Business & Economic Statistics*, 17, 74-90.

Lechner, M. (2001a): “Identification and estimation of causal effects of multiple treatments under the conditional independence assumption”, in M. Lechner and F. Pfeiffer (eds., 2001), *Econometric Evaluation of Active Labour Market Policies*, Heidelberg: Physica, 43-58.

Lechner, M. (2001b): “A note on the common support problem in applied evaluation studies”, *Discussion Paper 2001-01, Department of Economics, University of St. Gallen.*

Lechner, M. (2002a): “Programme Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labour Market Policies”, *Review of Economics and Statistics*, 84, 205-220.

Lechner, M. (2002b): “Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods”, *Journal of the Royal Statistical Society, Series A, Statistics in Society,* 165, 59-82.

Lechner, M., and R. Miquel (2001): “A Potential Outcome Approach to Dynamic Programme Evaluation – Part I: Identification”, *Discussion paper 2001-07, Department of Economics, University of St. Gallen*; substantially revised 2004.

Li, Y. P., K. J. Propert, and P. R. Rosenbaum (2001): “Balanced Risk Set Matching”, *Journal of the American Statistical Association*, 96, 870-882.

Miquel, R. (2003): "Identification of Effects of Dynamic Treatments", PhD Thesis, University of St. Gallen.

Nevo, A. (2003): "Using Weights to Adjust for Sample Selection When Auxiliary Information is Available", *Journal of Business & Economic Statistics*, 21, 43-52.

Robins, J. M. (1986): “A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect.” *Mathematical Modelling*, 7:1393-1512, with 1987 Errata to “A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect.” *Computers and Mathematics with Applications*, 14:917-921; 1987 Addendum to “A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect.” *Computers and Mathematics with Applications*, 14:923-945; and 1987 Errata to “Addendum to ‘A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect’.” *Computers and Mathematics with Applications*, 18:477.

Robins, J. M. (1997): "Marginal Structural Models", *Proceedings of the American Statistical Association*, Section on Bayesian Statistical Science, 1-10.

Robins, J. M. (1998): "Structural Nested Failure Time Models", in P.K. Anderson, N. Keiding (section editors), *Survival Analysis*, in P. Armitage, T. Colton (eds.), *The Encyclopaedia of Biostatistics*, 4372-4389, Chichester, UK: Wiley.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995): "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data", *Journal of the American Statistical Association*, 90, 106-121.

Robins, J. M., and A. Rotnitzky (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data", *Journal of the American Statistical Association*, 90, 122-129.

Rosenbaum, P. R. (1987): "Model-Based Direct Adjustment", *Journal of the American Statistical Association*, 82, 387-394.

Rosenbaum, P. R., and D. B. Rubin (1983): “The Central Role of the Propensity Score in Observational Studies for Causal Effects”, *Biometrika*, 70, 41-50.

Rosenbaum, P. R., and D. B. Rubin (1984): “Reducing the Bias in Observational Studies Using Subclassification on the Propensity Score”, *Journal of the American Statistical Association*, 79, 516-524.

Rosenbaum, P. R., and D. B. Rubin (1985): “Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score”, *American Statistician*, 39, 33-38.

Roy, A. D. (1951): “Some Thoughts on the Distribution of Earnings”, *Oxford Economic Papers*, 3, 135-146.

Rubin, D. B. (1973): “Matching to Remove Bias in Observational Studies”, *Biometrics*, 29, 159-183.

Rubin, D. B. (1974): “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies”, *Journal of Educational Psychology*, 66, 688-701.

Rubin, D. B. (1977): “Assignment to a Treatment Group on the Basis of a Covariate”, *Journal of Educational Statistics*, 2, 1-26.

Sianesi, B. (2001): “An Evaluation of the Active Labour Market Programmes in Sweden”, IFAU working paper 2001:5.

Smith, J., and P. Todd (2001): “Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods”, *American Economic Review, Papers & Proceedings*, 112-118.

Wooldridge, J. M. (2002): "Inverse Probability Weighted M-estimators for Sample Selection, Attrition, and Stratification, *cemmap discussion paper*, 11/02.

Zhao, Z. (2000): "Using Matching to Estimate Treatment Effects: Data Requirements, Sensitivity, and an Application", *mimeo*.

Appendix A: Additional considerations for multiple treatments and many periods

Since in many cases there may be more than two states in any period, this appendix informally discusses considerations important in applications and that are not addressed in the main body of the text. There is the issue for the propensity scores whether for each period one needs to take account for example for the event of not participating in s_2^k conditional on participating in s_1^k could mean any particular of the other different states. However, since in each step the independence assumption relates only to a binary comparison, e.g. $Y_2^{\underline{s}_2^k} \coprod \underline{1}(S_2 = s_2^k) | S_1 = s_1^k, \underline{X}_1 = \underline{x}_1$, and $Y_2^{s_1^k} \coprod \underline{1}(S_1 = s_1^k) | S_1 \in \{s_1^j, s_1^k\}, X_0 = x_0$ (s_1^j being the target population as before), it is obvious that the conditional probabilities of not participating in the event of interest conditional on the history is enough, and the exact composition of the alternative does not play any role.[33] This means that for the matching step in period 2 $P(S_2 = s_2^k | S_1 = s_1^k, \underline{X}_1 = \underline{x}_1)$, $P[S_1 = s_1^k | X_0 = x_0,\ S_1 \in \{s_1^l, s_1^k\}]$ and the matching step in period 1 $P[S_1 = s_1^k | X_0 = x_0,\ S_1 \in \{s_1^l, s_1^k\}]$ may be used. The multiple treatment feature of the problem does not add to the dimension of the propensity scores.

[33] Note that the same argument is developed by Imbens (2000) and Lechner (2001a) to show that in the static multiple treatment models conditioning on appropriate one-dimensional scores is sufficient.

Appendix B: The matching protocol used in the empirical application

Table B.1: A sequential matching estimator for $\theta_t^{\underline{s}_2^1,\underline{s}_2^0}(s_1^1)$ *based on propensity scores*

Step 0: Sample reduction		Delete all observation not belonging to s_1^1, $\underline{s}_2^1$, or $\underline{s}_2^0$
Step A: Match $\underline{s}_2^0 = (s_1^0, s_2^0)$ to s_1^1	A.1.0	Define a weight $w_i^{\underline{s}_2^0} = 0$ for every obs. in $\underline{s}_2^0$.
$(E(Y_t^{\underline{s}_2^0} \mid S_1 = s_1^1))$	A.1.P	Estimate a probit for $P(S_1 = s_1^0 \mid \underline{X}_0 = \underline{x}_0)$ → $p^{s_1^0}(\underline{x}_{0,i}) =: p_i^{s_1^0}$
	A.1.CS	Delete all obs. of s_1^1 with lower or higher values of $p_i^{s_1^0}$ than obs. in s_1^0
	A.1.M	For every obs. in s_1^1 not deleted in A.1.S find the obs. in s_1^0 that is closest in terms of $p_i^{s_1^0}$ (a match).
	A.1.C	For the matched obs. keep the value of $p_i^{s_1^0}$ of the obs. in s_1^1 they have been matched to. Some obs. in s_1^0 may appear many times in this matched sample.
	A.2.R	Define a sample of obs. in s_1^0.
	A.2.P	Estimate a probit for $P(S_2 = s_2^0 \mid S_1 = s_1^0, \underline{X}_1 = \underline{x}_1)$ → $p^{s_2^0 \mid s_1^0}(\underline{x}_{1,i}) =: p_i^{s_2^0 \mid s_1^0}$
	A.2.CS	Delete all obs. of the matched control sample of s_1^1 (defined in A.1.C) with lower or higher values of $p_i^{s_1^0}$ and $p_i^{s_2^0 \mid s_1^0}$ than obs. in $\underline{s}_2^0$.
	A.2.M	For every obs. in the matched control sample of s_1^1 not deleted in A.2.CS find an obs. in $\underline{s}_2^0$ that is closest in terms of $p_i^{s_2^0 \mid s_1^0}$ and $p_i^{s_1^0}$ using the Mahalanobis metric (covariance computed in s_1^1). Every time an obs. in $\underline{s}_2^0$ is matched, its weight $w_i^{\underline{s}_2^0}$ is increased by 1.
Step B: Match $\underline{s}_2^1 = (s_1^1, s_2^1)$ to s_1^1	B.1.0	Define a weight $w_i^{\underline{s}_2^1} = 0$ for every obs. in $\underline{s}_2^1$.
$(E(Y_t^{\underline{s}_2^1} \mid S_1 = s_1^1)$	B.2.R	Reduce sample to participants in s_1^1.
	B.2.P	Estimate a probit for $P(S_2 = s_2^1 \mid S_1 = s_1^1, \underline{X}_1 = \underline{x}_1)$ → $p^{s_2^1 \mid s_1^1}(\underline{x}_{1,i}) =: p_i^{s_2^1 \mid s_1^1}$
	B.2.CS	Delete all obs. of s_1^1 with lower or higher values of $p_i^{s_2^1 \mid s_1^1}$ than obs. in $\underline{s}_2^1$.
	B.2.M	For every obs. in s_1^1 not deleted in B.2.CS find the member of $\underline{s}_2^1$ that is closest in terms of $p_i^{s_2^1 \mid s_1^1}$. Every time an obs. in $\underline{s}_2^1$ is matched, its weight $w_i^{\underline{s}_2^1}$ is increased by 1.
Step C: Joint common support	C.1	Reduce $w_i^{\underline{s}_2^1}$ by 1 for every obs. *i* matched to an obs. in s_1^1 deleted in A.1.CS or A.2.CS.
	C.2	Reduce $w_i^{\underline{s}_2^0}$ by 1 for every obs. *i* matched to an obs. in s_1^1 deleted in B.2.CS.
Step D: Estimation of $\theta_t^{\underline{s}_2^1,\underline{s}_2^0}(s_1^1)$	D.1	$\hat{\theta}_t^{\underline{s}_2^1,\underline{s}_2^0}(s_1^1) = \frac{1}{\sum_{i \in \underline{s}_2^1} w_i^{\underline{s}_2^1}} \sum_{i \in \underline{s}_2^1} w_i^{\underline{s}_2^1} y_i - \frac{1}{\sum_{i \in \underline{s}_2^0} w_i^{\underline{s}_2^0}} \sum_{i \in \underline{s}_2^1} w_i^{\underline{s}_2^0} y_i$
	D.2	$\hat{V}ar(\hat{\theta}_t^{\underline{s}_2^1,\underline{s}_2^0}(s_1^1)) = \frac{\sum_{i \in \underline{s}_2^1} (w_i^{\underline{s}_2^1})^2 \, \hat{V}ar(Y_t \mid S = \underline{s}_2^1)}{(\sum_{i \in \underline{s}_2^1} w_i^{\underline{s}_2^1})^2} + \frac{\sum_{i \in \underline{s}_2^0} (w_i^{\underline{s}_2^0})^2 \, \hat{V}ar(Y_t \mid S = \underline{s}_2^0)}{(\sum_{i \in \underline{s}_2^0} w_i^{\underline{s}_2^0})^2}$ $\hat{V}ar(Y_t \mid S = \underline{s}_2) = \frac{1}{N^{\underline{s}_2}} \sum_{i \in \underline{s}_2} (y_i - \overline{y}_t^{\underline{s}_2})^2$, $\overline{y}_t^{\underline{s}_2} = \frac{1}{N^{\underline{s}_2}} \sum_{i \in \underline{s}_2} y_{ti}$, $N^{\underline{s}_2} = \sum_i \underline{1}(\underline{s}_{2,i} = \underline{s}_2)$

t > 1.

Appendix C: The sequential inverse probability-weighted estimators

In this appendix 3 examples show that the SIPW estimators proposed in Section 3 are indeed estimating the desired quantities. For simplicity, it is assumed that the probabilities are estimated consistently and with enough regularity such that the following exposition based on true probabilities holds asymptotically with estimated probabilities as well.

a) The SIPW estimator for $E(Y_2^{\underline{s}_2^k})$ is given great detail:

$$E\frac{1}{N}\sum_{i=1}^{N}\frac{y_{2,i}\underline{1}(s_{1,i}=s_1^k)\underline{1}(s_{2,i}=s_2^k)}{P(S_2=s_2^k\,|\,S_1=s_1^k,\underline{X}_1=\underline{x}_{1,i})P(S_1=s_1^k\,|\,X_0=x_{0,i})}=$$

$$=E\frac{Y_2^{\underline{s}_2^k}\underline{1}(S_1=s_1^k)\underline{1}(S_2=s_2^k)}{P(S_2=s_2^k\,|\,S_1=s_1^k,\underline{X}_1=\underline{x}_1)P(S_1=s_1^k\,|\,X_0=x_0)}=$$

$$=\underset{X_0}{E}\frac{E[\frac{Y_2^{\underline{s}_2^k}\underline{1}(S_1=s_1^k)\underline{1}(S_2=s_2^k)}{P(S_2=s_2^k\,|\,S_1=s_1^k,\underline{X}_1=\underline{x}_1)}\,|\,X_0=x_0]}{P(S_1=s_1^k\,|\,X_0=x_0)}=$$

$$=\underset{X_0}{E}\frac{E[\frac{Y_2^{\underline{s}_2^k}\underline{1}(S_2=s_2^k)}{P(S_2=s_2^k\,|\,S_1=s_1^k,\underline{X}_1=\underline{x}_1)}\,|\,S_1=s_1^k,X_0=x_0]P(S_1=s_1^k\,|\,X_0=x_0)}{P(S_1=s_1^k\,|\,X_0=x_0)}=$$

$$=\underset{X_0}{E}\underset{X_1|X_0,S_1}{E}\frac{E[Y_2^{\underline{s}_2^k}\underline{1}(S_2=s_2^k)\,|\,S_1=s_1^k,\underline{X}_1=\underline{x}_1]}{P(S_2=s_2^k\,|\,S_1=s_1^k,\underline{X}_1=\underline{x}_1)}=$$

$$=\underset{X_0}{E}\underset{X_1|X_0,S_1}{E}\frac{E[Y_2^{\underline{s}_2^k}\,|\,\underline{S}_2=\underline{s}_2^k,\underline{X}_1=\underline{x}_1]P(S_2=s_2^k\,|\,S_1=s_1^k,\underline{X}_1=\underline{x}_1)}{P(S_2=s_2^k\,|\,S_1=s_1^k,\underline{X}_1=\underline{x}_1)}=$$

$$=\underset{X_0}{E}\underset{X_1|X_0,S_1}{E}E[Y_2^{\underline{s}_2^k}\,|\,\underline{S}_2=\underline{s}_2^k,\underline{X}_1=\underline{x}_1]=$$

$$=\underset{X_0}{E}\underset{X_1|X_0,S_1}{E}E[Y_2^{\underline{s}_2^k}\,|\,S_1=s_1^k,\underline{X}_1=\underline{x}_1]=\underset{X_0}{E}E[Y_2^{\underline{s}_2^k}\,|\,S_1=s_1^k,X_0=x_0]=\underset{X_0}{E}E[Y_2^{\underline{s}_2^k}\,|\,X_0=x_0]=E(Y_2^{\underline{s}_2^k}).$$

b) The SIPW estimator for $E(Y_2^{\underline{s}_2^k}\,|\,S_1=\underline{s}_1^k)$ is the same IPW estimator as for static treatment effects:

$$E\frac{1}{N^{s_1^k}}\sum_{i\in s_1^k}\frac{y_{2,i}\underline{1}(s_{2,i}=s_2^k)}{P(S_2=s_2^k\,|\,S_1=s_1^k,\underline{X}_1=\underline{x}_{1,i})}=$$

$$= E\left[\frac{Y_2^{\underline{s}_2^k}\underline{1}(S_2 = s_2^k)}{P(S_2 = s_2^k \mid S_1 = s_1^k, \underline{X}_1 = \underline{x}_1)} \mid S_1 = s_1^k\right] = \ldots =$$

$$= \underset{\underline{X}_1|S_1}{E}\left[\frac{E(Y_2^{\underline{s}_2^k} \mid \underline{S}_2 = \underline{s}_2^k, \underline{X}_1 = \underline{x}_1)P(S_2 = s_2^k \mid S_1 = s_1^k, \underline{X}_1 = \underline{x}_1)}{P(S_2 = s_2^k \mid S_1 = s_1^k, \underline{X}_1 = \underline{x}_1)}\right] = \ldots =$$

$$= \underset{\underline{X}_1|S_1}{E}\left[E(Y_2^{\underline{s}_2^k} \mid \underline{S}_2 = \underline{s}_2^k, \underline{X}_1 = \underline{x}_{1,i})\right] = \underset{\underline{X}_1|S_1}{E}\left[E(Y_2^{\underline{s}_2^k} \mid S_1 = s_1^k, \underline{X}_1 = \underline{x}_{1,i})\right] = E(Y_2^{\underline{s}_2^k} \mid S_1 = s_1^k)\,.$$

c) The SIPW estimators for $E(Y_2^{\underline{s}_2^k} \mid S_1 = s_1^j)$ (of which $E(Y_2^{\underline{s}_2^k} \mid S_1 = s_1^k)$ is a special case) and $E(Y_2^{\underline{s}_2^k} \mid S_1 = \underline{s}_1^j)$ are obtained by using the same steps as in a) and noting that $\frac{N}{N^{s_1^j}}$ and $\frac{N}{N^{\underline{s}_2^j}}$ converge to $\frac{1}{P(S_1 = s_1^j)}$ and $\frac{1}{P(\underline{S}_2 = \underline{s}_2^j)}$. Furthermore, Bayes law gives the connection between the conditional and unconditional counterfactuals. For $E(Y_2^{\underline{s}_2^k} \mid S_1 = s_1^j)$, for example, using $f(x_0) = \frac{f(x_0 \mid S_1 = s_1^j)P(S_1 = s_1^j)}{P(S_1 = s_1^j \mid X_0 = x_0)}$ leads to the desired result:

$$\underset{X_0}{E}\left\{E[Y_2^{\underline{s}_2^k} \mid S_1 = s_1^k, X_0 = x_0]\frac{P(S_1 = s_1^j \mid X_0 = x_0)}{P(S_1 = s_1^j)}\right\} =$$

$$= \underset{X_0}{E}\left\{E[Y_2^{\underline{s}_2^k} \mid S_1 = s_1^j, X_0 = x_0]\frac{P(S_1 = s_1^j \mid X_0 = x_0)}{P(S_1 = s_1^j)}\right\} =$$

$$= \underset{X_0|S_1 = s_1^j}{E}\left\{E[Y_2^{\underline{s}_2^k} \mid S_1 = s_1^j, X_0 = x_0]\frac{P(S_1 = s_1^j \mid X_0 = x_0)}{P(S_1 = s_1^j)}\frac{P(S_1 = s_1^j)}{P(S_1 = s_1^j \mid X_0 = x_0)}\right\} =$$

$$= \underset{X_0|S_1 = s_1^j}{E} E[Y_2^{\underline{s}_2^k} \mid S_1 = s_1^j, X_0 = x_0] = E[Y_2^{\underline{s}_2^k} \mid S_1 = s_1^j]\,.$$